Essays on the economics of decision making

Doctoral Thesis

Presented to the Faculty of Management, Economics and Social Sciences at
the University of Fribourg (Switzerland)

by

Anis NASSAR
Develier, JU

in fulfillment of the requirements for the degree of
Doctor of Economics and Social Sciences (Dr. rer. pol.)

Accepted by the Faculty of Management, Economics
and Social Sciences on September 27, 2021 at the proposal of

Prof. Dr. Berno Buechel (first referee)
Prof. Dr. Holger Herz (chairman of the committee)
Prof. Dr. Antoine Mandel (second referee)
Dr. Romain Espinosa (external expert)

Fribourg (Switzerland), 2021

The Faculty of Management, Economics and Social Sciences at the University of Fribourg (Switzerland) neither approves nor disapproves the opinions expressed in a doctoral thesis. They are to be considered those of the author. (Faculty council decision 23.01.1990)

Acknowledgments

Some decisions seem to be obvious. They follow a series of logical steps and rational processes, as if they were the only possible outcome. Starting a Ph. D. in Economics, after four years in the private sector marketing razors and epilators, was not one of those decisions. Somehow, the knowledge I acquired about beards, body hair, and the best ways to remove them in a pain-free manner, would not prove to be useful ground to build my research on. However, getting a better understanding of how people reach decisions, even surprising ones, is the central theme of this thesis. So at least, that is fitting.

Four years after this puzzling decision, I can only reflect on all that I have learned and realize how indebted I am to each of my colleagues for the skills and knowledge they enabled me to develop. First, I would like to thank my supervisor, Berno Buechel. Working under his supervision and collaborating with him has been a pleasure. I benefited thoroughly from his guidance and experience. He has always put his trust in me and his door was constantly open. I am extremely grateful for his investment in my work and in me as a person.

I would also like to thank all of my colleagues at the University of Fribourg. Professors, post-docs and fellow Ph. D. candidates, who have always been available to discuss research and share valuable insights. They also to made this whole experience an enjoyable one. I am sorry I skipped so many lunches, I will make it up to you with afternoon beers. I would particularly like to mention Chris and Elisa, with whom I shared a few CUSO events and their many baby foot games, as well as a the first chapter of this thesis. Ximena, Anna, Leander, Patricia, Stefan, Fanyuan, Christian, Jana, Martin, Holger, Mark, Christelle and many others, thanks for the drinks at Cyclo, the coffees, the yoga and simply for the good times we shared. Ruth and Andrea, thanks for always finding a solution to every issue I came up with.

The other colleagues, friends and frères I would like to thank are my Gerzensee class. I am so glad we got to go the castle[1] together before the corona-life. I learned so much from you, including how to "auuuuh", how to excel at football and how to swim across an entire lake. Elio, Jerem, Matthias, Nico, Max, Juliette, Evert, Flavia, Florian, Laura, Lorenz, Hyacinthe, Radu, Giulia, Kyungbo, Benjamin, Tabea, Anja and anyone I missed, thanks for this unforgettable year!

[1] Yes, they send us to a castle with three course meals and a private lake to learn about economics from Professors at the top of their field. It's a good deal.

Et un merci spécial à Romain dans la langue de Molière (mais sans écriture inclusive, évitons la polémique). Travailler avec toi, c'était simplement le bonheur. J'en ai presque regretté la publication et que le projet se soit achevé. Merci aussi d'avoir ranimé mon espoir pour la cause animale. Je serai toujours admiratif de ta détermination et de tous les projets que tu entreprends, dans le monde scientifique et politique, pour améliorer la condition animale.

Finally, I thank the members of the committee, Holger Herz, Antoine Mandel, Berno Buechel, and Romain Espinosa for having accepted to invest their time in the evaluation of my thesis and for the comments they may provide.

Contents

List of figures

List of tables

Introduction

Human decision making is at the center of economics. Every issue that we are facing as a society, from the simplest to the most complex, will be addressed more or less adequately depending on our ability to make the right decisions.

Economists have tried to model and understand decision making since the early days of the discipline. Famously, Mill (1874) introduced a hypothetical subject with limited and well-defined objectives. This *homo economicus* made rational decisions driven by his four principal interests: accumulation, luxury, leisure, and procreation. On one side, this approach that limits the range of needs and motives of agents is well-suited for broad economic analysis, as trying to account for all the psychological make-up of human beings risks indeterminacy (Persky, 1995). On the other side, critics accused homo economicus of having overly simplistic motives and of being too rational, rendering him irrelevant in real-life situations (Drucker, 1939; Leibenstein, 1976). A century later, homo economicus is still alive and utilized when the simplification he provides is appropriate. A labor-leisure trade-off, for example, still characterizes the range of choices of agents in certain models. However, homo economomicus is now also accompanied by “flesh-and-blood” counterparts. In the 1960s, as collaboration between psychologists and economists accelerated, behaviors that did not match normative expectations were identified. Namely, light was shed on complex motives and on critical flaws in reasoning, which led the deviations between actual behaviors and the normative model to be fundamental, wide-spread and systemic (Tversky and Kahneman, 1986). Decisions are made by men and women. Their humanity could no longer be ignored.

New, complex motives have hence been investigated and measured, particularly in the confined environments of theoretic models and laboratory experiments. For example, it is now established that some people also have altruistic considerations when making decisions. They cooperate in prisoner’s dilemma, even when there is no possibility to build a reputation, and where defection is the unique dominant strategy (Andreoni and Miller, 1993). In dictator games, a fraction of participants willingly sacrifice part of their payoff, with no possibility of any monetary compensation, simply to donate to other participants or to charities (Eckel and Grossman, 1996). Unselfish behaviors have also been observed in public goods games, trust games, or gift exchanges (Andreoni et al., 2010). Similarly, people care not only about others’ absolute well-being, but also about concepts of fairness, with regards of the equity of outcomes, or with regards to adequately rewarding effort (Fehr and Schmidt, 2006). These unselfish interests even extend to non-human animals (Lusk and Norwood, 2012). It could be argued that these other-regarding behaviors are actually selfish if people get a mental reward for acting pro-socially, such as an improved self-image (Bénabou and Tirole, 2006). Nonetheless, under either interpretation, people are ready to forego payment to satisfy their either "true" or selfish altruism, far from the considerations of the simple homo economicus. Complex, intricate motives cannot be overlooked as they guide the economic choices of every decision maker.

Because humans are also bounded in their rationality, even considering every possible motive perfectly would not be sufficient to explain how decisions are reached. One might simply not have the technical capability to process all the parameters required to make the optimal choice. Interestingly, even if one has sufficient technical capability, sub-optimal decision making can still occur. Reasoning indeed requires substantial cognitive effort. To reduce the load of this effort, the human mind tends to employ different strategies, such as biases and heuristics (Tversky and Kahneman, 1974). Biases characterize our penchant to think in a certain way. Salience bias, for example, leads people to make decisions using the most prominent items or information, though they might be irrelevant (Bordalo et al., 2013). By contrast, heuristics distinguish the different mental shortcuts we take. Affect heuristics, for example, can bring people to make decisions based on how they feel (even when considering numeric issues), instead of using rational reasoning. The list of observed biases and heuristics is still growing today and their relevance has been established even in fields once thought to be governed by rationality, such as stock market investing (see Thaler and Ganser, 2015 for a discussion of different biases and heuristics). Importantly, these behavioral mechanisms may be purposefully exploited to influence decision making in a desired direction. Nudges that do not affect incentive structures or the set of possible choices may still alter decisions by relying on these unconscious pathways. Decisions and choices have to be considered not only through their expected outcome, but also as part of their context.

Economists thus developed various methods that allow the analysis of decisions in their context from different perspectives. For example, randomized controlled trials can isolate the causal effect of an intervention, or treatment, on a decision. Surveys allow the exploration of attitudes which influence decisions. Empirical studies investigate actual behaviors in their real contexts. And theoretic models propose a formalization of behaviors. A given method may be relevant depending on the situation and on the research question at hand, as the upcoming chapters illustrate.

This thesis contributes to the broader literature on decision making with a collection of independent essays. Each essay raises different questions about decision making and uses different methods to answer them. Complex motives and biases that influence decisions are measured, and their consequences are analyzed through the chapters using a randomized control trial, a survey, and a theoretic model. Though far from exhaustive, this thesis should thus provide the reader with a wide-ranging, but rigorous, impression of the economics of decisions making. In a laboratory randomized controlled trial, chapter 1 investigates the effect of a risk-salience nudge on participants' decisions to innovate. It shows that decisions to innovate can be altered without changing incentive structures, but simply using a behavioral nudge. Chapter 2 studies the various motives that lead people to accept public policies aiming to decrease the consumption of certain food items. An online survey is used to collect participants' perceptions. It concludes that decisions to accept (and to vote for) public food policies are correlated with complex underlying motives related to perceptions of the policies and the issues they are meant to solve. Chapter 3 proposes a theoretic model of decisions to share (mis)information, taking into account asymmetries in sharing behaviors when sharing true or false information. It postulates that these asymmetries can lead a society to be misinformed.

Both chapter 1 and 2 rely on data that has been generated for the chapters presented in this thesis, respectively in the laboratory of the University of Strasbourg, and online with citizens of the United Kingdom using the Prolific platform. They share meticulous pre-planning, including clearance from independent ethics committees, as well as pre-registration of the hypothesis developed in the American Economic Association's registry for randomized controlled trials. For chapter 1 the rich literature linking innovation and risk taking as well as salience and risk perception allowed the development of precise, testable hypotheses. This led to the choice of running a randomized control trial to test said hypotheses. For chapter 2, however, the literature that investigates factors leading people to accept public food policies is both younger and scarcer. As a consequence, a survey was ran to shed light on these factors. In a distinct study, one could then test the causal effect of selected factors on policy acceptability to confirm that they are not only correlated with policy acceptability, but that they can also be leveraged to increase policy acceptability. Finally for chapter 3, the study of behaviors in social networks in experiments is extremely complex due to the sheer size of the participant pool needed. To accurately simulate a social network structure, with its various groups and heterogeneous relationships, one may need dozens of participants who would then generate only one observation per network. Additionally, studying the topic misinformation with real participants raises ethical concerns as even if they are told an information is false, it can still be remembered as true (Johnson and Seifert, 1994). For these reasons, chapter 3 relies on mathematical models to define behaviors, and numeric simulations to illustrate them. The model developed shares with both previous chapters the assumptions of behaviors that are not perfectly rational because of human limitation in technical capabilities and biases.

The abstracts below provide an overview of each chapter.

Chapter 1: Nudging innovation: the effect of risk salience

Decisions to innovate are a fuel of economic growth. In this chapter, we investigate a nudge designed to foster innovative behavior by changing risk perceptions. In the lab, participants run a virtual lemonade stand. They can either exploit a given strategy or innovate and explore new ones. Their business choices generate respective profits and the subjects receive a performance-based payoff, making innovating risky. To make this risk more salient, we introduce periodic reporting of profits and expect participants to be less innovative. To draw attention away from the risk, we introduce reporting of strategies and expect participants to be more innovative. We find these nudges to affect behavior through the channel of risk attitudes, as only risk-averse participants are affected. However, both treatments inhibited innovation compared to a control with no reporting. We argue this is due to both types of reporting inducing an increased evaluation of risks. We thus encourage further research of risk salience as a tool to foster innovation and recommend practitioners to tailor interventions and their evaluation with care before implementation, as they can backfire.

Chapter 2: The Acceptability of Food Policies

Our diets, increasingly rich in sugar, fats an animal products are at the center of an environmental and public health crisis. The implementation of public food policies is thus expected to increase in the coming decade, raising the question of citizens' choices to accept or reject them. In this chapter, we propose and test a model of food policy acceptability. The model is structured in four levels: government, topic, policy, and individual. We focus on two levels that are actionable for policy-makers: the topic and policy levels. We assess nine factors using a first online survey with 600 UK nationals and replicate our results in a second survey with 588 participants. Our results suggest that three factors have a positive effect on acceptability at the topic level: awareness of the issue, the legitimacy of state intervention, and social norms. At the policy level, we report a positive effect of the policy's expected effectiveness, its appropriate targeting of consumers, and the perceived support of the majority. On the other hand, more coercive interventions and those generating inequalities are judged to be less acceptable. Additionally, we report an interaction between awareness and coerciveness on acceptability. Participants who are aware of the issue were more likely to support coercive policies. We also find evidence for a trade-off between coerciveness, effectiveness, and acceptability, as more coercive measures are considered more effective, but less acceptable by participants. Our findings offer policy-makers, nutrition experts, and advocates for healthier and more sustainable diets a new and integrated understanding of the underlying factors that determine food policy acceptability.

Chapter 3: Misinformation due to asymmetric information sharing

Information is a core component of optimal decision making. Misinformation, which is heavily shared in social networks, can induce people and societies to reach undesirable outcomes. In this chapter, we introduce a model of social learning in which agents share true and false information with different decays and to different networks of people. Our results establish that these asymmetries, thus far largely ignored in theory despite being empirically established, govern the long-run beliefs of a society. We derive a single threshold condition under which misinformation prevails. Misinformation is more likely when false information decays less, and when the false information network is locally denser, as measured by the largest eigenvalue of its adjacency matrix. Under these conditions, all agents guess the wrong state; a result that the literature otherwise only reaches in the presence of forceful (or stubborn) agents. Additionally, we measure speed of convergence and show that agents who are more central in the false information network are more prone to be misinformed. We illustrate our results using numerical simulations that incorporate societies segmented into groups, and derive policy implications that center on human information sharing behavior.

Note to the reader

The three chapters of this thesis all aim to contribute to the broader literature on decision making. However, they are stand-alone essays that can be read separately.

Chapter 1

Nudging innovation: the effect of risk salience[1]

1.1 Introduction

Innovation is required in many contemporary jobs and is a driver of firm performance and even survival (Shalley et al., 2000; Unsworth, 2001). Stimulating innovation is thus one of the highest concerns of CEO's (Rudis, 2004).[2] Innovation is inherently a risky endeavour. The literature documents an association between risk attitudes and innovative behavior (Lazear, 2005; Hall and Woodward, 2010; Hvide and Panos, 2014; Koudstaal et al., 2016; Kerr et al., 2017; Hudja and Woods, 2019). Hence, exploiting the possibility to manipulate the *risk* or *risk perceptions* is a promising starting point to tailor interventions designed to promote innovation. However, the literature has so far focused on mechanisms that manipulate the underlying *risk*. In a laboratory experiment, Ederer and Manso (2013) document innovation-inducing effects of contracts that tolerate early failure while rewarding late-stage performance. But changing the incentive schemes within a firm can both be costly and organizationally demanding.

We investigate whether innovation can be induced by channeling *risk perceptions*. Specifically, we ask whether innovative behavior can be nudged by altering the salience of risk, without changing monetary incentives. Adapting the design from Ederer and Manso

[1] This chapter is based on the homonymous working paper written in collaboration with Christian Zihlmann and Elisa Matthewes. The experiment was pre-registered through the AEA RCT registry (ID: AEARCTR-0005214), and approved by the ethics committee of the University of Fribourg. The authors gratefully acknowledge financial support from the Swiss National Bank fund of the Department of Economics of the University of Fribourg.

[2] To understand which kind of interventions can effectively promote innovative behavior, various determinants of innovation have already been identified and studied. A rich literature investigates, for example, the relationships between innovation and different types of leadership (Shalley and Gilson, 2004; Hughes et al., 2018), personality traits (Amabile, 1996) or incentives (Kohn, 1993; Amabile, 1996). Incentives can improve performance in innovative tasks, depending on the incentive scheme (Ederer and Manso, 2013), the need of hierarchy in team structure (Englmaier et al., 2018), or the type of creative task at hand (Charness and Grieco, 2014).

(2013), we conduct a controlled laboratory experiment: Participants run a virtual lemonade stand over 20 periods. The profit of the virtual stand determines the compensation paid to the participants. Thus, they are incentivized to maximize profits. Since the subjects do not know the profits associated with each of the available choices, they face a trade-off between exploration and exploitation – a core feature of the innovative process (March, 1991): Subjects can either fine-tune well-known strategies (exploitation) or explore untested strategies (exploration) that are thus associated with more risk.

To make the risk of innovating either more or less salient, we introduce two behavioral interventions that are designed to foster innovative behavior. Specifically, we implement a reporting mechanism that happens every three periods. In the profit treatment, participants report their profits for each of the past three periods. The periodical reporting of the profits is ought to increase the salience of the risk. In the strategy treatment, participants report their strategy for the past three periods. This intervention is designed to focus participants on the strategic choice variables and, thereby, intends to make profits, and with it the risk of reduced profits, less salient. In the control group, participants do not report. Hence, we hypothesize that participants in the strategy treatment engage in more innovative behavior while those in the profit treatment are the least innovative. We also hypothesize that the effect of our intervention will be more pronounced among the risk-averse participants.

We find that our interventions affect innovative behaviors through the channel of risk attitudes, as hypothesized: Only risk-averse participants display different patterns of innovation conditional on the treatment. Also in line with our hypothesis, we report that risk-averse participants are more attentive to the profit and explore less when assigned to the profit treatment than when assigned to the control group. However, contrarily to our expectations, the strategy intervention backfired and was detrimental to innovation. Subjects in the strategy treatment do not provide more attention to the strategic variables and participants in this treatment also engage in less exploration than control participants, for a given level of effort. Although we cannot completely exclude that reputational concerns or limited attention matter in our decision environment, these explanatory approaches are unable to explain some key results: First, reputational concerns do not obviously explain why only the more risk-averse participants react to the treatments. Also, maintaining a reputation in our anonymous setting is costly. Hence, we consider it unlikely to be the driving mechanism. Further, we rule out the conjecture that subjects reporting their strategy shift time and effort from a more productive (e.g. information-processing) to a rather unproductive domain, that is writing reports without an instrumental value: We find that the less risk-averse subjects actually decreased their effort and that the more risk-averse, strategy-reporting types innovate less compared to the control group when controlling for the effort level. Thus, we argue that the backfiring intervention is likely a consequence of both reporting treatments inducing an increased perception of risk compared to no reporting.

Our study extends on three main strands of existing research. First, we augment the evidence about fostering innovation through the channel of risk attitudes. Some authors have investigated which incentive structures can successfully induce exploratory behavior. These studies manipulate the underlying *risk* that subjects are facing. Closest to our study, Ederer and Manso (2013) find that incentive schemes that tolerate early failure can induce innovative behavior. Moreover, the authors report that this effect is mainly driven by risk-averse participants.[3] This confirms theoretical predictions by Manso (2011) who shows that the optimal incentive scheme that motivates innovation will tolerate or even reward early failure. This is because innovation involves the exploration of untested approaches that are likely to fail, and thus, standard pay-for-performance schemes that punish failure will be detrimental to innovation. Multiple other studies investigate how different incentive schemes affect creativity. Overall, they find only small differences between standard incentivizations like flat fees, linear payments or tournament incentives (Erat and Gneezy, 2015; Charness and Grieco, 2018).[4] However, Knight et al. (2003) show that also non-monetary interventions can reduce risk and foster innovative behavior. They suggest that education leads to more innovation adoption from farmers because it provides them with new skills and reduces uncertainty. We do not manipulate the riskiness of the task. Instead, we design a mechanism that alters the *perception of risk*. We thus propose and investigate a novel approach to stimulate innovative behavior. In line with the previous literature, we find only risk-averse subjects reacting to the induced treatments.

Second, we add to the literature on changing *risk perceptions*. While risk attitudes are commonly considered as stable personality traits and thus as more challenging to manipulate, influencing risk perceptions is suggested to be more effective in changing risky choice behavior (Sitkin and Weingart, 1995; Weber and Milliman, 1997; Pennings and Wansink, 2004). Risk perceptions are influenced by various factors, e.g. stress (Sobkow et al., 2016), graphical visualizations or framing (Lévy-Garboua et al., 2012; Tombu and Mandel, 2015). This can be exploited to (re-)direct behavior (Eppler and Aeschimann, 2009).[5] One wide field for applications is health behavior. The efficacy of various nudges targeting risk perceptions has been extensively investigated in this context (Gerrard et al., 1999; Ferrer and Klein, 2015): For example, Myers (2014) finds that media health messages can change smokers risk perceptions or Banerjee et al. (2021) show that the accuracy of risk perception can be increased by communicating infection risks via raw numbers. Further, how risks are presented is shown to affect risk perceptions in investment settings (Diacon and Hasseldine, 2007; Cohn et al., 2015; Holzmeister et al., 2020). Overall, nudges seem to be effective in influencing risk perceptions, however, this has not yet been exploited to increase innovative behavior. Hence, we contribute to

[3] Such schemes include an exploration contract that is tolerant of early failure but rewards late-stage performance or golden parachutes, that guarantee a certain level of compensation if participants fail in their innovation.

[4] Note that creativity is not an element of the exploitation-exploration trade-off that we study. However, as it shares common aspects, and namely risk-taking, we consider these findings relevant. For perspective on the difference between innovation and creativity, see Hughes et al. (2018).

[5] These approaches affect risk perceptions by either influencing how much people understand the risk (cognitive dimension) or how they feel about the risk (emotional dimension) (Loewenstein et al., 2001).

this interdisciplinary field of literature by investigating a new application of a nudge targeting risk perceptions.

Third, we provide a novel perspective on the use of salience as a tool to alter innovative behavior. Considering that people are limited in the attention that they can allocate to information when making decisions (DellaVigna, 2009), varying the salience of certain information affects the decision-making process. Köszegi and Szeidl (2013) and Bordalo et al. (2013) suggest that agents overweight certain parameters of a decision, simply because they are more salient. Salience nudges have thus been shown effective to influence decision-making in various and heterogeneous contexts such as looking before crossing the road, choosing meal options, or selecting the right pod size when farming sea weed (Thaler and Sunstein, 2009; Hanna et al., 2014; Kurz, 2018). However, these contexts have in common that they have a well-defined goal that is known ex-ante: making pedestrians look right before crossing the street in London, influencing students to chose more vegetarian options, or leading farmers to the optimal pod size. The innovation process in firms, however, implies uncertainty. It follows that it is impossible to make the ex-ante unknown optimal outcome of the innovation process salient. We thus do not nudge our participants to the optimal strategy, but rather use the nudge to alter risk perceptions and encourage innovation. This is novel and closer to what can effectively be implemented in the context of organizations aiming to promote innovative behavior.

Our results confirm the importance of accounting for risk perceptions for practitioners and scholars aiming to promote innovation. We encourage further investigation of salience as a tool for affecting innovative behavior, while warning about a possible backlash effect. This suggests a prudent investigation of behavioral nudges, ideally in controlled environments, before expanding their implementation.

The remainder of this chapter is structured as follows. Section 1.2 presents the experimental design and introduces our hypotheses. Section 1.3 reports our results. Section 1.4 provides a discussion and section 1.5 concludes.

1.2 Experimental design

The experimental task was adapted from Ederer and Manso (2013). Subjects had to solve a task in which they were facing a trade-off between exploration and exploitation: Participants managed a virtual lemonade stand. Over 20 experimental periods, participants had to make decisions on multiple parameters such as the recipe of the lemonade (sugar and lemonade content, color), the location of the lemonade stand and the price of a cup of lemonade. The possible combinations of these choice variables amounts to 23’522’994 combinations. All participants were compensated according to their realized profits such that their aim was to maximize the profit of the fictional lemonade stand, and, thereby, their own earnings. The payoff was determined by a standard pay-for-performance scheme: Participants were paid 50% of the profits they generated during all 20 periods. Participants faced uncertainty since they did not know the profits associated with each of the available choices. However, they did receive a default strategy, i.e. the choices and the associated profit of an imaginary previous manager. The default strategy was not the most profitable strategy.

After each period, participants were informed about the profit of their implemented choices. Additionally, they received a brief customer feedback. The feedback was an informative binary feedback: The computer randomly selected one of the three continuous choice variables (price, lemon or sugar content) and, then, provided a truthful feedback if the value of the selected variable is too high or too low compared to the optimal value (e.g. "Your lemonade is too sweet"). Subjects received no feedback regarding the location and color of the lemonade. Consequently, the feedback was only informative conditional on the chosen location and lemonade color.

The task is characterized by an exploration-exploitation trade-off: subjects can choose to either fine-tune the default strategy – yielding a profit similar to the previous manager (exploitation) – or experiment with new strategies, taking the associated risk of failure but also the chance of success (exploration). The parameters are designed in a way that exploration will increase chances to identify the strategy that leads to the global maximum while exploitation rather leads to local maxima. The parameters to calculate the profits of the lemonade stand in our experiment are one-to-one adapted from Ederer and Manso (2013).[6]

1.2.1 Treatments

To address our research question, whether innovative behavior can be nudged by a salience intervention, we integrated a reporting stage into the original game: In periods 3, 6, 9 and 12, subjects were requested to submit a report. The focus of these reports was exogenously varied, inducing an attention shift through making a specific aspect of the game salient, and consequently, another aspect of the game less salient. We used a between-subject design with three different groups: First, the control group in which subjects were not asked to report. Second, participants in the profit treatment were requested to report their profits. Third, the strategy treatment requested subjects to report their strategy. Note that the incentive structure between all treatments was identical: The reports were not payment-relevant, i.e. one's reports were not shared with another subject or with the experimenter while playing the game. This was common knowledge for the participants. Hence, the treatment groups only varied in the content of the reports. Controlling for incentive effects ensures that any differences between the treatment groups are causally based on the interventions.

Control Group: Subjects assigned to the control group were not requested to submit a report.

Profit Treatment: After the periods 3, 6, 9 and 12, subjects were requested to report the profits they made within the last three periods. Along with the wording *"Please report your profits of the last three periods."*, subjects faced an entry mask, where they

[6] A detailed description of the parametrization can be found in appendix 1.B.2.

needed to enter the profits of each of the last three periods. The timing and description of the required reports was communicated by the instructions before the start of the business game.

Strategy Treatment: After the periods 3, 6, 9 and 12, subjects were requested to report the strategy they followed within the last three periods. Along with the wording *"Please describe your strategy in the last three periods. Why did you choose this strategy?"*, they faced a free form text field. The timing and description of the required reports was thoroughly communicated by the instructions before the start of the business game.

The treatments served the purpose of shifting attention by making a specific aspect of the game more salient, the strategy or the profit. By manipulating the salient part of the task, we expected to influence the participants judgement and decisions.

1.2.2 Pre-registered hypotheses

Ederer and Manso (2013) document that more risk-averse individuals exhibit significantly less innovative behavior in exploration-exploitation tasks than less risk-averse agents. They promote the use of risk-tolerant incentive schemes to increase explorative behavior. While Ederer and Manso (2013) vary the underlying *risk* of the task by using different incentive schemes, we intended to induce more (or less) risk-tolerance by manipulating the *risk perceptions* of the subjects. Based on their findings, we expected to observe more exploratory behavior for agents with lower risk perceptions and the opposite for agents with higher risk perceptions.

Theoretical and experimental literature on salience suggests that shifting attention away from or towards a specific feature, can smoothly guide decision-making processes (e.g. Thaler and Sunstein, 2009; Bordalo et al., 2013). Besides, various studies indicate the effectiveness of framing and nudging in directing risk perceptions (e.g. Holzmeister et al., 2020; Banerjee et al., 2021). Based on this evidence, we expected to influence participants' risk perceptions, and consequent exploratory behaviors, by manipulating the salient part of a task through the content of the reports.

For subjects assigned to the strategy treatment, the salient aspect of the report is the strategy. We thus expected to shift their attention away from the profit. Because decreasing profits is the risk that can inhibit participants' exploration, focusing them away from it is expected to decrease risk perceptions. These participants were thus expected to explore more.

Hypothesis 1a. *Subjects in the strategy treatment explore more than subjects in the control group.*

Conversely, being required to regularly report your previous profits, makes the monetary aspect of the game more salient – and with it the risk to decrease profits. Participants in the profit treatment were thus expected to have an increased perception of risk and to explore less.

Hypothesis 1b. *Subjects in the profit treatment explore less than subjects in the control group.*

The main hypothesis is that our treatment affects behavior and results in higher (lower) exploration behavior for the strategy (profit) treatment. We subsequently investigated the mechanism driving this effect. We posit that a shift of attention to the salient aspect (the strategy, respectively the profit) takes place while the control group serves as a baseline.

Hypothesis 2a. *Subjects in the strategy treatment provide more attention to the strategy than the control group.*

Hypothesis 2b. *Subjects in the profit treatment provide more attention to the profit than the control group.*

We expect that the effect of risk salience on exploratory behavior works through the channel of risk attitudes. An intervention changing the perceived risk should affect risk-averse individuals more strongly than risk-neutral participants. The more risk-averse a participant, the more her innovative behavior should be inhibited by an increased perception of risk. We thus hypothesize heterogeneity in the treatment effect and that the treatment affects specifically risk-averse participants. This line of reasoning goes in the same direction as Ederer and Manso (2013) who find making the incentive structure more risk-tolerant leads to a behavioral change particularly among the risk-averse subjects.

Hypothesis 3a. *Risk-averse subjects assigned to the profit treatment explore less than risk-averse subjects in the control group.*

Hypothesis 3b. *Risk-averse subjects assigned to the strategy treatment explore more than risk-averse subjects in the control group.*

1.2.3 Sampling

We employed a sequential analysis plan, following the method outlined in Lakens (2014) based on the key outcome variable, the profit obtained in the final period.[7] This plan specified that our data collection will be terminated at $t = 0.5$ – that is half of the required sample size according to our power calculation – when the observed effect size is smaller than our smallest effect size of interest (SESOI), which was set at Cohen's $d = 0.3875$. Since our main effect sizes are lower than the pre-determined SESOI ($d < 0.3875$), we followed the scenario described in our sequential analysis plan and terminated the data collection after collecting 90 observations. The respective results will be described in the following section. A description of the sequential analysis approach and our underlying power analysis can be found in the appendix 1.D. If our expected

[7] In the final period, profit-maximizing subjects should stick to the most profitable strategy that they discovered during all previous periods. Hence, the profit in the final period constitutes a suitable proxy for exploratory behavior in this experiment (see Ederer and Manso, 2013).

effect size of $d = 0.5$ came into force, we would require 180 observations to reach statistical significance at the 5%-level (with 80% power). Because the actual effect size is lower than the hypothesized effect size of $d = 0.5$ and since we stopped data collection by adhering to the pre-registered procedure at $t = 0.5$ (90 observations), statistical significance is not attainable.

1.2.4 Procedures

We collected experimental data in January 2020 at the laboratory of the University of Strasbourg (Laboratoire d'Économie Expérimentale de Strasbourg LEES). The experiment was programmed with oTree (Chen et al., 2016) and conducted in French. Each of our four sessions lasted approximately 60 minutes. We used experimental currency units called Thalers with an exchange rate of 1:100. All subjects in the laboratory received a fixed show-up fee of 2€, and in addition a performance-based variable payoff. Overall, the average payoff was 15€. Subjects were randomly assigned to the treatment and control groups, constituting the exogenous variation in this study. The random assignment was performed within-session in order to mitigate potential session-specific effects. At the end of each session, we elicited demographics and risk preferences of the subjects (Falk et al., 2018). The sample consists of 90 subjects, i.e. 30 subjects for each treatment group.

1.3 Results

In the following, we present the results of our study. Each subject is treated as one observation and all standard errors are clustered at the individual level.

1.3.1 Performance and explorative behavior

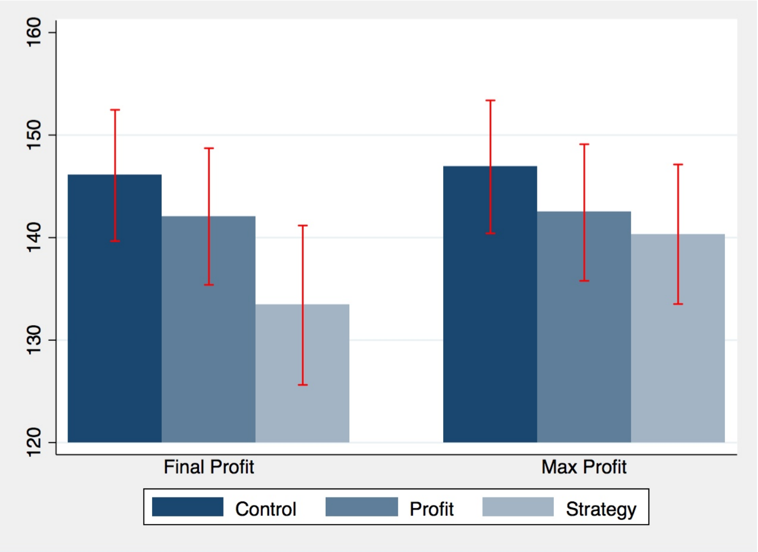
First, we focus on the effect of our treatments on exploration. We collected several outcome variables that serve as proxies for exploratory behavior, see also Ederer and Manso (2013) for a thorough discussion. In the following, we will focus on i) the profit realized in the final period and ii) the maximum profit over all periods. Other proxy variables are analyzed in the appendix 1.A.1. The results remain qualitatively unchanged.

Figure 1.1 compares the means of the final and the maximum profit between the treatments. As hypothesized, we find empirical evidence that reporting the profits decreases profits and, with it, exploratory behavior. However, the effect size is small ($d = 0.11$) and lower than the smallest effect size of interest.[8] Thus, this effect is statistically not significant. Further, in contrast to our hypothesis, the data suggest that subjects assigned to the strategy treatment do not realize higher profits than subjects assigned to the control group: the control group earns on average the highest final period profit (146 thalers, profit treatment: 142 thalers, strategy treatment: 133 thalers) and maximum

[8] For this effect to be statistically significant, we would need to collect 1250 observations each in the control group and the profit treatment.

profit of all periods (146 thalers, profit treatment: 142 thalers, strategy treatment: 140 thalers). Thus, against our hypothesis, reporting the strategy does not increase, but decrease profits. The effect size is $d = 0.32$[9] and thus larger than in the profit treatment, but again lower than the SESOI threshold and consequently, not significant.[10]

Figure 1.1: Means of the final and the maximal profits.



Notes: The figure reports the means of the final and the maximum profit for each of the treatments. Error bars indicate standard errors of the mean.

With the caveat of statistical insignificance due to small effect sizes, we conclude that reporting either the profits or the strategy decreased explorative behavior: we report the expected negative effect for the profit treatment and, in stark contrast to our hypothesis, also find a negative effect for the strategy treatment.

Result 1a. *Subjects in the strategy treatment seem to explore less than subjects in the control group.*

Result 1b. *Subjects in the profit treatment seem to explore less than subjects in the control group.*

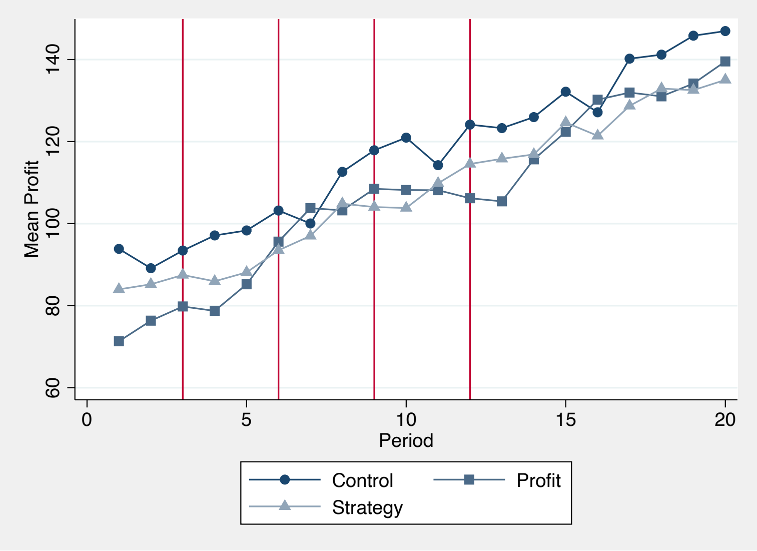
We observe that the mean overall profit is the highest in the control group (2323 thalers), followed by the strategy treatment (2151 thalers) and the profit treatment (2143 thalers).

[9] For this effect to be statistically significant, we would need to collect 150 observations each in the control group and the strategy treatment.

[10] Two-sided t-tests: p-values for the final profit are 0.6661 (control and profit treatment), 0.2140 (control and strategy treatment) and 0.4017 (profit and strategy treatment). For the maximum profit p-values are 0.6334, 0.4869 and 0.8246. Similarly, Mann-Whitney U-test, two-sided: p-values of 0.7394, 0.2804 and 0.4964 for final profit; p-values of 0.6843, 0.5249 and 0.8418 for maximum profit.

Figure 1.2 shows the evolution of the mean profit in all three treatment groups. The red bars represent periods that precede a reporting screen for the treatment groups. Note that the average profit in the control group is higher than in the reporting treatments in nearly all periods and already before subjects needed to report the first time in period 3. This indicates that subjects in the reporting treatments are already exhibiting a treatment effect in anticipation of future reportings. Indeed, subjects in the profit treatment significantly earn less profit in Period 1 ($p < .01$) than the control group. For the strategy treatment, the difference does not reach statistical significance. We will discuss the implications of a treatment effect already in the first period further in Section 1.4.

Figure 1.2: Evolution of profits over time.

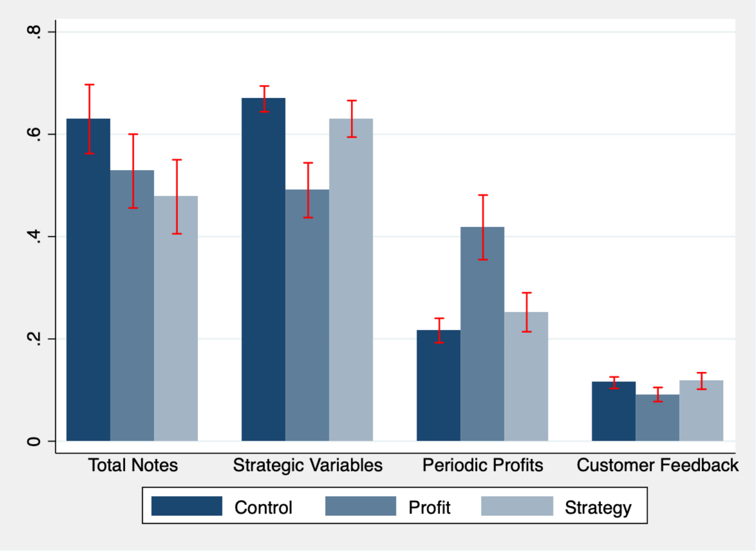


1.3.2 Attention

According to our hypothesis, the treatment affects behavior because it shifts attention: In the profit treatment, the salience of the profits is increased while in the strategy treatment it is the salience of the strategic choice variables. We construct a measure based on a notes sheet that subjects could fill out voluntarily. We compute the proportion of filled out fields for each subject and, then, derive how many notes each subject took on the strategic variable choices, the periodic profit and the customer feedback relative to the number of total notes taken.

Figure 1.3 compares the percentage of notes that are taken and illustrates what the subjects across the treatments mainly focused on. It becomes evident that the profit treatment subjects provide less attention to the strategic variables compared to the control group (t-test: $p < 0.01$), but more to the period profits (t-test: $p < 0.01$). Both findings are in line with our hypothesis: Through the reporting mechanism, we successfully shifted the attention of participants in the profit treatment towards the profit. For participants in the strategy treatment, we expect the opposite pattern, namely a shift away from the profit towards the strategy. Contrary to our hypothesis, we observe the same pattern also for the strategy treatment. Yet, the effects do not reach statistical significance at conventional levels.

Figure 1.3: Note-taking behavior.



Notes: The figure reports the proportion of total notes taken and, further, the means of subjects' notes of strategic decision variables, the periodic profits and feedback relative to their total notes taken for each of the treatments. Error bars indicate standard errors of the mean.

Result 2a. *Subjects in the profit treatment provide significantly more attention to the profit than the control group.*

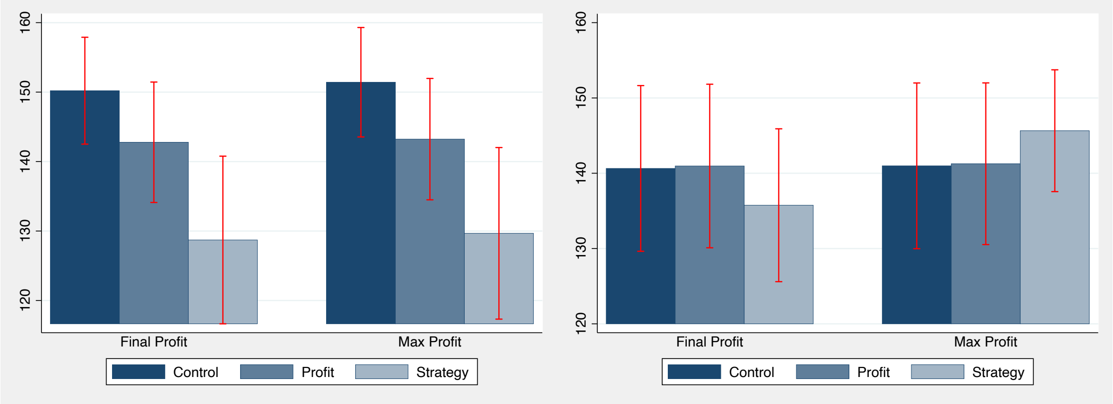
Result 2b. *Subjects in the strategy treatment do not provide more attention to the strategy than subjects in the control group.*

1.3.3 Heterogeneity

To test hypotheses 3a and 3b, we elicited the subjects' risk preferences based on the staircase elicitation method by Falk et al. (2018). We split our sample of participants at the median level of risk aversion and classify the participants into less and more risk-averse types.[11] Then, we compare the treatment groups to the control group separately for each type with respect to the final profit and to the maximum profit of all periods.[12]

Interestingly, the results look very different between more and less risk-averse subjects: Figure 1.4a shows that our treatments do have a negative effect on more risk-averse subjects. That is, requiring more risk-averse subjects to report *reduces* their performance. The effect is most pronounced for the strategy treatment: When more risk-averse subjects are requested to report their strategy, they reduce their profit by nearly 22 thalers, compared to the control group. This is a decrease of approximately 15%, resulting in a relatively large effect size (Cohen's $d = 0.62$).[13] This effect also holds for the profit treatment, however, the effect as well as its statistical significance are less pronounced.

Figure 1.4: Means of the final and the maximal profits by risk preferences.



(a) More risk-averse subjects

(b) Less risk-averse subjects

Notes: The figure reports the means of the final and the maximum profit for each of the treatments, for more and less risk-averse subjects separately. Error bars indicate standard errors of the mean.

Figure 1.4b depicts that the treatment does not lead to a measurable change of behavior for the less risk-averse subjects: They all perform equally well, regardless of the assigned group. The average treatment effect thus seems to be driven by the more risk-averse subjects: only these participants react to the treatment. This is in line with our pre-specified hypothesis that the treatment effect goes through the channel of risk aversion. The detrimental effect for the profit treatment is also in line with our hypothesis, while

[11] Median risk aversion, on an index from -3 to +3, amounts to -.03 in our sample, with a mean of .02. Thus, participants in our sample are at median very close to risk-neutrality.

[12] We elicited risk aversion after the experimental game, i.e. the lemonade stand task. Thus, the treatment conditions could potentially have impacted risk preferences. Nevertheless, we do not observe significant differences for risk-preferences among the three groups.

[13] Which results in low p-values conditional on our sample size: t-test: p=0.11, U-test: p=0.14)

Table 1.1: Heterogeneous treatment effect.

	Final Profit		*Maximal Profit*		*Final Location*		*Exploration Phases*	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Profit	-2.967	0.539	-3.290	-0.570	-0.275	-0.153	-0.055	-0.023
	(9.94)	(9.61)	(9.37)	(9.20)	(0.38)	(0.39)	(0.10)	(0.10)
Strategy	-13.075	-8.708	-7.592	-4.203	-0.506	-0.370	0.010	0.048
	(10.03)	(9.74)	(9.46)	(9.33)	(0.38)	(0.39)	(0.10)	(0.10)
Risk Aversion	10.775	9.521	11.666	10.694	0.796**	0.736*	0.321***	0.320***
	(9.03)	(8.67)	(8.51)	(8.30)	(0.39)	(0.38)	(0.09)	(0.09)
Profit× Risk Aversion	-6.979	-11.193	-7.961	-11.231	-0.775*	-0.878**	-0.338***	-0.383***
	(11.14)	(10.78)	(10.49)	(10.32)	(0.45)	(0.44)	(0.11)	(0.11)
Strategy× Risk Aversion	-22.613*	-22.138*	-28.083**	-27.715**	-1.113**	-1.118**	-0.400***	-0.404***
	(13.30)	(12.76)	(12.54)	(12.21)	(0.52)	(0.51)	(0.13)	(0.13)
Effort (Notes taken)		30.641***		23.771**		0.937**		0.258**
		(10.60)		(10.15)		(0.42)		(0.10)
Constant	144.634***	125.503***	145.356***	130.515***	0.896***	0.341	2.072***	1.905***
	(7.06)	(9.47)	(6.66)	(9.07)	(0.29)	(0.38)	(0.07)	(0.10)
N	90	90	90	90	90	90	90	90
R^2	0.056	0.142	0.066	0.124				
Adjusted R^2	-0.000	0.080	0.011	0.061				

Notes: The dependent variables are different proxies for exploration: the final profit (OLS regressions), the maximal profit (OLS regressions), the location chosen by the subject in the final round (Probit regressions) and the longest duration of an exploration phase (Poisson regressions). Risk Aversion is the subject-specific degree of risk aversion. Effort (Notes taken) proxies effort through the total number of filled out fields on the notes sheet. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

for the strategy treatment, we find the opposite effect of what we expected: risk-averse subjects decrease their exploratory behavior if they are required to report their strategy.

These findings are corroborated by regression analysis, employing the raw continuous measurement of risk preferences instead of a median split. Table 1.1 shows that both treatments interacted with risk-aversion yielding a significantly negative coefficient. Thus, compared to subjects in the control group, participants in the treatment groups reduce exploration with higher risk-aversion. The effect is particularly pronounced in the strategy treatment with a significance level below 5%. In short, the higher the risk-aversion, the stronger is the negative reaction to the treatment.

Result 3a. *The more risk-averse participants assigned to the profit treatment explore less than the more risk-averse subjects in the control group.*

Result 3b. *The more risk-averse participants assigned to the strategy treatment explore less than the more risk-averse subjects in the control group.*

Result 3c. *No treatment effect is observed on the less risk-averse participants who perform equally well regardless of the assigned group.*

1.4 Discussion

Our experimental findings at the interim analysis are only partly in line with the preregistered hypotheses. As expected, subjects in the profit treatment explore less than subjects in the control group.[14] Further, as hypothesized, we observe a prevalence of the treatment effect on more risk-averse subjects. However, in contrast to our hypotheses, subjects in the strategy treatment do not explore more but *less* than subjects in the control group who face no obligation to report. In this section, we elaborate and discuss potential explanations for this unexpected result. The potential channels are sorted by plausibility in an ascending order.

Hawthorne effect. The Hawthorne effect refers to a change of behavior among the participants of a study in response of their awareness of being observed. What speaks against this explanation is that only risk-averse subjects reacted to our intervention. Also, recent research that employs sophisticated empirical techniques casts serious doubt about the very existence of the Hawthorne effect (Levitt and List, 2011).

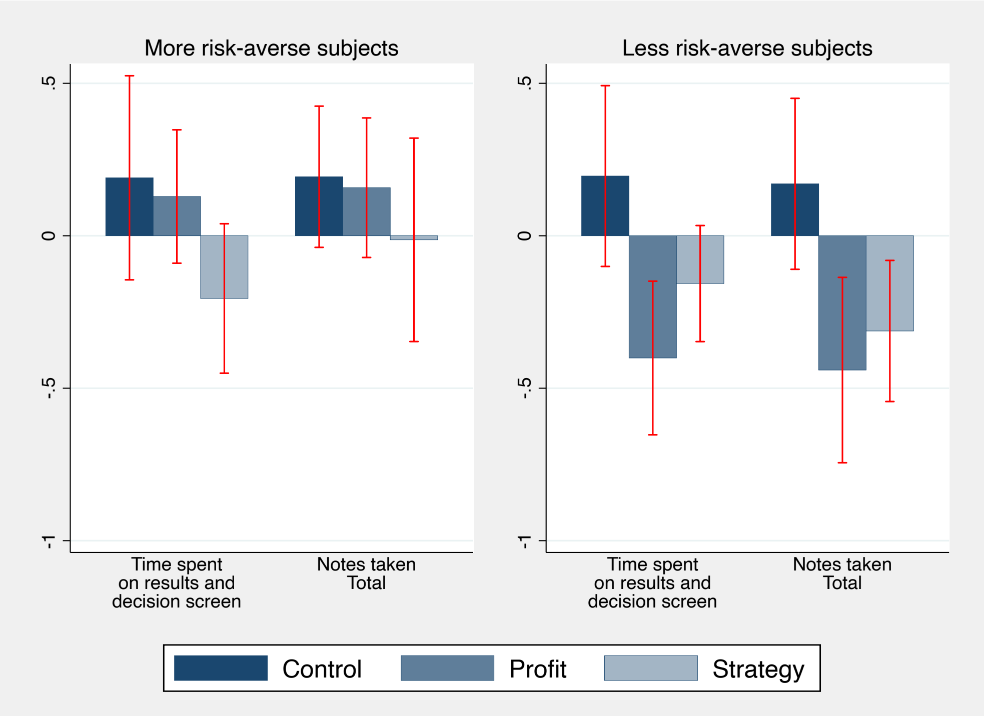
Reputational concerns. Suppose subjects feel reputational concerns when they need to report the strategies or the profits multiple times. Such reputational concerns may arise internally (from themselves) or externally (from the experimenter). The literature hypothesizes that such concerns lead to conservative behavior, i.e. less innovative and more risk-averse decisions (Scharfstein and Stein, 1990; Prendergast and Stole, 1996; Holmström, 1999; Suurmond et al., 2004). As a result, participants may experience disutility of acting inconsistently: They do not change their strategy too often but consistently apply slight variations of the initial strategy, which results in less exploration. Yet, multiple arguments speak against this explanation: First, it is unclear why in particular risk-averse subjects would experience such reputational concerns. But we observe the treatment effect to happen among risk-averse subjects. We also find that subjects in the strategy treatment are not significantly more risk-averse than subjects in the control group, suggesting that the interventions did not alter the risk preferences of our participants. Second, as Figure 1.2 depicts, the treatment effect can already be observed in the first period, before subjects even needed to report once. Third, through adhering consistently to a strategy in order to maintain their reputation, subjects will forgo potential profit. Thus, while we cannot fully rule out reputational concerns, we note that if it is the driving force, participants in our experiment would take on substantial costs for maintaining their reputation.

Limited attention. In addition to taking less notes, we also observe that participants in the strategy treatment spend less time on their decision-making and on analyzing their performance (see Figure 1.A.6 in the appendix). At first glance, this positive correlation between effort and exploration suggests that subjects in our strategy treatment might have simply exerted less effort for productive subtasks (such as note-taking, information-processing and decision-making), resulting in a hampered performance. This seems to be a reasonable argument since requiring participants to report their strategy is clearly effort-consuming. If subjects have a limited attention span and can only exert a certain fixed level of effort, then participants in the strategy treatment need to trade off time

[14] As mentioned previously, the effect is lower than expected and below the SESOI. The reason for the low effect size is that only some subjects react to the treatment induction, namely the rather risk-averse types.

and effort that they would otherwise invest in the decision-making process for writing the reports. While recent literature documents that people exhibit a limited span of effort (Gabaix, 2019), our data do not support this line of reasoning: we find that only the more risk-averse subjects react to the treatment and explore less, however, the less risk-averse participants are the ones that decrease their effort (see Figure 1.5). Moreover, if limited effort was a reasonable explanation, the exploration level should be the same across treatment groups for a given level of effort. Thus, the production function of explorative behavior – with effort as the input and exploration as the output – should be identical for all subjects, regardless of the treatment condition. Yet, since we observe the more risk-averse type reacting to the treatment but only the less risk-averse type exerting less effort, the production function of the more risk-averse type in the strategy treatment must be impaired. Plotting the production function for more and less risk-averse subjects indeed shows that the production function for the more risk-averse type is altered (see Figure 1.A.7 in the appendix): We find that conditional on the effort level, strategy-reporting participants initiate less exploration phases and yield lower profits. Regression analyses underline this finding: In Table 1.1 we investigate if the heterogeneous treatment effect persists when we control for the effort level. The interaction term is significant at the 5%-level and remains significant at the same level when controlling for effort. The coefficient of the interaction term remains stable. Hence, the lower exploration level observed in the strategy treatment is not associated with a reduction of effort, but with an impairment of the production function.

Figure 1.5: Effort exerted by type.



Notes: This figure displays the measured effort through i) time spent on the results and decision screen ii) the total of filled out fields in the notes table sheet. The variables are standardized. Error bars are displayed in red.

Backfiring nudge: Shifting attention *towards* risk. Lastly, the underlying behavioral mechanism might be that our intervention backfired. In the profit treatment, we arguably shifted attention to the profit and with it, to the risky aspect of the game. Consequently, as hypothesized, subjects explore less. In the strategy treatment, we intended to shift attention to the strategic variables, making risk less salient. However, the strategic choice variables come along with and may be non-separable from the profit. Thus, requiring subjects to report their strategy may have focused them on the profit, too: Because subjects in the strategy treatment reported their choices (and with it, their behavior) in an open form field, the risk of losing money may have been more salient than in the control group. Hence, we might have unintentionally nudged subjects in the strategy treatment to focus on the risky aspect of the business game.[15] If risk becomes more

[15] As Figure 1.2 depicts, the treatment effect is already visible before the first reporting took place: From period 1 on, the control group is almost first-order stochastic dominating the strategy and profit treatment. The mere awareness of needing to report may have been enough to increase the salience of risk.

salient, the performance of more risk-averse types may be inhibited.[16] Consequently, i) their production function should be impaired and ii) the treatment effect should be more pronounced among risk-averse participants. We find evidence for both these conjectures. Consequently, we deem this channel to be the most plausible.

1.5 Conclusion

Identifying and evaluating measures that effectively foster innovative behavior is highly relevant for leading decision-makers in organizations (Rudis, 2004). We experimentally investigate whether salience can serve as a tool to nudge exploratory behavior. First, we find that making risk more salient reduces exploration, as hypothesized. Second, contrary to our hypothesis, we observe the same negative effect on exploration for subjects for whom risk was arguably less salient. Therefore, the data reveal that our nudge backfires. We discuss potential reasons and find it most probable that our intervention, aimed at making the risk less salient, likely makes it more salient. Third, we find the effects of the salience nudge to be particularly pronounced among the more risk-averse participants. Subjects that are rather risk tolerant do not react to the intervention. This finding aligns with Ederer and Manso (2013). Thus, we demonstrate that purely behavioral interventions that only change the salience of risk, but do not alter monetary incentives, *are* able to alter explorative behavior.

We apply a rigorous methodological approach by following a pre-registered sequential analysis plan (Lakens, 2014). Since the effects for both groups are lower than the smallest effect size of interest defined in the sequential analysis plan, we adhered to our pre-determined procedure and stopped the data collection after the interim analysis. Consequently, our sample size is not large enough to make reliable inferences given the low effect sizes we observe and our results should be interpreted with prudence.

Still, our results have important implications for practitioners and researchers alike. First, we show that purely behavioral nudges that do not change incentives *are* able to affect innovative behavior by guiding risk perceptions. Second, even nudges that are carefully derived from the existing literature can turn out to be ineffective and may even backfire. Hence, interventions need to be carefully tailored and require meticulous evaluation before implementation, ideally through experimental methods (Banerjee et al., 2017, Sunstein, 2017). This is especially important since nudges have become increasingly popular due to their simplicity and alleged effectiveness (Reisch and Sunstein, 2016, Benartzi et al., 2017, Sunstein et al., 2019). However, DellaVigna and Linos (2020) find that RCTs conducted in so-called nudge units show substantially lower effect sizes than RCTs in published academic studies. The authors show that publication bias can

[16] Cognitively, this could work through stress: Making the risk more salient is likely to induce stress for the risk-averse subjects. Stress can in turn impair cognitive performance, demonstrated by psychological and neurological research (Lupien et al., 2007, Matthews et al., 2000, Schoofs et al., 2008). Such a channel is well possible and we cannot rule it out. As a matter of fact, rather than a separate explanation, we consider it to be a plausible and potential cognitive mechanism for the channel discussed in this paragraph.

account for the full difference. Also, Camerer et al. (2016) find that replicated studies have a 33% lower effect size than the original studies. By adhering to rigorous methods such as pre-registration and the publication of papers unconditional of the results, effect sizes will better reflect reality compared to the past, in which effect sizes and statistical significance are likely inflated (Olken, 2015).

By providing support for the essential role played by risk preferences in innovation, our study contributes to the literature on innovation and entrepreneurship. We are the first to demonstrate that purely behavioral nudges can affect exploratory behavior. This opens up interesting paths for future research since behavioral interventions that cost-effectively foster exploration could be promising approaches to help practitioners in making organizations more innovative. For example, manipulating the customer feedback rather than a reporting mechanism, might be an interesting candidate to test.

Appendix

Chapter 1 Appendix

1.A Results

1.A.1 Performance and explorative behavior

We investigate different proxies for explorative behavior, in line with Ederer and Manso (2013). Table 1.A.1 provides an overview of all measured outcome variables.

Some of these variables reflect choices or are constructed based on choices by the subjects and proxy their explorative behavior: the variables based on the location choice indicate whether the subject detected the profit-maximizing location. This is impossible when exclusively following the customer feedback. Further, the higher the standard deviation for the continuous variables (sugar, lemon, price) is, the more explorative the subject behaved. Instead, the outcome variables with respect to profits mirror how these choices are translated into payoffs. Since the business game is designed such that explorative behavior increases the chance of finding a profit-increasing strategy, those outcome variables should be closely correlated. This can be seen by Table 1.A.2.

Table 1.A.1: Overview of proxies for explorative behavior.

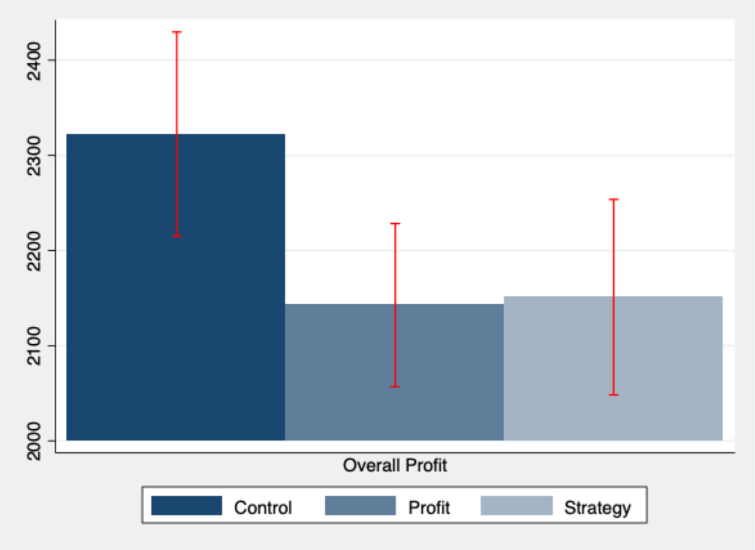
Variable	Description
final_profit	Profit in final round, continuous variable (min:0,max: 199.1).
max_profit	Highest profit in all rounds, continuous variable (min:0,max: 199.1).
overall_profit	Sum of total profit of all 20 periods, continuous variable (min:0,max: 3982).
final_location	Final location chosen, categorical variable (School, Business, Stadium).
location_non-default	Constructed variable. Count of chosen non-default locations, i.e. non-Business locations. Discrete variable (min:0, max: 20)
max_exploration_phase	Constructed variable. Longest duration of an exploration phase. An exploration phase starts when subjects choose a location other than the default location suggested by the previous manager. An explorative phase is defined as ending when a subject switches back to the default location or when a subject does not change location and lemonade color and also does not change lemon content, sugar content and price by more than 0.25 units. Discrete variable (min:0, max: 20). Adapted 1:1 from EM.
duration_exploration_phase	Constructed variable. Total duration of all exploration phases. Discrete variable (min:0, max: 20)
std_dev_sugar	Constructed variable. Standard deviation for sugar choices over all rounds. Continuous variable.
std_dev_lemon	Constructed variable. Standard deviation for lemon choices over all rounds. Continuous variable.
std_dev_price	Constructed variable. Standard deviation for price choices over all rounds. Continuous variable.
average_std_dev	Constructed variable. The average subject-specific standard deviation of strategy choices for the three continuous variables sugar, price, lemon. Continuous variable.

Table 1.A.2: Spearman cross-correlation table of exploration outcome measures.

Variables	finalprofit	maxprofit	overallprofit	final__loc__binary	loc__non-def	expl__phase__maxdur	expl__phase__totdur	sd__sugar__1-10	sd__sugar__11-20	sd__lemon__1-10	sd__lemon__11-20	sd__price__1-10	sd__price__11-20	sd__choices__1-10	sd__choices__11-20
finalprofit	1.000														
maxprofit	0.9759	1.000													
	0.000														
overallprofit	0.9057	0.8849	1.000												
	0.000	0.000													
final__loc__binary	0.7661	0.7385	0.7379	1.000											
	0.000	0.000	0.000												
loc__non-def	0.7414	0.7398	0.7976	0.7417	1.000										
	0.000	0.000	0.000	0.000											
expl__phase__maxdur	0.5948	0.6265	0.561	0.53	0.6373	1.000									
	0.000	0.000	0.000	0.000	0.000										
expl__phase__totdur	0.7105	0.7365	0.6668	0.6621	0.78	0.8447	1.000								
	0.000	0.000	0.000	0.000	0.000	0.000									
sd__sugar__1-10	0.3501	0.354	0.3469	0.2708	0.2526	0.42	0.3211	1.000							
	0.0007	0.0006	0.0008	0.0098	0.0163	0.000	0.002								
sd__sugar__11-20	0.1565	0.1908	0.0678	0.1731	0.0898	0.2411	0.2858	0.0744	1.000						
	0.1407	0.0716	0.5255	0.1027	0.4001	0.022	0.0063	0.486							
sd__lemon__1-10	0.3827	0.382	0.3232	0.282	0.1787	0.2807	0.1973	0.4654	0.1184	1.000					
	0.0002	0.0002	0.0019	0.0071	0.0919	0.0074	0.0623	0.000	0.2664						
sd__lemon__11-20	0.3859	0.4099	0.2915	0.3927	0.2923	0.3651	0.4654	0.2855	0.4309	0.2593	1.000				
	0.0002	0.0001	0.0053	0.0001	0.0052	0.0004	0.000	0.0064	0.000	0.0136					
sd__price__1-10	0.3692	0.404	0.2295	0.3056	0.2304	0.3379	0.3536	0.1681	0.2825	0.2455	0.3087	1.000			
	0.0003	0.0001	0.0296	0.0034	0.0289	0.0011	0.0006	0.1133	0.007	0.0197	0.0031				
sd__price__11-20	0.3748	0.4211	0.1951	0.2727	0.1418	0.286	0.3913	0.1525	0.3483	0.2084	0.3267	0.4113	1.000		
	0.0003	0.000	0.0653	0.0093	0.1825	0.0063	0.0001	0.1514	0.0008	0.0487	0.0017	0.0001			
sd__choices__1-10	0.3228	0.3509	0.334	0.1311	0.1557	0.2363	0.2124	0.4962	0.1365	0.4954	0.2001	0.2161	0.1471	1.000	
	0.0019	0.0007	0.0013	0.2183	0.1429	0.0249	0.0444	0.000	0.1997	0.000	0.0586	0.0408	0.1665		
sd__choices__11-20	0.6152	0.6263	0.6248	0.2486	0.392	0.3907	0.442	0.4483	0.0882	0.4197	0.3367	0.1404	0.1712	0.7068	1.000
	0.000	0.000	0.000	0.0182	0.0001	0.0001	0.000	0.000	0.4083	0.000	0.0012	0.187	0.1067	0.000	

Results with respect to the final profit and the maximum profit were already discussed in Section 1.3.1. In line with these discussed results, the mean overall profit in the control group is higher than in both reporting treatments, as shown by Figure 1.A.1.

Figure 1.A.1: Mean of overall profit by treatments.



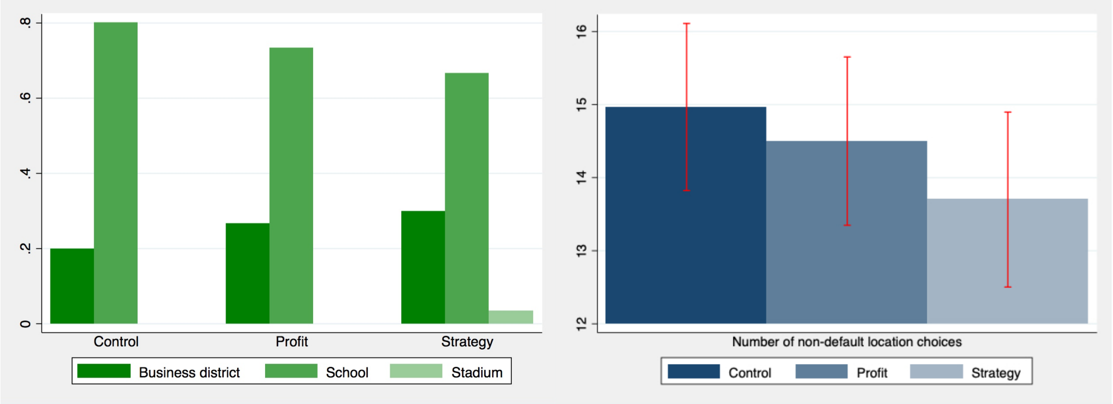
Analyzing the location chosen in the final round of the game shows the same pattern. There are three possible locations, with the school being the optimal one. As Figure 1.A.2a depicts, around 80% of subjects in the control group chose the optimal location in the final round. This proportion is higher than in the profit (appr. 73%) and strategy treatment (appr. 66%).[17]

Results with respect to the exploration phases further support our finding: the control group explores more than the profit and the strategy treatment. However, these differences are, again, not significant.

Lastly, Figure 1.A.4 shows that also the results for all outcome variables with respect to the standard deviations are in line with the previous results: the standard deviation in the control group is slightly higher than in the reporting treatments. As expected, the choices within the first ten rounds of the business game vary more than the choices in the last ten rounds. This early exploration is intuitive since the individual can profit from her findings for a longer time horizon than later stage explorative activities.

[17] These difference-in-means are not statistically significant (two-sided Mann-Whitney U-tests: p=0.5449 for control group vs. profit treatment, p=0.2469 for control group vs. strategy treatment).

Figure 1.A.2: Location measures.



(a) Proportion of subjects by Location in final round.

(b) Number of times the non-default location was chosen.

To further analyze the explorative behavior of our subjects, we compare the average subject-specific standard deviation of the profits. The variability of profits in all treatments is higher in periods 1-10 (see Figure 1.A.5). However, the standard deviation of the profits is not different between the treatments.[18]

1.A.2 Attention

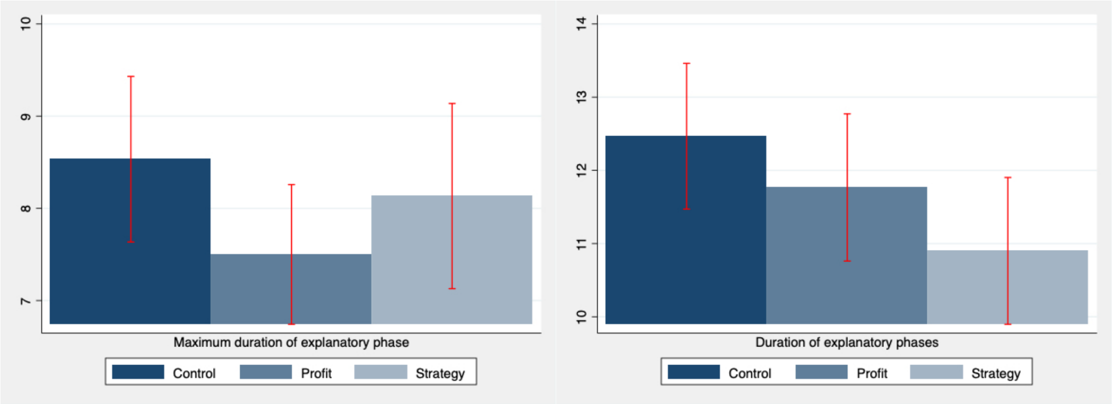
We analyze different proxies for exerted effort, again, in line with Ederer and Manso (2013). Table 1.A.3 provides an overview of all measured proxies. The respective results are discussed in Section 1.3.2.

Table 1.A.3: Overview of proxies for effort.

Variable	Description
total_time_decision	Time elapsed for all 20 Decision-Screens, continuous variable.
focus_time_decision	Focus time elapsed for all 20 Decision-Screens, continuous variable.
total_time_result	Time elapsed for all 20 Result-Screens, continuous variable.
focus_time_result	Focus time elapsed for all 20 Result-Screens, continuous variable.
total_time_reporting	Time elapsed for all 20 Reporting-Screens. Treatment groups only, continuous variable.
focus_time_reporting	Focus time elapsed for all 20 Reporting-Screens. Treatment groups only, continuous variable.
total_notes	Proportion of filled out fields in notes sheet, continuous variable.
notes_strategic	Proportion of notes with respect to strategic variables relative to total_notes, continuous variable.
notes_profit	Proportion of notes with respect to profits relative to total_notes, continuous variable.
notes_feedback	Proportion of notes with respect to customer feedback relative to total_notes, continuous variable.

[18] Two-sided Mann-Whitney U-tests: p=0.8941 for periods 1-10 and p=1.000 for periods 11-20 between control group and profit treatment. p=0.5444 for periods 1-10 and p=0.1833 for periods 11-20 between

Figure 1.A.3: Exploration phase measures.



(a) Maximum of all exploration phases

(b) Duration of all exploration phases

Notes: The figure reports the means of the maximum length and of the duration of all exploration phases. Error bars indicate standard errors of the mean.

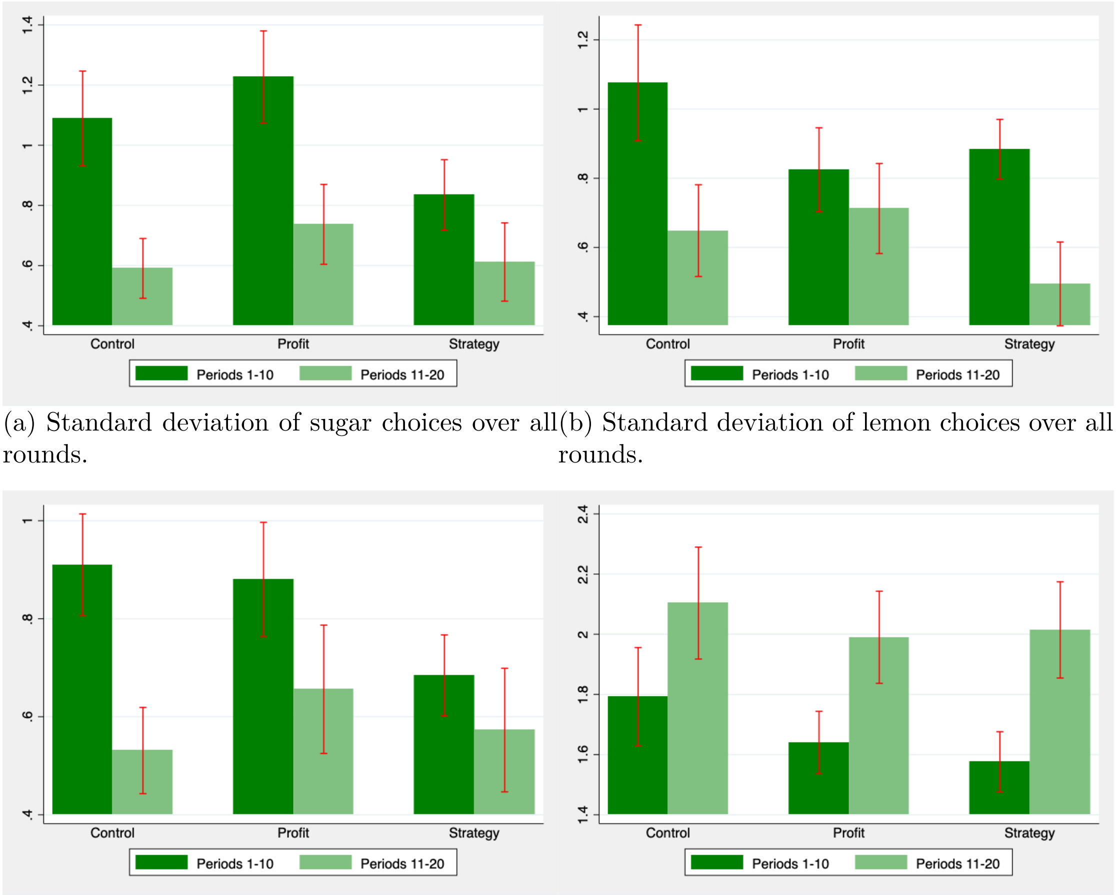
1.A.3 Discussion

If limited effort were the valid explanation for our findings, the production function of exploration should still be identical for all subjects. That is, for a given amount of effort, the exploration level should be the same across treatment groups. The reason is that according to the limited effort explanation, participants exert less effort for productive tasks because they need to exert more effort for the (arguably) unproductive reporting.

However, we find that participants in the Strategy treatment explore less conditional on the level of productive effort exerted (measured by time spent in front of the decision and results screen). In other words, the production function of the risk-averse type is impaired. This is best visible in Figure 1.A.8 that uses time spent on the reporting and the decision screens as a proxy for effort and the number of times a participant enters into an exploration phase as the outcome variable representing exploration behavior. The figure shows that the production function for the risk-averse type is indeed altered: Conditional on the effort level, treated participants initiate less exploration phases. Further, we observe that the production function of the less risk-averse type is not much different across treatments. Similar patterns are observed when measuring exploration through the final location (Figure 1.A.8) and final profit (Figure 1.A.9).

control group and strategy treatment and p=0.4598 for periods 1-10 and p=0.3142 for periods 11-20 between profit and strategy treatment.

Figure 1.A.4: Standard deviation measures.



(a) Standard deviation of sugar choices over all rounds.

(b) Standard deviation of lemon choices over all rounds.

(c) Standard deviation of price choices over all rounds.

(d) Standard deviation of all choices over all rounds.

1.B Experimental design

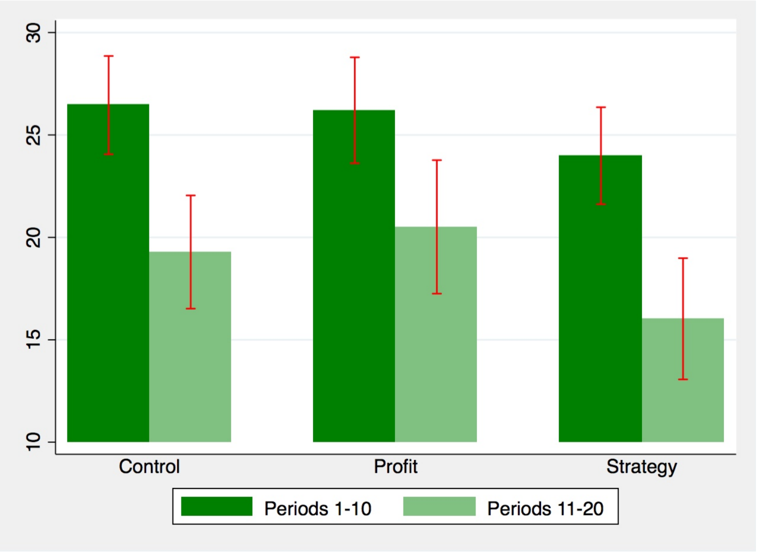
1.B.1 Instructions

Welcome

You are now taking part in a scientific study. Please read the following instructions carefully. Everything that you need to know in order to participate in this experiment is explained below. Should you have any difficulties in understanding these instructions, please notify us. We will answer your questions at your cubicle. During the course of the experiment you can earn money. The amount that you earn during the experiment depends on your decisions. All the gains that you make during the course of the experiment will be exchanged into cash at the end of the experiment.

The exchange rate will be: 100 thaler = 1 EUR.

Figure 1.A.5: Standard deviation of realized profits.



The experiment is divided into 20 periods. In each period you have to make decisions, which you will enter on a computer screen. The decisions you make and the amount of money you earn will not be made known to the other participants - only you will know them. At the end of the experiment, you will be requested to respond to survey questions. Please note that communication between participants is strictly prohibited during the study. Communication between participants and unnecessary interference with computers will lead to the exclusion from the study. In case you have any questions don't hesitate to ask us.

Procedures

In this experiment, you will take on the role of an individual running a lemonade stand. There will be 20 periods in which you will have to make decisions on how to run the business. These decisions will involve the location of the stand, the sugar and lemon content and the lemonade color and price. The decisions you make in one period, will be the default choices for the next period. At the end of each period, you will learn what profits you made during that period. You will also hear some customer reactions that may help you with your choices in the following periods.

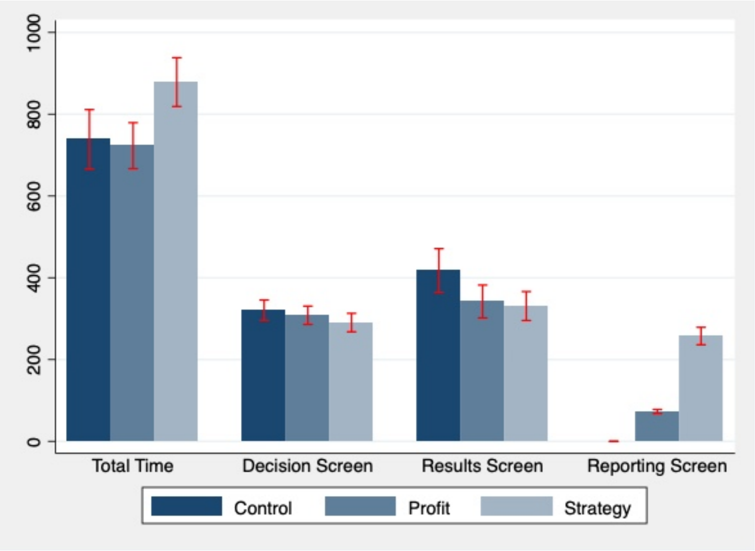
Letter from the Previous Manager

The previous manager of the lemonade stand has left you guidelines on how to run the business. The letter from the previous manager is the following:

Dear X,

I have enclosed the following guidelines that you may find helpful in running your lemonade stand. These guidelines are based on my previous experience running this stand. When running my business, I followed these basic guidelines:

Figure 1.A.6: Effort measured by time elapsed.



Notes: The figure reports the means of the subjects time elapsed for the total time, the time spent at the decision screen, at the results & feedback screen, and at the reporting screen, respectively.

Location: Business District
Sugar Content: 5.2%
Lemon Content: 7.0%
Lemonade Color: Green
Price: 8.2 thaler

With these choices, I was able to make an average profit of about 85 thaler per period. I have experimented with alternative choices of sugar and lemon content, as well as lemonade color and price. The above choices were the ones I found to be the best. I have not experimented with alternative choices of location though. They may require very different strategies.

Regards, Previous Manager

Note that in the first period, these choices will appear as defaults.

Compensation

Your compensation will be based on the profits you make with your lemonade stand. You will get paid 50% of your total lemonade stand profits during the 20 periods of the experiment.

For example, if your total profits during the 20 periods of the experiment were 1700 thaler, you will earn 850 thaler, worth 8.50 EUR.

In addition, you will earn the show-up fee of 2 EUR.

Report

Table 1.A.4: Regressions.

	Dependent variables:					
	Overall Time	*Decision Time*	*Results Time*	*Strategy Notes*	*Profit Notes*	*Feedback Notes*
	(1)	(2)	(3)	(4)	(5)	(6)
Profit Treatment	-17.982	-12.933	-77.952	-0.147	0.068	-0.103
	(90.19)	(33.36)	(62.17)	(0.11)	(0.09)	(0.11)
Strategy Treatment	144.199	-27.883	-80.104	-0.167	-0.123	-0.083
	(92.52)	(34.22)	(63.78)	(0.11)	(0.09)	(0.12)
Gender	33.602	-4.345	21.356	0.167*	0.204***	0.104
	(75.42)	(27.90)	(51.99)	(0.09)	(0.08)	(0.09)
Riskaversion	33.238	9.500	39.170	0.092	0.094	0.027
	(77.35)	(28.61)	(53.32)	(0.09)	(0.08)	(0.10)
Constant	704.299***	317.549***	385.747***	0.486***	0.642***	0.471***
	(81.23)	(30.05)	(55.99)	(0.10)	(0.08)	(0.10)
N	90	90	90	90	90	90
R^2	0.046	0.011	0.036	0.092	0.152	0.027
Adjusted R^2	0.001	-0.035	-0.009	0.049	0.112	-0.019

Level of significance: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Displayed only if assigned to *profit treatment*: Please report your profits of the last three periods.
Displayed only if assigned to *strategy treatment*: Please describe your strategy in the last three periods. Why did you choose this strategy?

1.B.2 Parametrization

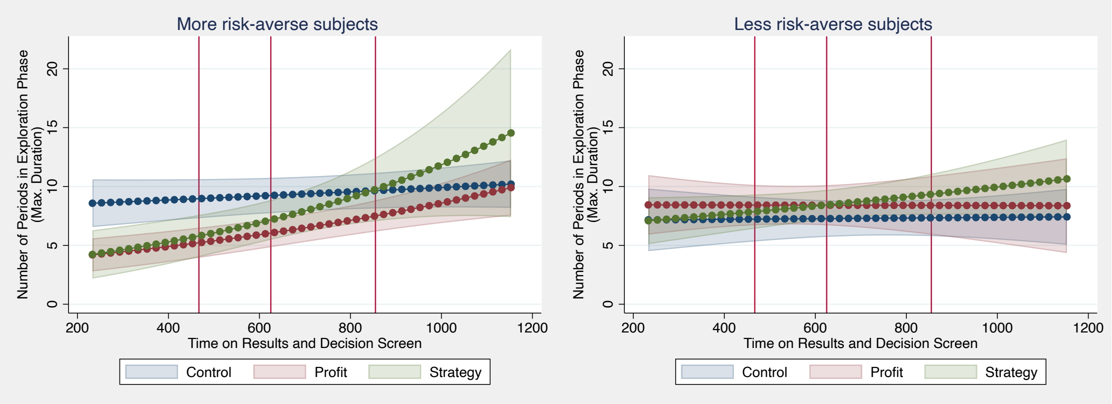
We adopted the experimental parameters from Ederer and Manso (2013). The participants could make following choices:

- Location = {Business District, School, Stadium}
- Sugar Content = {0,0.1,...,20}
- Lemon Content = {0,0.1,...,20}
- Lemonade Color = {Green, Pink}
- Price = {0,0.1,...,10}

The optimal product mix in each location is shown in table 1.B.1.

For the profit calculation in each location, a linear penalty function was implemented. So if the participant did not implement the optimal choices, she was penalized according to the values summarized in table 1.B.2. In each location, the penalty factors represented in the table are associated with a deviation of one unit for each of the variables. Note that we implemented a minimum of 0, i.e. participants could not earn negative profit.

Figure 1.A.7: Production functions (exploration regressed on effort).



Notes: This figure displays the production function by subject types, split by median risk-aversion into less and more risk-averse. The figure plots the following regression specification: $Explorationphase = Treatment + Timespent + Treatment \times Timespent$. Quartiles of the distribution of $Timespent$ are displayed in red.

Table 1.B.1: Optimal product mix, by location.

	Business District	School	Stadium
Sugar	1.5%	9.5%	5.5%
Lemon	7.5%	1.5%	5.5%
Lemonade Color	Green	Pink	Green
Price	7.5	2.5	7.5
Maximum Profit	100	200	60

1.C Elicitation of control variables

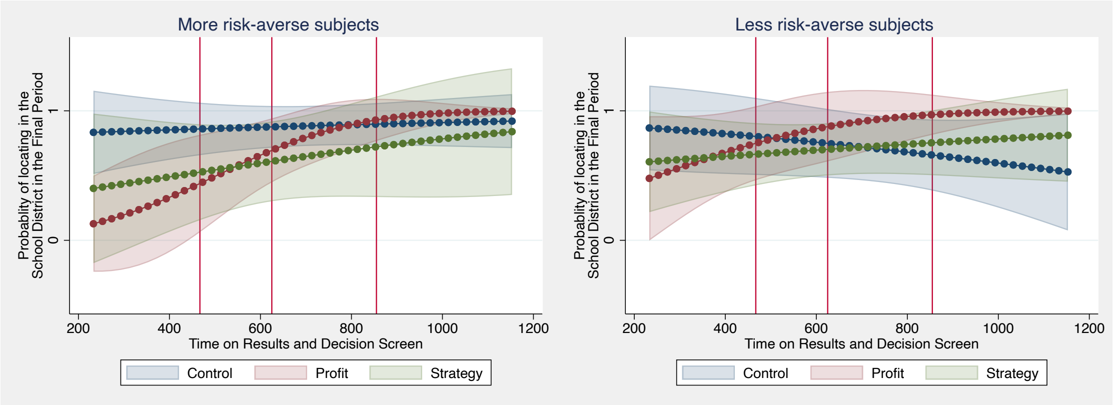
1.C.1 Risk preference

For the elicitation of risk preferences, we relied on the staircase elicitation method developed by Falk et al. (2018). We adapted their method without any changes. In the following, the wording for the first decision is summarized, with the subsequent question following the same wording:

Please imagine the following situation: You can choose between a sure payment of a particular amount of money, or a draw, where you would have an equal chance of getting 300 EUR or getting nothing. We will present to you five different situations. The draw is the same in all situations. The sure payment is different in every situation.

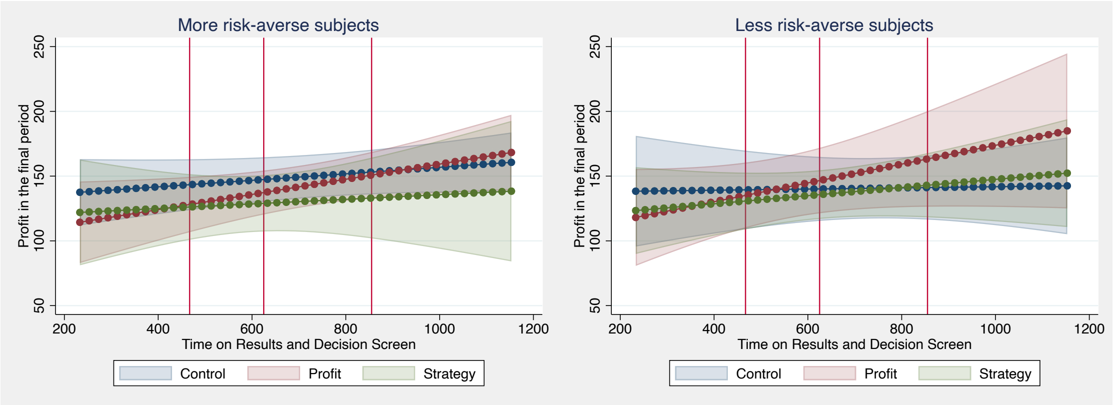
What would you prefer: a draw with a 50 percent chance of receiving 300 USD, and the same 50 percent chance of receiving nothing, or the amount of 160 USD as a sure payment?

Figure 1.A.8: Production functions (final location school regressed on effort).



Notes: This figure displays the production function by subject types, split by median risk-aversion into less and more risk-averse. The figure plots the following regression specification: $Finallocation = Treatment + Timespent + Treatment \times Timespent$. Quartiles of the distribution of Time on Results and Decision Screen are displayed in red.

Figure 1.A.9: Production functions (final profit regressed on effort).



Notes: This figure displays the production function by subject types, split by median risk-aversion into less and more risk-averse. The figure plots the following regression specification: $Finalprofit = Treatment + Timespent + Treatment \times Timespent$. Quartiles of the distribution of Time on Results and Decision Screen are displayed in red.

Table 1.B.2: Penalty factors, by location.

	Business District	School	Stadium
Sugar	3	6	0.5
Lemon	3	6	0.5
Lemonade Color	20	60	0.5
Price	3	6	0.5

1.C.2 Demographics

What is your gender?

What is your age?

What is the highest level of education you have completed or the highest degree you have received?

What is your academic field?

What is your nationality? (If more than one apply, select the one you feel is most representative for you).

In what country did you grow up? (If you grew up in more than one country, please indicate the country you lived the longest while growing up).

In what country do you currently reside?

Do you identify yourself with any of the following religions? ['Christianity', 'Judaism', 'Islam', 'Buddhism', 'Hinduism', 'Other', 'No religion / Atheism']

1.D Sequential analysis

Our sequential analysis plan followed the summarized procedure outlined below. It concerned the key outcome variable, the profit in the final round. The assessment on how to proceed after the first stage was based on a hypothesis test conducted with a two sample t-test.

Based on an expected effect size of Cohen's $d = 0.5$, a power analysis indicated that for a two-sided test with an alpha of 0.05, a desired statistical power of 0.8, and two looks using a linear spending function, a total of 180 participants is needed (60 per group). If the expected difference is significant at the first interim analysis (after 90 participants or time = 0.5, with an alpha boundary of 0.025) the data collection will be terminated. The data collection will also be terminated if the observed effect size is smaller than the smallest effect size of interest, which is set at $d = 0.3875$ based on the researcher's willingness to collect at most 300 participants for this study, and the fact that with one interim analysis, 300 participants provide 0.8 power to detect an effect of $d = 0.3875$. If the interim analysis reveals an effect size larger than 0.5, but while $p > 0.025$, the data collection will be continued until 60 participants per group have been collected. If the effect size lies between the smallest effect size of interest ($d = 0.3875$) and the expected effect size ($d = 0.5$), the planned sample size will be increased based on a conditional power analysis to achieve a power of 0.9 (or to a maximum of 100 participants per group, or 300 participants in total). The second analysis is performed at an alpha boundary of 0.0358.

In the original paper of Ederer and Manso (2013), subjects in the pay-for-performance contract yield on average a profit in the last period of 111 thalers. Because our control group is mimics one-to-one the pay-for-performance group in Ederer and Manso (2013), we assume that our control group will yield 111 thalers in the last period too. Furthermore, based on a small test conducted in September 2019, we expect a standard deviation of $SD = 40$. It is our best assumption that the standard deviation is equal for all three groups.

Next, we define the smallest detectable effect size of interest. Based on practical limitations, namely budget restrictions, we are willing to collect at most 300 observations in total. To identify an effect with 80% power when comparing the control with one of the two treatment groups, with one interim look at the data, the effect must be at least 15.5 thalers large, translating in a Cohen's $d = 0.3875$. We deem such an effect also from a real-life perspective as appropriate - implementing a new reporting policy comes along with costs, and thus, the beneficial or detrimental effect should be large enough to be of practical relevance.

Next, we elaborate on the expected effect. In Ederer and Manso (2013), the exploration contract yields in the final period on average a profit of 140. This leads to a profit difference between pay-for-performance and exploration contract of $29(= 140 - 111)$. With a $SD = 40$, this yields a Cohen's $d = 0.725$. Most probably, our effects will be lower since our treatment interventions are based on a behavioural mechanism but do not change monetary incentives, as this is the case in Ederer and Manso (2013). As a best-estimate, we expect our effect to be 30% (or 9 thalers) lower than in their study. Therefore, we estimate that subjects in the strategy treatment will yield a lower profit in the last period than their exploration contract, and we estimate this to be at 131. This yields an effect size of $20(= 131 - 111)$, or a Cohen's $d = 0.50$. For the profit treatment, we cannot base our estimates on a previous study due the lack of comparable alternatives. However, we expect the profit treatment to perform similarly as the strategy treatment (but in the opposite direction, of course). Consequently, we adopt the same Cohen's $d = 0.5$ for the profit treatment. Based on the expected effect size of $d = 0.5$, with power 0.8 and an alpha of 0.05, we obtain a sample size $n = 60$ per group or $n = 180$ in total (after both looks). We will have a first look at time $= 0.5$, that is when 90 subjects are collected.

For controlling type 1 error rates, we use a linear spending function (power family function), as outlined in Jennison and Turnbull (2001). The alpha of 0.05 for a single look is adjusted for sequential analyses, namely for two looks using a linear spending function, yielding a nominal alpha of 0.0586. Thus, planning on analyzing the data at two different stages of the experiment, that is with one interim analysis, we formulate the analysis plan outlined in the main body of the text.

Chapter 2

The acceptability of food policies[1]

2.1 Introduction

Population growth, combined with global changes in diets that are increasingly rich in sugars, refined fats, and animal-based products, is putting both our environment and public health under great stress. Food production is responsible for more than 25% of our greenhouse gas emissions (Gerber et al., 2013), occupies close to half of all habitable land (FAO, 2020), and is the main driver of deforestation of tropical forests from the Amazon to South-East Asia (Steinfeld et al., 2006; Vijay et al., 2016). Unhealthy diets are also responsible for a greater risk of morbidity and mortality than unsafe sex, alcohol, drug, and tobacco use combined (Willett et al., 2019). On top of environmental and human health concerns, the expected increasing demand for animal-based proteins in the coming years is likely to consolidate intensive farming, which severely deteriorates animal welfare (Springmann et al., 2016). A dietary transition is thus one of humanity's great challenges.

Recent works have shown that healthier and more sustainable diets can efficiently mitigate these issues, namely by reducing sugar consumption and shifting to plant-based proteins and unsaturated oils (Tilman and Clark, 2014; Poore and Nemecek, 2018; Willett et al., 2019). However, even though private action in this direction is producing positive results in some developed countries,[2] decentralized and spontaneous dietary changes are likely to fall short in addressing this global challenge. Numerous consumers still underestimate the social impact of their diets (Macdiarmid et al., 2016), and a significant proportion of them refuse to acknowledge the consequences of their consumption (Rothgerber, 2014; Espinosa and Stoop, 2021). Aware consumers might also be reluctant

[1] This chapter is based on the homonymous article written in collaboration with Romain Espinoa and published in *Nutrients* as Espinosa and Nassar (2021). The survey received the approval of the ethics committee of the Centre de Recherche en Economie et Management and the replication was registered in the AEA RCT Registry (ID: AEARCTR-0006429). The authors gratefully acknowledge the funding by ANR grant number ANR-19-CE21-0005-01.

[2] For instance, the consumption of meat per capita in France decreased by 12% between 2007 and 2016 and the proportion of self-declared *flexitarians* increased from 25% in 2015 to 34% in 2017 (Espinosa, 2019).

to change consumption habits that are part of their social identity and to which they are attached (Vartanian et al., 2007; Graça et al., 2015). Even consumers who are effectively willing to change their diets might have difficulties following through, especially given that most food-related decisions are made unconsciously (Wansink and Sobal, 2007).

Appeals for public interventions supporting food transitions are therefore accumulating (WHO & FAO, 2003). Authorities should, nonetheless, be cautious in their design and implementation, as policies regulating food that are not accepted can backfire (Stok et al., 2014; Ungar et al., 2015). Ensuring the acceptability of a public policy is therefore critical for its success. Widely accepted policies are also more likely to be implemented in the first place. Policies that benefit from a large popular support are indeed more likely to be enacted in direct democracies through referenda or in indirect democracies, in which political competition leads politicians to support popular interventions (Cullerton et al., 2016).

The objective of this paper is to offer a better understanding of the different factors that are actionable for policy-makers to increase the acceptability of policies regulating food. To that aim, we propose and test a level-based model of food policy acceptability. The governmental level includes macroscopic factors that influence the acceptability of *any* governmental intervention. The topic level adds factors that relate to the acceptability of public interventions *for a specific topic*. Citizens might have preferences for interventions depending on what the policy concerns (e.g., it is acceptable to regulate products containing cage eggs). The policy level corresponds to factors that influence the acceptability of a specific policy for a given topic (e.g., it is acceptable to have a GBP 0.1 tax on products that contain cage eggs). Finally, the individual level includes demographics and personal characteristics that may influence acceptability perceptions. In this study, we focus on the topic and policy levels. We restrict our attention to these two levels as we seek to understand what drives the variation of acceptability for different interventions across topics and policies for a given government and population.

In this work, we asked about 1200 UK nationals about their perceptions of different policies regulating a given food item (sugar, palm oil, or battery-cage eggs). The policies are composed of two education interventions (information campaigns, labeling) and four increasingly coercive market interventions (three levels of taxation and market withdrawal of the targeted product). Regarding the topic-level factors, we found that participants were more likely to support a food policy when they were more aware of the issue at stake, when they believed that it was legitimate to have collective rules regulating the product under scrutiny, and when there were strong social norms regarding the necessity to reduce the consumption of said product. As far as policy-level factors are concerned, we observed that policies were more popular when they were seen as more effective and when they targeted the appropriate group of consumers. On the contrary, consumers found coercive interventions and policies that generated inequalities less acceptable. Our results show the existence of a trade-off between coerciveness, effectiveness, and acceptability, as participants judged more coercive policies to be more effective, but also less acceptable. These conclusions can help guide policy-makers in the design of policies supporting a dietary transition.

The remainder of this chapter is structured as follows. In section 2.2, we introduce the factors influencing acceptability, relate them to the literature, and structure them in levels. The survey is laid out in section 2.3. We present the results in section 2.4 and discuss them in section 2.5.

2.2 A model of food policy acceptability

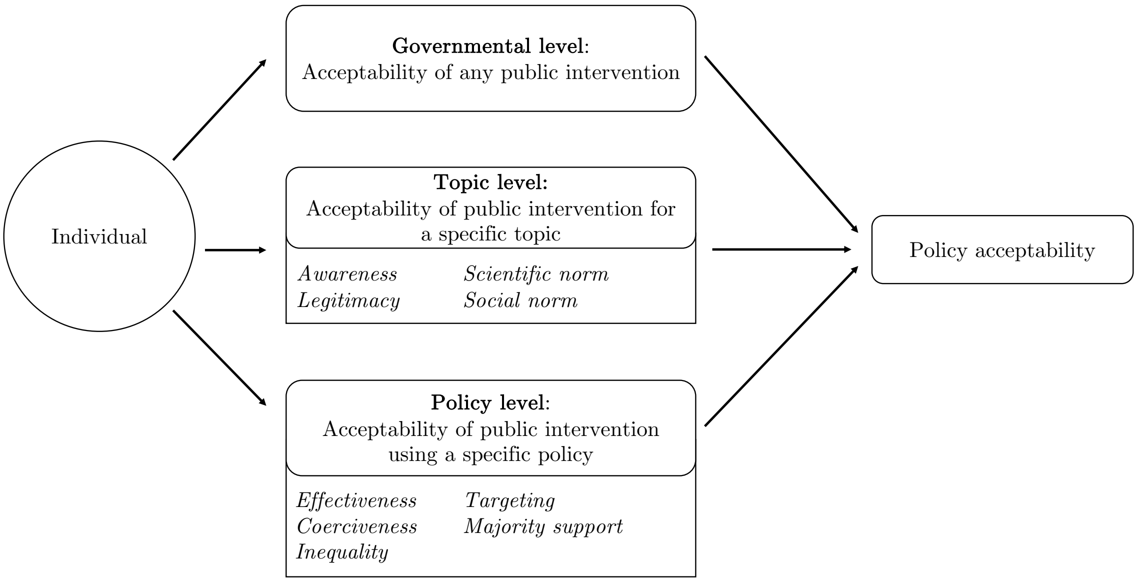
The interest in food policy acceptability has grown in the past decades, mainly following the increasing prevalence of obesity in developed countries (Gortmaker et al., 2011). The resulting literature has aimed at understanding the acceptability of food policies targeting obesity (Mazzocchi et al., 2015; Reynolds et al., 2019) and, more precisely, the support or aversion for taxation of sugar-sweetened beverages (see Eykelenboom et al., 2019 for a review). The latest works on nutrition have shown a growing interest in the new dietary challenges faced by developed countries, such as the reduction of animal-based food consumption and the increase in plant-based protein intakes (Willett et al., 2019). In addition, growing environmental and animal-welfare concerns further call for the enforcement of new food policies. This highlights the need for an integrated model of food policy acceptability that allows an understanding of the underlying factors that determine the acceptability of any food policy for any reason for intervention.

We propose a structuring of these factors in a multilevel model of food policy acceptability, which we relate to the existing literature, which often considers them in isolation (Stok et al., 2014; Bos et al., 2015; Reynolds et al., 2020). In contrast, systems-like approaches that consider issues with a complex breadth of interconnected causes, interactions, and effects have gained relevance in recent years and have helped us to better understand critical food-related issues, such as obesity or unhealthy eating (Hammond, 2009; Lee et al., 2017; Barnhill et al., 2018). Scholars are also increasingly recognizing the importance of such comprehensive approaches for accurately assessing the (multiple) effects of a given policy and calling for their systematic use for the promotion and the evaluation of food policies (Rutter et al., 2017; Moore et al., 2019; Evans, 2020; Fanzo et al., 2020). Our work contributes to the development of such approaches by proposing a multilevel framework of food policy acceptability.

In our model, the governmental level includes factors that relate to perceptions about the government itself, such as the level of (mis-)trust in the government or the level of corruption, which can influence the acceptability of any policy proposed. At the topic level are factors that affect acceptability of an intervention due to a certain reason, such has limiting deforestation or preventing obesity. At the policy level are factors that determine acceptability of a specific type of policy. Finally, the individual level includes demographics, political leaning, and other personal characteristics. Although previous works showed that individual-level factors may influence acceptability, such as women finding interventions generally more acceptable (Diepeveen et al., 2013; Hagmann et al., 2018), some evidence suggests that they have a limited influence compared to topic- and policy-level factors (Mazzocchi et al., 2015).

In this paper, we focus on the roles of the topic and policy levels in the acceptability of public interventions and how they vary across topics and types of policies for a given government and population. Our ultimate objective is to identify factors that are actionable by policy-makers to maximize the chances of a successful dietary transition. The model and the factors that were tested in the online survey are summarized in Figure 2.2.1 and Table 2.2.1.

Figure 2.2.1: A model of food policy acceptability.



2.2.1 Topic-level factors

The acceptability of a specific policy might depend on the underlying issue calling for governmental intervention, i.e., the topic of the policy. Previous research found that the level of awareness of the issue at stake is a key determinant of policy acceptability. Bos et al. (2013) showed, in semi-structured interviews, that awareness of the problems of obesity leads to higher acceptability of different types of policies promoting healthier foods. These findings concur with the results of increasing acceptability linked to awareness of policies regarding other heath topics, such as smoking and drinking (Diepeveen et al., 2013), or environmental policies, such as energy use (Steg et al., 2005).

Scientific research is regularly used as a basis for the implementation of new policies, but does it affect public acceptability? Policy-makers in the USA and New Zealand have suggested that more scientific evidence and/or a larger spread of scientific results could increase the public support for taxes on sugar-sweetened beverages (Signal et al., 2018;

Table 2.2.1: Summary of factors.

Factor	Description
Awareness	The high consumption of (sugar | palm oil | cage eggs) causes serious problems for society.
Legitimacy	It is legitimate to have collective rules that govern the consumption of (sugar | palm oil | cage eggs).
Social norm	It is commonly accepted that (sugar | palm oil | cage eggs) consumption should be reduced.
Scientific norm	We consume more (sugar | palm oil | cage eggs) in our society than recommended by the (most recent scientific work | the most recent environmental scientific work | most recent scientific work on preserving animal welfare).
Effectiveness	The measure is effective in reducing the consumption of (sugar | palm oil | cage eggs).
Coerciveness	The measure is coercive.
Inequality	The measure will increase social inequalities.
Targeting	The measure will affect the appropriate group of consumers and producers.
Majority support	A majority of citizens would agree to implementing the measure.

Purtle et al., 2018), as there are still vast discrepancies between the established scientific consensus and public beliefs (Eykelenboom et al., 2019).

Social norms have also been established as a determining factor of the acceptability of public policies. Stok et al. (2014) reported that an intervention aimed at increasing fruit intake is more accepted if participants are informed that trying to increase fruit intake is the norm in their peer group. A positive influence of perceived social norms was also reported in terms of the acceptability of transport pricing policies (Schade and Schlag, 2003).

Last, views about the boundary between decisions that should remain inherently private and those that should be regulated by the state may influence the acceptability of public policies. As the body of evidence emphasizing the link between diets and chronic diseases has grown, governments have started taking action to influence diets (Traill et al., 2014). In 2014, Mexico famously introduced a tax on sugar-sweetened beverages (Colchero et al., 2017). Since then, voices have risen to protest against what they consider to be privacy-invading policies, as illustrated in Washington, where citizens voted to prohibit new taxes on food grocery items in 2018.[3] Several authors have also shown that, with respect to obesity, the ascription of responsibility for personal choices or environmental reasons influences the perception of the legitimacy of governmental

[3] Archive of the ballot available at: `http://web.archive.org/web/20191226010045/https://ballotpedia.org/Washington_Initiative_1634,_Prohibit_Local_Taxes_on_Groceries_Measure_(2018)`

intervention and, with it, the support for food policies (Niederdeppe et al., 2011; Bos et al., 2013; Mazzocchi et al., 2015). The stronger the belief that the individual is responsible, the smaller the legitimacy of governmental intervention and the smaller the acceptability of public policies.

2.2.2 Policy-level factors

Despite being favorable to public intervention on a particular topic, citizens might consider a specific policy choice unacceptable. The perceived fairness of a policy has been consistently reported to increase its acceptability for both health interventions and environmental ones (Bamberg and Rölle, 2003; Bos et al., 2015). However, some researchers have questioned the relevance of explaining policy acceptability with fairness judgements, as the two notions of fairness and acceptability can be synonyms (Steg and Schuitema, 2007). Consequently, we do not include fairness as a determinant of policy acceptability per se, but investigate two elements that determine opinions about the fairness of a policy: the types of individuals impacted by the policy and the inequalities expected to result from its implementation. It has indeed been established that the acceptability of nutrition policies depends on the extent to which certain groups are targeted, with higher acceptability for key groups, such as children and teenagers (for a review, see Caraher and Cowburn, 2005). In addition, policies that are expected to worsen inequalities, i.e., those that will disproportionately impact citizens with low income, are expected to be less acceptable (Bos et al., 2013).

Moreover, citizens are also more prone to accepting interventions that are perceived as effective (Bos et al., 2013, 2015). In a recent review, Reynolds et al. (2020) reported that communicating the effectiveness of health-related policies is causal to increased public support. Similarly, the belief that a policy is supported by the majority improves its acceptability (De Groot and Schuitema, 2012). Participants are expected to be more willing to adapt their behaviors when they believe that a policy is effective and supported by their peers.

Restricting freedom of choice, i.e., the coerciveness of the policy, is, on the other hand, expected to negatively impact acceptability judgements. This effect has been documented for dietary interventions and for different health-related behaviors (Diepeveen et al., 2013; Mazzocchi et al., 2015).

2.3 Survey

We investigated the role of the nine aforementioned factors in the determination of the acceptability of public interventions using an online survey.

2.3.1 Topics

We explored the determinants of policy acceptability for dietary interventions on three products that address the main challenges that food systems currently face: sugar (health), palm oil (environment), and cage eggs (animal welfare).

Sugar consumption has been shown to be a major driver of obesity and is associated with greater health risks (Ludwig et al., 2001; Yang et al., 2014). This has led an increasing number of governments to undertake actions to limit sugar intake, such as sugary drink taxes (e.g., Australia, France, Portugal, Mexico, India). The production of palm oil significantly contributes to the increasing deforestation and represents a great threat to biodiversity and climate change (Koh et al., 2011). These problems have been acknowledged by the EU, which discussed its partial ban for certain uses (Russel, M., 2020). Cage eggs deliver the worst living conditions for egg-laying hens, with higher mortality and wound rates (Dikmen et al., 2016). In 2020, 39% of the egg production still originated from cage-bound hens in the UK[4], while they are already banned in different countries.

2.3.2 Online survey

We designed an online survey to elicit the factors that determine the acceptability of food policies. We adapted the survey to each of the three topics (between-subject design). The three questionnaires are displayed in the supplementary materials. Each questionnaire is made up of four parts.

First, we displayed six policies that could be implemented to regulate the consumption of snacks that contain the product targeted by the public policy. In the Sugar treatment, the public intervention focuses on snacks that have a high sugar content (more than 33 g of sugar per 100 g). In the Palm and Eggs treatments, the policies target snacks that contain palm oil or cage eggs, respectively. The six interventions are similar across topics: introducing a label to identify the targeted products, setting up an information campaign to educate consumers about the social impact of the products, introducing taxes of GBP 0.10, 0.30, or 0.50 for the targeted 30 g individual snacks, and removing the targeted snacks from the market (see table 2.3.1).

Second, on a seven-point Likert scale, participants were required to indicate the extent to which they agreed with a list of statements regarding topic-level factors. The statements concerned the legitimacy of having collective rules on the consumption of the targeted product (*legitimacy*), the perception of the problems generated by the consumption of the targeted product (*awareness*), the social norms about whether the consumption of the targeted product should be reduced (*social norm*), and whether the product was over-consumed compared to the latest scientific recommendations (*scientific norm*).

Third, we asked participants about the degree to which they agreed with a second list of statements regarding policy-level factors for each of the six public interventions listed at the beginning of the survey (label, information campaign, GBP 0.10 tax, GBP 0.30 tax, GBP 0.50 tax, withdrawal from the market). Using a seven-point Likert scale,

[4] United Kingdom department for Environment, food and rural affairs - Egg Statistics – Quarter 4, 2020. Archived at `http://web.archive.org/web/20210705092224/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/981924/eggs-statsnotice-29apr21.pdf`

Table 2.3.1: Summary of policies.

Policy	Description
Information campaign	Set up information campaigns to inform consumers about the impact of (sugar | palm oil | cage eggs) on (health | environment | animal welfare) and society.
Label	Display labels on snacks with (high sugar content | palm oil | cage eggs).
Tax10	Tax the snacks with (high sugar content | palm oil | cage eggs) by GBP 0.10 (for a 30 g individual snack, such as a cereal bar).
Tax30	Tax the snacks with (high sugar content | palm oil | cage eggs) by GBP 0.30 (for a 30 g individual snack such as a cereal bar).
Tax50	Tax the snacks with (high sugar content | palm oil | cage eggs) by GBP 0.50 (for a 30 g individual snack such as a cereal bar).
Withdrawal	Remove the snacks with (high sugar content | palm oil | cage eggs) from the market.

participants were required to report whether they agreed that the intervention was effective in reducing the consumption of the targeted product (*effectiveness*), targeted the appropriate group of consumers (*targeting*), was coercive (*coerciveness*), was acceptable for the participant (*acceptability*), if majority of citizens would agree with its implementation (*majority*), and if it would increase inequalities in society (*inequality*). As an attention check, the questionnaire displayed the *effectiveness* question twice (once in the first position and once in the second-lowest position).

Last, we asked participants whether they would vote "In favor" or "Against" the implementation (*vote*) of each of the proposed interventions individually (compared to the status quo, where nothing is done).

2.4 Results

2.4.1 Sample

We ran the survey on the online platform Prolific in February and early March 2020.[5] To be eligible for the study, participants had to be born in the UK, to be UK nationals, and to have English as their native language. About 36,000 participants on the platform fulfilled our selection criteria (checked in early 2021) and were considered as active (i.e., in the last 90 days). All Prolific participants who fulfilled the above selection criteria could take part in our study until it reached the maximum number of participants. We defined an exclusion rule prior to the survey with Prolific by asking the *effectiveness*

[5] Prolific is an online platform similar to Amazon MTurk, where people can subscribe to complete tasks for payment. Unlike Amazon MTurk, Prolific is mostly used for research purposes, and previous works concluded that Prolific yields better data than other online survey platforms (e.g., less dishonest participants, higher success rate in attention checks) (Peer et al., 2017).

question twice and removing participants who gave significantly different answers to the same series of questions.[6] We retained the answers to the first *effectiveness* question in the analysis.

In total, 600 participants completed the study and passed the attention check (200 respondents per topic). As it was not the focus of this study, we did not ask for demographics during the survey to avoid cross-contamination, but still exported the data previously gathered by Prolific. We were able to retrieve the demographics for 198 participants in SUGAR, 191 participants in PALM, and 193 participants in EGGS (some participants revoked their consent to transmit the data to the researchers). We further asked Prolific for data on the BMI scores.

Descriptive statistics of the three samples are displayed in Table 2.4.1. We observed no statistical differences across topics regarding the sample composition: All Cohen's d statistics were below 0.12, and none of the mean comparison tests (t-tests or proportion tests) rejected the null hypothesis. Participants were, on average, 35 years old (M = 34.67, SD = 11.45), mostly female (M = 0.73, SD = 0.44), and had a job (M = 0.72, SD = 0.45). One out of five respondents was a student (M = 0.20, SD = 0.40). The distribution of BMI scores was concentrated between 20 and 30 (51.92%). About one out of twenty participants was underweight (BMI < 20: 4.90%), and one out of five participants could be classified as obese (BMI > 30: 19.58%). In addition, one out of five participants refused to report or did not know their BMI (21.50%). Compared to the overall UK population, the sample was younger and more feminine and had a higher share of students.

2.4.2 Descriptive statistics

We begin by discussing the policy *acceptability* scores (Figure 2.4.1). First, we observed very high and similar levels of acceptability across the three topics for *labels* and *information campaigns*. Second, the level of *acceptability* decreased with the degree of *coerciveness*. We observed the highest *acceptability* scores for *labels* and *information campaigns*, followed by *tax10*. The lowest *acceptability* scores were for *withdraw* in SUGAR and for *tax50* in PALM and EGGS [7] Third, interventions regulating palm oil and cage eggs displayed similar *acceptability* levels, but were more accepted than those related to sugar consumption (*tax10, tax30, tax50, withdraw*).

The findings for the hypothetical votes (Figure 2.4.2) were very similar to those of *acceptability*, and *vote* was strongly correlated ($\hat{\rho} = 0.65$, $p < 0.001$, $N = 3600$, pooled data). The multivariate analysis below clusters the observations at the individual level to take into account the repeated observations in the data). We observed similar patterns, and some of the above figures are even more salient. The highest shares of votes were

[6] We computed the average absolute deviation in answers given to the *effectiveness* questions for the six policies under consideration: $\frac{1}{6}\sum_{j=1}^{6} \text{abs}(\text{eff}_{1j} - \text{eff}_{2j})$. Answers could take values between 1 and 7. We dropped participants whose average absolute deviations were greater than 2 (30 participants in SUGAR, 32 in PALM OIL, and 24 in EGGS).

[7] We address the differences between topics in the Discussion.

Table 2.4.1: Descriptive statistics for Demographics.

	Descriptive Statistics				Effect size (Cohen's d) and mean comparison (p-values)		
	All	Sugar	Palm	Eggs	Sugar = Palm	Sugar = Eggs	Palm = Eggs
Age	34.67	35.021	35.164	33.826	$d = 0.012$	$d = 0.112$	$d = 0.113$
	(11.45)	(10.694)	(12.973)	(10.549)	$p = 0.906$	$p = 0.272$	$p = 0.271$
Female	0.73	0.736	0.709	0.742	$d = 0.06$	$d = 0.014$	$d = 0.074$
	(0.44)	(0.442)	(0.455)	(0.439)	$p = 0.559$	$p = 0.887$	$p = 0.47$
Student	0.20	0.197	0.201	0.2	$d = 0.01$	$d = 0.008$	$d = 0.003$
	(0.40)	(0.399)	(0.402)	(0.401)	$p = 0.919$	$p = 0.939$	$p = 0.979$
Job	0.72	0.699	0.72	0.732	$d = 0.044$	$d = 0.071$	$d = 0.027$
	(0.45)	(0.46)	(0.45)	(0.444)	$p = 0.665$	$p = 0.486$	$p = 0.793$
BMI < 20	4.90%	3.63%	5.82%	5.26%			
$20 \geq$ BMI ≥ 24.9	26.92%	28.50%	23.28%	28.95%			
$25 \geq$ BMI ≥ 29.9	25.00%	23.83%	25.93%	25.26%			
$30 \geq$ BMI ≥ 34.9	11.36%	10.88%	13.23%	10.00%		$\chi^2 = 11.26$	
$35 \geq$ BMI ≥ 39.9	4.37%	5.18%	5.29%	2.63%		p =0.666	
$40 \geq$ BMI	3.85%	3.11%	2.12%	6.32%			
Refused to share	21.50%	22.28%	22.75%	19.47%			
BMI missing	2.10%	2.59%	1.59%	2.11%			
N	572	193	189	190			

Notes: (1) The figures here are the empirical means, with standard deviations in parentheses. (2) Absolute Cohen's d values are reported. (3) Two-group mean comparison tests correspond to t-tests for Age and to proportion tests for Female, Student, and Job. (4) The total sample contains 600 participants. The figures show the descriptive statistics for the entire sample. Because of the missing values in some demographics, the final sample used for the regressions consists of 193 complete data for Sugar, 189 for Palm Oil, and 190 for Eggs.

found in *label* and *information campaign*, followed by *tax10*. Here, the lowest share of votes was also for *withdrawal* for Sugar and for *tax50* for Palm and Eggs. Moreover, only 20% of the participants supported the withdrawal of the products with high sugar content compared to 60% for palm oil and cage eggs. Last, a majority of respondents rejected the highest level of taxation for palm oil and cage eggs, but they were more than 60% likely to accept the withdrawal from the market.

Figure 2.4.1: Policy acceptability: averages and 95% confidence intervals (spikes).

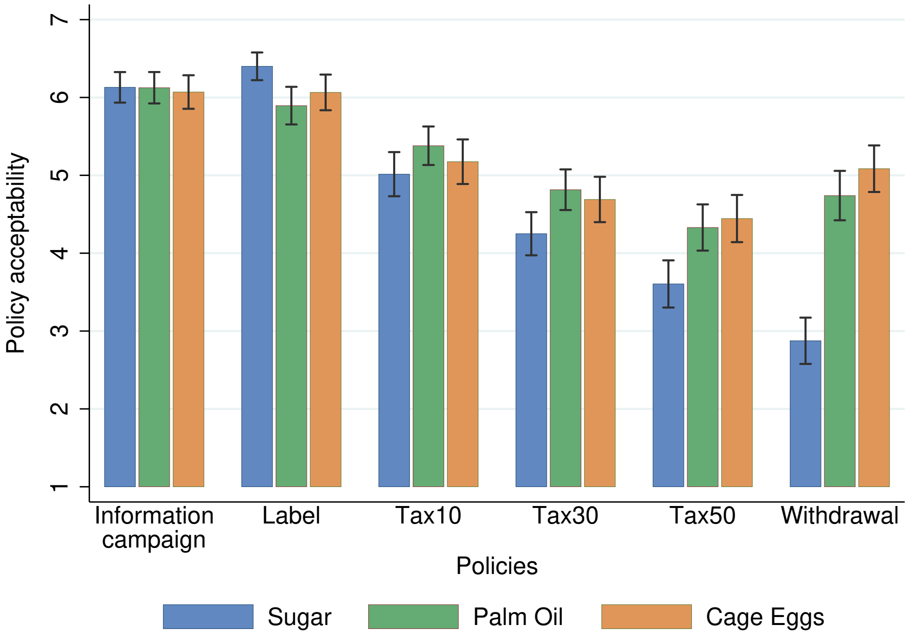
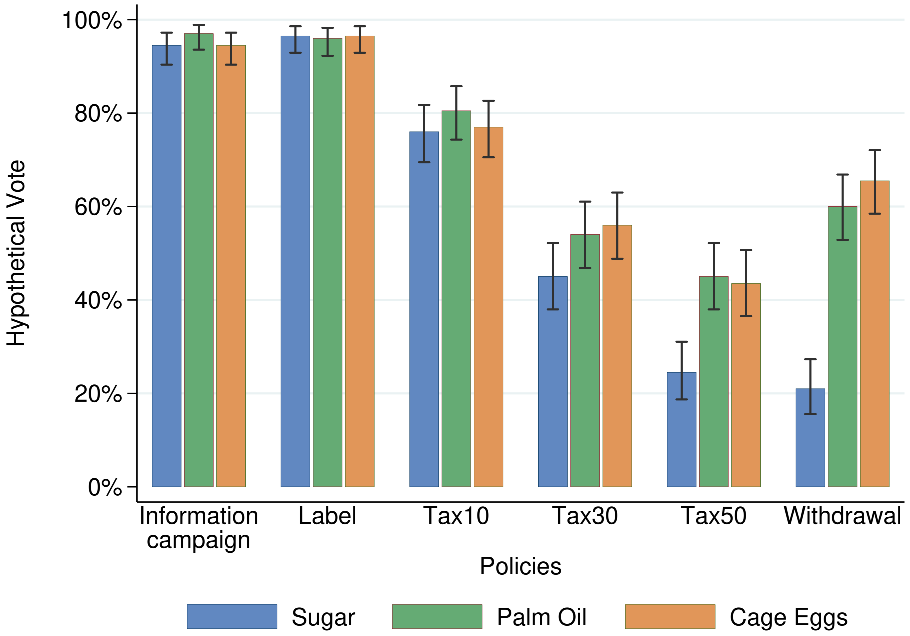


Figure 2.4.2: Hypothetical votes: averages and 95% confidence intervals (spikes).



We now consider the topic-level factors, whose descriptive statistics are displayed in Table 2.4.2. First, we can see that for the three topics under scrutiny, at the aggregate level, participants tended to agree that (i) it is legitimate to intervene (*legitimacy*), (ii) there are negative effects associated with the targeted product (*awareness*), (iii) the product is over-consumed compared to scientific recommendations (*scientific norm*), and (iv) it is commonly accepted that consumption should be reduced (*social norm*). Indeed, all averages were greater than four points (the middle of the scale). Second, we can see that the intervention of the state was as legitimate for Palm as for Eggs ($p = 0.577$), but it was significantly lower for Sugar ($p < 0.001$). On the contrary, the perception of a *scientific norm* calling for a reduction in consumption was significantly greater for Sugar than for Palm Oil ($p = 0.001$) and Eggs ($p = 0.002$), and was similar for the two latter ($p = 0.939$). Consumers displayed the largest *awareness* of negative effects for sugar consumption (M = 6.04, SD = 1.13), followed by palm oil (M = 5.44, SD = 1.33) and cage eggs (M = 4.58, SD = 1.54). Similarly, people were more likely to consider that the product was over-consumed for sugar (M = 6.37, SD = 0.98) than for palm oil (M = 5.58, SD = 1.23) and cage eggs (M = 5.28, SD = 1.28).

Table 2.4.2: Descriptive statistics for topic level factors.

	Descriptive Statistics			Effect size (Cohen's d) and Wilcoxon rank-sum tests (p-values)		
	Sugar	Palm	Eggs	Sugar = Palm	Sugar = Eggs	Palm = Eggs
Legitimacy	5.13	5.74	5.81	d=0.397	d=0.425	d=0.056
	(1.73)	(1.25)	(1.43)	p =0.001	p <0.001	p =0.193
Awareness	6.04	5.44	4.58	d=0.486	d=1.084	d=0.600
	(1.13)	(1.33)	(1.54)	p <0.001	p <0.001	p <0.001
Scientific norm	6.09	5.68	5.69	d=0.342	d=0.314	d=0.008
	(1.17)	(1.23)	(1.37)	p <0.001	p =0.002	p =0.573
Social norm	6.37	5.58	5.28	d=0.71	d=0.955	d=0.239
	(0.98)	(1.23)	(1.28)	p <0.001	p <0.001	p =0.019
N	200	200	200			

Notes: (1) The figures for the descriptive statistics are the empirical means, with standard deviations in parentheses. (2) Absolute Cohen's d values are reported. (3) All respondents are considered here (i.e., 200 per treatment), including participants with missing demographics (which are excluded from the regression analyses below).

We analyzed the relationship between the topic-level factors and acceptability. To do so, we computed the average of the *acceptability* scores given to each of the six policies. We then correlated it with the topic-level factors (see Table 2.A.1 in Appendix 2.A). We observed positive and statistically significant correlations between the four topic-level factors and acceptability. *Legitimacy* had the strongest correlation ($\hat{\rho} = 0.449$). *Awareness*, *scientific norm*, and *social norm* also displayed positive and significant correlations close to $\hat{\rho} = 0.21$.

Regarding the policy-level items, we observed significant differences across topics (see Table 2.4.3). Intervening against sugar consumption was perceived as less effective ($p < 0.001$), less legitimate ($p < 0.001$), and as having a lower social support ($p < 0.005$) than interventions regulating palm oil and cage eggs. However, we did not find significant differences across topics in the average level of perceived coerciveness and the risks of generating inequalities. Regarding Palm and Eggs, we only found a weak statistical difference regarding the effectiveness of the policies ($p = 0.061$): Regulating the consumption of cage eggs in snacks was expected to be more effective than for palm oil.

All policy-level items significantly correlated with policy acceptability (pooled correlation, $p < 0.001$, $N = 3,600$). Public interventions were more likely to be accepted when a respondent believed that a majority of citizens would support it ($\hat{\rho} = 0.576$), when they were perceived as effective ($\hat{\rho} = 0.122$), and when they were thought to target the appropriate group of consumers ($\hat{\rho} = 0.112$). On the contrary, policies were less likely to be accepted when they were seen as more coercive ($\hat{\rho} = -0.134$) and as generating more inequalities ($\hat{\rho} = -0.222$).

Table 2.4.3: Descriptive statistics for policy level factors averaged over the six policies.

	Descriptive Statistics			Effect size (Cohen's d) and Wilcoxon rank-sum tests (p-values)		
	Sugar	Palm	Eggs	Sugar = Palm	Sugar = Eggs	Palm = Eggs
Effective	4.35	4.70	4.88	d =0.366	d =0.541	d =0.188
	(0.98)	(0.93)	(0.98)	p <0.001	p <0.001	p =0.037
Targeting	4.08	4.52	4.68	d =0.421	d =0.557	d =0.142
	(1.07)	(1.06)	(1.08)	p <0.001	p <0.001	p =0.119
Coercive	4.09	3.98	4.10	d =0.098	d =0.007	d =0.102
	(1.11)	(1.26)	(1.20)	p =0.489	p =0.835	p =0.375
Majority	4.03	4.36	4.30	d =0.35	d =0.287	d =0.049
	(0.90)	(0.98)	(1.04)	p =0.001	p =0.008	p =0.591
Inequality	3.21	3.21	3.21	d <0.001	d <0.001	d <0.001
	(1.25)	(1.23)	(1.34)	p =0.93	p =0.94	p =0.994
N	200	200	200			

Notes: (1) The figures for the descriptive statistics are the empirical means, with standard deviations in parentheses. (2) Absolute Cohen's d values are reported. (3) All respondents are considered here (i.e., 200 per treatment), including participants with missing demographics (which are excluded from the regression analyses below).

2.4.3 Multivariate analysis

To identify the determinants of the acceptability of food policies, we estimated the following mixed linear model on *acceptability* and *vote*:

$$y_{ik} = \alpha_0 + X_i\alpha_1 + Z_{ik}\alpha_2 + \mu_i + \phi_k + \epsilon_{ik} \tag{2.1}$$

where y_{ik} is either the *acceptability* or *vote* of individual i for policy k, X_i is the vector of variables defined at the individual level (topic-level factors: *legitimacy, awareness, scientific norm*, and *social norm*) and demographics (gender, employment status, student status, age, BMI), and Z_{ik} is the vector of variables defined at the policy–individual level (policy-level factors: *effectiveness, targeting, coerciveness, majority, inequality*). The terms μ_i are normally distributed individual effects, ϕ_k are (fixed) policy effects, and ϵ_{ik} are idiosyncratic error terms. The pooled regression further included dummy variables for the topics. The results presented below are robust to non-linear specifications (ordered probit and probit specifications, see Table 2.A.3 in Appendix 2.A).

Estimates of the *acceptability* and *vote* are displayed in Tables 2.4.4 and 2.4.5, respectively. We observed that *acceptability* and *vote* correlated with similar factors. The more one believes that it is legitimate to have collective rules for the topic under scrutiny (*legitimacy*), the more acceptable public interventions are. Policies that are expected to be more effective (*effectiveness*), to target the appropriate group of consumers (*targeting*), and to be accepted by the a majority of fellow citizens (*majority*) are also more likely to be accepted. On the contrary, coercive policies and those that are perceived as a source of inequalities in the society are less likely to be supported (*coerciveness* and *inequality*, respectively). On the other hand, the *scientific norm* did not significantly correlate with the level of *acceptability* in the multivariate analysis. Two variables showed weaker evidence. *Social norm* positively and significantly correlated with *acceptability*, but only partially with *vote* (significant for the pooled data and EGGS, but not for others). The second variable that showed weaker associations was *awareness.* It positively and significantly correlated with *vote* in all specifications, but not for *acceptability* (the association was significant only in SUGAR), which we discuss in the next section.

Since we estimated a linear model in which all variables had the same range of values, we can compare the intensity of the contribution of each factor to the policy acceptability. The most influential factor is the expected majority support for a specific policy: An additional point in *majority* is associated with an increase of 0.5 points in *acceptability* and an increase of eight percentage points in *vote. Legitimacy* is the second-largest contributor to *acceptability* (+0.26) and *vote* (+4 pp). The policy's effectiveness is estimated to be the third-largest contributor (*individual acceptability*: +0.17, *vote*: +3 pp). Perceived increases in inequalities and coercion have overall similar negative impacts on *individual acceptability* (−0.07) and *vote* (−1.5 and −1.3 pp respectively). The appropriate targeting of the policy has a similar but positive impact on *acceptability* (+0.12) and *vote* (+1.3 pp).

Last, we ran a backward model selection analysis to understand which set of variables best explained *acceptability* and *vote.* To do so, we started with the full models (first column of Tables 2.4.4 and 2.4.5) and removed the least significant variable. We then re-estimated our model with the new subset of variables and repeated the exclusion process. We stopped when all variables were significant at the 5% level. We computed the AIC and BIC scores at each step of the process. Control variables, individual random effects, and policy fixed effects were maintained in all specifications. The results suggest that a seven-component model fits the data best for *acceptability* compared to an eight-component model for *vote.* The variable *scientific norm* was excluded for the two dependent variables, and *awareness* was rejected for *acceptability*, but accepted for *vote.*

Table 2.4.4: Regression of policy acceptability scores.

	All	Sugar	Palm Oil	Eggs
Legitimacy	0.227***	0.174***	0.212***	0.300***
	(0.0329)	(0.0422)	(0.0738)	(0.0647)
Awareness	0.0338	0.167**	0.0298	-0.0490
	(0.0382)	(0.0676)	(0.0701)	(0.0634)
Scientific norm	-0.00493	-0.0589	0.000393	0.0426
	(0.0369)	(0.0556)	(0.0693)	(0.0691)
Social norm	0.133***	0.0485	0.147**	0.136**
	(0.0392)	(0.0661)	(0.0729)	(0.0672)
Effective	0.169***	0.172***	0.137***	0.156***
	(0.0203)	(0.0314)	(0.0373)	(0.0359)
Targeting	0.122***	0.105***	0.114***	0.156***
	(0.0197)	(0.0309)	(0.0346)	(0.0355)
Coercive	-0.0704***	-0.111***	-0.0355	-0.0349
	(0.0168)	(0.0262)	(0.0286)	(0.0309)
Majority	0.498***	0.511***	0.467***	0.435***
	(0.0165)	(0.0277)	(0.0272)	(0.0308)
Inequalities	-0.0707***	-0.0252	-0.0961***	-0.111***
	(0.0159)	(0.0248)	(0.0272)	(0.0307)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	No	No	No
Number of individuals	572	193	189	190
Number of policies	6	6	6	6
Log-likelihood	−6243.19	−2034.42	−2012.94	−2104.98
Observations	3,432	1,158	1,134	1,140

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

2.4.4 Replication: confirmatory analysis

In order to test the robustness of our results, we proceeded to a statistical replication (see Duvendack et al., 2017). To do so, we pre-registered the above findings (AEARCTR-0006429) and collected new data. We specified in the pre-registration that we would invite 220 individuals to participate in our study for each topic. We further committed to using the same recruitment platform (Prolific). We applied the same screening criteria and collected the data on 16 September 2020. We used the same exclusion rule as for the main analysis, which we also specified in the pre-registration protocol. The new sample of participants was statistically similar regarding age, student status, and job (see Table

Table 2.4.5: Regression of hypothetical votes in favor of the policies.

	All	Sugar	Palm Oil	Eggs
Legitimacy	0.0314***	0.0193**	0.0266*	0.0484***
	(0.00630)	(0.00863)	(0.0155)	(0.0106)
Awareness	0.0239***	0.0362***	0.0285*	0.0174*
	(0.00732)	(0.0138)	(0.0148)	(0.0104)
Scientific norm	0.000300	-0.00829	0.00793	-0.00503
	(0.00706)	(0.0114)	(0.0146)	(0.0113)
Social norm	0.0165**	-0.00762	0.00874	0.0326***
	(0.00750)	(0.0135)	(0.0153)	(0.0110)
Effective	0.0301***	0.0282***	0.0281***	0.0311***
	(0.00484)	(0.00751)	(0.00922)	(0.00842)
Targeting	0.0132***	0.0169**	0.0104	0.0140*
	(0.00467)	(0.00734)	(0.00854)	(0.00816)
Coercive	-0.0129***	-0.0232***	-0.00579	-0.00387
	(0.00390)	(0.00616)	(0.00695)	(0.00688)
Majority	0.0785***	0.0773***	0.0703***	0.0752***
	(0.00389)	(0.00660)	(0.00670)	(0.00695)
Inequalities	-0.0152***	-0.00634	-0.0191***	-0.0224***
	(0.00368)	(0.00583)	(0.00665)	(0.00662)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	No	No	No
Number of individuals	572	193	189	190
Number of policies	6	6	6	6
Log-likelihood	−1292.22	−374.26	−420.45	−430.00
Observations	3,432	1,158	1,134	1,140

Notes: (1)The figures displayed are the estimated coefficients, standard errors in brackets. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

2.A.6 in Appendix 2.A). Participants in this second study were slightly less female (63% vs. 73%) and had slightly lower BMI scores (BMI < 30: 58.4% vs. 54.5%).

We pre-registered that all variables but *scientific norm* were significantly associated with the acceptability of food policies. The results are displayed in Table 2.4.6 and show that we obtained similar results in the replication study (Tables 2.A.7 and 2.A.8 in Appendix 2.A show the full results). For the *acceptability* score, all variables but *scientific norm* were significant in the pooled analysis. Regarding *vote*, we obtained a similar pattern, except for the *social norm*, which was not significant. Finally, we can note that the average levels of *acceptability* showed similar patterns to those obtained previously.

We reproduced Figures 2.4.1 and 2.4.2 with the new data in the supplementary figures Appendix 2.B.

Table 2.4.6: Results of the replication study: original and replicated estimates.

	Acceptability		Hypothetical vote	
	Original	Replication	Original	Replication
Legitimacy	0.227***	0.170***	0.0314***	0.0177**
	(0.0329)	(0.0423)	(0.00630)	(0.00797)
Awareness	0.0338	0.151***	0.0239***	0.0391***
	(0.0382)	(0.0482)	(0.00732)	(0.00906)
Scientific norm	-0.00493	0.0268	0.000300	0.0148*
	(0.0369)	(0.0441)	(0.00706)	(0.00828)
Social norm	0.133***	0.134***	0.0165**	0.0138
	(0.0392)	(0.0510)	(0.00750)	(0.00957)
Effective	0.169***	0.218***	0.0301***	0.0363***
	(0.0203)	(0.0188)	(0.00484)	(0.00466)
Targeting	0.122***	0.110***	0.0132***	0.0139***
	(0.0197)	(0.0178)	(0.00467)	(0.00438)
Coercive	-0.0704***	-0.101***	-0.0129***	-0.0185***
	(0.0168)	(0.0159)	(0.00390)	(0.00383)
Majority	0.498***	0.467***	0.0785***	0.0685***
	(0.0165)	(0.0176)	(0.00389)	(0.00429)
Inequalities	-0.0707***	-0.0488***	-0.0152***	-0.0136***
	(0.0159)	(0.0162)	(0.00368)	(0.00389)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	Yes	Yes	Yes
Number of individuals	572	588	572	588
Number of policies	6	6	6	6
Log-likelihood	−6243.19	−6357.29	−1292.22	−1418.71
Observations	3,432	3,528	3,432	3,528

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

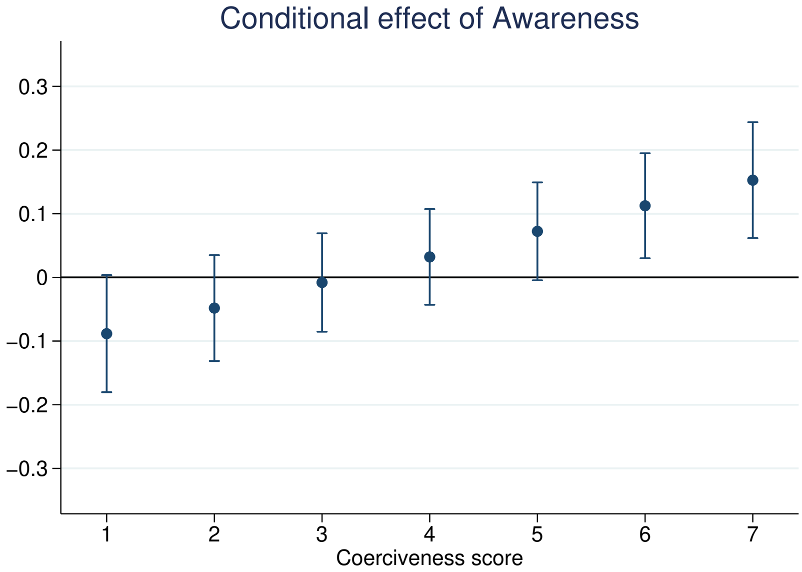
2.5 Discussion

2.5.1 Factors

All but one (scientific norm) of the nine factors tested correlate with acceptability judgments or hypothetical voting behavior. At the topic level, the perceived legitimacy of having collective rules to regulate the product at stake and favorable social norms are associated with stronger support for public intervention. At the narrower policy level, policies that are expected to have majority support, that target the appropriate group of consumers, that do not generate inequalities, and that are less coercive benefit from a larger acceptability. Our results concur with previous conclusions on food and health policy acceptability (Diepeveen et al., 2013; Mazzocchi et al., 2015; Reynolds et al., 2019), as well as results from other fields (Schade and Schlag, 2003; Steg and Schuitema, 2007; De Groot and Schuitema, 2012). Most importantly, they also validate the construct of these underlying factors having an effect at both the topic and policy levels and that this effect is consistent and valid for different topics of food interventions.

Surprisingly, in the first study, *awareness* did not significantly correlate with *acceptability* in the multivariate analysis. We explored the possibility that the effect of awareness on policy acceptability depends on the type of policy. To do so, we regressed the acceptability scores and hypothetical votes for each type of policy separately (Tables 2.A.4 and 2.A.5 in the Appendix 2.A). We observed that *awareness* is the only factor that displayed opposite and significant associations with *acceptability* and *vote*, depending on the policy. Aware consumers are more likely to support coercive measures, such as *tax50* or *withdrawal*, but they are also significantly less likely to support the implementation of *labels* or *information campaigns*. A possible explanation is that consumers who are aware of the problem do not consider labels coercive enough to address the problem and thus oppose their implementation. As a result, pooled regressions would lead to an average null effect of *awareness*, which would fail to take into account the underlying dynamic. To confirm this hypothesis, we investigated whether the impact of *awareness* depends on the coercion level. We ran a pooled regression similar to those presented in the previous section, but added an interaction term between *awareness* and *coerciveness*. The estimated relationship between *awareness* and *acceptability* is displayed in Figure 2.5.1 and confirms the aforementioned hypothesis: Awareness increases the acceptability of policies only if they are sufficiently coercive. This result was also pre-registered and successfully replicated (see Figure 2.B.3).

Figure 2.5.1: Conditional effect of awareness on policy acceptability.



Notes: Results of a linear regression of acceptability with policy fixed effects, topic fixed effects individual random effects, all identified acceptability factors, demographics, and an interaction term between awareness and coerciveness.

Our work also considered factors that affect the fairness of public interventions. Asking participants whether a policy is fair in their view is rather uninformative, as many of them might consider fair and acceptable as synonyms. To address this concern, we investigated two important dimensions that relate to fairness: whether the proposed policy targets the appropriate group of consumers and whether it generates inequalities in society. These two dimensions cover important aspects of fairness discussed by psychologists and economists, while ruling out the possibility of participants confounding the factors with the acceptability outcome (Alesina and Angeletos, 2005). Both inequality and appropriate targeting were reported to significantly correlate with acceptability judgements, confirming what other authors proposed (Caraher and Cowburn, 2005; Bos et al., 2013), and supporting their relevance for future research.

Noticeably, *scientific norm* was the only factor for which we found no significant correlation with acceptability. Participants who agreed that consumption of the targeted product was higher than the recommendation of most recent scientific works did not consider policies to address the issue more acceptable. The benefits of using the scientific norm to support dietary changes therefore seemed limited. Importantly, we did not provide participants with informative scientific evidence, which might have yielded different results. Our results can only be interpreted as the use of scientific work as a form of authority. This result could be understood in the context of previous studies that have shown that simply telling people what to do is ineffective and can trigger reactance (Stok et al., 2014).

2.5.2 Policies

Our survey included six policies: labeling, information campaigns, taxing at low, intermediate, or high levels, and withdrawing the products from the market. Among these policies, we observed that labels and information campaigns benefit from a large acceptability: More than 90% of the participants would support the implementation of labels or information campaigns for the three topics under scrutiny. Policies that are not coercive and consist mainly of informing the population without changing the choice structure appear to be considered overwhelmingly acceptable, as reported by Mazzocchi et al. (2015) for food and by Diepeveen et al. (2013) for other health policies. Citizens are supportive of policies designed to give them the best tools to make informed decisions. On the contrary, the intermediate and high taxation levels showed significantly lower support, with the average level of acceptability decreasing with taxation intensity. It was equal to 5.19 for *tax10* versus 4.13 for *tax50* (t-test: $p < 0.001$).

This trend corroborates the established trade-off that public authorities face between coerciveness and popular support. Coercive measures are appealing to policy-makers, as they are more likely to be effective, which participants also acknowledge. *Tax10* was, for example, perceived as less coercive than *tax50* (t-test: $p < 0.001$), but also as less effective (t-test: $p < 0.001$). Across policies and topics, we observed a strong correlation between reported coerciveness and reported effectiveness ($\hat{\rho} = 0.427$, $p < 0.001$). Policy-makers must therefore choose the appropriate level of coercion that maximizes the change in behaviors while maintaining a sufficiently high level of acceptability for the population. In this paper, we did not investigate the actual capacity of a policy to change behaviors. However, it is worth noting that recent evidence indicates that certain information interventions, which are considered widely acceptable, can affect food consumption behaviors in the short and mid term (Jalil et al., 2020). When possible, it is also up to the policy-maker to consider nudges, which can, under certain circumstances, overcome the trade-off between coerciveness and effectiveness. Promising research has also already started in this direction (Garnett et al., 2019; Hansen et al., 2019).

2.5.3 Topics

We observed similar patterns in the *acceptability* and *vote* scores across topics. A notable exception was the change in acceptability between *tax50* and *withdrawal*. A priori, one could reasonably expect a lower acceptability for *withdrawal* than for *tax50*, as removing the possibility to buy the product should be judged as more coercive than taxing it. However, this decrease in acceptability was observed for the SUGAR survey only, and people were more likely to accept the withdrawal than the highest level of taxation for PALM and EGGS. While participants indeed perceived the withdrawal as more coercive than taxation for all topics, we observed that their views diverged regarding its *effectiveness*: The effectiveness gain of shifting from high taxation to withdrawal was smaller for Sugar than for Palm and Eggs (Table 2.A.2). In addition, participants were more likely to think that a majority of citizens would support the withdrawal than for *tax50* in Palm and Eggs. Regarding sugar, we observed the opposite: Participants expected a larger social consensus for *tax50* than for the withdrawal. We observed a similar pattern in the confirmatory analysis.

These discrepancies can be explained by the fact that sugar differs from palm oil and cage eggs in terms of its impact. The negative consequences of consuming sugar are mostly borne by consumers themselves (increased risk of mortality or diseases), while they are mainly borne by others in the cases of palm oil (deforestation) and cage eggs (animals). Consuming sugar is thus more likely to be perceived as a personal choice, unlike palm oil and cage egg consumption, which can be seen as mainly a social issue. This idea is supported by the data, as we observed a lower legitimacy for interventions regarding sugar. In this case, coercive measures for sugar could be considered as a way to protect consumers from themselves, which can be perceived as paternalistic, while they would be considered as a way to protect others in the cases of cage eggs and palm oil.

2.5.4 Future directions

Future research can build on this work in three main ways. First, our study focused on two levels of a four-level food policy acceptability model, as we excluded perceptions of the central government, judging them to be hardly actionable for policy-makers, and did not seek more individual information from our participants than what was already available. Future researchers would still benefit from incorporating these macro- and microscopic factors into their analyses. Cross-country comparisons could, for example, provide interesting insights on the variation of the effect of such factors in countries with different institutions and social capital.

Second, our research did not specifically measure the behavioral costs of policies. Policies with similar designs and pursuing the same objectives might still have very different behavioral implications. For example, citizens might have different acceptability judgements regarding the banning of meat compared to the banning of palm oil, even though both policies would help prevent deforestation. Other factors than those that we identified here might come into play. These could include citizens' personal preferences, such as attachment to a certain part of their diet (Graça et al., 2015), or the ease of substitution with other products.

Third, our work reports correlations between factors and acceptability judgments, which does not allow us to rule out reverse causality. Motivated beliefs (see Bénabou, 2015) can, for example, lead a participant who dislikes a policy to judge it as ineffective. Similarly, the false consensus effect can influence perceptions of the majority's opinion (Ross et al., 1977). This would lead to a strong correlation between expected majority support and personal judgment of acceptability, which we observed. Importantly, a key element for the policy-maker is to determine the possibility of using the identified factors as tools to increase the acceptability of a policy. Future research should therefore intend to determine the causal relationships between factors and acceptability, as well as the most effective factors for influencing acceptability. Influencing topic-level factors might generate horizontal spillovers (i.e., increasing the acceptability for all types of policies for a given topic), while influencing policy-specific factors might generate vertical spillovers (i.e., increasing the acceptability of similar policies for different topics). Identifying spillovers could be of great interest for policy-makers and advocates of healthier and more sustainable diets to tailor their communication strategies and identify key leverages. Horizontal spillovers could be valuable for nutrition experts when they seek to improve the general acceptability of interventions for one specific topic with a possible range of policy options. Vertical spillovers would be useful, for instance, for policy-makers that seek to internalize externalities in food consumption with taxation across different topics.

2.5.5 Limitations

Future works could investigate the robustness of our results in terms of two dimensions. First, the participants in our study were not representative of the UK population. As discussed in Section 2.4.1, our participants were indeed younger and more feminine than the average UK population, and are more likely to be students. We cannot rule out the possibility that other types of participants would weigh the factors differently when they consider the acceptability of food policies. Second, the framework of our experiment is relatively general, as is the case with most experiments, and participants could view policies in a different way when they arise in a specific context. For instance, the support for the taxation of highly sweetened snacks could depend on other pre-existing policies. Replicating these surveys before and after the actual implementation of food policies could thus provide new perspectives.

Appendix

Chapter 2 Appendix

2.A Supplemtary tables

Table 2.A.1: Correlation between topic level factors and average acceptability.

	Acceptability (PCA)			
	SUGAR	PALM	EGGS	ALL
Legitimacy	0.469	0.427	0.396	0.449
	[p<0.001]	[p<0.001]	[p<0.001]	[p<0.001]
Awareness	0.389	0.326	0.235	0.207
	[p<0.001]	[p<0.001]	p<0.001	[p<0.001]
Scientific norm	0.127	0.312	0.308	0.220
	[p=0.074]	[p<0.001]	[p<0.001]	[p<0.001]
Social norm	0.299	0.396	0.228	0.210
	[p<0.001]	[p<0.001]	[p=0.001]	[p<0.001]

Notes: (1) The figures here are the estimated correlation coefficients, with p-values in brackets. (2) Acceptability obtained by taking the average of the acceptability scores over all policies.

Table 2.A.2: Summary statistics of tax 50 and withdrawal.

	Tax 50			Withdrawal		
	Sugar	PalmOil	Eggs	Sugar	PalmOil	Eggs
Acceptability	3.61	4.33	4.45	2.88	4.74	5.09
Vote	.25	.45	.44	.21	.6	.66
Effectiveness	4.99	5.34	5.39	5.36	6.06	6.32
Targeting	4.54	5.09	5.11	5.3	5.82	6.08
Coerciveness	4.83	4.59	4.79	5.32	5.05	5.31
Majority	2.37	3.05	2.8	2	3.13	3.34
Inequalities	4.76	4.55	4.35	3.18	3.18	3.71

Notes: The figures here are the empirical means.

Table 2.A.3: Robustness check: linear and non-linear specifications.

	Acceptability		Hypothetical vote	
	Linear (Original)	Ordered Probit	Linear (Original)	Probit
Legitimacy	0.227***	0.225***	0.0314***	0.182***
	(0.0329)	(0.0342)	(0.00630)	(0.0367)
Awareness	0.0338	0.0410	0.0239***	0.136***
	(0.0382)	(0.0397)	(0.00732)	(0.0430)
Scientific norm	-0.00493	-0.0136	0.000300	0.000192
	(0.0369)	(0.0382)	(0.00706)	(0.0406)
Social norm	0.133***	0.170***	0.0165**	0.0994**
	(0.0392)	(0.0409)	(0.00750)	(0.0434)
Effective	0.169***	0.168***	0.0301***	0.198***
	(0.0203)	(0.0189)	(0.00484)	(0.0309)
Targeting	0.122***	0.113***	0.0132***	0.0885***
	(0.0197)	(0.0184)	(0.00467)	(0.0288)
Coercive	-0.0704***	-0.0824***	-0.0129***	-0.0861***
	(0.0168)	(0.0162)	(0.00390)	(0.0243)
Majority	0.498***	0.425***	0.0785***	0.385***
	(0.0165)	(0.0161)	(0.00389)	(0.0240)
Inequalities	-0.0707***	-0.0801***	-0.0152***	-0.0842***
	(0.0159)	(0.0145)	(0.00368)	(0.0206)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	Yes	Yes	Yes
Number of individuals	572	572	572	572
Number of policies	6	6	6	6
Log-likelihood	−6243.19	−4657.26	−1292.22	−1214.10
Observations	3,432	3,432	3,432	3,432

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

Table 2.A.4: Regression of acceptability by policy.

	Label	InfoCamp	Tax 10	Tax 30	Tax 50	Withdrawal
Legitimacy	0.0906**	0.0821**	0.369***	0.350***	0.387***	0.118**
	(0.0420)	(0.0404)	(0.0555)	(0.0553)	(0.0581)	(0.0573)
Awareness	-0.120**	-0.0247	-0.103	-0.00459	0.112*	0.342***
	(0.0490)	(0.0470)	(0.0640)	(0.0636)	(0.0670)	(0.0659)
Scientific norm	-0.0412	0.0548	-0.0187	-0.0289	-0.0303	0.0757
	(0.0478)	(0.0456)	(0.0615)	(0.0614)	(0.0645)	(0.0637)
Social norm	0.0718	0.124**	0.119*	0.169***	0.164**	0.171**
	(0.0505)	(0.0486)	(0.0660)	(0.0651)	(0.0686)	(0.0682)
Effective	0.102**	0.202***	0.187***	0.119*	0.248***	0.195***
	(0.0410)	(0.0457)	(0.0545)	(0.0616)	(0.0565)	(0.0536)
Targeting	0.0944**	0.00719	0.166***	0.176***	0.102**	0.127**
	(0.0389)	(0.0406)	(0.0541)	(0.0557)	(0.0498)	(0.0541)
Coercive	-0.0492	-0.0243	-0.0354	-0.0830*	-0.0922**	-0.0945**
	(0.0333)	(0.0287)	(0.0477)	(0.0479)	(0.0460)	(0.0376)
Majority	0.474***	0.419***	0.330***	0.446***	0.456***	0.476***
	(0.0377)	(0.0365)	(0.0424)	(0.0440)	(0.0439)	(0.0413)
Inequalities	-0.107***	-0.0872***	-0.145***	-0.101**	-0.168***	-0.0279
	(0.0379)	(0.0327)	(0.0435)	(0.0404)	(0.0384)	(0.0335)
Demographics	Yes	Yes	Yes	Yes	Yes	Yes
Topic FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	572	572	572	572	572	572
R^2	0.383	0.350	0.318	0.359	0.406	0.521

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

Table 2.A.5: Regression of hypothetical vote by policy.

	Label	InfoCamp	Tax 10	Tax 30	Tax 50	Withdrawal
Legitimacy	0.00775	0.00558	0.0544***	0.0566***	0.0538***	0.0215
	(0.00586)	(0.00648)	(0.0124)	(0.0150)	(0.0141)	(0.0133)
Awareness	-0.0135**	-0.0135*	0.00814	0.0402**	0.0445***	0.0672***
	(0.00684)	(0.00753)	(0.0143)	(0.0172)	(0.0162)	(0.0153)
Scientific norm	-0.00261	-0.00120	0.00898	-0.000777	-0.000688	0.00647
	(0.00667)	(0.00731)	(0.0138)	(0.0167)	(0.0156)	(0.0148)
Social norm	0.0115	0.0140*	0.0248*	0.0145	0.0253	0.0330**
	(0.00704)	(0.00779)	(0.0148)	(0.0177)	(0.0166)	(0.0158)
Effective	0.0120**	0.0189**	0.0435***	0.0404**	0.0369***	0.0255**
	(0.00572)	(0.00732)	(0.0122)	(0.0167)	(0.0137)	(0.0124)
Targeting	0.00122	0.00264	0.00937	0.0113	0.00736	0.0215*
	(0.00542)	(0.00651)	(0.0121)	(0.0151)	(0.0121)	(0.0125)
Coercive	0.00543	0.00216	0.00566	-0.0145	-0.0241**	-0.0274***
	(0.00464)	(0.00460)	(0.0107)	(0.0130)	(0.0111)	(0.00872)
Majority	0.0281***	0.0377***	0.0641***	0.102***	0.0927***	0.0878***
	(0.00526)	(0.00584)	(0.00952)	(0.0119)	(0.0106)	(0.00957)
Inequalities	-0.0118**	0.00679	-0.0185*	-0.0287***	-0.0322***	-0.0181**
	(0.00529)	(0.00523)	(0.00976)	(0.0110)	(0.00931)	(0.00776)
Demographics	Yes	Yes	Yes	Yes	Yes	Yes
Topic FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	572	572	572	572	572	572
R^2	0.138	0.151	0.239	0.254	0.296	0.415

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

Table 2.A.6: Sample comparisons on demographics.

	Descriptive statistics		Effect size (Cohen's d) and mean comparison (p-Values)
	FIRST STUDY	SECOND STUDY	FIRST = SECOND
Age	34.67	35.81	d =0.093
	(11.45)	(12.94)	p =0.113
Female	0.73	0.63	d =0.215
	(0.44)	(0.48)	p <0.001
Student	0.20	0.20	d =0.012
	(0.40)	(0.40)	p =0.839
Job	0.72	0.64	d =0.173
	(0.45)	(0.48)	p =0.003
BMI $<$ 20	4.90%	6.80%	
20 $\geq$ BMI $\geq$ 24.9	26.92%	31.63%	
25 $\geq$ BMI $\geq$29.9	25.00%	20.07%	
30 $\geq$ BMI $\geq$ 34.9	11.36%	8.50%	$\chi^2 = 16.33$
35 $\geq$ BMI $\geq$ 39.9	4.37%	3.91%	$p = 0.022$
40 $\geq$ BMI	3.85%	1.70%	
Don't say	21.50%	25.34%	
BMI missing	2.10%	2.04%	
N	572	588	

Notes: (1) The figures for the descriptive statistics are the empirical means, with standard deviations in parentheses. (2) Absolute Cohen's d values are reported. (3) Two-group mean comparison tests correspond to t-tests for Age and to proportion tests for Female, Student and Job.

Table 2.A.7: Regression of individual acceptability - replication.

	All	Sugar	Palm Oil	Eggs
Legitimacy	0.170***	0.242***	-0.0114	0.155*
	(0.0423)	(0.0539)	(0.0911)	(0.0895)
Awareness	0.151***	0.108	0.222**	0.165**
	(0.0482)	(0.102)	(0.0923)	(0.0754)
Scientific norm	0.0268	0.0239	0.0889	-0.00999
	(0.0441)	(0.0983)	(0.0719)	(0.0716)
Social norm	0.134***	-0.0328	0.219***	0.110
	(0.0510)	(0.116)	(0.0815)	(0.0817)
Effective	0.218***	0.186***	0.188***	0.234***
	(0.0188)	(0.0311)	(0.0318)	(0.0342)
Targeting	0.110***	0.126***	0.137***	0.109***
	(0.0178)	(0.0306)	(0.0296)	(0.0318)
Coercive	-0.101***	-0.107***	-0.0921***	-0.0620**
	(0.0159)	(0.0280)	(0.0262)	(0.0280)
Majority	0.467***	0.556***	0.391***	0.414***
	(0.0176)	(0.0337)	(0.0295)	(0.0296)
Inequalities	-0.0488***	-0.0424	-0.0641**	-0.0758***
	(0.0162)	(0.0283)	(0.0272)	(0.0289)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	No	No	No
Number of individuals	588	192	203	193
Number of policies	6	6	6	6
Log-likelihood	-6357.29	-2054.31	-2160.74	-2073.94
Observations	3,528	1,152	1,218	1,158

Notes: (1)The figures here are the estimated coefficients, with standard errors in parentheses. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

Table 2.A.8: Regression of vote - replication.

	All	Sugar	Palm Oil	Eggs
Legitimacy	0.0177**	0.0351***	-0.00898	0.000990
	(0.00797)	(0.00959)	(0.0172)	(0.0182)
Awareness	0.0391***	0.0239	0.0627***	0.0337**
	(0.00906)	(0.0181)	(0.0174)	(0.0153)
Scientific norm	0.0148*	0.00535	0.0248*	0.00985
	(0.00828)	(0.0173)	(0.0135)	(0.0145)
Social norm	0.0138	0.00386	0.00818	0.0217
	(0.00957)	(0.0205)	(0.0153)	(0.0166)
Effective	0.0363***	0.0300***	0.0246***	0.0474***
	(0.00466)	(0.00732)	(0.00836)	(0.00858)
Targeting	0.0139***	0.00850	0.0200***	0.0201**
	(0.00438)	(0.00719)	(0.00767)	(0.00794)
Coercive	-0.0185***	-0.00918	-0.0168**	-0.0236***
	(0.00383)	(0.00644)	(0.00663)	(0.00689)
Majority	0.0685***	0.0679***	0.0592***	0.0679***
	(0.00429)	(0.00784)	(0.00751)	(0.00730)
Inequalities	-0.0136***	-0.0187***	-0.0177***	-0.0105
	(0.00389)	(0.00648)	(0.00677)	(0.00705)
Demographics	Yes	Yes	Yes	Yes
Individual RE	Yes	Yes	Yes	Yes
Policy FE	Yes	Yes	Yes	Yes
Topic FE	Yes	No	No	No
Number of individuals	588	192	203	193
Number of policies	6	6	6	6
Log-likelihood	−1418.71	−393.14	−513.71	−463.37
Observations	3,528	1,152	1,218	1,158

Notes: (1)The figures displayed are the estimated coefficients, standard errors in brackets. (2) * significant at 10%, ** significant at 5%, *** significant at 1%. (3) Controls include: age, gender, student status, job, and body mass index.

2.B Supplementary figures

Figure 2.B.1: Policy acceptability - replication.

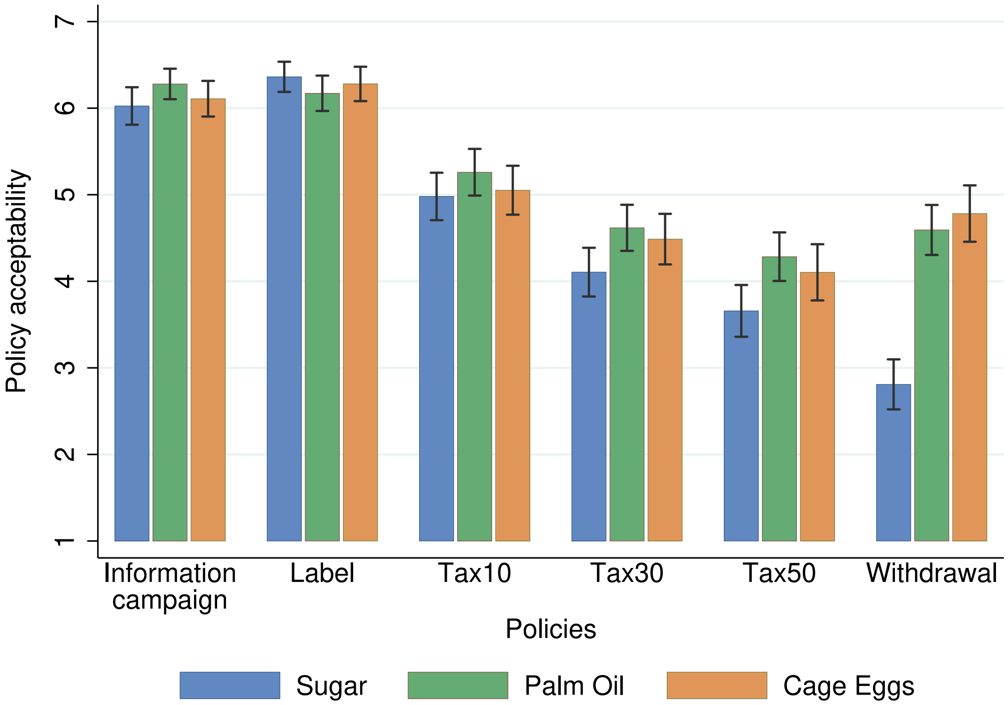


Figure 2.B.2: Hypothetical votes - replication.

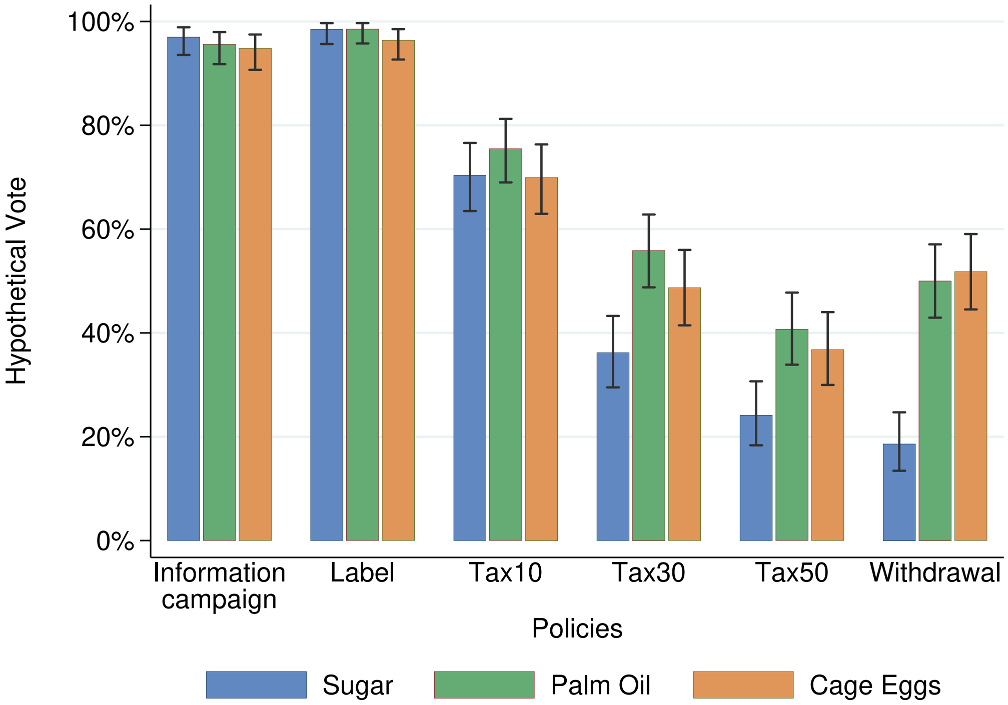
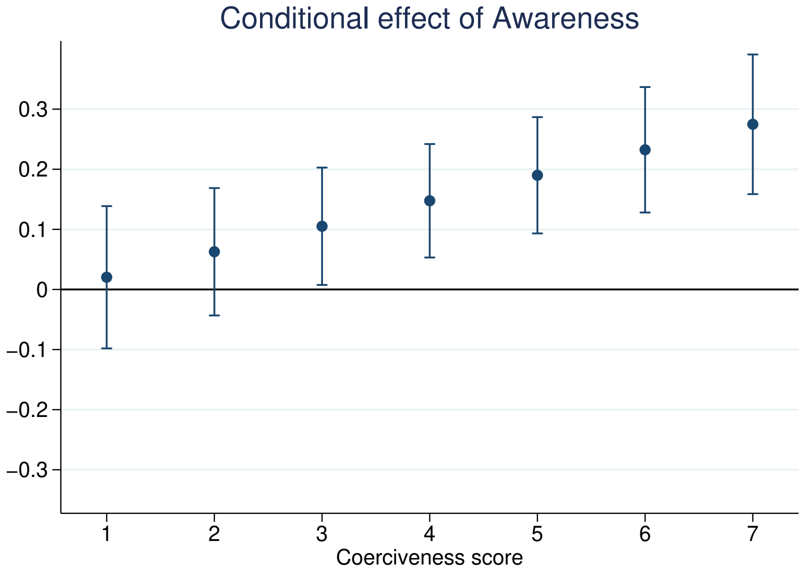


Figure 2.B.3: Conditional effect of awareness on policy acceptability - replication.



Note: Results of a linear regression of acceptability with policy fixed-effects, topic fixed-effects individual random-effects, all identified acceptability factors, demographics, and an interaction term between awareness and coerciveness.

2.C Experimental design

Questionnaire for sugar

In this questionnaire, we are interested in the consumption of snacks that can be bought in supermarkets or in grocery stores. These snacks vary in different aspects, and in particular in the sugar content. We consider here that products have a "high sugar content" if sugar exceeds one-third of the product (i.e., more than 33g of sugar out of 100g of product).

We are interested in six measures that could be implemented to regulate the country's sugar intake through the consumption of snacks:

- Measure 1: display labels on snacks with high sugar content.
- Measure 2: tax the snacks with high sugar content by £0.10 (for a 30g individual snack such as a cereal bar).
- Measure 3: tax the snacks with high sugar content by £0.30 (for a 30g individual snack such as a cereal bar).
- Measure 4: tax the snacks with high sugar content by £0.50 (for a 30g individual snack such as a cereal bar).
- Measure 5: remove the snacks with high sugar content from the market.
- Measure 6: set up information campaigns to inform consumers about the impact of sugar on health and society.

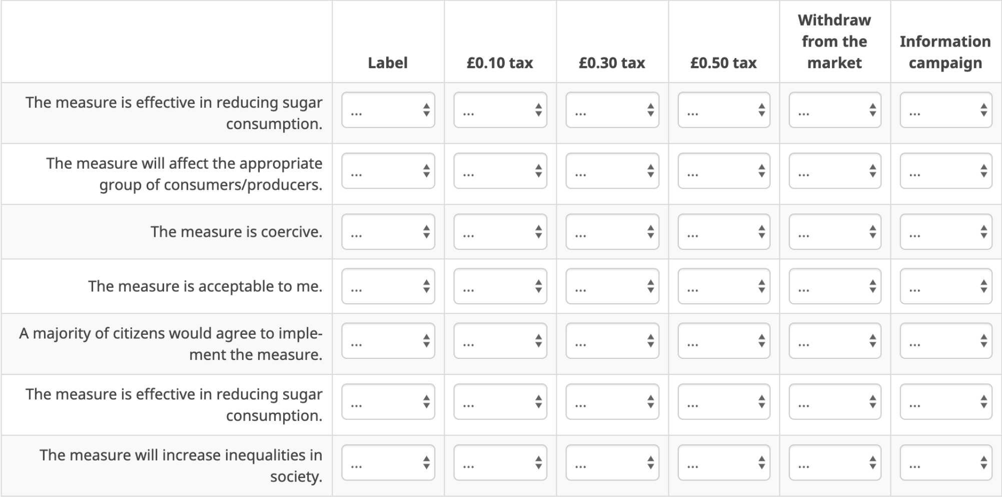
You will face below a list of statements. You are asked to indicate to which extent you agree with each of the statements on a scale ranging from 1 (completely disagree) to 7 (completely agree).

- It is legitimate to have collective rules that govern the consumption of sugar.
- The high consumption of sugar causes serious problems for society.
- As a whole, it is commonly accepted that sugar consumption should be reduced.
- As a whole, we eat more sugar in our society than recommended by the most recent scientific work.

You will now face the second series of statements (in lines) which will apply this time to several measures designed to decrease the consumption of high sugar content products (in columns). You are asked to indicate to what extent you agree, between 1 and 7, with each of the statements for each of the measures.

Remember that you can choose from the following options:

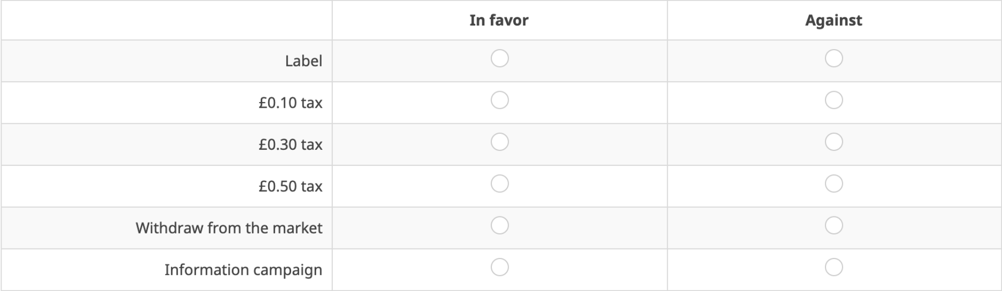
- 1 - Totally disagree
- 2
- 3
- 4 - Indifferent
- 5
- 6
- 7 - Totally agree



	Label	£0.10 tax	£0.30 tax	£0.50 tax	Withdraw from the market	Information campaign
The measure is effective in reducing sugar consumption.	...	...	...	...	...	...
The measure will affect the appropriate group of consumers/producers.	...	...	...	...	...	...
The measure is coercive.	...	...	...	...	...	...
The measure is acceptable to me.	...	...	...	...	...	...
A majority of citizens would agree to implement the measure.	...	...	...	...	...	...
The measure is effective in reducing sugar consumption.	...	...	...	...	...	...
The measure will increase inequalities in society.	...	...	...	...	...	...

Now imagine that you have the opportunity to vote for each of the measures presented independently. You must indicate, for each of them, if you would support the implementation ("In favor"), or if you would prefer to do nothing ("Against").

Would you be in favor or against the implementation of the measure?



	In favor	Against
Label	○	○
£0.10 tax	○	○
£0.30 tax	○	○
£0.50 tax	○	○
Withdraw from the market	○	○
Information campaign	○	○

Questionnaire for palm oil

In this questionnaire, we are interested in the consumption of snacks that can be bought in supermarkets or in grocery stores. These snacks vary in different aspects, and in particular in the presence of palm oil in the ingredients. Some products contain palm oil while others don't.

We are interested in six measures that could be implemented to regulate the country's palm oil consumption through the consumption of snacks:

- Measure 1: display labels on snacks that contain palm oil.
- Measure 2: tax the snacks that contain palm oil by £0.10 (for a 30g individual snack such as a cereal bar).
- Measure 3: tax the snacks that contain palm oil by £0.30 (for a 30g individual snack such as a cereal bar).
- Measure 4: tax the snacks that contain palm oil by £0.50 (for a 30g individual snack such as a cereal bar).
- Measure 5: remove the snacks that contain palm oil from the market.
- Measure 6: set up information campaigns to inform consumers about the impact of palm oil on the environment and society.

You will face below a list of statements. You are asked to indicate to which extent you agree with each of the statements on a scale ranging from 1 (completely disagree) to 7 (completely agree).

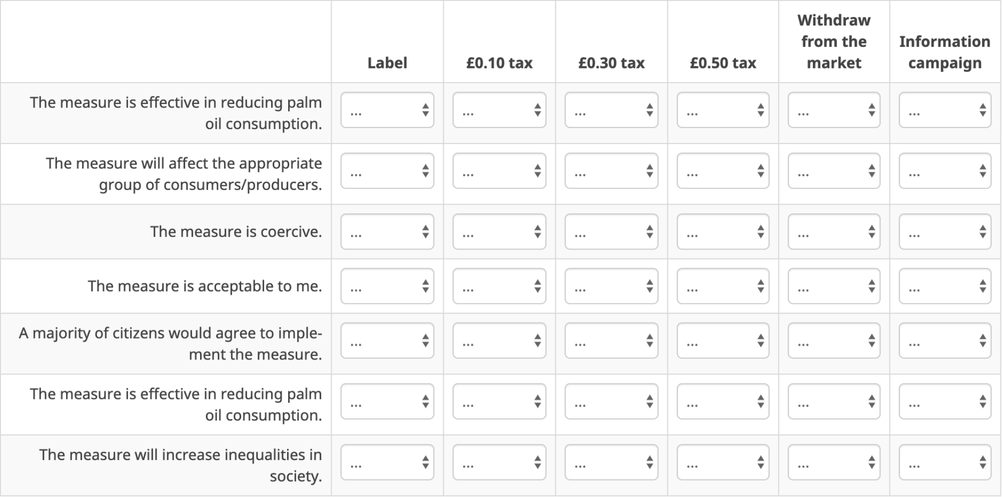
- It is legitimate to have collective rules that govern the consumption of palm oil.
- The high consumption of palm oil causes serious problems for society.
- As a whole, it is commonly accepted that palm oil consumption should be reduced.
- As a whole, we consume more palm oil in our society than recommended by the most recent environmental scientific work.

You will now face the second series of statements (in lines) which will apply this time to several measures designed to decrease the consumption of high sugar content products (in columns). You are asked to indicate to what extent you agree, between 1 and 7, with each of the statements for each of the measures.

Remember that you can choose from the following options:

- 1 - Totally disagree

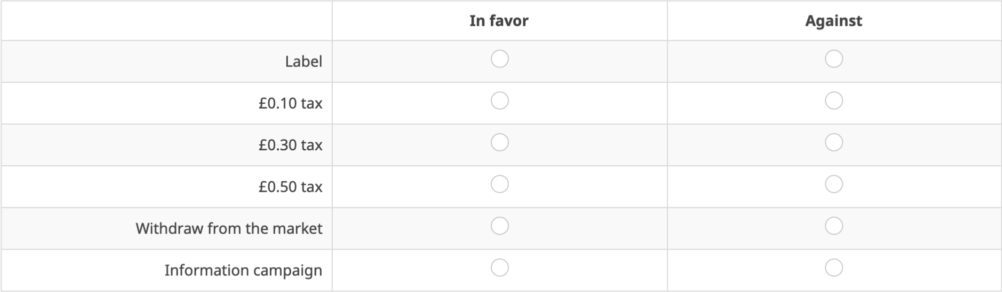
- 2
- 3
- 4 - Indifferent
- 5
- 6
- 7 - Totally agree



	Label	£0.10 tax	£0.30 tax	£0.50 tax	Withdraw from the market	Information campaign
The measure is effective in reducing palm oil consumption.	...	...	...	...	...	...
The measure will affect the appropriate group of consumers/producers.	...	...	...	...	...	...
The measure is coercive.	...	...	...	...	...	...
The measure is acceptable to me.	...	...	...	...	...	...
A majority of citizens would agree to implement the measure.	...	...	...	...	...	...
The measure is effective in reducing palm oil consumption.	...	...	...	...	...	...
The measure will increase inequalities in society.	...	...	...	...	...	...

Now imagine that you have the opportunity to vote for each of the measures presented independently. You must indicate, for each of them, if you would support the implementation ("In favor"), or if you would prefer to do nothing ("Against").

Would you be in favor or against the implementation of the measure?



	In favor	Against
Label	○	○
£0.10 tax	○	○
£0.30 tax	○	○
£0.50 tax	○	○
Withdraw from the market	○	○
Information campaign	○	○

Questionnaire for eggs

In this questionnaire, we are interested in the consumption of snacks that can be bought in supermarkets or in grocery stores. These snacks vary in different aspects, and in particular in the presence of cage-eggs in the ingredients. Some products contain eggs from laying hens kept in battery-cages while others don't.

We are interested in six measures that could be implemented to regulate the country's cage-eggs consumption through the consumption of snacks:

- Measure 1: display labels on snacks that contain cage-eggs.
- Measure 2: tax the snacks that contain cage-eggs by £0.10 (for a 30g individual snack such as a cereal bar).
- Measure 3: tax the snacks that contain cage-eggs by £0.30 (for a 30g individual snack such as a cereal bar).
- Measure 4: tax the snacks that contain cage-eggs by £0.50 (for a 30g individual snack such as a cereal bar).
- Measure 5: remove the snacks that contain cage-eggs from the market.
- Measure 6: set up information campaigns to inform consumers about the impact of cage-eggs on animal welfare and society.

You will face below a list of statements. You are asked to indicate to which extent you agree with each of the statements on a scale ranging from 1 (completely disagree) to 7 (completely agree).

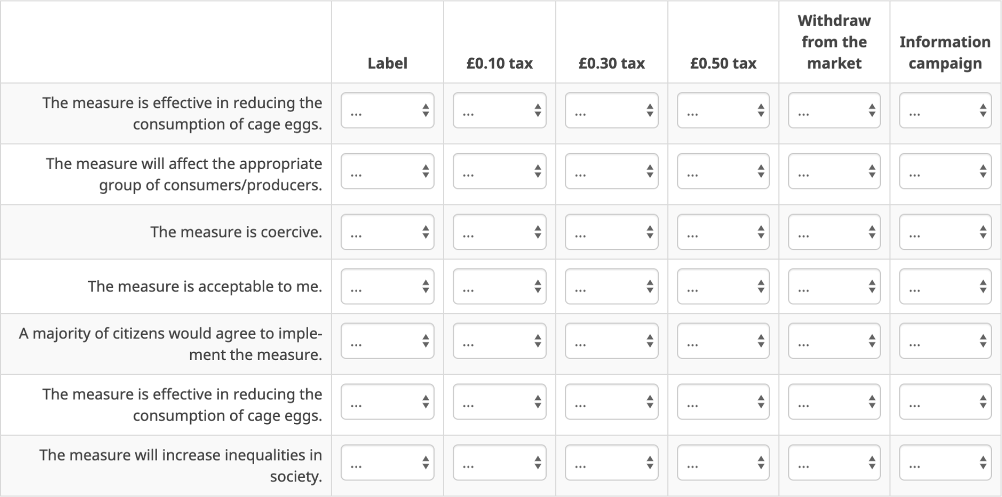
- It is legitimate to have collective rules that govern the consumption of cage-eggs.
- The high consumption of cage-eggs causes serious problems for society.
- As a whole, it is commonly accepted that the consumption of cage-eggs should be reduced.
- As a whole, we eat more cage-eggs in our society than recommended by the most recent scientific work to preserve animal welfare.

You will now face the second series of statements (in lines) which will apply this time to several measures designed to decrease the consumption of high sugar content products (in columns). You are asked to indicate to what extent you agree, between 1 and 7, with each of the statements for each of the measures.

Remember that you can choose from the following options:

- 1 - Totally disagree

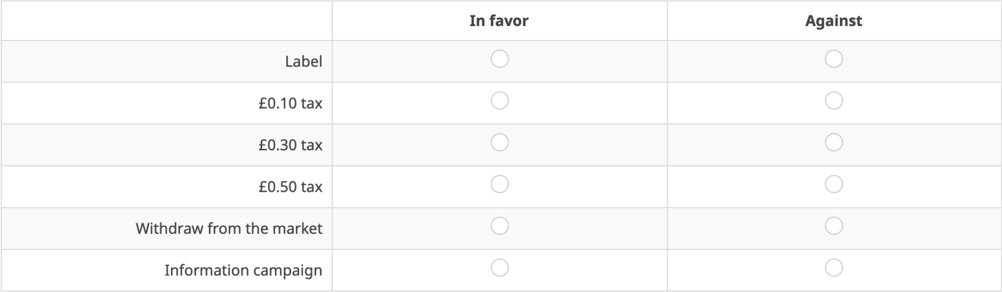
- 2
- 3
- 4 - Indifferent
- 5
- 6
- 7 - Totally agree



	Label	£0.10 tax	£0.30 tax	£0.50 tax	Withdraw from the market	Information campaign
The measure is effective in reducing the consumption of cage eggs.	...	...	...	...	...	...
The measure will affect the appropriate group of consumers/producers.	...	...	...	...	...	...
The measure is coercive.	...	...	...	...	...	...
The measure is acceptable to me.	...	...	...	...	...	...
A majority of citizens would agree to implement the measure.	...	...	...	...	...	...
The measure is effective in reducing the consumption of cage eggs.	...	...	...	...	...	...
The measure will increase inequalities in society.	...	...	...	...	...	...

Now imagine that you have the opportunity to vote for each of the measures presented independently. You must indicate, for each of them, if you would support the implementation ("In favor"), or if you would prefer to do nothing ("Against").

Would you be in favor or against the implementation of the measure?



	In favor	Against
Label	○	○
£0.10 tax	○	○
£0.30 tax	○	○
£0.50 tax	○	○
Withdraw from the market	○	○
Information campaign	○	○

Chapter 3

Misinformation due to asymmetric information sharing[1]

3.1 Introduction

Misinformation is thriving. This is problematic because it can lead people to inaccurate beliefs. These beliefs may then generate various socially harmful outcomes, such as lower vaccination coverage or dangerous political developments (Burki, 2019; Pennycook and Rand, 2021). Practitioners and academics are thus increasingly voicing their concern and investigating solutions to stop or curb the spread of misinformation (Lazer et al., 2018; European Commission, 2018).

To evaluate adequate policies in the fight against misinformation, the way it spreads and how it competes with true information has to be properly understood. An important dynamic, we argue, is that people do not share true and false information in the same manner. Two significant asymmetries that have been measured in social media sharing behaviours support this idea. First, false information tends to be shared further in a network than true information. Importantly, this asymmetry is not driven by bots but by humans (Vosoughi et al., 2018), even though people of all political orientation agree that accuracy is the most important criterion to consider before sharing information (Pennycook et al., 2021). We refer to this kind of asymmetry as *decay asymmetry*: true and false information need not be shared to the same extent. Second, true and false information is shared more or less heavily in different parts of a given network. Some agents are heavily connected in networks that share false information, while others are more connected in networks over which true information travels (Del Vicario et al., 2016; Zollo et al., 2017; Johnson et al., 2020). We refer to this kind of asymmetry as *network asymmetry*: true and false information need not be shared with the same people.

In this paper, we propose the first model of social learning that admits such asymmetries. We investigate their consequences on long-run beliefs and on speed of convergence

[1] This chapter is based on the homonymous working paper, written in collaboration with Berno Buechel, Stefan Klößner and Fanyuan Meng.

to a belief. Our model builds on the literature on social learning, where agents repeatedly share binary signals (see, e.g. Golub and Sadler, 2016, for a survey). We introduce decay asymmetry and network asymmetry, meaning that agents can share true and false information with different decay and with different people. Our results show that both decay and network asymmetries can cause a society to go from being able to discover the true state to being misinformed in the long run. By distinguishing between networks in which true information and false information is shared, we establish a threshold condition that determines which state a society converges to in the long run. The condition compares the product of decay factor and largest eigenvalue between the true and the false information sharing network. Noticeably, the long-run outcome in most cases does not depend on the initial distribution of signals (as long as there is at least one signal of each type) and is thus governed solely by sharing behaviors and the structure of the two networks. Additionally, we show that the speed of convergence to the long-run belief also depends on the ratio between the products of respective decay factors and largest eigenvalues. The smaller the difference between said products, the larger the half-life to convergence. When the difference gets close to zero, the half-life explodes, implying that the speed of convergence is particularly low when there are only slight asymmetries. We further show that agents can be ordered according to their ratio of eigenvector centralities in the two different networks. Those who are relatively more central in the false information sharing network are more prone to be misinformed. We then extend the model to allow agents to have pair-specific relationships, which accommodates heterogeneous subgroups, directed networks and idiosyncratic learning, and show that the threshold condition, speed of convergence and ordering by eigenvector centralities results still hold.

We illustrate the properties of our model using numeric simulations. The simulations highlight the effect of both the number of links in a given network and of the distribution of these links on the long-run state. Additional links in the false information network intuitively always favor reaching the misinformed state. However, for a given number of links, smaller denser groups, or echo chambers, have a (much) stronger influence than sparser larger groups. We thus show that the required decay asymmetry favoring true information needed to compensate misinformation gets disproportionately large with the existence of misinformation echo chambers.

To counter misinformation, our model suggests two main avenues. The first works on decay asymmetry, the second on network asymmetry. Regarding decay, measures to make true information shared more and false information shared less have to be considered. These can include both changes to the information itself (such as making true information more shareable, easier to understand) or to the behaviors (such as educating people to recognize false information to avoid sharing it further). Regarding the second, measures to make the true information network denser and the false information network sparser have to be considered in order to increase the ratio of largest eigenvalues in favor of the true information sharing network. Particularly, breaking links of highly connected echo chambers over which false information travels will disproportionately favor reaching the true state in the long run.

The remainder of thi chapter is structured as follows. Section 3.2 relates it to the literature. We introduce the model in section 3.3 and study the special case of symmetry as a benchmark in section 3.4. Section 3.5 presents results and an extension. We illustrate our results using simulations in section 3.6. A discussion and policy implications are provided in section 3.7.

3.2 Related Literature

Our paper belongs to the literature on social learning (or opinion dynamics), where agents repeatedly learn from their neighbors in a social network. If agents were fully Bayesian, then perfect information aggregation would occur in any connected network (DeMarzo et al., 2003; Mueller-Frank, 2013). However, Bayesian rationality is very demanding when it comes to learning in a network structure. As experimental studies reveal, actual behaviors are less often consistent with Bayesian learning than they are with simpler updating rules such as repeated averaging (Corazzini et al., 2012; Friedkin and Bullo, 2017; Grimm and Mengel, 2018; Chandrasekhar et al., 2020). The classic model of repeated linear updating (DeGroot, 1974) has been studied intensively and it has been extended in many interesting directions (Friedkin and Johnsen, 1990; DeMarzo et al., 2003; Golub and Jackson, 2010, 2012; Buechel et al., 2015; Grabisch et al., 2019; Banerjee et al., 2019). A main focus is whether the long-run consensus opinion optimally aggregates the initially dispersed pieces of information. This is satisfied, at least asymptotically, if there are no individuals with excessive influence on the others (Golub and Jackson, 2010); or if there is at least one agent who is perfectly Bayesian (Mueller-Frank, 2014); or if agents additionally receive a private signal in *each* period and treat this signal in a Bayesian manner (Jadbabaie et al., 2012). More generally, there are (relatively weak) conditions on non-Bayesian learning which are sufficient for learning the true underlying state in the long run (Molavi et al., 2018).

However, none of these models is able to capture asymmetric treatment of signals. Symmetric treatment of signals is related to so-called “label neutrality,” which is an underlying assumption of the models in the literature and in fact a characterizing feature of the DeGroot model (Molavi et al., 2018). We contribute, to our best knowledge, the first model of social learning that relaxes label neutrality. We do so by addressing network or decay asymmetry in signal sharing. As we show, the consequences are drastic as either decay or network asymmetry favoring negative signals can be sufficient for misinformation. That is, not only can the society’s long-run belief be a suboptimal aggregation of the initial signals, but all members of the society would guess the wrong state and hence make the wrong decision with high probability. Similarly drastic conclusions are known in the literature only in the presence of forceful (or biased or stubborn) agents who, at some point, do not learn at all but still heavily influence the other agents (Acemoglu et al., 2010; Grabisch et al., 2018; Azzimonti and Fernandes, 2018; Rusinowska and Taalaibekova, 2019; Della Lena, 2019; Sikder et al., 2020). The policy implications that can be drawn from these models rather suggest acting on forceful agents and on fighting disinformation, which is quite different from ours. Complementary to these insights, we argue that in the absence of forceful agents, the way signals are shared is crucial for avoiding misinformation.

Sharing behavior and signal accumulation in our model is closest to Sikder et al. (2020), in which agents repeatedly share binary signals in a rather naïve way but with Bayesian-style updating. These authors show that the introduction of confirmation bias leads to polarization. While Sikder et al. (2020) mostly focus on regular graphs, we first show for any connected network that misinformation in this baseline model is bounded. We generalize their baseline model to any network structure and by introducing decay and network asymmetries depending on the type of information shared. These asymmetries are closest to Taalaibekova (2020, Chapter 3.3). In one variation of her dissertation, she introduces the concept of “optimists” and “pessimists” who receive more positive signals, respectively negative signals, without being aware of it. Despite receiving signals in a biased manner, these agents still treat different signals in the same way, while our agents treat true and false information asymmetrically. Moreover, decay in a framework of naïve learning, which marries the DeGroot model with the bounded confidence model, has been introduced recently by Grabisch et al. (2021). While in their paper social influence decays over time, in our model information decays.

3.3 A Model of Asymmetric Signal Sharing

Ingredients. There are n agents $N = \{1, ..., n\}$ who talk about a binary issue. Classic examples include whether the NASA has landed on the moon, whether there is human induced climate change, or whether vaccines cause autism; new examples are popping up every day. To model uncertainty, nature draws the true state with a commonly known prior probability and then each agent receives a signal. Specifically, the true state $\theta \in \{0, 1\}$ is drawn with prior probability $b^0 = P(\theta = 1) \in (0, 1)$. Each agent i independently receives signal $s_i \in \{0, 1\}$ which matches the true state with probability ρ, i.e. $P(s_i = 1|\theta = 1) = P(s_i = 0|\theta = 0) = \rho \in (\frac{1}{2}, 1)$. Conditional on the state, the signals are independent. As a convention, we let the true state be $\theta = 1$ and call 1 a *positive* signal and 0 as a *negative* signal.[2]

Time is discrete $t = 0, 1, 2, ...$ At time $t = 0$, the initial signals are received. At each time step $t > 0$, agents communicate with their neighbors in a social network. We model this communication activity in a way that admits asymmetric treatment of positive and negative signals. We introduce decay asymmetry and network asymmetry.

When a signal is passed on from one agent to the next, it decays by $\delta^+ \in (0, 1]$ if it is positive and by $\delta^- \in (0, 1]$ if it is negative. Decay can be either (i) due to a sender who only shares a fraction δ^+ of her signals, (ii) due to the communication channel on

[2] This is only known to the modeler and, of course, not to the agents. Agents thus do not consciously recognize a signal to be true or false, but will treat them differently. This might be due to, for example, the difference in emotional content carried by true and false information as suggested by Vosoughi et al. (2018).

which a fraction $1-\delta^+$ gets lost, or (iii) due to the recipient who discounts the received signals by δ^+.[3] If, for any reason, $\delta^+ \neq \delta^-$, there is *decay asymmetry.*

Positive signals are shared in a network (N, A^+), negative signals in a network (N, A^-), where A^+ and A^- are both symmetric $n \times n$ matrices with entries 0 or 1, i.e. adjacency matrices representing each an undirected unweighted network. We assume that these networks are connected, which implies that the matrices are irreducible. Of course, these two networks can be highly related to each other.[4] If $A^+ \neq A^-$, there is *network asymmetry.*

Signal Accumulation. Given these ingredients, we can now formulate how signals are shared and accumulated. Let $N_i^+(t)$ denote number of positive signals of node i at time t and let $N^+(t)$ be the $(n \times 1)$ vector. The law of motion for positive signals is

$$N^+(t) = (I + \delta^+ A^+) N^+(t-1). \tag{3.1}$$

Hence, the number of agent i's positive signals in a given period is the number of positive signals i held in the previous period plus the number of positive signals that i's neighbors in network A^+ held in the previous period discounted by δ^+. For the negative signals the law of motion and the notation is fully analogous. $N^-(t)$ is the vector of negative signals and the law of motion is $N^-(t) = (I + \delta^- A^-)N^-(t-1)$. Let $N^+(0)$ and $N^-(0)$ be the vectors of initial signals, i.e. $N_i^+(0) = s_i$ and $N_i^-(0) = 1 - s_i$. Given the initial signals and the law of motion, we can compute the number of signals of any agent at any time. The signal accumulation for positive signals is

$$N^+(t) = (I + \delta^+ A^+)^t N^+(0). \tag{3.2}$$

Technically, the entries in the matrix $(I + \delta^+ A^+)^t$ can be considered as the number of walks of length t or smaller, when the network A^+ is augmented by self-loops (ones on the diagonal); thereby each walk is discounted by δ^+ at any step, except when using a self-loop. A less technical interpretation is that agents share all the positive signals that they have with their neighbors in the positive signals sharing network; and whenever signals are passed on only a fraction δ^+ fully arrives at the recipient.

Beliefs and Signal Mixes. In this model, beliefs are formed similar to Bayes' rule with a crucial difference: the signals are taken at face value, i.e. treated as independent, ignoring that many accumulated signals are merely repetitions of the same initial signals.[5]

[3] In Appendix 3.C.2, we show that all three interpretations can be explicitly modeled and are captured in a reduced form in our model with one parameter δ^+. In an extension of the model, we will study pair-specific decay factors, which admit heterogeneity in behavior of senders, communication channels, or recipients (Section 3.5.4).

[4] In several scenarios we investigate, there is a connected network A that is a subnetwork of both A^+ and A^-.

[5] This kind of behavioral mistake is also known as "persuasion bias" and provides a justification for the DeGroot model (DeMarzo et al., 2003).

Let $k_i(t) := N_i^+(t) - N_i^-(t)$ be the *signal difference* of agent i at time t. Then i's *belief* is

$$b_i(t) = P(\theta = 1|k_i(t)) = \frac{\rho^{k_i(t)} * b^0}{\rho^{k_i(t)} * b^0 + (1-\rho)^{k_i(t)} * (1-b^0)}, \tag{3.3}$$

where b^0 is the prior belief.[6] The evolution of beliefs $b_i(t)$ is not as easy to study as that of *signal mixes*

$$x_i(t) := \frac{N_i^+(t)}{N_i^+(t) + N_i^-(t)}, \tag{3.4}$$

i.e. the fraction of positive signals that i holds. Again, we denote vectors by $b(t)$, $k(t)$ and $x(t)$.

To measure misinformation, we assess whether an agent's belief tends to the correct or to the wrong state. We consider an agent i to be *misinformed* if her belief satisfies $b_i(t) < 0.5$, while she is not misinformed if her belief satisfies $b_i(t) > 0.5$. (Agents with $b_i(t) = 0.5$ are considered as something in between.) In most cases, signal mixes determine whether an agent is misinformed, as Lemma 1 assures.

Lemma 1. *Suppose that either $b^0 = 0.5$ or $t \gg 0$ holds (or both), i.e. either the prior belief is one half or sufficient time has elapsed (or both). If an agent's signal mix satisfies $x_i(t) < 0.5$, then she is misinformed, i.e. $b_i(t) < 0.5$. If an agent's signal mix satisfies $x_i(t) > 0.5$, then she is not misinformed, i.e. $b_i(t) > 0.5$.*

Lemma 1 means that, apart from very special cases, it is sufficient to know an agent's signal mix $x_i(t)$ to determine whether she is misinformed. One exception are signal mixes that converge to exactly one half, i.e. $\lim_{t\to\infty} x_i(t) = 0.5$, as they can lead to nasty behavior of beliefs.[7] Based on Lemma 1, we use an agent's signal mix $x_i(t)$ as a proxy for misinformation on the individual level.

3.4 Benchmark: misinformation under symmetry

To assess the effects of decay and network asymmetry, a natural benchmark is, of course, the special case of *symmetry*: $\delta^+ = \delta^- =: \delta$ and $A^+ = A^- =: A$. This is a generalization of the baseline model introduced in Sikder et al. (2020), which our model nests for $\delta = 1$. In the original article by Sikder et al. (2020), the focus is on regular networks (N, A), that are characterized by each node having the same degree.[8] Regular networks happen to minimize the probability of misinformation, as our Proposition 1 below implies. More

[6] This formula can be derived from Bayes' rule. It applies to symmetric binomial signals that are independently drawn.

[7] Various examples showcasing the behaviour of beliefs and best guesses can be found in Appendix 3.C.3. For instance, $k_i(t)$, which determines the belief, might diverge to $-\infty$, to ∞, converge to 0, converge to some other number, or even not converge at all, while $x_i(t)$ always converges.

[8] Sikder et al. (2020) find polarization in an extension of their model, in which they introduce agents with confirmation bias.

importantly, this result shows that the probability of misinformation is always bounded under symmetry.

Throughout the analysis, the adjacency spectrum will play an important role. Let λ_1 be the largest eigenvalue of A.[9] Let $c = (c_1, ..., c_n)^\top$ be its corresponding eigenvector, normalized such that its components sum to one, i.e. $Ac = \lambda_1 c$ and $\sum_i c_i = 1$.[10] Entries in the eigenvector c are a measure of *eigenvector centrality* (Bonacich, 1972; Friedkin, 1991).

Proposition 1 (Symmetry). *Under symmetry, the long-run signal mix is a convex combination of the initial signals s_j with weights according to eigenvector centrality, i.e. for all i, $\lim\limits_{t\to\infty} x_i(t) = \sum_{j=1}^n c_j s_j$. Hence, misinformation prevails if $\sum_{j=1}^n c_j s_j < 0.5$ and the probability of misinformation is bounded from above by* 0.5*.*

The proofs of this proposition and all others are collected in Appendix 3.A. Proposition 1 means that misinformation under symmetry only occurs due to an unlucky distribution of signals. It depends on the relative influence of each agent on the long-run opinions, as measured by her entry in the largest eigenvector. In the best case, all agents, who are by assumption equally well-informed, are equally influential.[11] This is satisfied in regular graphs, in which by definition every agent has the same degree. Then the long-run signal mix of every agent exactly reflects the initial signal distribution, i.e. $\lim\limits_{t\to\infty} x_i(t) = \frac{1}{n}\sum_{j=1}^n s_j$, which is just the mean of the initial signals. Hence, under symmetry and when the network is regular, naïve agents who accumulate the same signals over and over again can fare equally well as Bayesians would (as we discuss in Example 3.C.1 in Appendix 3.C.1). In the worst case, there is a group of agents who are overly influential (as illustrated in Example 3.C.2 in Appendix 3.C.1).

Since misinformation in the worst case can be substantial, the question arises whether this is also true for more realistic networks. We therefore simulate probabilities of misinformation for two classes of random graphs: Erdös-Rényi (ER), in which the probability of every link is fixed (p), and Barabasi-Albert (BA), where in every step of constructing the network, m new links are created. The expected degree distribution of BA networks is scale-free, a feature that many real social networks share (Albert and Barabási, 2002; Strogatz, 2001). Since the degree distribution in ER random graphs is more uniform than in BA random graphs, we expect the probability of misinformation to be higher in the latter, when selecting parameters (p and m) such that the two random networks have the same expected density.

Figure 3.4.1 illustrates the simulation results. As expected, misinformation tends to be lower in ER as compared to BA random graphs. The lowest line is the probability of misinformation in a regular graph (from Example 3.C.1) while the highest line is the

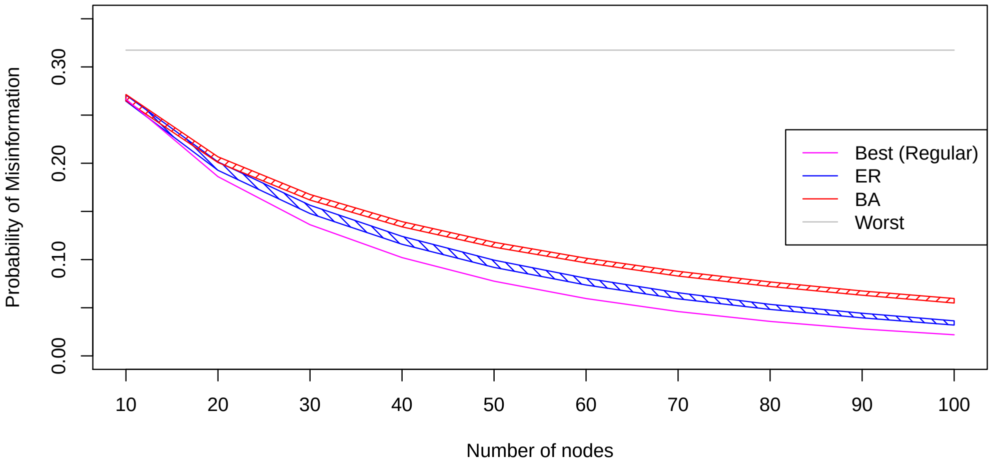
[9] More precisely, λ_1 is the largest positive eigenvalue of A and other eigenvalues of A might exist which are as large in absolute value as λ_1. For ease of notation, we usually omit 'positive' when addressing largest positive eigenvalues.

[10] Recall that by assumption the adjacency matrix is symmetric. Hence, there is no need to distinguish between left-hand and right-hand eigenvectors.

[11] More generally, the requirement is a balance between idiosyncratic signal precision and social influence (see, e.g. Buechel et al., 2015).

probability of misinformation in the worst network (from Example 3.C.2), where a small group of agents is extremely influential. Misinformation for the two classes of random graphs is closer to the regular graph's. In particular, misinformation decreases with network size n. The thickness of the two lines represents the variation of misinformation that covers 50% of all simulation runs (from 25th to 75th percentile). For large enough networks, the probability of misinformation is small.

Figure 3.4.1: Misinformation under symmetry: comparing different network structures.



Notes: Signal precision $\rho = 0.6$. Number of nodes $n = 10, ..., 100$ on the x-axis. Random graph parameters are set such that asymptotic average degree is 6 in these simulation runs. 1000 simulation runs per class of random network of a given size. Thickness of corresponding line represents variation covering 50% of all outcomes (from 25th to 75th percentile).

The results in this section closely resemble those of the common DeGroot model of naïve learning. In Appendix 3.C.5, we describe the similarities and differences of these two models and their results in detail. Importantly, our model of naïve learning keeps track of all signals, while in the DeGroot model positive and negative signals are mingled into opinions. This has the consequence that asymmetries in signal sharing cannot be introduced into the DeGroot model. Moreover, some measures against misinformation can only be assessed in a model that keeps track of the initial signals.

3.5 Results

3.5.1 Key result

We now study how misinformation in the long run is affected by asymmetric treatment of signals. Decay asymmetry is captured simply by its two parameters δ^+ and δ^-. For network asymmetry, the adjacency spectrum of the two matrices A^+ and A^- matters. Let λ_1^+ be the largest eigenvalue of A^+. Let $c^+ = (c_1^+, ..., c_n^+)^\top$ be its corresponding eigenvector, normalized such that its components sum to one, i.e. $A^+c^+ = \lambda_1^+c^+$ and $\sum_i c_i^+ = 1$. And likewise λ_1^- and c^- are the largest eigenvalue and its normalized eigenvector of A^-, i.e. $A^-c^- = \lambda_1^-c^-$ and $\sum_i c_i^- = 1$. Entries in the eigenvectors c^+, c^- are the respective eigenvector centrality.

Proposition 2 (Key Result)**.** *Suppose that the initial distribution of signals contains at least one positive and at least one negative signal.*

1. *If $\delta^+\lambda_1^+ < \delta^-\lambda_1^-$, then for all i and large t:*

$$x_i(t) \approx \frac{c_i^+}{c_i^-}\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t \frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2} \frac{\sum_{j=1}^n c_j^+ s_j}{1-\sum_{j=1}^n c_j^- s_j}$$

such that $\lim_{t\to\infty} x_i(t) = 0$. Hence, misinformation prevails.

2. *If $\delta^+\lambda_1^+ > \delta^-\lambda_1^-$, then for all i and large t:*

$$x_i(t) \approx 1 - \frac{c_i^-}{c_i^+}\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^t \frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2} \frac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{j=1}^n c_j^+ s_j}$$

such that $\lim_{t\to\infty} x_i(t) = 1$. Hence, misinformation vanishes.

3. *If $\delta^+\lambda_1^+ = \delta^-\lambda_1^-$, then for all i:*

$$\lim_{t\to\infty} x_i(t) = \frac{1}{1+\dfrac{c_i^-}{c_i^+}\dfrac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\dfrac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{j=1}^n c_j^+ s_j}} \in (0,1).$$

Hence, long-run misinformation depends on eigenvector centralities and the signal distribution.

Let us first discuss the "big picture". The proposition states that the combination of decay factor and largest eigenvalue is crucial for the level of misinformation in the long run: the condition $\delta^+\lambda_1^+ \lesseqgtr \delta^-\lambda_1^-$ determines whether all signal mixes and therefore all beliefs will converge to 0 (Case 1) or whether they will converge to 1 (Case 2). Intuitively, the case with full misinformation, Case 1, is rather reached when the decay factor for positive signals δ^+ is low compared to the decay factor with negative signals δ^-, which means that positive signals are shared to a lower extent; and when the eigenvalue

λ_1^+ of the positive signal sharing network A^+ is low compared to λ_1^-, which has the interpretation that the agents are generally better connected in the negative signals sharing network A^-, as discussed below. The case distinction relies on the product of these two factors: $\delta^+\lambda_1^+ \lesseqgtr \delta^-\lambda_1^-$; or $\frac{\delta^+}{\delta^-} \lesseqgtr \frac{\lambda_1^-}{\lambda_1^+}$. Hence, decay factor and largest eigenvalue can only compensate each other to some extent. Case 3, $\delta^+\lambda_1^+ = \delta^-\lambda_1^-$, is the knife-edge case, in which these two products coincide. In that case it is possible that some agents are misinformed in the limit while others are not, as we will demonstrate below (in Example 1).

In the limit, the initial distribution of signals does virtually not matter in Cases 1 and 2. Given that there is initially at least one positive and one negative signal, either positive or negative signals fully dominate in the long run. Hence, all beliefs $b_i(t)$ in Case 1 converge to 0 such that all agents in the society are misinformed, independent of the signal distribution and independent of the prior belief b^0. Likewise, all beliefs $b_i(t)$ in Case 2 converge to 1 such that no agent in the society is misinformed. For Case 3, the limit is less trivial. We can first observe that the long-run signal mix $x_i(t)$ is increasing in all s_j, i.e. with increasing number of initial signals being correct, agents' signal mixes and therefore also their beliefs get closer to the truth.

Comparing this key result to the symmetric benchmark draws three main conclusions. First, the symmetric benchmark is nested in Case 3. Second, while misinformation is bounded in the symmetric benchmark, introducing asymmetry induces full misinformation, in which every agent is misinformed, if the conditions of Case 1 are satisfied. Third, while in the symmetric benchmark the common decay factor δ did not affect the long-run opinions, decay factors δ^+ and δ^- are crucial for the case distinction under asymmetry.

Let us now discuss the terms that determine the asymptotic behavior in Cases 1 and 2 of Proposition 2. In Case 1, there are four factors:

$$x_i(t) \approx \underbrace{\frac{c_i^+}{c_i^-}}_{\text{centrality ratio}} \cdot \underbrace{\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t}_{\text{exponential decay}} \cdot \underbrace{\frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2}}_{\text{normalizing constant}} \cdot \underbrace{\frac{\sum_{j=1}^n c_j^+ s_j}{1-\sum_{j=1}^n c_j^- s_j}}_{\text{signal averages}} \tag{3.5}$$

The first is agent-specific, the three latter factors are common for all agents in a given society. The first factor shows that agent i's characteristics enter her asymptotic signal mix $x_i(t)$ via her ratio of eigenvector centrality in the positive signals sharing network over her centrality in the negative signals sharing network: $\frac{c_i^+}{c_i^-}$. We will refer to this as i's *centrality ratio*. The second factor shows the exponential decay process, which depends on the (information) decay factors δ^+ and δ^-, as well as on the the largest eigenvalues of the two networks. The next factor is the ratio of the sums of squared centralities. This can be considered as a normalizing constant. The last factor is determined by weighted averages of the initial signals, whereas the weights are the network centralities. We can observe that this factor is increasing in s_j, reflecting the fact that an agent's signal mix $x_i(t)$ becomes closer to 1 when any agent's signal s_j flips from false information ($s_j = 0$) to true information ($s_j = 1$). The centrality ratio will determine opinion diversity, the exponential decay will determine speed of convergence, the weighted signal averages will determine the levels of the signal mixes. The factor decomposition is analogous for Case 2. In Case 3, opinion diversity is also determined by the centrality ratio, whereas speed of convergence is determined differently from Cases 1 and 2, as further discussed below.

3.5.2 Implications

Through the main result, we can derive implications regarding the state that is reached in the long run, the potential diversity of opinions or consensus and how long it takes to reach the long-run state.

Long-run misinformation. Suppose there is decay asymmetry, while the networks are symmetric, i.e. $\delta^+ \neq \delta^-$ and $A^+ = A^-$. Then the crucial condition for long-run misinformation, $\delta^+\lambda_1^+ \lesseqgtr \delta^-\lambda_1^-$ (Proposition 2), becomes $\delta^+ \lesseqgtr \delta^-$. Now, if it is true that negative signals exhibit less decay than positive signals, i.e. $\delta^+ < \delta^-$, then we are in Case 1 of Proposition 2 and misinformation prevails. Hence, already slight asymmetry in decay fully changes the long-run outcome from bounded misinformation, which we have observed in the case of symmetry, to full misinformation. However, it may take time until this long-run consensus on the false opinion is reached.

Suppose there is decay symmetry, but network asymmetry, i.e. $\delta^+ = \delta^-$ and $A^+ \neq A^-$. Then the crucial condition of Proposition 2, $\delta^+\lambda_1^+ \lesseqgtr \delta^-\lambda_1^-$, becomes $\lambda_1^+ \lesseqgtr \lambda_1^-$. As argued before, the largest eigenvalues capture how well-connected the agents are in some sense. This interpretation becomes more specific in special cases. Suppose the networks A^+ and A^- are the same, apart from some links in A^- that are not present in A^+, e.g. because some agents use these additional channels to share only negative signals. We write $A^+ \subset A^-$ to denote that the positive signals sharing network is a subgraph of the negative signals sharing network (i.e. for each entry, $a_{ij}^+ \leq a_{ij}^-$, and for some entry $a_{ij}^+ < a_{ij}^-$). Then it holds that $\lambda_1^+ < \lambda_1^-$. As a consequence, we are in Case 1 of Proposition 2 and misinformation prevails. Using some links to only share negative signals has hence drastic consequences for long-run misinformation. Likewise, if $A^+ \supset A^-$, then $\lambda_1^+ > \lambda_1^-$ and we are in Case 2 such that misinformation vanishes. In general, however, the positive signals sharing network need not be a subnetwork of the negative signals sharing network, nor the other way around.

Finally, let us admit both decay and network asymmetry, i.e. $\delta^+ \neq \delta^-$ and $A^+ \neq A^-$. Clearly, when both kinds of asymmetry point to the same direction, for instance, when negative signals decay less ($\delta^+ < \delta^-$) and the negative signals sharing network is better connected ($\lambda_1^+ < \lambda_1^-$), then the results follow from Proposition 2: Since $\delta^+\lambda_1^+ < \delta^-\lambda_1^-$, misinformation prevails. Moreover, this setting boosts the fraction $\log\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)$ (i.e. the exponential decay factor of Case 2), which measures the speed of convergence. And analogously for the opposite setting. The more interesting question is: to which extent decay and network asymmetry can compensate each other when they point to opposite directions? We illustrate such cases in Section 3.6.

Opinion diversity. Proposition 2 already indicates that individual differences between agent's beliefs are driven by their centrality ratio $\frac{c_i^+}{c_i^-}$. The following corollary establishes this relation, considering two agents' ratios of positive over negative signals, $\frac{N_i^+(t)}{N_i^-(t)} \Big/ \frac{N_j^+(t)}{N_j^-(t)}$.[12] Clearly, if an agent's ratio of positive over negative signals is above another agent's ratio, then her signal mix and belief are closer to 1 and hence closer to the truth.

Corollary 1 (Centrality Ratios and Opinion Diversity)**.** *Suppose that the initial distribution of signals contains at least one positive and at least one negative signal. Then the ratio of two agents' ratios of positive over negative signals converges to these agents' ratio of centrality ratios, i.e.*

$$\lim_{t\to\infty} \frac{N_i^+(t)}{N_i^-(t)} \Big/ \frac{N_j^+(t)}{N_j^-(t)} = \frac{c_i^+}{c_i^-} \Big/ \frac{c_j^+}{c_j^-}. \tag{3.6}$$

Hence, an agent i with higher centrality ratio than another agent j has a higher asymptotic signal mix and belief, i.e. if $\frac{c_i^+}{c_i^-} > \frac{c_j^+}{c_j^-}$, then for large t, $x_i(t) > x_j(t)$ and $b_i(t) > b_j(t)$.

Corollary 1 applies to all three cases of Proposition 2. It says that agents with relatively high centrality in the negative signals network are more prone to be misinformed. Moreover, if some agent is misinformed in the long run, then all agents with lower centrality ratios must be as well. Similarly, if an agent is well-informed in the long run, then all agents with higher centrality ratios must be as well.

Minimal opinion diversity is given in all networks where $\frac{c_i^+}{c_i^-}$ is constant across agents, which holds in particular, if $A^+ = A^-$ (which we discuss in the next paragraph). Strong opinion diversity is given in networks where these ratios differ strongly across agents. For example, consider two star networks (N, A^+) and (N, A^-) with different centers. The ratio of one center is $\frac{c_i^+}{c_i^-} = \frac{\sqrt{n-1}}{1}$, the ratio of the other center $\frac{c_j^+}{c_j^-} = \frac{1}{\sqrt{n-1}}$. Hence, by Corollary 1, $\frac{N_i^+(t)}{N_i^-(t)} \Big/ \frac{N_j^+(t)}{N_j^-(t)}$ converges to $n-1$. That is, agent i has asymptotically $n-1$ times more positive signals over negative signals than agent j. Again, this holds in all of the three cases, even if the absolute differences vanish for large t.

In terms of opinion diversity, the Corollary 1 means that, even if beliefs converge to consensus in Cases 1 and 2, they are ordered by their centrality ratios. In Case 3, this order also holds and opinion diversity even persists in the limit, as Example 1 below shows. We have seen in a special case of Case 3, the symmetry benchmark (Section 3.4), that consensus may also emerge in Case 3. This was due to the absence of network asymmetry. Indeed, $A^+ = A^-$, implies that an agents i's centrality c_i^+ in the

[12] Notice that agent i's ratio of positive over negative signals, $\frac{N_i^+(t)}{N_i^-(t)}$, can equivalently be written as $\frac{x_i(t)}{1-x_i(t)}$.

positive signals sharing network equals her centrality c_i^- in the negative signals sharing network. Hence, for every agent i, the ratio of network centralities is $\frac{c_i^+}{c_i^-} = 1$. Therefore, Corollary 1 implies that all agents approach the same signal mix, opinions converge to consensus in the absence of network asymmetry.

Speed of convergence. The motivation for studying misinformation is that agents make decisions, which may be based on inaccurate information. If the point of decision making is not sufficiently far in the future, then the short or medium term opinion dynamics matter. Speed of convergence can also be measured with the help of Proposition 2. For instance, in Case 1, the speed of convergence is governed by the speed that the exponential decay factor $\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t$ converges to 0 (see also Eq. (3.5)). Looking for the half-life, as it is standard for exponential decay processes, we define $t_{1/2}$ as the number of periods it takes for this quantity to fall to one half of its initial value. We get $t_{1/2} = \log(0.5)/\log\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)$ in Case 1. Analogously, in Case 2 half-time $t_{1/2}$ equals $\log(0.5)/\log\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)$. In Case 3, things are more complicated as not only the largest eigenvalues are relevant asymptotically.[13] The case dependent half-life can then be generalized as:

$$t_{1/2} = \frac{\log(0.5)}{\log(\tau)}, \quad \text{with} \tag{3.7}$$

$$\tau := \begin{cases} \frac{1+\min\{\delta^+\lambda_1^+,\delta^-\lambda_1^-\}}{1+\max\{\delta^+\lambda_1^+,\delta^-\lambda_1^-\}}), & \text{if } \delta^+\lambda_1^+ \neq \delta^-\lambda_1^- \text{ (Cases 1, 2)} \\ \max\left\{\max\left\{\frac{|1+\delta^+\lambda_i^+|}{1+\delta^+\lambda_1^+}, i=2,\ldots,n\right\}, \max\left\{\frac{|1+\delta^-\lambda_i^-|}{1+\delta^-\lambda_1^-}, i=2,\ldots,n\right\}\right\}, & \text{if } \delta^+\lambda_1^+ = \delta^-\lambda_1^- \end{cases}$$

In the first two cases, half-life will be large when $\delta^+\lambda_1^+$ and $\delta^-\lambda_1^-$ are close to each other, i.e. when we are close to Case 3. In some sense, Case 3 can be considered as unlikely because it is a special case of the parameter space. Still, it is important to study this case for at least two reasons. First, in the absence of asymmetry – the special case that has been studied in the literature – we are in Case 3, as it was discussed in Section 3.4. Second, under certain conditions, opinions in the short and medium term are well approximated by the analysis of Case 3, even if this does not hold in the long run.[14]

3.5.3 Illustration of key result and implications

Example 1 illustrates the case distinction, the signal mix dynamics, the speed of convergence implied by Proposition 2, as well as the opinion diversity implied by Corollary 1.

[13] The exponential decay factor can be derived from the representations for $N^+(t)$ and $N^-(t)$ given in Equations (3.A.2) and (3.A.3).

[14] Conditions are discussed in Appendix 3.C.4.

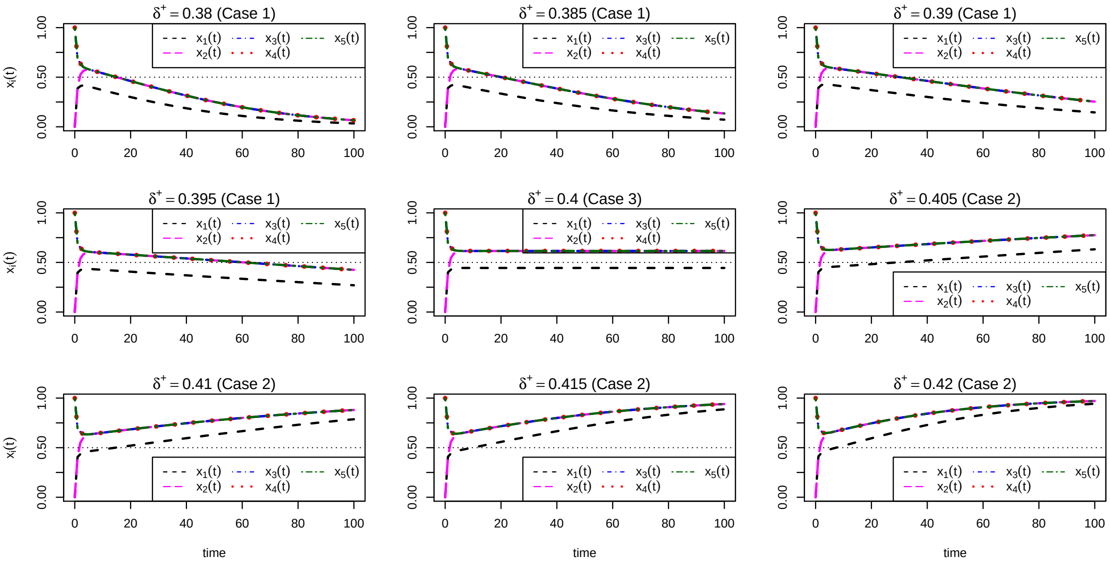
Example 1. *Consider five agents* $N = \{1, 2, 3, 4, 5\}$*. Let the prior be* $b^0 = 0.5$*. The positive signals sharing network* (N, A^+) *is the complete network, i.e.* $a^+_{ij} = 1$ *for all* $i \neq j$*. The negative signals sharing network* (N, A^-) *is a star network with agent 1 at the center, i.e.* $a^-_{1j} = a^-_{j1} = 1$ *for all* $j = 2, 3, 4, 5$ *and* $a^-_{ij} = 0$ *else. For illustration purpose, one can think of this example as five colleagues who discuss scientific publications with each other, and discuss conspiracy theories with only one person.*

We fix the decay factor of negative signals to be $\delta^- = 0.8$*, whereas we vary the decay factor of the positive signals. Proposition 2 distinguishes three cases based on the condition* $\delta^+\lambda^+_1 \lesseqgtr \delta^-\lambda^-_1$*. In the complete network, we have* $\lambda^+_1 = n - 1 = 4$*. In the star network, we have* $\lambda^-_1 = \sqrt{n-1} = 2$*. Hence, we are in Case 3 of Proposition 2 if* $\delta^+ = 0.4$*, in Case 1 if* $\delta^+ < 0.4$*, and in Case 2 if* $\delta^+ > 0.4$ *(indeed,* $\delta^+\lambda^+_1 = 0.4 \cdot 4 = 0.8 \cdot 2 = \delta^-\lambda^-_1$ *yields Case 3). The corresponding dynamics of signal mixes is illustrated in Figure 3.5.1 for the signal distribution* $s = (0, 0, 1, 1, 1)$ *and for values of parameter* δ^+ *close to the equality threshold of* 0.4*. The first four panels belong to Case 1 of Proposition 2, where signal mixes converge to* 0 *and misinformation prevails. The last four panels belong to Case 2, where signal mixes converge to* 1 *and misinformation vanishes. The middle panel induces Case 3 by setting* $\delta^+ = 0.4$*. Interestingly, in that case Agent 1's belief converges to* 0 *as* $x_1(\infty) < 0.5$*, whereas the other agent's beliefs converge to* 1 *as* $x_i(\infty) > 0.5$ *for* $i = 2, 3, 4, 5$*.*[15] *Hence, only the center of this star network, is misinformed in the limit (which holds for any initial distribution of signals that has at least one positive and one negative signal).*

To illustrate Corollary 1 we compute eigenvector centralities: $c^+_{1,2,3,4,5} = \frac{1}{5}$*,* $c^-_1 = \frac{1}{3}$ *and* $c^-_{2,3,4,5} = \frac{1}{6}$*. The ratios* $\frac{c^+_i}{c^-_i}$ *of relative importance in the positive network relative to the negative network for the four agents are hence* $(\frac{3}{5}, \frac{6}{5}, \frac{6}{5}, \frac{6}{5}, \frac{6}{5})$*. Thus, the first agent has a smaller centrality ratio than all other agents. The reason is that this agent, as the center of the star network, is better connected in the negative signals sharing network than the others, while all agents are equally well-connected in the positive signals sharing network, the complete network. By Corollary 1, this implies that this agent's signal mix will asymptotically always be closer to* 0 *than those of the other agents:* $x_1(t) < x_i(t)$*, for* $i = 2, 3, 4, 5$*. In all panels of Figure 3.5.1 this can be observed as Agent* 1*'s signal mix is below the others' signal mix, even if they converge to* 1 *or* 0*. Finally, observe that convergence is slower the closer the parameter* δ^+ *to the critical value that induces Case 3. Figure 3.5.2 shows half-life for all the values of parameter* δ^+*. The red line is half-life in Case 1, the blue line is half-life in Case 2, the black dot is half-life in Case 3, i.e. when* $\delta^+ = 0.4$*. Half-life "explodes" near this value, but is reasonably low for lower and higher values.*

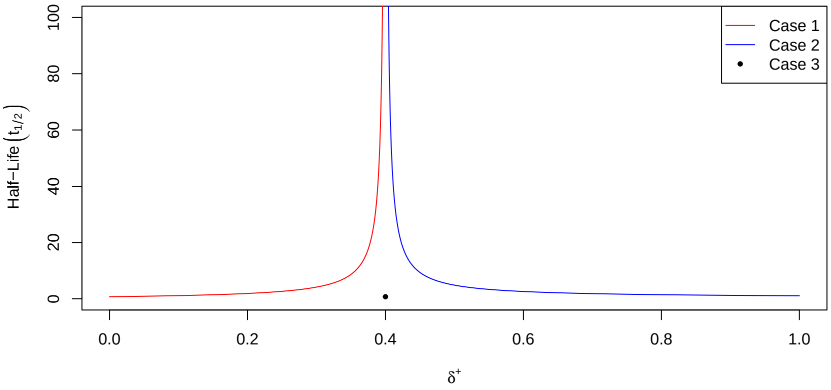
[15] Indeed, if the signal mix is above 0.5, the difference between positive and negative signals $k_i(t)$ grows such that the belief converges to 1, as more and more signals are acquired.

Figure 3.5.1: Illustration of key result.



Notes: Dynamics of signal mixes in Example 1 over time. Recall that signal mixes below 0.5 indicate misinformation. The first four panels with $\delta^+ < 0.4$ belong to Case 1 of Proposition 2. The middle panel with $\delta^+ = 0.4$ yields Case 3. The last four panels with $\delta^+ > 0.4$ belong to Case 2.

Figure 3.5.2: Speed of convergence in example 1.



Notes: Half-life in Example 1 for different values of δ^+. Setting $\delta^+ = 0.4$ yields Case 3 such that the black dot represents the relevant half-life. $\delta^+ < 0.4$ induces misinformation in the long-run, with half-life in red. Conversely, with $\delta^+ > 0.4$ misinformation vanishes in the long-run, with half-life in blue. Convergence is mostly fast as half-life is mostly low. Nearby the critical value of $\delta^+ = 0.4$ half-life “explodes”, meaning speed of convergence is extremely low.

3.5.4 Extension: heterogeneous relations

In this section, we extend the model by defining decay factors that are pair-specific. Formally, we consider two weighted graphs (N, M^+) and (N, M^-), where M^+ and M^- are $n \times n$ matrices with entries $m_{ij}^+ = \delta_{ij}^+ \in (0,1]$ or $m_{ij}^+ = 0$ and $m_{ij}^- = \delta_{ij}^- \in (0,1]$ or $m_{ij}^- = 0$. Like in the baseline model, a positive entry $\delta_{ij}^+ > 0$ is the decay of information for the pair ij, i.e. the fraction of positive signals that agent i receives when communicating with agent j. Signal accumulation becomes $N^+(t) = (I + M^+)N^+(t-1)$ for positive signals and likewise $N^-(t) = (I + M^-)N^-(t-1)$ for negative signals. In all other aspects, the extended model is defined as the baseline model, introduced in Section 3.3. In particular, we assume again that both networks are strongly connected.

This extension makes the model more flexible since it can accommodate many forms of heterogeneity. First, it allows for every pair i, j to have a different quality of communication and hence a pair-specific decay factor. Second, networks can be directed such that some agent i receives signals from another agent j, while j does not receive signals of i, i.e. $m_{ij}^+ = \delta_{ij} > 0$ and $m_{ji}^+ = 0$. Third, it allows an agent i to face a stronger decay as a receiver of information, i.e. $\delta_{ij} < \delta_{kj}$ for all j, k linked to i. This can incorporate in particular differences in discounting of received signals. Fourth, it allows for some agent i to face a stronger decay as a sender of information, i.e. $\delta_{ji} < \delta_{jk}$ for and all j, k linked to i. This could incorporate in particular differences in which a fraction of signals is shared. These variations can be applied at the individual level to single agents, or more broadly to groups of agents, as we will discuss in the next section.

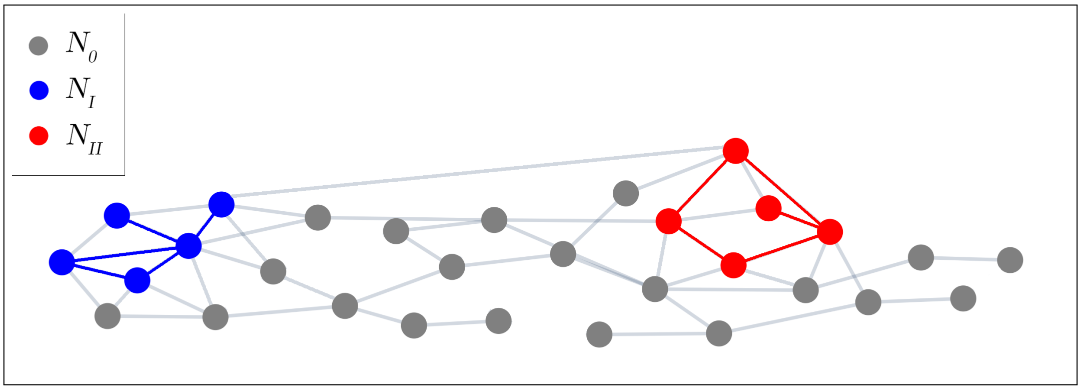
Even though this extension carries a rich array of new applications, the key result, Proposition 2, as well as its Corollary 1, neatly generalize. A noteworthy difference is that left and right eigenvectors do not coincide. The right eigenvectors determine the relations between the asymptotic signal mixes of agents, while the left eigenvectors determine the effect of agents' initial signals on the asymptotic signal mix, as Proposition 3.B.1 and Corollary 2 in the appendix demonstrate. The generalization shows that it is not essential that the network is undirected or that discounting is similar for different agents, while symmetry or asymmetry between positive and negative networks is crucial.

3.6 The effect of groups

To explore and illustrate how misinformation depends on the different asymmetries we introduced, we apply our model to group structures. The existence of groups, which are more connected amongst themselves and have similar sharing behaviors is an established feature of social networks and it has been identified as a source of the spread of misinformation (Zollo et al., 2017; Quattrociocchi et al., 2016). First, we illustrate network asymmetry and how a more connected group can determine the long run state. Second, we show that not only the number of links but also how they are distributed matters, namely that echo chambers generate a disproportional advantage. Third, we put in perspective what network asymmetry would be needed to compensate for a given decay asymmetry.

To this end, we use the following set up. Let a society be partitioned into three groups $N = \{N_0, N_I, N_{II}\}$ of sizes n_0, n_I, n_{II}. Members of N_I are more connected with each other in the positive signal sharing network, while N_{II} members are more connected with each other in the negative signal sharing network. Technically, we first generate a connected network (N, A) with α links. Then, we generate α_I^+ additional links in A^+ between members of group N_I. Finally, we construct A^+ as the union of the links of A and the additionally generated links. Likewise, we generate α_{II}^- additional links in A^- between members of group N_{II}. For illustration, one can think of N_I members as belonging to a social media group dedicated to scientific information (e.g. regarding vaccines), members of N_{II} belonging to a group dedicated to (e.g. vaccine) conspiracies and members of N_0 as the majority that belongs to no such group. The parameter space for such a society can be summarized by $(n_0, n_I, n_{II}, \alpha, \alpha_I^+, \alpha_{II}^-)$. Figure 3.6.1 illustrates such a society (with only 30 nodes for ease of readability) with $(n_0, n_I, n_{II}, \alpha, \alpha_I^+, \alpha_{II}^-) = (20, 5, 5, 40, 5, 5)$.

Figure 3.6.1: Illustration of a society with groups.



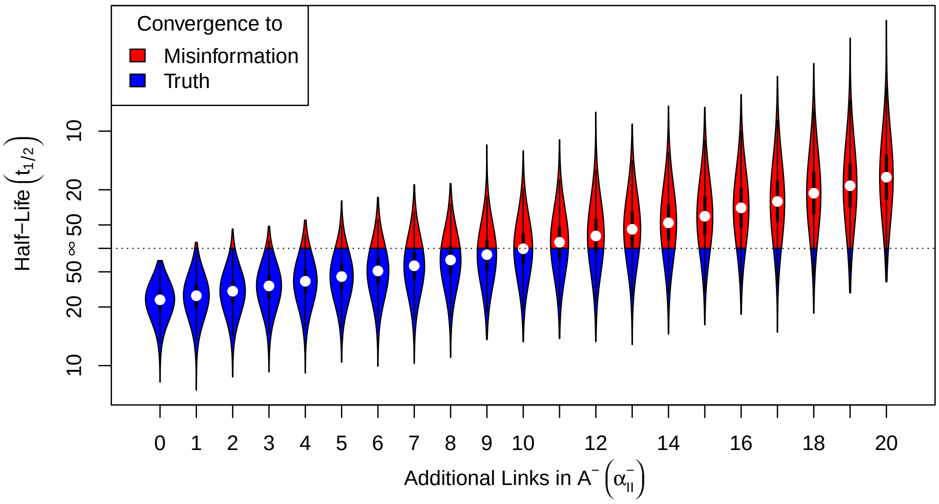
Notes: Members of N_0 and α common links are in grey. Members of N_I and their α_I^+ additional links are blue. Members of N_{II} and their α_{II}^- additional links are red. Positive signals travel over the A^+ network, which consists of the grey and blue links. Negative signals travel over the A^- network, which consists of the grey and red links.

3.6.1 Network asymmetry: more connected groups

To illustrate the effect of additional links, we use a society of 100 people in a connected ER random graph. These people are connected with a total of $\alpha = 250$ randomly generated links in the A network, meaning the average degree is 5. We then include the group structure and generate additional random links in the A^+ and A^- with the following parameters: $(n_0, n_I, n_{II}, \alpha, \alpha_I^+, \alpha_{II}^-) = (60, 20, 20, 250, 10, \alpha_{II}^-)$ with $\alpha_{II}^- \in [0, 20]$, i.e. Groups I and II consist of 20 agents each. Group I has 10 additional positive links. Group II has between 0 and 20 additional negative links.

The long-run signal mix as well as the speed of convergence to that signal mix depending on α_{II}^{-} is illustrated in Figure 3.6.2. As one would expect, the probability of this society being misinformed in the long run increases with α_{II}^{-}, i.e. the more negative links added to Group II. We also see that the larger the difference of number of links between the two networks, the faster the speed of convergence. We observe however that it is possible to have misinformation in the long run with values of $\alpha_{II}^{-} \leq \alpha_{I}^{+} = 10$. Societies that are more connected in the positive signal sharing network tend to reach the true state but can, with a lower probability, still converge to a misinformed state in the long run. Given the density of the A, A^{+} and A^{-} network is always exactly the same for a fixed set of parameters, it must then be that this fluctuation is a consequence of the randomness of the link generation process. We can also observe this with the violins getting thinner the more links are added. As more links are created, the randomness in the process gets increasingly influential in determining the outcome of the convergence, leading to higher variance around the median. We conclude that not only the number of links in a society, but also how they are distributed affects the long-run belief, as we further investigate in the next section.

Figure 3.6.2: Network asymmetry: the effect of additional links.



Notes: Results based on 10'000 replications per setting. Effect of varying α_{II}^{-} on misinformation and speed of convergence measured by half-life. α_{I}^{+} is constant and equal to 10. The networks are symmetric in expectations for $\alpha_{II}^{-} = 10$. For each violin, the red area represents the probability of misinformation for a given set of parameters. White dots represent the median, central black bars are the interquartile range and the thickness of the violin are density estimates.

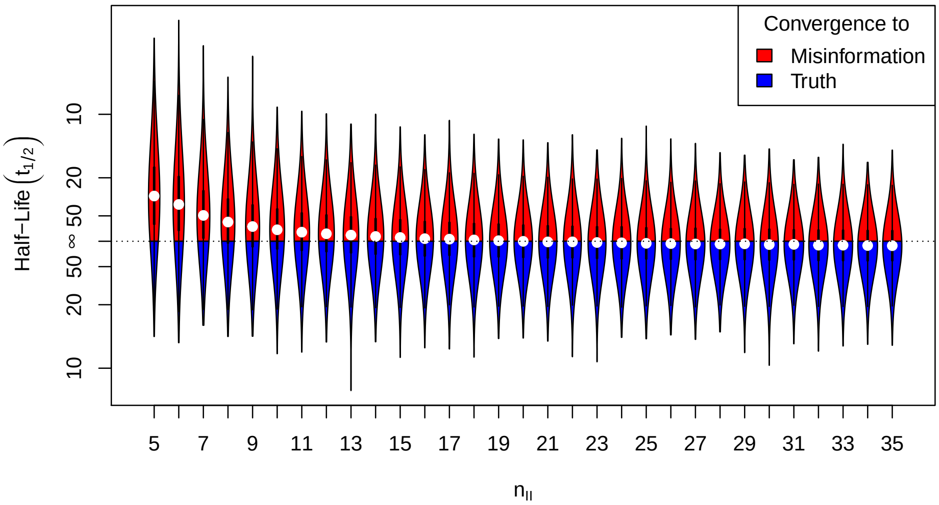
3.6.2 Network asymmetry: groups as echo chambers

To illustrate that a link can have more or less influence on the long-run belief of a society, we start from the same society as in the previous example. 100 people are in a connected ER random graph with an average degree of 5 in the A network. The parameters are the following: $(n_0, n_I, n_{II}, \alpha, \alpha_I^+, \alpha_{II}^-) = (80 - n_{II}, 20, n_{II}, 250, 10, 10)$ with $n_{II} \in [5, 35]$, i.e. Group I is of size 20, Group II's size ranges from 5 to 35, both groups have 10 additional links, which are added to their respective positive or negative networks. α_I^+ and n_I are fixed. Consequently all properties of the A^+ networks are in expectations the same across the parameter space and replications. However, despite a fixed α_{II}^-, the local density of A^- varies significantly depending on the size of n_{II}. When $n_{II} = 5$ a small clique, i.e. a fully connected subgroup is generated. We call such a group an echo chamber. When $n_{II} = 35$, the group is larger and the additional network between them is sparser. Figure 3.6.3 illustrates the effect of the distribution of the links in A^- on misinformation. For a fixed number of links, denser groups (i.e. for smaller values of n_{II}) tend to induce misinformation in the long run. This effect is stronger when generating highly connected groups as the difference in probability of misinformation between $n_{II} = 5$ and $n_{II} = 20$ is much more pronounced than between $n_{II} = 20$ and $n_{II} = 35$. The reason is that the high local density, works as an echo chamber, in which the signals of these agents are amplified.

3.6.3 Decay asymmetry vs. network asymmetry

Finally, to illustrate how additional links in the positive signal sharing network A^+ can compensate for decay favoring negative signals (or inversely), we start from the same society as in the two previous cases. 100 people are in a connected ER random graph with an average degree of 5 in the A network. We then add the group structure with the following parameters, which now include the respective decays δ^+ and δ^-: $(n_0, n_I, n_{II}, \alpha, \alpha_I^+, \alpha_{II}^-, \delta^+, \delta^-) = (100 - n_I, n_I, 0, 250, \alpha_I^+, 0, \delta^+, 1)$. We thus vary the asymmetry between decays, favoring negative signals given $\delta^- = 1$ and $\delta^+ \in [0, 1]$. The asymmetry between networks is also varied, favoring the positive signal sharing network, which is the only one to receive additional links. As we have established, how the additional α_I^+ links are added will influence the state reached by this society. We thus compare two link generation processes: random and echo chamber. In the random link generation process, we first add one link at random in A^+, which generates network asymmetry. We then compute what decay asymmetry it exactly compensates. We then add another link in A^+ at random, and again, compute the new decay asymmetry it compensates, and so on. In the echo chamber process, each additional link is created with the intent of generating an increasingly large echo chamber. The first additional link is generated between two random nodes, the second between these two nodes and a third. The third link will either increase the connectedness of the first three nodes, or, if all links are already present, create a link between a fourth node and the first three, and so on. At each step, we compute what decay asymmetry is compensated.

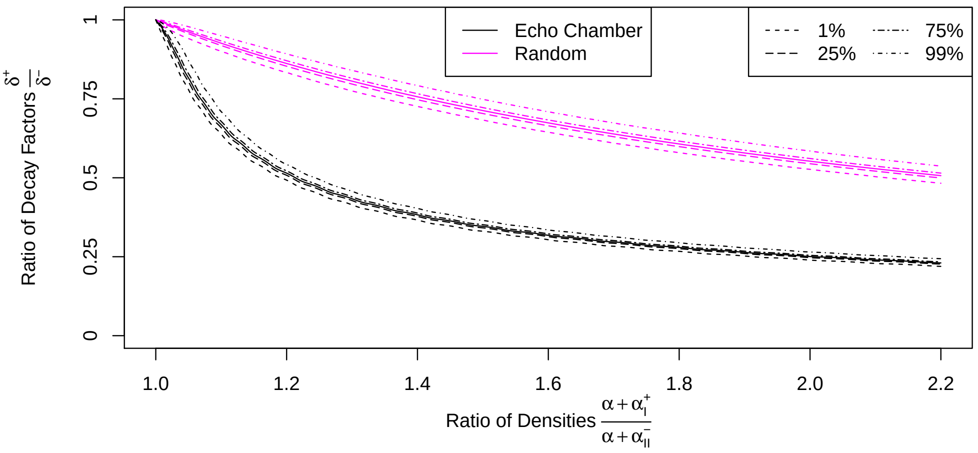
Figure 3.6.3: Network asymmetry: the echo chamber effect.



Notes: Results based on 10'000 replications per setting. Effect of varying local density while keeping number of links constant on misinformation and speed of convergence. α_I^+ and α_{II}^- are constant and equal to 10, meaning both group get 10 additional links in respectively the positive and negative network. The smaller n_{II}, the larger the local density, with a complete echo chamber at $n = 5$. The networks are symmetric in expectations for $n_{II} = 20$. For each violin, the red area represents the probability of misinformation for a given set of parameters. White dots represent the median, central black bars are the interquartile range and the thickness of the violin are density estimates.

In Figure 3.6.4, we show the ratio of links between A^+ and A^- needed to exactly compensate different levels of decay asymmetry, using the two different link generation processes. For example, if positive signals face twice the decay of negative signals, i.e. $\frac{\delta^+}{\delta^-} = \frac{0.5}{1} = 0.5$, then it takes more than twice as many random links in the positive network than in the negative network to compensate it, as the purple lines are around $\frac{\alpha+\alpha_I^+}{\alpha+\alpha_{II}^-} = \frac{250+\alpha_I^+}{250} \approx 2.2$. In contrast, for the same compensation it takes only about 20% more links if the new links are arranged in an echo chamber (as the black lines are roughly at ratio of densities $\frac{\alpha+\alpha_I^+}{\alpha+\alpha_{II}^-} = \frac{250+\alpha_I^+}{250} \approx 1.2$.) Consequently, if the ratio of densities is between these two values and still positive signals face twice the decay of negative signals, there is most likely misinformation when the positive links were added at random, while there is most likely convergence to the truth when these links are arranged as an echo chamber.

Figure 3.6.4: Compensating decay asymmetry with network asymmetry.



Notes: Results based on 10'000 replications per setting. Each line represents a different additional link generation process. The lines draw the median ratios of decay asymmetry that exactly compensate the ratio of links between the A^+ and A^- networks, generating Case 3 of Proposition 2. For parameters at the lower left of a line (with constant link generation process), a society is misinformed. For parameters at the upper right of a line, a society reaches the true state.

3.7 Discussion

3.7.1 Policy Implications

From our analysis, we can extract counter-measures in the fight against misinformation. We organize them into four categories.

Slowing decay of true information ($\delta^{+} \uparrow$). Recently, scholars have reported that the emotional content of true and false information is different. Scientific communication tends to generate more neutral and positive emotional reaction than unfounded conspiracies (Zollo et al., 2015). Similarly, false information in different topics such as business, politics, or technology tend to generate more surprise, fear and disgust than true information on the same topics. Vosoughi et al. (2018) argue that these emotional differences could explain why false information travels further. This implies that we have to consider how the emotional content of true information makes an impact on the receiver and how it will help or hamper its ability to travel. Easily forgettable information will decay faster.

Similarly, the medium has to be considered. The question of how to best get scientific information to the general public is not new (Miller, 2001), and has arguably still to be answered. Most scientific publications still exist only in article form, and are often not accessible to the general public. If they are, their complexity regularly constitutes an impenetrable barrier. In the meantime, false information is created in ways that is easily shareable online via social media or messaging apps such as images, videos and news-like article. Providers of scientific information are thus not always fighting with the appropriate tools and should consider how to tailor their message in a way that allows it to keeps its informative integrity, while being easily shareable.

Accelerating decay of false information ($\delta^{-} \downarrow$). Educating citizens appears to be an important way to accelerate the decay of false information. This education can take different forms. For example, educating to recognize false information, and training to question one's own information sharing behaviors. Mo and Mortensen (2019) have shown that people with higher information literacy scores are significantly better at identifying false information. Information literacy emphasizes the ability to navigate and locate information, to recognize opinion pieces or to search databases (Livingstone et al., 2008). Providing citizens opportunities to develop these skills is thus critical. On top, allowing people to reflect on their existing sharing behavior can prove effective, without even building new skills. Fazio (2020) reports that simply asking participants to pause and consider the veracity of a headline significantly decreases their intention to share false information. Pausing did not affect intentions to share true information. Similarly, priming the importance of accuracy in the mind of participants decreases their intent to share misinformation (Pennycook et al., 2021). Having well equipped citizens, who are able to recognize false information, is thus a necessity to slow down its spread. They should however not be left alone to carry this responsibility, simply due to the sheer amount and complexity of information to verify. Institutions are calling for the support of trusted fact-checkers which should facilitate the identification of false information for the end user (European Commission, 2018).

There is room for social media platforms to integrate these ways to accelerate the decay of false information. May it be by the integration of fact-checkers and subsequent flags in feeds or via introducing mechanisms that nudge users to reflect on what they share. How to best implement it is then still to be determined and tested.

Improving the true information sharing network ($\lambda_1^+ \uparrow$). The intuitive network policy to favor true information would be to increase the density of the true information sharing network. That is, adding links to said network. However, what exactly constitutes a *link* in an online social network is less clear than it seems in the models of social learning. In the context of this paper, what matters are links through which information can travel from one person to another, which happens online mainly through a timeline or feed, depending on which platform is used. The information one receives, is then the result of conscious choices of the user, through the people and pages she links with, and recommendations from algorithms that are based on previous behaviors and other undisclosed factors. This implies that out of all the possible pieces of information to display based on a user's full network, the algorithm influences which information to display first (or at all). The feed algorithm thus activates or silences existing links. On top, the algorithm can generate new links and display information from nodes that a user was not connected to, e.g. when advertising. Feed algorithms can hence generate new links, and do it. Social media platforms thus have the capacity to both favor displaying trusted sources of information when the link exists, or create it if it does not.

Weakening the false information sharing network ($\lambda_1^- \downarrow$). Finally, making the network over which false information is shared sparser favors reaching the true state. Our results in this respect concur with the theoretical results of (Törnberg, 2018) and the empirical observations of (Johnson et al., 2020); the organization in tight clusters, or echo chambers, of groups that spread misinformation increase their influence on the long-run beliefs. Breaking these echo chambers thus appears as one of the most efficient way to fight misinformation. However, silencing or breaking links that users have generated consciously raises ethical concerns, which are outside of the scope of this paper.

3.7.2 Limitations

Separating decay and signal type. The model assumes decay to be fully dependent on signal type. Agents share all true, and respectively false, signals in the same manner. More realistically, decay would be idiosyncratic to the signal and only partly correlated to signal type. Negative signals might generally be more shareable and hence decay less, but some positive signals might also be highly shareable.

Extending our model to allow for idiosyncratic decay factors would generalize our main condition $\delta^+\lambda_1^+ \lesseqgtr \delta^-\lambda_1^-$ to considering not the common decay factors, but the maximally realized decay factors in this equation, i.e. the highest decay among all positive signals and the highest decay among all negative signals. While gaining realism, this extension would not change the comparative-statics of decay asymmetry and network asymmetry.

Accounting for several networks. We assume the existence of two networks over which are shared true or false signals exclusively. In reality, there exists a complex overlay of different online networks, messaging apps, live social networks, etc. One can thus be connected with one another through many links. We partly cover this possibility in the extension in Section 3.5.4 as one can interpret the weighted matrix to reflect the strength of connections between agents across all existing networks collapsed into one number. However, another viable option would be extending the model to allow for multiple links between agents.

3.7.3 Conclusion

We have investigated asymmetric sharing of true and false information as a source of misinformation. Our results established that decay and network asymmetries determine the long-run state a society reaches. Though it is clear from our discussion that there are no silver bullets in the fight against misinformation, we can conclude that countermeasures should consider signal-dependent sharing behaviors. This aspect has not yet been studied in the (theoretical) literature on social learning, but may deserve more attention in the future.

Appendix

Chapter 3 Appendix

3.A Proofs

3.A.1 Proof of lemma 1

First of all, due to the definitions of $k_i(t) = N_i^+(t) - N_i^-(t)$ and $x_i(t) = \frac{N_i^+(t)}{N_i^+(t)+N_i^-(t)}$, it is easy to see that $k_i(t) = 2\left(x_i(t) - \frac{1}{2}\right)\left(N_i^+(t) + N_i^-(t)\right)$. Thus, Equation (3.3) can be rewritten as

$$b_i(t) = \frac{1}{1 + \frac{1-b^0}{b^0}\left(\frac{1-\rho}{\rho}\right)^{k_i(t)}} = \frac{1}{1 + \frac{1-b^0}{b^0}\left(\frac{1-\rho}{\rho}\right)^{2\left(x_i(t)-\frac{1}{2}\right)\left(N_i^+(t)+N_i^-(t)\right)}}$$

Thus, $b_i(t)$ is larger (smaller) than 0.5 if and only if $\frac{1-b^0}{b^0}\left(\frac{1-\rho}{\rho}\right)^{2\left(x_i(t)-\frac{1}{2}\right)\left(N_i^+(t)+N_i^-(t)\right)}$ is smaller (larger) than 1. For $b^0 = 0.5$, this is equivalent to $x_i(t) > 0.5$ ($x_i(t) < 0.5$), due to $\frac{1-\rho}{\rho}$ being smaller than 1 and $N_i^+(t) + N_i^-(t)$ being positive. As for $t \gg 0$, $N_i^+(t) + N_i^-(t)$ will be large,[16] $b_i(t)$ will be larger than 0.5 if $x_i(t) > 0.5$, and smaller than 0.5 if $x_i(t)$ is smaller than 0.5.

3.A.2 Proof of proposition 1

To prove Proposition 1 we apply Proposition 2, which is proven below. Using the result of Case 3 of Proposition 2 and $c^+ = c^- = c$, we find that

$$\begin{aligned}\lim_{t\to\infty} x_i(t) &= \frac{1}{1 + \frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{j=1}^n c_j^+ s_j}} = \frac{1}{1 + \frac{1-\sum_{j=1}^n c_j s_j}{\sum_{j=1}^n c_j s_j}} \\ &= \frac{1}{\frac{\sum_{j=1}^n c_j s_j + 1 - \sum_{j=1}^n c_j s_j}{\sum_{j=1}^n c_j s_j}} = \sum_{j=1}^n c_j s_j.\end{aligned}$$

[16] This follows i.a. from the representations for $N^+(t)$ and $N^-(t)$ given in Equations (3.A.2) and (3.A.3).

We now show that the probability of misinformation is bounded by 0.5, i.e.

$$P(x_i(\infty) < 0.5) + 0.5P(x_i(\infty) = 0.5) \leq 0.5 \text{ for all } i.$$

This amounts to showing that

$$P\left(\sum_{j=1}^{n} c_j s_j < \frac{1}{2}\right) + \frac{1}{2}P\left(\sum_{j=1}^{n} c_j s_j = \frac{1}{2}\right) \leq \frac{1}{2}, \tag{3.A.1}$$

as $x_i(\infty) = \sum_{j=1}^{n} c_j s_j$ for all i. In order to prove Equation (3.A.1), we define the following quantities: $p_l(\rho) := P\left(\sum_{j=1}^{n} c_j s_j < \frac{1}{2}\right)$, $p_m(\rho) := P\left(\sum_{j=1}^{n} c_j s_j = \frac{1}{2}\right)$, and $p_u(\rho) := P\left(\sum_{j=1}^{n} c_j s_j > \frac{1}{2}\right)$, which obviously sum to unity: $p_l(\rho) + p_m(\rho) + p_u(\rho) = 1$ for all ρ. Using these, Equation (3.A.1) can be restated as $p_l(\rho) + \frac{1}{2}p_m(\rho) \leq \frac{1}{2}$, which is equivalent to $2p_l(\rho) + p_m(\rho) \leq 1 = p_l(\rho) + p_m(\rho) + p_u(\rho)$, which is in turn equivalent to $p_l(\rho) \leq p_u(\rho)$. This part of the proof can therefore be completed by showing that indeed $p_l(\rho) \leq p_u(\rho)$ for all $\rho > \frac{1}{2}$. Due to $\sum_{j=1}^{n} c_j(1-s_j) = \sum_{j=1}^{n} c_j - \sum_{j=1}^{n} c_j s_j = 1 - \sum_{j=1}^{n} c_j s_j$, the condition $\sum_{j=1}^{n} c_j s_j < \frac{1}{2}$ is equivalent to $\sum_{j=1}^{n} c_j(1-s_j) > \frac{1}{2}$. As $1 - s_j$ takes the values 0 and 1 with probabilities ρ and $1 - \rho$, respectively, we find that

$$p_l(\rho) = P\left(\sum_{j=1}^{n} c_j s_j < \frac{1}{2}\right) = P\left(\sum_{j=1}^{n} c_j(1 - s_j) > \frac{1}{2}\right) = p_u(1-\rho) \leq p_u(\rho),$$

where the inequality at the end holds true because $1 - \rho < \rho$, due to $\rho > \frac{1}{2}$. We thus have shown that $p_l(\rho) \leq p_u(\rho)$, concluding the proof.

3.A.3 Proof of proposition 2

To begin the proof, we use the eigendecompositions of the real symmetric matrices A^+ and A^-, writing them as $A^+ = Q^+\Lambda^+\left(Q^+\right)^\top$ and $A^- = Q^-\Lambda^-\left(Q^-\right)^\top$, respectively, with Q^+ and Q^- being orthogonal matrices whose columns are eigenvectors of A^+ and A^-, and Λ^+ and Λ^- being diagonal matrices whose entries are the eigenvalues of A^+ and A^-. From this, we have that $I + \delta^+A^+ = Q^+\left(I + \delta^+\Lambda^+\right)\left(Q^+\right)^\top$ and $I + \delta^-A^- = Q^-\left(I + \delta^-\Lambda^-\right)\left(Q^-\right)^\top$ as well as $\left(I + \delta^+A^+\right)^t = Q^+\left(I + \delta^+\Lambda^+\right)^t\left(Q^+\right)^\top$ and $\left(I + \delta^-A^-\right)^t = Q^-\left(I + \delta^-\Lambda^-\right)^t\left(Q^-\right)^\top$ for all t. Overall, this delivers

$$\left(I + \delta^+A^+\right)^t = \sum_{i=1}^{n}\left(1 + \delta^+\lambda_i^+\right)^t q_i^+\left(q_i^+\right)^\top, \; \left(I + \delta^-A^-\right)^t = \sum_{i=1}^{n}\left(1 + \delta^-\lambda_i^-\right)^t q_i^-\left(q_i^-\right)^\top,$$

with q_i^+ and q_i^- ($i = 1, \ldots, n$) denoting the eigenvectors of A^+ and A^-, respectively. Denoting the vector of initial signals by s, we thus get:

$$N^+(t) = \left(I + \delta^+A^+\right)^t N^+(0) = \left(I + \delta^+A^+\right)^t s = \sum_{j=1}^{n}\left(1 + \delta^+\lambda_j^+\right)^t q_j^+\left(q_j^+\right)^\top s,$$

$$N^-(t) = \left(I + \delta^- A^-\right)^t N^-(0) = \left(I + \delta^+ A^+\right)^t (\mathbb{1} - s) = \sum_{j=1}^{n} \left(1 + \delta^- \lambda_j^-\right)^t q_j^- \left(q_j^-\right)^\top (\mathbb{1} - s).$$

From this, we get for the numbers of positive signals at time t,

$$N^+(t) = \left(1 + \delta^+ \lambda_1^+\right)^t \left(q_1^+ \left(q_1^+\right)^\top s + \sum_{j=2}^{n} \left(\frac{1 + \delta^+ \lambda_j^+}{1 + \delta^+ \lambda_1^+} \right)^t q_j^+ \left(q_j^+\right)^\top s \right), \tag{3.A.2}$$

and for the numbers of negative signals at time t,

$$N^-(t) = \left(1 + \delta^- \lambda_1^-\right)^t \left(q_1^- \left(q_1^-\right)^\top (\mathbb{1} - s) + \sum_{j=1}^{n} \left(\frac{1 + \delta^- \lambda_j^-}{1 + \delta^- \lambda_1^-} \right)^t q_j^- \left(q_j^-\right)^\top (\mathbb{1} - s) \right). \tag{3.A.3}$$

Due to Perron-Frobenius theory, it is clear that $\dfrac{1 + \delta^+ \lambda_j^+}{1 + \delta^+ \lambda_1^+}$ and $\dfrac{1 + \delta^- \lambda_j^-}{1 + \delta^- \lambda_1^-}$ are both smaller than 1 for $j = 2, \ldots, n$, implying

$$\left(1 + \delta^+ \lambda_1^+\right)^{-t} N^+(t) \overset{t \to \infty}{\longrightarrow} q_1^+ \left(q_1^+\right)^\top s, \; \left(1 + \delta^+ \lambda_1^-\right)^{-t} N^-(t) \overset{t \to \infty}{\longrightarrow} q_1^- \left(q_1^-\right)^\top (\mathbb{1} - s).$$

Now, using that $q_1^+ = \dfrac{c^+}{||c^+||} = \dfrac{c^+}{\sqrt{\sum_{k=1}^{n} \left(c_k^+\right)^2}}$ and $q_1^- = \dfrac{c^-}{||c^-||} = \dfrac{c^-}{\sqrt{\sum_{k=1}^{n} \left(c_k^-\right)^2}}$, we finally get:

$$\left(1 + \delta^+ \lambda_1^+\right)^{-t} N^+(t) \overset{t \to \infty}{\longrightarrow} c^+ \frac{\left(c^+\right)^\top s}{\sum_{k=1}^{n} \left(c_k^+\right)^2} = c^+ \frac{\sum_{j=1}^{n} c_j^+ s_j}{\sum_{k=1}^{n} \left(c_k^+\right)^2}, \tag{3.A.4}$$

$$\left(1 + \delta^- \lambda_1^-\right)^{-t} N^-(t) \overset{t \to \infty}{\longrightarrow} c^- \frac{\left(c^-\right)^\top (\mathbb{1} - s)}{\sum_{k=1}^{n} \left(c_k^-\right)^2} = c^- \frac{\sum_{j=1}^{n} c_j^- (1 - s_j)}{\sum_{k=1}^{n} \left(c_k^-\right)^2} = c^- \frac{1 - \sum_{j=1}^{n} c_j^- s_j}{\sum_{k=1}^{n} \left(c_k^-\right)^2}. \tag{3.A.5}$$

With these general results at hand, we can address the three cases considered in Proposition 2.

1. For $\left(\dfrac{1 + \delta^+ \lambda_1^+}{1 + \delta^- \lambda_1^-}\right)^{-t} x_i(t)$, we get when $\delta^+ \lambda_1^+ < \delta^- \lambda_1^-$:

$$\begin{aligned} \left(\frac{1 + \delta^+ \lambda_1^+}{1 + \delta^- \lambda_1^-}\right)^{-t} x_i(t) &= \left(\frac{1 + \delta^+ \lambda_1^+}{1 + \delta^- \lambda_1^-}\right)^{-t} \frac{N_i^+(t)}{N_i^+(t) + N_i^-(t)} \\ &= \frac{\left(1 + \delta^+ \lambda_1^+\right)^{-t} N_i^+(t)}{\left(1 + \delta^- \lambda_1^-\right)^{-t} \left(N_i^+(t) + N_i^-(t)\right)} \end{aligned}$$

$$= \frac{\left(1+\delta^+\lambda_1^+\right)^{-t} N_i^+(t)}{\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t \left(1+\delta^+\lambda_1^+\right)^{-t} N_i^+(t) + \left(1+\delta^-\lambda_1^-\right)^{-t} N_i^-(t)}$$

$$\xrightarrow{t\to\infty} \frac{c_i^+ \frac{\sum\limits_{j=1}^n c_j^+ s_j}{\sum\limits_{k=1}^n \left(c_k^+\right)^2}}{c_i^- \frac{1-\sum\limits_{j=1}^n c_j^- s_j}{\sum\limits_{k=1}^n \left(c_k^-\right)^2}} = \frac{c_i^+}{c_i^-} \frac{\sum\limits_{k=1}^n (c_k^-)^2}{\sum\limits_{k=1}^n (c_k^+)^2} \frac{\sum\limits_{j=1}^n c_j^+ s_j}{1-\sum\limits_{j=1}^n c_j^- s_j}.$$

2. For $\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^{-t}(1-x_i(t))$, we get when $\delta^+\lambda_1^+ > \delta^-\lambda_1^-$:

$$\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^{-t}(1-x_i(t)) = \left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^{-t} \frac{N_i^-(t)}{N_i^+(t)+N_i^-(t)}$$

$$= \frac{\left(1+\delta^-\lambda_1^-\right)^{-t} N_i^-(t)}{\left(1+\delta^+\lambda_1^+\right)^{-t}\left(N_i^+(t)+N_i^-(t)\right)}$$

$$= \frac{\left(1+\delta^-\lambda_1^-\right)^{-t} N_i^-(t)}{\left(1+\delta^+\lambda_1^+\right)^{-t} N_i^+(t) + \left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^t \left(1+\delta^-\lambda_1^-\right)^{-t} N_i^-(t)}$$

$$\xrightarrow{t\to\infty} \frac{c_i^- \frac{1-\sum\limits_{j=1}^n c_j^- s_j}{\sum\limits_{k=1}^n \left(c_k^-\right)^2}}{c_i^+ \frac{\sum\limits_{j=1}^n c_j^+ s_j}{\sum\limits_{k=1}^n \left(c_k^+\right)^2}} = \frac{c_i^-}{c_i^+} \frac{\sum\limits_{k=1}^n (c_k^+)^2}{\sum\limits_{k=1}^n (c_k^-)^2} \frac{1-\sum\limits_{j=1}^n c_j^- s_j}{\sum\limits_{j=1}^n c_j^+ s_j}.$$

3. Finally, when $\delta^+\lambda_1^+ = \delta^-\lambda_1^-$, we get:

$$x_i(t) = \frac{\left(1+\delta^+\lambda_1^+\right)^{-t} N_i^+(t)}{\left(1+\delta^+\lambda_1^+\right)^{-t} N_i^+(t) + \left(1+\delta^-\lambda_1^-\right)^{-t} N_i^-(t)}$$

$$\xrightarrow{t\to\infty} \frac{c_i^+ \frac{\sum\limits_{j=1}^n c_j^+ s_j}{\sum\limits_{k=1}^n \left(c_k^+\right)^2}}{c_i^+ \frac{\sum\limits_{j=1}^n c_j^+ s_j}{\sum\limits_{k=1}^n \left(c_k^+\right)^2} + c_i^- \frac{1-\sum\limits_{j=1}^n c_j^- s_j}{\sum\limits_{k=1}^n \left(c_k^-\right)^2}} = \frac{1}{1+\frac{c_i^-}{c_i^+}\frac{\sum\limits_{k=1}^n (c_k^+)^2}{\sum\limits_{k=1}^n (c_k^-)^2}\frac{1-\sum\limits_{j=1}^n c_j^- s_j}{\sum\limits_{j=1}^n c_j^+ s_j}}.$$

3.A.4 Proof of corollary 1

First of all, notice that $\frac{N_i^+(t)}{N_i^-(t)}\Big/\frac{N_j^+(t)}{N_j^-(t)}$ can be written as $\frac{x_i(t)(1-x_j(t))}{x_j(t)(1-x_i(t))}$. We thus have to prove that $\lim\limits_{t\to\infty}\frac{x_i(t)(1-x_j(t))}{x_j(t)(1-x_i(t))} = \frac{c_i^+ c_j^-}{c_i^- c_j^+}$, which we will do by successively tackling the three cases given in Proposition 2. For Case 1, we know that $\delta^+\lambda_1^+ < \delta^-\lambda_1^-$ and

$$x_i(t)\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^t \xrightarrow{t\to\infty} \frac{c_i^+}{c_i^-}\frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2}\frac{\sum_{l=1}^n c_l^+ s_l}{1-\sum_{l=1}^n c_l^- s_l}.$$

as well as $x_i(t) \xrightarrow{t\to\infty} 0$ for all i. Trivially, thus, $1-x_i(t)$ and $1-x_j(t)$ each converge to 1. For

$$\frac{x_i(t)}{x_j(t)} = \frac{x_i(t)\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^t}{x_j(t)\left(\frac{1+\delta^-\lambda_1^-}{1+\delta^+\lambda_1^+}\right)^t},$$

we then find that it converges to

$$\frac{\frac{c_i^+}{c_i^-}\frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2}\frac{\sum_{l=1}^n c_l^+ s_l}{1-\sum_{l=1}^n c_l^- s_l}}{\frac{c_j^+}{c_j^-}\frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2}\frac{\sum_{l=1}^n c_l^+ s_l}{1-\sum_{l=1}^n c_l^- s_l}} = \frac{\frac{c_i^+}{c_i^-}}{\frac{c_j^+}{c_j^-}},$$

which proves the assertion for Case 1.

For Case 2, we know that $\delta^+\lambda_1^+ > \delta^-\lambda_1^-$ and

$$(1-x_i(t))\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t \xrightarrow{t\to\infty} \frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{\sum_{l=1}^n c_l^- s_l}{1-\sum_{l=1}^n c_l^+ s_l}.$$

as well as $x_i(t) \xrightarrow{t\to\infty} 1$ for all i. Trivially, thus, $x_i(t)$ and $x_j(t)$ each converge to 1. For

$$\frac{1-x_j(t)}{1-x_i(t)} = \frac{(1-x_j(t))\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t}{(1-x_i(t))\left(\frac{1+\delta^+\lambda_1^+}{1+\delta^-\lambda_1^-}\right)^t},$$

we then find that it converges to

$$\frac{\frac{c_j^-}{c_j^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{\sum_{l=1}^n c_l^- s_l}{1-\sum_{l=1}^n c_l^+ s_l}}{\frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{\sum_{l=1}^n c_l^- s_l}{1-\sum_{l=1}^n c_l^+ s_l}} = \frac{\frac{c_i^+}{c_i^-}}{\frac{c_j^+}{c_j^-}},$$

which proves the assertion for Case 2.

Finally, for the Case 3, we know that for all i

$$\lim_{t\to\infty} x_i(t) = \frac{1}{1 + \frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{1-\sum_{l=1}^n c_l^- s_l}{\sum_{l=1}^n c_l^+ s_l}} \in (0,1).$$

This immediately implies that

$$\lim_{t\to\infty} 1 - x_i(t) = \frac{\frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{l=1}^n c_l^+ s_l}}{1 + \frac{c_i^-}{c_i^+}\frac{\sum_{k=1}^n (c_k^+)^2}{\sum_{k=1}^n (c_k^-)^2}\frac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{l=1}^n c_l^+ s_l}} \text{ as well as}$$

$$\lim_{t\to\infty} \frac{x_i(t)}{1-x_i(t)} = \frac{c_i^+}{c_i^-}\frac{\sum_{k=1}^n (c_k^-)^2}{\sum_{k=1}^n (c_k^+)^2}\frac{\sum_{l=1}^n c_l^+ s_l}{1-\sum_{l=1}^n c_l^- s_l}$$

for all i, from which it easily follows that $\lim_{t\to\infty} \frac{x_i(t)(1-x_j(t))}{x_j(t)(1-x_i(t))} = \frac{c_i^+ c_j^-}{c_i^- c_j^+}$.

3.B Extension: heterogeneous relations

3.B.1 Extended key result

Denote by $\lambda_1(M^+)$ the largest eigenvalue of matrix M^+ and denote by c^+, d^+ the corresponding right and left eigenvector, normalized such that $\sum_{j=1}^n c_j^+ = 1 = \sum_{j=1}^n d_j^+$.[17] Likewise, let $\lambda_1(M^-)$ be the largest eigenvalue of matrix M^- and denote by c^-, d^- the corresponding normalized right and left eigenvector. Notice that these eigenvalues and eigenvectors now contain information not only about network asymmetry, but also about decay asymmetries, as the weights δ_{ij}^+ and δ_{ij}^- have already entered the matrices M^+ and M^-. When these matrices are considered as weighted networks, c^+ and c^- are called eigenvector centrality or right-hand eigenvector centrality of M^+ and M^- (Bonacich, 1987), while d^+ and d^- can be called left-hand eigenvector centrality (e.g. Golub and Sadler, 2016).

Proposition 3.B.1 (Extended Key Result)**.** *Suppose that the initial distribution of signals contains at least one positive and at least one negative signal.*

[17] The two eigenvectors c^+ and d^+ coincide in the special case that the matrix M^+ is symmetric.

1. *If $\lambda_1(M^+) < \lambda_1(M^-)$, then for all i and large t*

$$x_i(t) \approx \frac{c_i^+}{c_i^-} \left(\frac{1+\lambda_1(M^+)}{1+\lambda_1(M^-)} \right)^t \frac{\sum_{k=1}^n c_k^- d_k^-}{\sum_{k=1}^n c_k^+ d_k^+} \frac{\sum_{j=1}^n d_j^+ s_j}{1-\sum_{j=1}^n d_j^- s_j}$$

such that $\lim_{t\to\infty} x_i(t) = 0$. Hence, misinformation prevails.

2. *If $\lambda_1(M^+) > \lambda_1(M^-)$, then for all i and large t:*

$$x_i(t) \approx 1 - \frac{c_i^-}{c_i^+} \left(\frac{1+\lambda_1(M^-)}{1+\lambda_1(M^+)} \right)^t \frac{\sum_{k=1}^n c_k^+ d_k^+}{\sum_{k=1}^n c_k^- d_k^-} \frac{1-\sum_{j=1}^n d_j^- s_j}{\sum_{j=1}^n d_j^+ s_j}$$

such that $\lim_{t\to\infty} x_i(t) = 1$. Hence, misinformation vanishes.

3. *If $\lambda_1(M^+) = \lambda_1(M^-)$, then for all i:*

$$\lim_{t\to\infty} x_i(t) = \frac{1}{1 + \frac{c_i^-}{c_i^+} \frac{\sum_{k=1}^n c_k^+ d_k^+}{\sum_{k=1}^n c_k^- d_k^-} \frac{1-\sum_{j=1}^n d_j^- s_j}{\sum_{j=1}^n d_j^+ s_j}}.$$

Hence, long-run misinformation depends on the signal distribution and on the eigenvectors that correspond to the largest eigenvalues.

Proposition 3.B.1 first of all shows that the key result obtained in our baseline model (Proposition 2) is robust to the broad generalization. Second, the crucial condition is now expressed in terms of the largest eigenvalues $\lambda_1(M^+)$ and $\lambda_1(M^-)$. Again there are special cases in which it is clear which of those is larger. For instance, suppose every element of matrix M^+ is weakly smaller than every element of M^-, i.e. for all i, j, $m_{ij}^+ \leq m_{ij}^-$; and for some the inequality is strict. Then $\lambda_1(M^+) < \lambda_1(M^-)$ such that we are in Case 1 and misinformation prevails. Analogously, for the situation that M^+ is element-wise bigger than M^-. Similarly, if $M^+ = M^-$, we are clearly in Case 3 such that misinformation depends on the distribution of initial signals and on the eigenvectors corresponding to the largest eigenvalue. Then, the eigenvectors for the positive and the negative signal sharing networks coincide: $c^+ = c^-$ and $d^+ = d^-$ and long-run misinformation can be expressed in terms of eigenvector centrality with respect to the weighted network $(N, M^+) = (N, M^-)$. Third, the dynamics in the three cases correspond to the dynamics characterized in the baseline model. For instance, the observation in Proposition 2 Case 3 that the signal mix is increasing in s_j translates into the same observation in Case 3 of Proposition 3.B.1. The speed of convergences can also be assessed like in the baseline model. For instance, $\lambda_1(M^+) < \lambda_1(M^-)$ (Case 1), then speed can be measured by $\log\left(\frac{1+\lambda_1(M^-)}{1+\lambda_1(M^+)}\right)$. Finally, the fact that decay and network asymmetry are mingled into single matrices M^+ and M^- makes it more difficult to disentangle their effects.

Corollary 2 (Centrality Ratios and Opinion Diversity)**.** *Suppose that the initial distribution of signals contains at least one positive and at least one negative signal. Then, the ratio of two agents' ratios of positive over negative signals converges to these agents' ratio of centrality ratios, i.e.*

$$\lim_{t\to\infty} \frac{N_i^+(t)}{N_i^-(t)} \Big/ \frac{N_j^+(t)}{N_j^-(t)} = \frac{c_i^+}{c_i^-} \Big/ \frac{c_j^+}{c_j^-}. \tag{3.B.1}$$

Hence, an agent i with higher centrality ratio than another agent j has a higher asymptotic signal mix and belief, i.e. if $\frac{c_i^+}{c_i^-} > \frac{c_j^+}{c_j^-}$, then for large t, $x_i(t) > x_j(t)$ and $b_i(t) > b_j(t)$.

Eigenvector centralities c and their ratios govern the relations of asymptotic signal mixes between agents.

The results under symmetry (Section 3.4), do not change in the extended model, i.e. if $\delta_{ij}^+ = \delta_{ij}^-$ and $a_{ij}^+ = a_{ij}^-$ for all i, j, then the results of Proposition 1 stay essentially unchanged. The only difference is that every appearance of (left and right) eigenvector centrality c, in Proposition 1 has to be replaced by d, the right eigenvector centrality. The important point of symmetry thus is symmetry with respect to positive and negative networks, but not symmetry of the matrices A or M: It is not essential that the network is undirected or that the discounting is symmetric in the sense that $\delta_{ij} = \delta_{ji}$. The only thing that matters is the symmetry between positive and negative networks.

3.B.2 Proof of proposition 3.B.1

In order to prove the assertions, we will show that

$$\Big(1 + \lambda_1(M^+)\Big)^{-t} N^+(t) \overset{t\to\infty}{\longrightarrow} c^+ \frac{\left(d^+\right)^\top s}{\sum\limits_{k=1}^{n} c_k^+ d_k^+} = c^+ \frac{\sum\limits_{j=1}^{n} d_j^+ s_j}{\sum\limits_{k=1}^{n} c_k^+ d_k^+}, \tag{3.B.2}$$

$$\left(1 + \lambda_1(M^-)\right)^{-t} N^-(t) \overset{t\to\infty}{\longrightarrow} c^- \frac{\left(d^-\right)^\top (\mathbb{1} - s)}{\sum\limits_{k=1}^{n} c_k^- d_k^-} = c^- \frac{1 - \sum\limits_{j=1}^{n} d_j^- s_j}{\sum\limits_{k=1}^{n} c_k^- d_k^-}. \tag{3.B.3}$$

With Equations (3.B.2) and (3.B.3) at hand, the assertions of Proposition 3.B.1 then follow from exactly the same arguments as those given in the proof of Proposition 2 after Equations (3.A.4) and (3.A.5).

As the essential parts of Equations (3.B.2) and (3.B.3) coincide, determining the limits of $\left(1 + \lambda_1(M^+)\right)^{-t} N^+(t)$ is completely analogous to determining the limit of $\left(1 + \lambda_1(M^-)\right)^{-t} N^-(t)$. Thus, we will do this in one sweep by looking at the limit of $\left(1 + \lambda_1(M)\right)^{-t} (I + M)^t$, where M and stands for M^+ and M^-, respectively. The proof will thus be complete when showing that

$$\left(1 + \lambda_1(M)\right)^{-t} (I + M)^t \overset{t\to\infty}{\longrightarrow} \frac{c\, d^\top}{c^\top d}, \tag{3.B.4}$$

where c (d) stands for c^+ and c^- (d^+ and d^-), respectively.[18] In order to prove Equation (3.B.4), we first rewrite M using its Jordan normal form: $M = SJS^{-1}$, where J is a block diagonal matrix

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{pmatrix}$$

formed of Jordan blocks J_i ($i = 1, \dots, p$), which are either scalars consisting of eigenvalues λ_i of M or have the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}.$$

Due to the network being strongly connected and M containing only non-negative entries, M is irreducible and Perron-Frobenius theory allows to infer that the spectral radius of M is a simple eigenvalue of M. We will assume without loss of generality that this value, $\lambda_1(M)$, corresponds to the matrix J_1. As $\lambda_1(M)$ is the spectral radius of M, we also know that $|\lambda_i| \leq \lambda_1(M)$ for all $i > 1$. From all this, by setting $\widetilde{\lambda}_i := \frac{1+\lambda_i}{1+\lambda_1(M)}$ we find that $\widetilde{M} := \frac{1}{1+\lambda_1(M)}(I + M) = S\widetilde{J}S^{-1}$, with

$$\widetilde{J} = \begin{pmatrix} 1 & & & \\ & \widetilde{J}_2 & & \\ & & \ddots & \\ & & & \widetilde{J}_p \end{pmatrix}, \; \widetilde{J}_i = \begin{pmatrix} \widetilde{\lambda}_i & 1 & & \\ & \widetilde{\lambda}_i & \ddots & \\ & & \ddots & 1 \\ & & & \widetilde{\lambda}_i \end{pmatrix} (i = 2, \dots, p).$$

Additionally, we know that $\widetilde{\lambda}_i < 1$ due to $|\lambda_i| \leq \lambda_1(M)$ and λ_i being different from $\lambda_1(M)$. Taking all this together, we find that

$$(1 + \lambda_1(M))^{-t} (I + M)^t = \widetilde{M}^t = S\widetilde{J}^t S^{-1} = S \begin{pmatrix} 1 & & & \\ & \widetilde{J}_2^t & & \\ & & \ddots & \\ & & & \widetilde{J}_p^t \end{pmatrix} S^{-1}.$$

With respect to the terms $\widetilde{J}_i^t$, it is well known (and easy to prove) that for large t

$$\widetilde{J}_i^t = \begin{pmatrix} \widetilde{\lambda}_i^t & \binom{t}{1}\widetilde{\lambda}_i^{t-1} & \binom{t}{2}\widetilde{\lambda}_i^{t-2} & \cdots \\ & \widetilde{\lambda}_i^t & \ddots & \vdots \\ & & \ddots & \binom{t}{1}\widetilde{\lambda}_i^{t-1} \\ & & & \widetilde{\lambda}_i^t \end{pmatrix},$$

[18] Additionally, Equation (3.B.4) implies that for Cases 1 and 2, the formulas given in the main text for speed of convergence remain valid in the generalized setting of Proposition 3.B.1.

implying that all $\widetilde{J}_i^t$ shrink to 0 due to $|\widetilde{\lambda}_i| < 1$, entailing that the rate of convergence of $\widetilde{M}^t$ is essentially being determined by $\max\{\widetilde{\lambda}_i : i = 2, \ldots, p\}$.[19] For the limit of $\widetilde{M}^t$, we thus have: $\widetilde{M}^t \overset{t\to\infty}{\longrightarrow} Se_1e_1^\top S^{-1}$. Setting $u := Se_1$ and $v := S^{-\top}e_1$, we can rewrite this as $\widetilde{M}^t \overset{t\to\infty}{\longrightarrow} u\,v^\top$. The following derivations show that u is a right eigenvector of M for $\lambda_1(M)$, while v is a corresponding left eigenvector:

$$Mu = MSe_1 = SJS^{-1}Se_1 = SJe_1 = S\lambda_1(M)e_1 = \lambda_1(M)u,$$

$$\begin{aligned} v^\top M &= \left(S^{-\top}e_1\right)^\top M = e_1^\top S^{-1}M = e_1^\top S^{-1}SJS^{-1} = e_1^\top JS^{-1} = \lambda_1(M)e_1^\top S^{-1} \\ &= \lambda_1(M)\left(S^{-\top}e_1\right)^\top = \lambda_1(M)v^\top. \end{aligned}$$

As the left and right eigenvectors of M for $\lambda_1(M)$ are unique up to multiplying by a constant, $uv^\top$ and $cd^\top$ differ only by a constant: $uv^\top = \alpha cd^\top$ for some constant α. Now, from $uv^\top uv^\top = Se_1e_1^\top S^{-1}Se_1e_1^\top S^{-1} = Se_1e_1^\top e_1e_1^\top S^{-1} = Se_1e_1^\top S^{-1} = uv^\top$, we find that $uv^\top uv^\top = \alpha cd^\top \alpha cd^\top$ must equal $uv^\top = \alpha cd^\top$, thus we have $\alpha^2 d^\top c = \alpha$, implying $\alpha = \frac{1}{d^\top c}$ and $\widetilde{M}^t \overset{t\to\infty}{\longrightarrow} \frac{cd^\top}{c^\top d}$, proving Equation (3.B.4) and concluding the proof.

3.B.3 Proof of corollary 2

The proof of this corollary is perfectly analogous to the one of Corollary 1, building on Proposition 3.B.1 instead of Proposition 2.

3.C Further appendices

3.C.1 Additional examples for the benchmark case of symmetry

To illustrate occurrence of misinformation in the setting of symmetry, we study two extreme examples: A regular graph in Example 3.C.1 and a network with a clique of five in Example 3.C.2.

Example 3.C.1 (Regular network)**.** *Consider network (N, A) that is connected and regular of degree k, i.e. every agent has exactly k links.*

Regularity of degree k implies that the largest eigenvalue is $\lambda_1 = k$ and eigenvector centrality is $c = (\frac{1}{n}, ..., \frac{1}{n})$. From Proposition 1, $\lim_{t\to\infty} x_i(t) = \frac{1}{n}\sum_{j=1}^n s_j$, which is just the mean of the initial signals. This is a remarkable observation: Under symmetry and when the network is regular, the long-run signal mix of every agent exactly reflects the initial signal distribution.

The only source of misinformation is hence that the initial draw of signals is "unlucky" (i.e. it happens to consist of many negative signals). For instance, let n be odd. Then the expected fraction of misinformed agents is $\sum_{r=0}^{\frac{n-1}{2}} \binom{n}{r}\rho^r(1-\rho)^{n-r}$, which equals

[19] This implies that the formulas for speed of convergence given in the main text still apply to the generalized setting for Case 3.

the probability that the minority of n independent signals is correct. To have concrete numerical examples, let the quality of each initial signal be $\rho = 0.6$. Then expected fraction of misinformed agents is 0.267 for $n = 9$ agents and 0.022 for $n = 99$ agents. Observe that the probability of misinformation in regular graphs goes to zero for growing n.

Observe finally the comparison to Bayesian learners. Suppose for a moment that all agents are proper Bayesian learners in the following sense: they account for the repetition of signals and form their beliefs according to Bayes' rule using each independent signal only once. In a connected network, these Bayesian learners will update until they have received each initial signal and then form their belief based on exactly the same signal mix as our much more naïve agents form in the long run when the network is regular (see, e.g., DeMarzo et al. (2003), Theorem 3).[20]

Example 3.C.2 (Network with clique of five)**.** *Consider the network (N, A) depicted in Figure 3.C.1. This network consists of $n = 10$ agents. Five of them, $1, ..., 5$, form a clique, i.e. the network restricted to these agents is complete; the others are arranged in a line.*

The normalized eigenvector corresponding to the largest eigenvalue is

$$c = (19.42\%, 18.41\%, 18.41\%, 18.41\%, 18.41\%, 5.12\%, 1.35\%, 0.36\%, 0.09\%, 0.02\%)$$

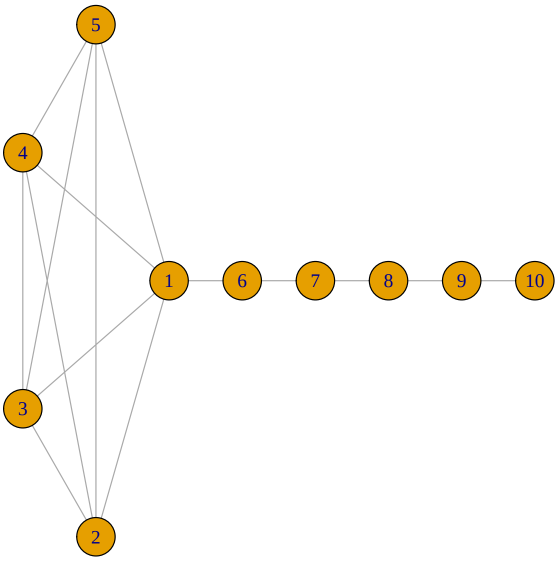
The nodes are labeled according to their entry in this eigenvector with 1 having the largest entry and 10 the lowest. Observe that the five members of the clique obtain the highest eigenvector entries. In fact, any three of their entries sum up to more than half of all entries. Hence, if it happens that at least three out of the five agents $1, ..., 5$ receive the wrong signal, we have $\sum_{j=1}^{n} c_j s_j < 0.5$ and hence misinformation prevails (by Proposition 1). The probability to have such a draw of signals and in fact the probability of misinformation is $(1-\rho)^5 + 5\rho(1-\rho)^4 + 10\rho^2(1-\rho)^3$, e.g. for $\rho = 0.6$, it is $Ef^{mis}(x(\infty)) = 0.31744$. There are many more such networks (with the same expected level of misinformation) for $n = 10$, but there is no network with higher probability of misinformation.

More importantly, we can construct networks with a clique of five and all others arranged in a line for all $n > 7$. The probability of misinformation is unchanged, as we checked for n up to 1'000 by using programming language R. The eigenvector centralities converge to $c_1 = 19.41919\%$ and $c_{2,...,5} = 18.40593\%$ for the members of the clique. Hence, misinformation still happens when at least three out of these five receive the wrong signal. Thus, we observe that as the number of nodes grow, the probability of misinformation need not go to zero, as there are networks with a substantial probability of misinformation.

The example shows that a small group of people who are well-connected among themselves may have a disproportional large influence on the long-run signal mixes and hence

[20] There is a difference in the belief due to the signal difference $k_i(t)$, which grows in our model and is constant among Bayesian learners, but the signal mix as well as the best guess of each agent is identical in the two models.

Figure 3.C.1: A network with a clique of five and all other agents arranged in a line.



Eigenvector centrality is:

$$c = (19.42\%, 18.41\%, 18.41\%, 18.41\%, 18.41\%, 5.12\%, 1.35\%, 0.36\%, 0.09\%, 0.02\%)$$

can be a cause for misinformation under symmetry. However, under symmetry the probability of misinformation is always bounded.

3.C.2 Different interpretations for the decay factor

With respect to the decay factors, there are in fact three different interpretations, as we will explain in more detail below. In a nutshell, these interpretations are (i) the sender only shares part of her signals, (ii) the communication channel does not transmit 100% of the signals, and (iii) the recipient discounts part of the received signals. In order to improve readability, we will in this subsecion omit all '+' and '-' superscripts, thus A may stand for A^+ and A^-, respectively, $N(t)$ may denote either the numbers of positive signals $N^+(t)$ or that of negative signals $N^-(t)$, δ will be either δ^+ or δ^- and so on.

In order to showcase all the explanations given above for the existence of decay factors, we might consider the following very general model: by $N_{(s)}(t)$, we denote the numbers of signals that agents send out to their neighbours, and we write $N_{(s)}(t) = \delta_{(s)}N(t)$ to model that agents do not communicate all their signals to their neighbors, with $\delta_{(s)} \in (0, 1]$ capturing the share of signals that agents are willing to transmit. We then denote by $N_{(t)}(t)$ the numbers of signals that are transmitted between the agents, and by modeling $N_{(t)}(t) = \delta_{(t)}AN_{(s)}(t)$, with $\delta_{(t)} \in (0, 1]$ describing the share of signals that are successfully transmitted by the communication channel. Finally, we use $N_{(p)}(t)$ to denote the numbers of signals that agents are actually processing when updating their signals from time t to $t + 1$. Here, by setting $N_{(p)}(t) = \delta_{(p)}N_{(t)}(t)$, the discounting of received signals by agents would be described by $\delta_{(p)} \in (0, 1]$. Taken together and defining $\delta := \delta_{(p)}\delta_{(t)}\delta_{(s)}$, agents process

$$N_{(p)}(t) = \delta_{(p)}N_{(t)}(t) = \delta_{(p)}\delta_{(t)}AN_{(s)}(t) = \delta_{(p)}\delta_{(t)}A\delta_{(s)}N(t) = \delta_{(p)}\delta_{(t)}\delta_{(s)}AN(t) = \delta AN(t), \tag{3.C.1}$$

which is exactly the formula that we use in our main model. By doing so, we are able to model any of the three interpretations, by setting two of the three factors to 1 and allowing only one to be smaller than 1: e.g., setting $\delta_{(p)} = 1$, $\delta_{(t)} = 1$, and $\delta_{(s)} < 1$ leads to a model where the decay factor δ captures that agents share only some part of the signals they receive. Furthermore, our model also allows variations where two or even all three effects are at play.

If some of the above phenomena are no longer homogeneous across agents, but agent-specific, we might preserve the general structure, but replace the scalar quantities $\delta_{(p)}$, $\delta_{(t)}$, and $\delta_{(s)}$ by matrices $\Delta_{(p)}$, $\Delta_{(t)}$, and $\Delta_{(s)}$. In this case, the equation describing sharing only parts of available signals becomes $N_{(s)}(t) := \Delta_{(s)}N(t)$, and $\Delta_{(s)}$ will be a diagonal matrix which captures the agent-specific factors describing which share of their signals agents do actually share. Similarly, the generalized equation for the discounting of received signals becomes $N_{(p)}(t) = \Delta_{(p)}N_{(t)}(t)$, with the diagonal matrix $\Delta_{(p)}$ capturing the agent-specific factors used for ignoring some part of the signals transmitted to the agents. Finally, with the non-diagonal matrix $\Delta_{(t)}$ whose entries ${\Delta_{(t)}}_{ij}$ determine the share of signals that are successfully transmitted from agent j to agent i, the equation for the transmitted signals becomes $N_{(t)}(t) := (\Delta_{(t)} \circ A)N_{(s)}(t)$, with $\Delta_{(t)} \circ A$ denoting the Hadamard product of $\Delta_{(t)}$ and A. Equation (3.C.1) then generalizes to

$$N_{(p)}(t) = \Delta_{(p)}N_{(t)}(t) = \Delta_{(p)} \left(\Delta_{(t)} \circ A\right) N_{(s)}(t) = \Delta_{(p)} \left(\Delta_{(t)} \circ A\right) \Delta_{(s)}N(t) \qquad (3.C.2)$$

Due to properties of diagonal matrices and the Hadamard product, $\Delta_{(p)} \left(\Delta_{(t)} \circ A\right) \Delta_{(s)}$ turns out to be identical to $\left(\Delta_{(p)}\Delta_{(t)}\Delta_{(s)}\right) \circ A$, implying that we finally have $N_{(p)}(t) = (\Delta \circ A)\, N(t)$, with $\Delta := \Delta_{(p)}\Delta_{(t)}\Delta_{(s)}$, which amounts to the formula we use in our extended model. The components Δ_{ij} thus simultaneously capture agent i's possible discounting (${\Delta_{(p)}}_i$), the communication between i and j not working properly (${\Delta_{(t)}}_{ij}$), and agent j not sharing all signals (${\Delta_{(s)}}_j$). Similar to above, our generalized model therefore also allows for one, two, or even all three of these effects being at play.

3.C.3 Examples for special dynamics

When an agent signal mix $x_i(t)$ converges to $\frac{1}{2}$, then beliefs might behave in a "nasty" way, as we illustrate in the examples below. For all examples, $b^0 = \frac{1}{2}$, $\rho = \frac{3}{5}$ and $\delta^+\lambda_1^+ = \delta^-\lambda_1^-$ (Case 3). In addition to the signal mix $x_i(t)$ and belief $b_i(t)$, we also report the best guess $g_i(t)$, which is the state of nature that agent i considers as most likely to be the true state at time t. When the agent is undecided (i.e. if $b_i(t) = \frac{1}{2}$), we define $g_i(t) = \frac{1}{2}$.

1. Beliefs and best guesses of all agents converge extremely fast to $\frac{1}{2}$:
   Let $A^+ = A^-$ be the complete network on $n = 4$ agents, $\delta^+ = \delta^- = 1$ (symmetry as in Section 3.4), and initial signal vector $s(0) = (0, 0, 1, 1)$. Then, for all $t > 0$:

$$\begin{aligned}
N^+(t) &= \frac{1}{2}4^t\mathbb{1},\\
N^-(t) &= \frac{1}{2}4^t\mathbb{1},\\
k(t) &= 0,\\
b(t) &= \frac{1}{2}\mathbb{1},\\
g(t) &= \frac{1}{2}\mathbb{1}.
\end{aligned}$$

2. All beliefs converge to $\frac{1}{2}$, best guesses are constant at 0 and 1 for two players each:
   Let $A^+ = A^-$ be the complete network on $n = 4$ agents, $\delta^+ = \delta^- = \frac{1}{2}$ (symmetry

as in Section 3.4), and initial signal vector $s(0) = (0,0,1,1)$. Then, for all $t > 0$:

$$\begin{aligned}
N^+(t) &= \frac{1}{2}\left(\frac{5}{2}\right)^t \mathbb{1} + \frac{1}{2}\left(\frac{1}{2}\right)^t (-1,-1,1,1)^\top,\\
N^-(t) &= \frac{1}{2}\left(\frac{5}{2}\right)^t \mathbb{1} - \frac{1}{2}\left(\frac{1}{2}\right)^t (-1,-1,1,1)^\top,\\
k(t) &= \left(\frac{1}{2}\right)^t (-1,-1,1,1)^\top \overset{t\to\infty}{\longrightarrow} 0,\\
b(t) &\overset{t\to\infty}{\longrightarrow} \frac{1}{2}\mathbb{1},\\
g(t) &= (0,0,1,1)^\top.
\end{aligned}$$

3. Beliefs of all agents alternate, best guess of all agents, but one, alternate:
Let $A^+ = A^-$ be the star network of $n = 5$ agents, $\delta^+ = \delta^- = 1$ (symmetry as in Section 3.4), and initial signal vector $s(0) = (0,0,1,1,1)$. Then, for all $t > 0$:

$$\begin{aligned}
N^+(t) &= \frac{1}{8}3^t\,(6,3,3,3,3)^\top + \frac{1}{4}(0,-3,1,1,1)^\top + \frac{1}{8}\,(-1)^t\,(-6,3,3,3,3)^\top,\\
N^-(t) &= \frac{1}{8}3^t\,(6,3,3,3,3)^\top - \frac{1}{4}(0,-3,1,1,1)^\top + \frac{1}{8}\,(-1)^t\,(2,-1,-1,-1,-1)^\top,\\
k(t) &= \frac{1}{2}(0,-3,1,1,1)^\top + \frac{1}{2}\,(-1)^t\,(-2,1,1,1,1)^\top,\\
&b(t) \text{ alternates between } \frac{1}{5}(2,2,3,3,3)^\top \text{ and } \left(\frac{3}{5},\frac{4}{13},\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)^\top,\\
&g(t) \text{ alternates between } (0,0,1,1,1)^\top \text{ and } \left(1,0,\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)^\top.
\end{aligned}$$

Notice that $g_2(t) = 0$ is constant, $g_1(t)$ alternates between 0 and 1, while $g_3(t)$, $g_4(t)$, $g_5(t)$ alternate indefinitely between 1 and $\frac{1}{2}$.

4. Beliefs of all agents converge to $\frac{1}{2}$, best guesses converge for some agents, but alternate for other agents:
Let A^+ be a network on $n = 5$ agents with agents 1 and 2 being connected to everyone and no separate connections between agents 3, 4, and 5, A^- the complete network, $\delta^+ = \frac{4}{5}$, $\delta^- = \frac{3}{5}$ (network and decay asymmetry), and initial signal vector

$s(0) = (0,0,1,1,1)$. Then, for all $t > 0$:

$$\begin{aligned}
N^{+}(t) &= \frac{1}{5}\left(\frac{17}{5}\right)^{t}(3,3,2,2,2)^{\top} + \frac{3}{5}\left(-\frac{3}{5}\right)^{t}(-1,-1,1,1,1)^{\top}, \\
N^{-}(t) &= \frac{2}{5}\left(\frac{17}{5}\right)^{t}\mathbb{1}^{\top} + \frac{1}{5}\left(\frac{2}{5}\right)^{t}(3,3,-2,-2,-2)^{\top}, \\
k(t) &= \frac{1}{5}\left(\frac{17}{5}\right)^{t}(1,1,0,0,0)^{\top} + \frac{3}{5}\left(-\frac{3}{5}\right)^{t}(-1,-1,1,1,1)^{\top} \\
&\quad - \frac{1}{5}\left(\frac{2}{5}\right)^{t}(3,3,-2,-2,-2)^{\top} \overset{t\to\infty}{\longrightarrow} (\infty,\infty,0,0,0)^{\top}, \\
b(t) &\overset{t\to\infty}{\longrightarrow} \left(1,1,\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)^{\top}, \\
g(t) &\text{ asymptotically alternates between } \mathbb{1}^{\top} \text{ and } (1,1,0,0,0)^{\top}.
\end{aligned}$$

Notice that $g_1(t) = g_2(t)$ converge to 1, while $g_3(t) = g_4(t) = g_5(t)$ alternate indefinitely between 0 and 1.

3.C.4 Conditions for Case 3 to Approximate Case 1 and 2

The first condition is that parameters are reasonably close to case 3, i.e. $\delta^{+}\lambda_1^{+}$ being close to $\delta^{-}\lambda_1^{-}$. We can observe this in Example 1 in Figure 3.5.1, where values of δ^{+} which are close to the critical value of 0.4 induce similar dynamics. There is however a second condition, which happens to be satisfied in Example 1. For explaining that condition, let us reconsider Equation 3.7 and define $\tau^{\text{Case 3}} := \max\{\frac{|1+\delta^{+}\lambda_i^{+}|}{1+\delta^{+}\lambda_1^{+}}, \frac{|1+\delta^{-}\lambda_i^{-}|}{1+\delta^{-}\lambda_1^{-}}, i = 2,\ldots,n\}$, where λ_i^{+} and λ_i^{-} denote all but the largest eigenvalues of A^{+} and A^{-}, respectively. The second condition is that $\tau^{\text{Case 3}}$ is substantially smaller than $\frac{1+\min\{\delta^{+}\lambda_1^{+},\delta^{-}\lambda_1^{-}\}}{1+\max\{\delta^{+}\lambda_1^{+},\delta^{-}\lambda_1^{-}\}}$. Intuitively, the first condition assures that the importance of the largest eigenvalues persists sufficiently long (they always matter in Case 3, but vanish in the long run of the two others) and the second condition assures that the importance of all other eigenvalues vanishes sufficiently fast. As a rule of thumb, when slow convergence occurs in Cases 1 or 2, resulting in large values for actual half-life $t_{1/2} = \log(0.5)/\log\left(\frac{1+\min\{\delta^{+}\lambda_1^{+},\delta^{-}\lambda_1^{-}\}}{1+\max\{\delta^{+}\lambda_1^{+},\delta^{-}\lambda_1^{-}\}}\right)$, and at the same time the half-life formula for Case 3, $\frac{\log(0.5)}{\log(\tau^{\text{Case 3}})}$ produces a much lower "pseudo half-life," then the formula given for Case 3 of Proposition 2 may provide a better approximation for the relevant misinformation in the short or medium term than the actual long-term limit of either 0 or 1.

3.C.5 Comparison with DeGroot model

In the following, we will discuss the similarities and differences of our model as compared to the classical DeGroot model, with respect to model set-up, conditions for convergence, reaching a consensus, and eigenvector centrality.

With respect to the set-up of our model, the most striking difference to the DeGroot model is that we separate the evolution of positive and of negative signals, while in the DeGroot model this distinction is not possible. However, the evolution of each type of signals (positive or negative) resembles the DeGroot model, in the sense that values at time $t+1$ are linear functions of values at time t. While the weights of these regression-type recursions are restricted to be non-negative in our model as well in the DeGroot model, our model for the signals' evolution does not require convex combinations, i.e. the weights do not have to sum up to unity, in contrast to the DeGroot case. For the special case of symmetry in the sense that $A^+ = A^- =: A$ as well as $\delta^+ = \delta^- =: \delta$ and denoting $I + \delta A$ by W, it is easy to see that $N^+(t) = W^t s$, $N^-(t) = W^t(\mathbb{1} - s)$, and $N(t) = W^t \mathbb{1}$, implying

$$x_i(t) = \frac{e_i^\top W^t s}{e_i^\top W^t \mathbb{1}}, \tag{3.C.3}$$

with e_i denoting the i-th unit vector. Furthermore, Equation (3.C.3) implies that the updating of the signal mixes $x(t)$ in our model may be written as

$$x(t) = \widetilde{W}(t)x(t-1) \tag{3.C.4}$$

with the row-stochastic matrices $\widetilde{W}(t)$ having entries $\widetilde{w}(t)_{ij} = w_{ij}\frac{\kappa_j(t-1)}{\kappa_i(t)}$, with w_{ij} denoting the entries of $W = I + \delta A$ and $\kappa(t) := W^t \mathbb{1}$.[21] Equation (3.C.4) thus is a representation of the updating process in our model as a generalized DeGroot model, where in general the updating matrices $\widetilde{W}(t)$ are not constant, but change over time. Furthermore, if the networks described by A are regular, then these updating matrices will in fact not depend on time, and the updating formula (3.C.4) will actually become a DeGroot model, in complete analogy to a special case of the model considered by Sikder et al. (2020).

With regard to conditions ensuring convergence, it is well-known that values converge in the DeGroot model if and only if every set of nodes is strongly connected and closure is also aperiodic. This is in fact very similar to our model, where we receive convergence by assuming that the whole society is strongly connected and by observing that the matrices W^+ and W^- are aperiodic. Aperiodicity of the matrices $W^+ = I + \delta^+ A^+$ and $W^- = I + \delta^- A^-$ is guaranteed, as they both involve the identity matrix I.

For the DeGroot model, there emerges a consensus whenever the model converges, where the consensus is a convex combination of the initial values. This is different for our model, which in contrast to the DeGroot model, allows for the signal mixes to converge to the truth (1) or to complete misinformation (0), depending on the networks' eigenvalues and the decay parameters. In addition, in our model, signal mixes will converge to consensus if $\delta^+\lambda^+ = \delta^-\lambda^-$ and $\frac{c_i^-}{c_i^+}$ does not depend on i: the latter condition is equivalent to $c^- = c^+$, i.e. in our main as well as in the extended model (Section 3.5.4), the networks must have identical centralities. This actually can happen even though A^+

[21] A similar representation of $x(t)$ also holds for the extended model, it can be developed by simply replacing all appearances of δA by M.

and A^- are different, with an example being the case of both networks being regular, but of different degrees.

With respect to eigenvector centrality in DeGroot models, the left-hand eigenvector determines the weights for the asymptotically emerging consensus (Jackson, 2010; Golub and Sadler, 2016). In our model, however, centralities only play this role when the relation $\delta^+\lambda_1^+ = \delta^-\lambda_1^-$ holds (i.e. in case 3). In this case, eigenvector centralities play two roles (in our main model): they influence the long-run signal mix of all agents commonly through $\frac{1-\sum_{j=1}^n c_j^- s_j}{\sum_{j=1}^n c_j^+ s_j}$, reflecting the initial signals' impact, and they influence the agents' individual long-run signal mixes through the centrality ratios $\frac{c_i^-}{c_i^+}$. In our extended model, the former role (the part common to all agents) is played by the left-hand eigenvectors d^+ and d^-, while the latter role (the individual part) is still played by the (right-hand) eigenvector centralities c^+ and c^-.

Bibliography

Acemoglu, D., Ozdaglar, A., and ParandehGheibi, A. (2010). Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.

Alesina, A. and Angeletos, G.-M. (2005). Fairness and redistribution. *American Economic Review*, 95(4):960–980.

Amabile, T. M. (1996). *Creativity in Context.* Westview Press, Boulder, CO, USA.

Andreoni, J., Harbaugh, W. T., and Vesterlund, L. (2010). Altruism in experiments. In *Behavioural and Experimental Economics*, pages 6–13. Springer.

Andreoni, J. and Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, 103(418):570–585.

Azzimonti, M. and Fernandes, M. (2018). Social Media Networks, Fake News, and Polarization. NBER Working Papers 24462, National Bureau of Economic Research, Inc.

Bamberg, S. and Rölle, D. (2003). Determinants of people's acceptability of pricing measures: replication and extension of a causal model. *Acceptability of Transport Pricing Strategies*, pages 235–248.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., and Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, 31(4):73–102.

Banerjee, A., Breza, E., Chandrasekhar, A. G., and Mobius, M. (2019). Naive Learning with Uninformed Agents. NBER Working Papers 25497, National Bureau of Economic Research, Inc.

Banerjee, R., Bhattacharya, J., and Majumdar, P. (2021). Exponential-growth prediction bias and compliance with safety measures related to covid 19. *Social Science & Medicine*, 268:113473.

Barnhill, A., Palmer, A., Weston, C. M., Brownell, K. D., Clancy, K., Economos, C. D., Gittelsohn, J., Hammond, R. A., Kumanyika, S., and Bennett, W. L. (2018). Grappling with complex food systems to reduce obesity: a us public health challenge. *Public Health Reports*, 133(1_suppl):44S–53S.

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'Économie Politique*, 125(5):665–685.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.

Bos, C., Lans, I. V. D., Van Rijnsoever, F., and Van Trijp, H. (2015). Consumer acceptance of population-level intervention strategies for healthy food choices: the role of perceived effectiveness and perceived fairness. *Nutrients*, 7(9):7842–7862.

Bos, C., Van der Lans, I. A., Van Rijnsoever, F. J., and Van Trijp, H. C. (2013). Understanding consumer acceptance of intervention strategies for healthy food choices: a qualitative study. *BMC Public Health*, 13(1):1073.

Buechel, B., Hellmann, T., and Klner, S. (2015). Opinion dynamics and wisdom under conformity. *Journal of Economic Dynamics and Control*, 52:240 – 257.

Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6):e258–e259.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Caraher, M. and Cowburn, G. (2005). Taxing food: implications for public health nutrition. *Public Health Nutrition*, 8(8):1242–1249.

Chandrasekhar, A. G., Larreguy, H., and Xandri, J. P. (2020). Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32.

Charness, G. and Grieco, D. (2014). Creativity and financial incentives. *University of California, Santa Barbara, Working Paper.*

Charness, G. and Grieco, D. (2018). Creativity and Incentives. *Journal of the European Economic Association*, 17(2):454–496.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Cohn, A., Engelmann, J., Fehr, E., and Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2):860–85.

Colchero, M. A., Rivera-Dommarco, J., Popkin, B. M., and Ng, S. W. (2017). In mexico, evidence of sustained consumer response two years after implementing a sugar-sweetened beverage tax. *Health Affairs*, 36(3):564–571.

Corazzini, L., Pavesi, F., Petrovich, B., and Stanca, L. (2012). Influential listeners: An experiment on persuasion bias in social networks. *European Economic Review*, 56(6):1276–1288.

Cullerton, K., Donnet, T., Lee, A., and Gallegos, D. (2016). Playing the policy game: a review of the barriers to and enablers of nutrition policy change. *Public Health Nutrition*, 19(14):2643–2653.

De Groot, J. I. and Schuitema, G. (2012). How to make the unpopular popular? Policy characteristics, social norms and the acceptability of environmental policies. *Environmental Science & Policy*, 19:100–107.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Della Lena, S. (2019). Non-Bayesian Social Learning and the Spread of Misinformation in Networks. Working Papers 2019:09, Department of Economics, University of Venice "Ca' Foscari".

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–72.

DellaVigna, S. and Linos, E. (2020). Rcts to scale: Comprehensive evidence from two nudge units. Working Paper 27594, National Bureau of Economic Research.

DeMarzo, P. M., Vayanos, D., and Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968.

Diacon, S. and Hasseldine, J. (2007). Framing effects and risk perception: The effect of prior performance presentation format on investment fund choice. *Journal of Economic Psychology*, 28(1):31 – 52.

Diepeveen, S., Ling, T., Suhrcke, M., Roland, M., and Marteau, T. M. (2013). Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis. *BMC Public Health*, 13(1):1–11.

Dikmen, B. Y., Ipek, A., Şahan, Ü., Petek, M., and Sözcü, A. (2016). Egg production and welfare of laying hens kept in different housing systems (conventional, enriched cage, and free range). *Poultry science*, 95(7):1564–1572.

Drucker, P. (1939). *The end of economic man: a study of the new totalitarianism.* John Day Co., New York, NY, USA.

Duvendack, M., Palmer-Jones, R., and Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, 107(5):46–51.

Eckel, C. C. and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, 16(2):181–191.

Ederer, F. and Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science*, 59(7):1496–1513.

Englmaier, F., Grimm, S., Schindler, D., and Schudy, S. (2018). The effect of incentives in non-routine analytical teams tasks - evidence from a field experiment. Working paper, Center for Economic Studies.

Eppler, M. J. and Aeschimann, M. (2009). A systematic framework for risk visualization in risk management and communication. *Risk Management*, 11(2):67–89.

Erat, S. and Gneezy, U. (2015). Incentives for creativity. *Experimental Economics*, 19:269–280.

Espinosa, R. (2019). L'éléphant dans la pièce. pour une approche économique de l'alimentation végétale et de la condition animale. *Revue d'Économie Politique*, 129(3):287–324.

Espinosa, R. and Nassar, A. (2021). The acceptability of food policies. *Nutrients*, 13(5):1483.

Espinosa, R. and Stoop, J. (2021). Do people really want to be informed? ex-ante evaluations of information-campaign effectiveness. *Experimental Economics*, pages 1–25.

European Commission (2018). Tackling online disinformation: a European approach. Technical Report 236, Brussels, Belgium.

Evans, C. E. L. (2020). Next steps for interventions targeting adolescent dietary behaviour. *Nutrients*, 12(1):190.

Eykelenboom, M., Van Stralen, M. M., Olthof, M. R., Schoonmade, L. J., Steenhuis, I. H., and Renders, C. M. (2019). Political and public acceptability of a sugar-sweetened beverages tax: a mixed-method systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):1–19.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Fanzo, J., Drewnowski, A., Blumberg, J., Miller, G., Kraemer, K., and Kennedy, E. (2020). Nutrients, foods, diets, people: promoting healthy eating. *Current Developments in Nutrition*, 4(6):nzaa069.

FAO (2020). FAO stats. Available online: `http://www.fao.org/faostat/en/#home` (accessed on 6 July 2020).

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, 1(2).

Fehr, E. and Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity*, 1:615–691.

Ferrer, R. A. and Klein, W. M. (2015). Risk perceptions and health behavior. *Current Opinion in Psychology*, 5:85 – 89.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology*, 96(6):1478–1504.

Friedkin, N. E. and Bullo, F. (2017). How truth wins in opinion dynamics along issue sequences. *Proceedings of the National Academy of Sciences*, 114(43):11380–11385.

Friedkin, N. E. and Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206.

Gabaix, X. (2019). Chapter 4 - behavioral inattention. *Handbook of Behavioral Economics: Applications and Foundations 2.*, 2:261 – 343.

Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., and Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences*, 116(42):20923–20929.

Gerber, P. J., Steinfeld, H., Henderson, B., Mottet, A., Opio, C., Dijkman, J., Falcucci, A., Tempio, G., et al. (2013). Tackling climate change through livestock: a global assessment of emissions and mitigation opportunities. Technical report, FAO, Rome, Italy.

Gerrard, M., Gibbons, F. X., and Reis-Bergan, M. (1999). The effect of risk communication on risk perceptions: the significance of individual differences. *JNCI Monographs*, 1999(25):94 – 100.

Golub, B. and Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338.

Golub, B. and Sadler, E. D. (2016). Learning in social networks. Working paper, Social Science Electronic Publishing.

Gortmaker, S. L., Swinburn, B. A., Levy, D., Carter, R., Mabry, P. L., Finegood, D. T., Huang, T., Marsh, T., and Moodie, M. L. (2011). Changing the future of obesity: science, policy, and action. *The Lancet*, 378(9793):838–847.

Grabisch, M., Mandel, A., and Rusinowska, A. (2021). On the design of public debate in social networks. Mimeo.

Grabisch, M., Mandel, A., Rusinowska, A., and Tanimura, E. (2018). Strategic influence in social networks. *Mathematics of Operations Research*, 43(1):29–50.

Grabisch, M., Poindron, A., and Rusinowska, A. (2019). A model of anonymous influence with anti-conformist agents. *Journal of Economic Dynamics and Control*, 109(C).

Graça, J., Calheiros, M. M., and Oliveira, A. (2015). Attached to meat?(un) willingness and intentions to adopt a more plant-based diet. *Appetite*, 95:113–125.

Grimm, V. and Mengel, F. (2018). An Experiment on Belief Formation in Networks. *Journal of the European Economic Association.*

Hagmann, D., Siegrist, M., and Hartmann, C. (2018). Taxes, labels, or nudges? public acceptance of various interventions designed to reduce sugar intake. *Food Policy*, 79:156–165.

Hall, R. E. and Woodward, S. E. (2010). The burden of the nondiversifiable risk of entrepreneurship. *American Economic Review*, 100(3):1163–94.

Hammond, R. A. (2009). Peer reviewed: complex systems modeling for obesity research. *Preventing Chronic Disease*, 6(3).

Hanna, R., Mullainathan, S., and Schwartzstein, J. (2014). Learning through noticing: Theory and evidence from a field experiment. *The Quarterly Journal of Economics*, 129(3):1311–1353.

Hansen, P. G., Schilling, M., and Malthesen, M. S. (2019). Nudging healthy and sustainable food choices: three randomized controlled field experiments using a vegetarian lunch-default as a normative signal. *Journal of Public Health.*

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

Holzmeister, F., Huber, J., Kirchler, M., Lindner, F., Weitzel, U., and Zeisberger, S. (2020). What drives risk perception? a global survey with financial professionals and laypeople. *Management Science*, 66(9):3977–4002.

Hudja, S. and Woods, D. (2019). Behavioral Bandits: Analyzing the Exploration Versus Exploitation Trade-Off in the Lab. Working Paper SSRN 3484498, Social Science Research Network.

Hughes, D. J., Lee, A., Tian, A. W., Newman, A., and Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly*, 29(5):549–569.

Hvide, H. K. and Panos, G. A. (2014). Risk tolerance and entrepreneurship. *Journal of Financial Economics*, 111(1):200–223.

Jackson, M. O. (2010). *Social and economic networks.* Princeton University Press, Princeton, NJ, USA.

Jadbabaie, A., Molavi, P., Sandroni, A., and Tahbaz-Salehi, A. (2012). Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225.

Jalil, A. J., Tasoff, J., and Bustamante, A. V. (2020). Eating to save the planet: Evidence from a randomized controlled trial using individual-level food purchase data. *Food Policy*, 95:101950.

Jennison, C. and Turnbull, B. W. (2001). Group sequential tests with outcome-dependent treatment assignment. *Sequential Analysis*, 20(4):209–234.

Johnson, H. M. and Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of experimental psychology: Learning, memory, and cognition*, 20(6):1420.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., Zheng, M., Manrique, P., Wuchty, S., and Lupu, Y. (2020). The online competition between pro-and anti-vaccination views. *Nature*, 582(7811):230–233.

Kerr, S. P., Kerr, W. R., and Xu, T. (2017). Personality traits of entrepreneurs: A review of recent literature. Working Paper 24097, National Bureau of Economic Research.

Knight, J., Weir, S., and Woldehanna, T. (2003). The role of education in facilitating risk-taking and innovation in agriculture. *The Journal of Development Studies*, 39(6):1–22.

Koh, L. P., Miettinen, J., Liew, S. C., and Ghazoul, J. (2011). Remotely sensed evidence of tropical peatland conversion to oil palm. *Proceedings of the National Academy of Sciences*, 108(12):5127–5132.

Kohn, A. (1993). *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise and Other Bribes.* Houghton Mifflin, Boston, MA, USA.

Köszegi, B. and Szeidl, A. (2013). A model of focusing in economic choice. *The Quarterly Journal of Economics*, 128(1):53–104.

Koudstaal, M., Sloof, R., and van Praag, M. (2016). Risk, uncertainty, and entrepreneurship: Evidence from a lab-in-the-field experiment. *Management Science*, 62(10):2897–2915.

Kurz, V. (2018). Nudging to reduce meat consumption: Immediate and persistent effects of an intervention at a university restaurant. *Journal of Environmental Economics and Management*, 90:317–341.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710.

Lazear, E. (2005). Entrepreneurship. *Journal of Labor Economics*, 23(4):649–680.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lee, B. Y., Bartsch, S. M., Mui, Y., Haidari, L. A., Spiker, M. L., and Gittelsohn, J. (2017). A systems approach to obesity. *Nutrition Reviews*, 75(suppl_1):94–106.

Leibenstein, H. (1976). *Beyond economic man.* Harvard university press, Cambridge, MA, USA.

Levitt, S. D. and List, J. A. (2011). Was there really a hawthorne effect at the hawthorne plant? an analysis of the original illumination experiments. *American Economic Journal: Applied Economics*, 3(1):224–238.

Livingstone, S., Van Couvering, E., Thumin, N., Coiro, J., Knobel, M., Lankshear, C., and Leu, D. (2008). Converging traditions of research on media and information literacies. *Handbook of Research on New Literacies*, pages 103–132.

Loewenstein, G., Weber, E., Hsee, C. K., and Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2):267–286.

Ludwig, D. S., Peterson, K. E., and Gortmaker, S. L. (2001). Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. *Lancet*, 357(9255):505–508.

Lupien, S. J., Maheu, F. S., Tu, M. T., Fiocco, A. J., and Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition*, 65(3):209–237.

Lusk, J. L. and Norwood, F. B. (2012). Speciesism, altruism and the economics of animal welfare. *European Review of Agricultural Economics*, 39(2):189–212.

Lévy-Garboua, L., Maafi, H., Masclet, D., and Terracol, A. (2012). Risk aversion and framing effects. *Experimental Economics*, 15:128–144.

Macdiarmid, J. I., Douglas, F., and Campbell, J. (2016). Eating like there's no tomorrow: Public awareness of the environmental impact of food and reluctance to eat less meat as part of a sustainable diet. *Appetite*, 96:487–493.

Manso, G. (2011). Motivating innovation. *The Journal of Finance*, 66(5):1823–1860.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87.

Matthews, G., Davies, D., Westerman, S., and & Stammers, R. (2000). *Human performance: Cognition, stress and individual differences.* East Sussex, UK: Psychology Press.

Mazzocchi, M., Cagnone, S., Bech-Larsen, T., Niedźwiedzka, B., Saba, A., Shankar, B., Verbeke, W., and Traill, W. B. (2015). What is the public appetite for healthy eating policies? Evidence from a cross-European survey. *Health Economics, Policy and Law*, 10(3):267–292.

Mill, J. S. (1874). *Essays on some unsettled questions of political economy.* Longmans, Green, Reader, and Dyer, Harlow, UK.

Miller, S. (2001). Public understanding of science at the crossroads. *Public Understanding of Science*, 10(1):115–120.

Mo, J. J. and Mortensen, T. (2019). Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American Behavioral Scientist.*

Molavi, P., Tahbaz-Salehi, A., and Jadbabaie, A. (2018). A theory of non-bayesian social learning. *Econometrica*, 86(2):445–490.

Moore, G. F., Evans, R. E., Hawkins, J., Littlecott, H., Melendez-Torres, G., Bonell, C., and Murphy, S. (2019). From complex social interventions to interventions in complex social systems: future directions and unresolved questions for intervention development and evaluation. *Evaluation*, 25(1):23–45.

Mueller-Frank, M. (2013). A general framework for rational learning in social networks. *Theoretical Economics*, 8(1):1–40.

Mueller-Frank, M. (2014). Does one Bayesian make a difference? *Journal of Economic Theory*, 154(C):423–452.

Myers, L. B. (2014). Changing smokers' risk perceptions – for better or worse? *Journal of Health Psychology*, 19(3):325–332.

Niederdeppe, J., Shapiro, M. A., and Porticella, N. (2011). Attributions of responsibility for obesity: Narrative communication reduces reactive counterarguing among liberals. *Human Communication Research*, 37(3):295–323.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Pennings, J. M. E. and Wansink, B. (2004). Channel contract behavior: The role of risk attitudes, risk perceptions, and channel members' market structures. *The Journal of Business*, 77(4):697–724.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Pennycook, G. and Rand, D. (2021). Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *The Harvard Kennedy School Misinformation Review*, 2.

Persky, J. (1995). The ethology of homo economicus. *Journal of Economic Perspectives*, 9(2):221–231.

Poore, J. and Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. *Science*, 360(6392):987–992.

Prendergast, C. and Stole, L. (1996). Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of Political Economy*, 104(6):1105–1134.

Purtle, J., Langellier, B., and Lê-Scherban, F. (2018). A case study of the philadelphia sugar-sweetened beverage tax policymaking process: implications for policy development and advocacy. *Journal of Public Health Management and Practice*, 24(1):4–8.

Quattrociocchi, W., Scala, A., and Sunstein, C. R. (2016). Echo chambers on facebook. Working Paper SRN 2795110, Social Science Research Network.

Reisch, L. A. and Sunstein, C. R. (2016). Do europeans like nudges? *Judgment and Decision Making*, 11(4):310–325.

Reynolds, J., Archer, S., Pilling, M. a., Kenny, M., Hollands, G. J., and Marteau, T. (2019). Public acceptability of nudging and taxing to reduce consumption of alcohol, tobacco, and food: A population-based survey experiment. *Social Science & Medicine*, 236:112395.

Reynolds, J., Stautz, K., Pilling, M., van der Linden, S., and Marteau, T. (2020). Communicating the effectiveness and ineffectiveness of government policies and their impact on public support: A systematic review with meta-analysis. *Royal Society Open Science*, 7(1):190522.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Rothgerber, H. (2014). Efforts to overcome vegetarian-induced dissonance among meat eaters. *Appetite*, 79:32–41.

Rudis, E. (2004). *CEO Challenge 2004: Perspectives and Analysis.* The Conference Board, New York, NY, USA.

Rusinowska, A. and Taalaibekova, A. (2019). Opinion formation and targeting when persuaders have extreme and centrist opinions. *Journal of Mathematical Economics*, 84(C):9–27.

Russel, M. (2020). Palm oil: Economic and environmental impacts. Technical report, European Parliament, Brussels, Belgium.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., Greaves, F., Harper, L., Hawe, P., Moore, L., et al. (2017). The need for a complex systems model of evidence for public health. *Lancet*, 390(10112):2602–2604.

Schade, J. and Schlag, B. (2003). Acceptability of urban transport pricing strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(1):45–61.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *American Economic Review*, 80(3):465–479.

Schoofs, D., Preuss, D., and Wolf, O. T. (2008). Psychosocial stress induces working memory impairments in an n-back paradigm. *Psychoneuroendocrinology*, 33:643–653.

Shalley, C. E. and Gilson, L. L. (2004). What leaders need to know: A review of social and contextual factors that can foster or hinder creativity. *The Leadership Quarterly*, 15:33–53.

Shalley, C. E., Gilson, L. L., and Blum, T. C. (2000). Matching creativity requirements and the work environment: Effects on satisfaction and intentions to leave. *The Academy of Management Journal*, 43(2):215–223.

Signal, L. N., Watts, C., Murphy, C., Eyles, H., and Ni Mhurchu, C. (2018). Appetite for health-related food taxes: New zealand stakeholder views. *Health promotion international*, 33(5):791–800.

Sikder, O., Smith, R. E., Vivo, P., and Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. *Scientific Reports*, 10(1):1–11.

Sitkin, S. B. and Weingart, L. R. (1995). Determinants of risky decision-making behavior: A test of the mediating role of risk perceptions and propensity. *Academy of Management Journal*, 38:1573–1592.

Sobkow, A., Traczyk, J., and Zaleskiewicz, T. (2016). The affective bases of risk perception: Negative feelings and stress mediate the relationship between mental imagery and risk perception. *Frontiers in Psychology*, page 932.

Springmann, M., Godfray, H. C. J., Rayner, M., and Scarborough, P. (2016). Analysis and valuation of the health and climate change cobenefits of dietary change. *Proceedings of the National Academy of Sciences*, 113(15):4146–4151.

Steg, L., Dreijerink, L., and Abrahamse, W. (2005). Factors influencing the acceptability of energy policies: A test of vbn theory. *Journal of environmental psychology*, 25(4):415–425.

Steg, L. and Schuitema, G. (2007). Behavioural responses to transport pricing: a theoretical analysis. *Threats to the quality of urban life from car traffic: Problems, causes, and solutions*, pages 347–366.

Steinfeld, H., Gerber, P., Wassenaar, T., Castel, V., Rosales, M., Rosales, M., and de Haan, C. (2006). Livestock's long shadow: environmental issues and options. Technical report, FAO, Rome, Italy.

Stok, F. M., De Ridder, D. T., De Vet, E., and De Wit, J. B. (2014). Don't tell me what i should do, but what others do: The influence of descriptive and injunctive peer norms on fruit consumption in adolescents. *British journal of health psychology*, 19(1):52–64.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, 1(1):4–25.

Sunstein, C. R., Reisch, L. A., and Kaiser, M. (2019). Trusting nudges? lessons from an international survey. *Journal of European Public Policy*, 26(10):1417–1443.

Suurmond, G., Swank, O. H., and Visser, B. (2004). On the bad reputation of reputational concerns. *Journal of Public Economics*, 88(12):2817 – 2838.

Taalaibekova, A. (2020). *Diffusion of opinions and innovations among limitedly forward-looking individuals*. PhD thesis, UCL-Université Catholique de Louvain.

Thaler, R. H. and Ganser, L. (2015). *Misbehaving: The making of behavioral economics*. WW Norton, New York, NY, USA.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, New York, NY, USA.

Tilman, D. and Clark, M. (2014). Global diets link environmental sustainability and human health. *Nature*, 515(7528):518–522.

Tombu, M. and Mandel, D. R. (2015). When does framing influence preferences, risk perceptions, and risk attitudes? the explicated valence account. *Journal of Behavioral Decision Making*, 28(5):464–476.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE*, 13(9):e0203958.

Traill, W. B., Mazzocchi, M., Shankar, B., and Hallam, D. (2014). Importance of government policies and other influences in transforming global diets. *Nutrition Reviews*, 72(9):591–604.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59(4 pt 2).

Ungar, N., Sieverding, M., Schweizer, F., and Stadnitski, T. (2015). Intervention-elicited reactance and its implications: Let me eat what i want. *Zeitschrift für Psychologie*, 223(4):247.

Unsworth, K. (2001). Unpacking creativity. *The Academy of Management Review*, 26(2):289–297.

Vartanian, L. R., Herman, C. P., and Polivy, J. (2007). Consumption stereotypes and impression management: How you are what you eat. *Appetite*, 48(3):265–277.

Vijay, V., Pimm, S. L., Jenkins, C. N., and Smith, S. J. (2016). The impacts of oil palm on recent deforestation and biodiversity loss. *PLOS ONE*, 11(7).

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wansink, B. and Sobal, J. (2007). Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior*, 39(1):106–123.

Weber, E. U. and Milliman, R. A. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science*, 43(2):123–144.

WHO & FAO (2003). Diet, nutrition, and the prevention of chronic diseases: report of a joint who/fao expert consultation. Technical report, WHO & FAO, Geneva, Switzerland.

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., et al. (2019). Food in the anthropocene: the eat–lancet commission on healthy diets from sustainable food systems. *Lancet*, 393(10170):447–492.

Yang, Q., Zhang, Z., Gregg, E. W., Flanders, W. D., Merritt, R., and Hu, F. B. (2014). Added sugar intake and cardiovascular diseases mortality among us adults. *JAMA Internal Medicine*, 174(4):516–524.

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., and Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, 12(7):e0181821.

Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLOS ONE*, 10(9):e0138740.