Counteracting Narratives: Evidence from An Online Experiment*

MANWEI LIU† and SILI ZHANG‡

April 15, 2025

Abstract

Can people counteract biased narratives with subsequent information? Using an online experiment where counteracting may have the best odds by design, we investigate this question by first randomly assigning subjects to read different narratives that contain the same facts, and then offering them the opportunity to acquire and process balanced arguments. We document three main findings. First, subjects shift their attitudes towards the standpoint of the randomly assigned narrative, knowing that the narrative is slanted and randomly assigned. Second, the opportunity to read additional arguments does not prompt subjects to counteract the persuasion effects of the initial narratives. Third, when evaluating subsequent arguments, participants find arguments aligned with the randomly assigned narrative more convincing. These findings remain qualitatively similar in additional treatments where the balanced arguments are provided two weeks after initial exposure to narratives. Only when we replace these arguments with the exact opposing narratives that subjects do not initially see are they able to fully counteract the effects. Taken together, our results highlight the importance of balanced and complete exposure at the outset in counteracting the influence of biased narratives.

Keywords: narratives, counteract, information acquisition, awareness

JEL Codes: C90, D83, D91

*We thank the editor and three anonymous reviewers for their many insightful comments and suggestions. We are grateful to Chiara Aina, Peter Andre, Davide Cantoni, Francesco Capozza, Yan Chen, Tilman Fries, Eline van der Heijden, Florian Englmaier, Matthias Lang, Michel Maréchal, Jan Potters, Klaus Schmidt, Simeon Schudy, Marta Serra-Garcia, Roberto Weber, Mirko Wiederholt, and participants in various seminars and conferences for the comments and suggestions. We thank Georg Weizsäcker for connecting our paper with the general public: a summary of the initial findings from this paper was shared with *BSE Insights* and *Researching Misunderstandings*. Financial support from the TiSEM CentERlab grant from the University of Tilburg and the LMUexcellent grant from the Ludwig Maximilian University of Munich is gratefully acknowledged. Liu is also supported by the National Natural Science Foundation of China (Grant No.72403124, No.72250710170, No.72373069). The experiment reported in this paper was approved by TiSEM Institutional Review Board at Tilburg University (IRB EXE 2020-009). The main experiment was pre-registered at AEA RCT registry as AEARCTR-0005595 (Liu and Zhang, 2020), and the additional long-run experiment and the mechanism experiment were pre-registered at AEA RCT registry as AEARCTR-0013582 (Liu and Zhang, 2024).

†Nanjing Audit University, Department of Economics, 21000 Nanjing, China. Email: liumanwei@nau.edu.cn.

‡LMU Munich, Department of Economics, 80539 Munich, Germany. Email: sili.zhang@econ.lmu.de.

1 Introduction

Ever since the birth of human society, much of the information communicated is not just fact, but narrative—qualitative interpretations of objective facts or events (Harari, 2015). By word choice, framing, and organizing events chronologically and logically, narratives can convey a particular, often biased, viewpoint without necessarily lying about the facts (Bursztyn et al., 2023; Shiller, 2017, 2020). Biased narratives have been shown to impact opinions and voting behavior (DellaVigna and Kaplan, 2007), the formation of echo chamber (Bakshy et al., 2015; Cinelli et al., 2021), attitudes towards redistribution (Alesina et al., 2023), and political polarization (Gentzkow and Shapiro, 2010; Iyengar and Westwood, 2015; Sunstein, 2017). In the face of these challenges to modern society, many advocate for raising awareness, assuming that once individuals recognize the use of biased narratives, they will be able to counteract them. However, this key assumption—that awareness naturally leads to action and better outcomes—remains an open and often overlooked question.

This paper investigates whether, and to what extent, people can counteract the influence of slanted narratives with subsequent information when they are fully aware of the slant—a situation in which *counteracting may have the best odds*.[1] Answers to these questions have important implications for how best to regulate information exchange in mass media, and for evaluating any policy proposals that rely crucially on individual sophistication, such as awareness campaigns. However, these questions are difficult to answer with observational data because narratives are often bundled with selective coverage or editorial endorsement, and exposure to narratives is often endogenous—what people read depends on what they believe. This is especially true in the political realm, which has received the most attention to date, where people tend to have strong and divided views to begin with. To address these challenges, we design an online experiment around a concrete topic, deliberately staying away from political contexts, that allows for the construction of narratives based on the same set of objective facts. Importantly, we randomly assign these constructed narratives to participants and make the subjects explicitly aware of the random assignment, and thus the existence of bias. Meanwhile, to further ensure the best chance of counteracting, we also minimize the role of prior opinions by combining topic selection in an auxiliary survey with a pre-screening process.

Our online experiment is built around a concrete topic in the non-political domain: the use of genetically modified (GMO) mosquitoes in disease control. This topic is selected using an auxiliary experiment so that the subjects have little knowledge about it and have no strong prior attitudes toward

[1] We use the words "bias" and "slant" interchangeably in this study without attaching any normative judgement to them. We will discuss the construction of normative benchmarks in more detail later in the context of our experiment.

it, which we further ensure by screening the participants' knowledge and attitudes. We then collect a set of published academic journals and news reports, and based on the same set of facts about the topic contained in these materials, we construct two versions of narratives, slanted toward two opposite sides: *pro-GMO* or *anti-GMO*. In this way, we also keep constant the source of information perceived by the subjects in the experiment. The two versions of the narratives generate contrasting impressions by interpreting the same objective facts or events in qualitatively different ways. Otherwise, the two versions of narratives are very similar. Importantly, we consider individuals with access to both narratives as the benchmark, as they receive a balanced perspective in the sense of Mullainathan and Shleifer (2005).

The design of our main experiment involves two stages. In the first stage, we randomly assign the constructed narratives between subjects and examine the effect of narratives on attitudes. Subjects read either the *pro-GMO* narrative, the *anti-GMO* narrative, or both versions in randomized order as a benchmark described above. Their attitudes towards the use of GMO are elicited using both a self-reported measure and an incentivized measure implemented as a donation, in order to establish the first stage effect of narrative persuasion. The effect of slanted narratives is always demonstrated relative to the benchmark group that has been exposed to both narratives.[2]

The second stage is information acquisition. We offer subjects the opportunity to read additional arguments made by laypeople on the Internet for both sides of the GMO issue. Our setting thus largely represents everyday situations where narrative persuasion matters, in which individuals form first impressions based on initial media coverage that is more comprehensive, before exchanging arguments with others or acquiring additional information in a more fragmented manner. Specifically, we encourage and incentivize argument acquisition by asking subjects to estimate the attitude of the benchmark group who have read all of the available arguments and rewarding them for accuracy. As a result, information acquisition has instrumental value, and it is in the subjects' interest to acquire all the available arguments that are monetarily costless.[3] To evaluate the effect of the opportunity to acquire additional arguments, we elicit the subjects' attitudes towards the use of GMO for a second time after the argument acquisition.

A key feature of our experiment is that we explicitly make subjects aware of the use of slanted narratives. All subjects are informed upfront that each of the narratives they receive is slanted towards one side of the issue, and that the assignment of narratives is completely random. Additionally, each

[2]In Section 2.2, we clarify that we are agnostic about the normative merits of this benchmark, and elaborate on why this benchmark can be considered as the best proxy for a narrative-free treatment when distilling facts turns out to be empirically challenging (e.g., Bursztyn et al., 2023).

[3]Admittedly, processing additional information takes time and mental effort, even if it is not financially costly. We will discuss this point in more detail and explicitly address the role of information processing costs in additional data collection.

narrative is labeled explicitly as either *pro-GMO* or *anti-GMO*. At the very end, to verify the extent to which participants are indeed aware of the impact of the slanted narratives, we elicit subjects' second-order beliefs—their estimates of the donations made by the benchmark group, who read both narratives, and by their counterparts, who read a narrative of the opposing side.

We begin by documenting that the initial exposure to narratives substantially influences people's attitudes, despite the fact that people are made aware of the slant and the random assignment of the narratives. Specifically, subjects who read a *pro-GMO* narrative are significantly more positive towards the use of GMO mosquitoes than those who read both narratives, whereas those who read an *anti-GMO* narrative are significantly more negative towards the use of GMO mosquitoes than those who read both narratives. The impact of these slanted narratives is also reflected in the incentivized measure, the allocation of donations, in a similar manner. This result aligns with existing research on the effectiveness of media persuasion. What is surprising is that we have demonstrated a similar effect in a setting where people are made to have full awareness.

Our main focus is on the effect of subsequent information, which addresses the question of whether people are able to counteract the impact of slanted narratives when they are fully aware of them. By comparing subjects' attitudes elicited before and after the acquisition of arguments, we find that the opportunity to read additional arguments *does not fully offset* subjects' attitudes shaped by randomly assigned slanted narratives at the beginning. In other words, subjects do not fully counteract the initial influence of slanted narratives, even in a situation where counteracting may have the best odds. This result is unlikely to be driven by a preference for consistency, as the distribution of attitudes before and after arguments is different.

Our next step is to investigate the potential mechanisms underlying subjects' failure to counteract the impact of slanted narratives. First, we observe a clear discrepancy in subjects' second-order beliefs about the benchmark group and the opposing group, suggesting that subjects indeed anticipate the effects of slanted narratives. This allows us to rule out the possibility that subjects are nevertheless unaware of the narrative persuasion effect. Second, we find that the majority of the subjects do not acquire all the available arguments, even though reading all the arguments indeed improves estimation accuracy and subjects are incentivized to do so. However, this incomplete argument acquisition alone cannot account for subjects' failure to counteract the bias, because in additional treatments in which all arguments are provided exogenously, the effect of slanted narratives again persists. Third, we find that subjects evaluate arguments that are aligned with the randomly assigned narrative more favorably. This behavioral mechanism provides a plausible explanation for why offering more balanced arguments

fails to narrow the gap between different sides. That is, the initial exposure to slanted narratives not only affects subjects' attitudes but also the way they process subsequent information. Finally, we provide suggestive evidence to rule out the possibility that arguments are (perceived as) unconvincing.[4]

To shed light on the long-run dynamics of the effect of narratives, thereby informing whether time helps to counteract, we conduct two additional treatments in which the balanced arguments are provided two weeks instead of immediately after the initial exposure to narratives. These two treatments are conducted with minimal design changes to assess the robustness to a doubling of the stakes, the role of extreme priors, and other potential explanations for the failure to counteract, which we will discuss later. Otherwise, the experimental protocol remains similar to the main experiment to ensure that the studies are comparable. The results in these additional treatments are qualitatively similar to those in the main treatments when considering the same subsample, although not all retain statistical significance.[5]

To further elucidate mechanisms related to the potential distinction between initial narratives and subsequent arguments,[6] we conduct three additional treatments in which the arguments in the second stage are replaced by the exact counter-narrative that subjects do not see in the first stage. The rest of the experiment follows the same modified protocol as in the long-run treatments for robustness. We again replicate the strong narrative persuasion in the first stage, but now subjects are able to fully counteract the effects after seeing the exact opposing narratives from the other side in the second stage. The results in this set of treatments thus provide an empirical justification for the choice of benchmark in our main experiment, as well as confirming the role of additional mechanisms such as information complementarity at play. Finally, we explore the role of more alternative mechanisms using explicit measures that we include in all mechanism treatments, such as the experimenter demand effect, rational inattention, and induced group identity, but find little evidence for them.

Taken together, we draw two broader lessons for counteracting the influence of narratives. First, exposure to slanted narratives can have lasting effects, rendering much of the subsequent exchange of arguments ineffective by also biasing the evaluation of those arguments. This is the case especially when people have modest views, e.g. when they are most uncertain. Second, while providing the exact opposing narratives is effective, this "boundary condition" is unlikely to be met in practice, since real-life situations that resemble such a point-to-point opposition scenario, where people are forced to directly

[4]We will further discuss the role of potential differences between arguments and narratives using the additional mechanism treatments. As we will describe later, in these additional mechanism treatments, we find that half of these arguments are perceived as more convincing than narratives.

[5]Note that these long-term results should be treated with caution due to the limited statistical power of the tests. We will discuss the power issue and compare the effect size of our results with other related work in Section 4.

[6]This distinction can be interpreted in different ways to explain our results. For example, some arguments may be complementary to the narratives, or the arguments may not directly address the narratives, etc.

juxtapose counter-narrative constructed as the exact opposite, are extremely rare.[7] Jointly, our findings highlight the importance of balanced and complete exposure at the outset, as opposed to counteracting the effects later. Awareness alone is certainly not enough.

This paper is connected to three sets of work. First, it adds to the nascent literature on narratives in economics (e.g. Andre et al., 2024; Morag and Loewenstein, 2024; Shiller, 2017) by studying how best to counteract narratives, thereby informing the dynamic property of narratives. Instead of imposing directed acyclic graph (DAG) structures (Eliaz and Spiegler, 2020) or other model-based structures (Schwartzstein and Sunderam, 2021) on narratives, we adopt a reduced-form definition of narratives—the qualitative interpretations of objective facts or events. Most contemporary empirical work focuses on the static persuasion effect of narratives in a variety of domains (Barron and Fries, 2024; Barron et al., 2021; Bénabou et al., 2018; Bursztyn et al., 2023; Harrs et al., 2021; Hillenbrand and Verrina, 2022; Kendall and Charles, 2024),[8] or the mechanism of why narratives can be more persuasive than statistics (Graeber et al., 2024). We contribute to this literature by investigating (i) whether narratives are still persuasive when people are explicitly made aware of the usage, and (ii) the dynamics of narrative persuasion, i.e., whether people can counteract the effects of narratives with subsequent information and over time. Our results highlight that awareness alone is not sufficient to counter narrative persuasion and help clarify the conditions under which narrative persuasion is more or less persistent.

Second, this paper is connected to the interdisciplinary literature on confirmation bias (e.g., Jones and Sugden, 2001; Klayman and Ha, 1987; Lord et al., 1979; Rabin and Schrag, 1999).[9] Existing work on confirmation bias typically takes the form of demonstrating belief polarization due to prior biased inferences about the same information on emotionally charged or contested issues, such as politics, where motivated or identity-driven beliefs can play a large role (e.g., Bauer et al., 2023; Iyengar and Hahn, 2009; Thaler, 2023). We add to this line of work in two ways. First, we focus on a case where priors are not pre-established, but are instead *induced* by narratives that subjects know to be randomly given in a relatively neutral context. Thus, what we find in the main experiments appears to be a novel form of confirmation bias, where subjects' evaluation of arguments is biased by the randomly assigned narratives instead of home-grown priors to which people are attached. This suggests that the pattern may be due to imperfect cognition or information processing, independent of motivations (Charness et al., 2021; Jehiel

[7] Even if they exist, most people are unlikely to actively seek them out. When asked what kind of information they would seek to maintain an unbiased perspective, only 26% of subjects choose comprehensive articles from the opposing side. In contrast, around 40% prefer comprehensive articles from their own side, and 27% prefer balanced arguments.

[8] The empirical literature of narratives itself thus connects to empirical work on media persuasion (DellaVigna and Gentzkow, 2010), which demonstrates that slanted news can persuade and affect people's attitudes and behaviors using survey experiments (Coppock et al., 2018) or quasi-experimental variations (Chiang and Knight, 2011; DellaVigna and Kaplan, 2007; Djourelova, 2023; Martin and Yurukoglu, 2017).

[9] See Nickerson (1998), Charness and Dave (2017), and Chapter 2.8 in Benjamin (2019) for a thorough review.

and Steiner, 2020; Pennycook and Rand, 2019). Second, our results in the opposing narrative treatments provide a boundary condition for confirmation bias when the subsequent information provides the *exact* opposite interpretation. By removing the room for open interpretation, opposing narratives eliminate confirmation bias. This provides empirical support for a mechanism of (non-motivated) confirmation bias and polarization that relies on information being open to interpretation (Fryer et al., 2019).

Lastly, by focusing on whether people are able to counteract the effects of narratives, this paper is related to the interdisciplinary work on continued influence of misinformation (e.g., Chan et al., 2017; Lewandowsky et al., 2012; Walter and Tukachinsky, 2020). Our results demonstrate a similar persistent effect and its boundary condition in settings without any misinformation, thus providing a rationale for why "pre-bunking" may work better than "de-bunking" (Ecker et al., 2022). Relatedly, this paper also connects to a growing applied literature that aims to mitigate polarization or to impact attitudes by providing cross-cutting information (Bail et al., 2018; Broockman and Kalla, 2022; Chen and Yang, 2019; Di Tella et al., 2021; Levy, 2021) or arguments from the other side during a debate (Schwardmann et al., 2022).[10] In contrast to existing studies that primarily focus on topics in already highly contested environments,[11] this paper takes a step back by setting up an apolitical environment in which people are not yet divided. Nevertheless, our results suggest that the provision of cross-cutting information is ineffective in counteracting light-touched narratives even in favorable conditions, which naturally casts a pessimistic light on its effect in more challenging scenarios. In this spirit, our results on people's inability to fully counteract narratives can be seen as complementary to Gonçalves et al. (2024), which finds in a stylized setting that the retraction—the counter signal to the previous signal—does not affect beliefs as much as the original information.

The remainder of this paper is organized as follows. Section 2 presents the experimental design. Section 3 presents the results of the main experiment. Section 4 examines the long-term effects of the narratives, and Section 5 explores the mechanisms associated with the opposing narratives. Finally, Section 6 discusses other mechanisms and then concludes.

[10] Evidence is mixed across existing studies. Bail et al. (2018) find that following Twitter bots with opposing political ideologies backfires, especially among conservatives. Di Tella et al. (2021) also find a backfire effect of counterattitudinal exchanges among people initiated within echo chambers. Levy (2021) looks at individuals' subsequent news consumption after being asked to subscribe to liberal or conservative news outlets. In his setting, while exposure to counter-attitudinal news reduces negative attitudes toward the opposing party, it has no effect on political opinions. Similarly, Schwardmann et al. (2022) finds no evidence of convergence in factual beliefs and attitudes. In a field experiment in China, Chen and Yang (2019) find that temporary free access to the uncensored Internet alone does not lead to more acquisition of cross-cutting information, but incentivizing information acquisition does matter and leads to substantial changes in citizens' attitudes. Broockman and Kalla (2022) find a strong positive effect of Fox News viewers watching CNN on political views in the short term, which largely dissipates after two months.

[11] In psychology, Brenner et al. (1996) focus on hypothetical legal disputes. They show that one-sided evidence leads to biased predictions and judgments, as in our first stage, but do not study subsequent counteractions.

2 Experimental Design

We aim to build our experiment around a concrete topic that allows for the construction of narratives based on the same set of objective facts, while *minimizing the role of prior opinions* in order to give counteracting the best chance. Section 2.1 describes an auxiliary survey and a pre-screening process that we use to achieve these goals. In particular, we first use the auxiliary survey to select the issue about which a representative subject knows little and does not have strong prior views—the use of genetically modified (GMO) mosquitoes in disease control. We then elicit subjects' prior opinions and knowledge about the issue in order to identify subjects who know little and do not have strong prior opinions about the issue. While we screen out subjects with strong priors and knowledge in the main experiment by design, we include these subjects in additional mechanism treatments to explicitly assess the role of priors. Section 2.2 describes the construction of narratives based on the pre-selected issue and our benchmarking principle, which serves as the basis for our experiment.

Section 2.3 describes the main treatments and additional mechanism treatments, all of which share a two-stage structure: (i) in stage I, subjects are randomly assigned to read different narratives we have constructed that contain the same facts, explicitly informed about the random assignment of different narratives; (ii) in stage II, subjects evaluate subsequent information with incentives to counteract. Treatment variations occur in stage II. Specifically, subjects in the main treatments *Endo* counteract by acquiring additional balanced arguments from both sides, whereas subjects in the *Exo* treatment counteract with exogenously provided arguments. Two additional sets of mechanism treatments, *Long* and *Oppo*, are added to investigate the long-term effects of narratives and to elucidate mechanisms. All treatments are pre-registered at AEA RCT registry.

We summarize all sets of treatments and discuss procedural details across the two waves of data collection in Section 2.4. Finally, we present our hypotheses in Section 2.5.

2.1 Issue Selection and Pre-Screening

Issue selection An ideal issue for the experiment meets the following criteria: i) subjects have limited knowledge about the issue; ii) subjects do not have strong prior opinions about the issue. The second criterion distinguishes our work from previous research that primarily focuses on settings that invoke motivated reasoning, such as political domains.

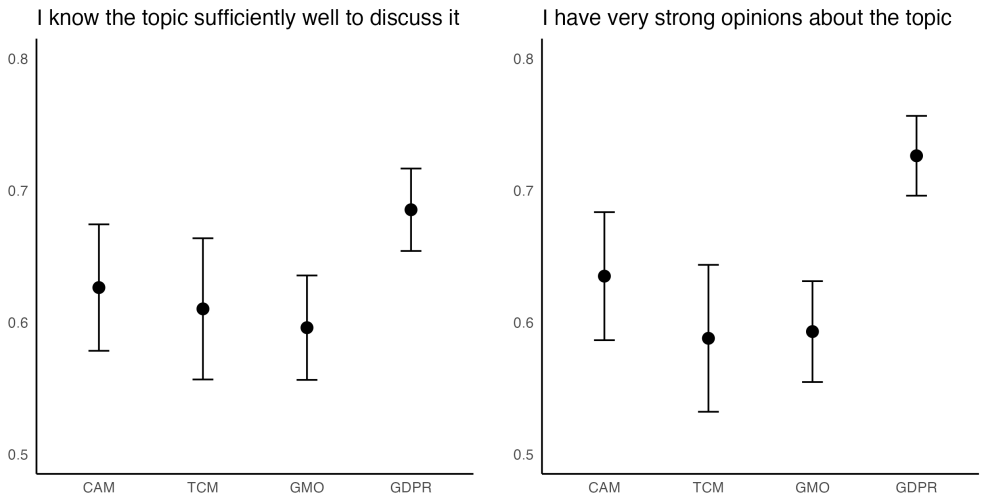
To choose an issue that best suits our needs, we conducted an auxiliary survey of 163 respondents using Amazon Mechanical Turk (henceforth Mturk). Below is the list of candidate topics in the survey.

For each of the topics, we provide a one-sentence description of the topic and then ask respondents to indicate their level of knowledge and their prior opinion on a seven-point Likert scale.

- Complementary and alternative medicine (CAM)
- Traditional Chinese medicine (TCM)
- Genetically modified (GMO) mosquitoes
- The General Data Protection Regulation (GDPR)

Figure 1 shows the average standardized level of knowledge and the average strength of prior opinions across topics. Among the four topics we tested, genetically modified (GMO) mosquitoes and traditional Chinese medicine (TCM) outperform the other two topics in meeting our two criteria. We choose GMO mosquitoes over TCM because the former has smaller variances in responses.

Figure 1: Knowledge and Prior across Topics



Notes: This figure presents subjects' level of knowledge and the strength of prior attitudes across topics (95% CI). The left panel shows the extent to which subjects agree with the statement "I know the topic sufficiently well to discuss it". The right panel shows the extent to which subjects agree with the statement "I have very strong opinions about the topic". Responses vary from "Strongly disagree" to "Strongly agree" and are standardized to [0,1].

Pre-screening and exclusion The pre-screening process takes place right before the main experiment. It aims to identify the relevant subsample of subjects who are not too knowledgeable about GMO mosquitoes and do not have strong pre-existing attitudes towards them, thereby reinforcing the goal of giving counteracting the best chance. We first briefly introduced the subjects to the topic, GMO mosquitoes. We then asked them: (1) how much do you know about genetically modified mosquitoes without searching online? and (2) what is your general attitude toward GMO mosquitoes? For the first

question, subjects state their answers on a five-point scale from "none at all" to "a great deal". For the second question, they choose from "extremely negative" to "extremely positive" on a five-point scale.

In the main experiments *Endo* and *Exo*, we screen out subjects who think they know "a great deal" about GMO mosquitoes and subjects who are either "extremely negative" or "extremely positive" about GMO mosquitoes already at the recruitment stage. In the second wave, the mechanism treatments *Long* and *Oppo*, while we perform the analysis on the same pre-screened subsample to ensure the comparability of our analysis, we do not exclude corresponding subjects from participating the study. This allows us to further assess the role of priors in individuals' ability to counteract. The pre-screened subsample is about 84% of the full sample, and the distribution of priors and knowledge about GMO mosquitoes is similar across the two waves of data collection, as shown in Figure 10 in Appendix D.

2.2 Construction of Narratives and the Benchmark

To construct slanted narratives, we narrow down the topic—GMO mosquitoes—to a more concrete issue: releasing GMO mosquitoes into the wild for disease control. Whether one supports or opposes such practices are two mutually exclusive sides of this issue. We collect a set of published academic journals and news reports about a field practice in Brazil. Based on the same set of events and facts mentioned in these papers and reports, we construct two articles slanted toward either the pro-GMO side or the anti-GMO side. Specifically, we use a combination of techniques such as word choice, framing, logical reorganization of events, etc., without introducing any new events, such that the two narratives offer different interpretations of the same events and facts. This strategy allows us to keep events and facts as constant as possible without having to isolate a set of facts first or take a stance on the precise definition of facts, which is an important point to which we come back when constructing a benchmark.

Both narratives can be found in Appendix A, and in particular Table 7 lists various examples of the facts used in the narratives, the interpretations of these facts by the narratives, and the specific techniques we use in constructing the narratives. Below is one of the examples.

- Fact: The field practice in Brazil reduced the local mosquito population by 90%.
- Pro-GMO narrative: "The practice was a huge success. The release of GMO mosquitoes worked tremendously well by reducing local mosquito populations by 90% during the trial."
- Anti-GMO narrative: "The release of GMO mosquitoes seemed to work. The local mosquito population was reduced substantially, yet, almost 10% of them remained active."
- Technique: Word choice and framing.

Given that we aim to investigate the effect of biased narratives and how people counteract them, it helps to first clarify what the "unbiased" or narrative-free benchmark is. While it is possible to slant a given set of facts in different directions, thereby adding narratives, it turns out to be challenging to reverse the process and create a narrative-free or an unbiased benchmark, for two reasons. First, narrative is a loose and non-binary concept on which there is no consensus yet. Fact-based narratives can vary in degree of narrativity depending not only on linguistic features, but also on domains and readers. Thus, it is almost impossible for a variety of readers to agree on what writing is narrative-free and what writing is not.[12] Second, in the developing field of GMO mosquitoes with many scientific nuances, we are not equipped with the expertise to determine the "unbiased" position by any simple and objective standard when experts in this field may even disagree. Therefore, we adopt a pragmatic approach without taking a normative stance, as in Mullainathan and Shleifer (2005), where "bias" is defined relative to a balanced benchmark group that has access to both narratives. Subjects who have access to both narratives could easily recognize how narratives are constructed and therefore are *as-if* in a narrative-free treatment.[13]

2.3 Treatment Variations: Stage I Narratives $\times$ Stage II Information

As previewed earlier, all treatments share a two-stage structure, where in stage I subjects are randomly exposed to slanted narratives that they know to be random, and in stage II subjects counteract the effect of narratives with additional information. Below we first describe variations in stage I and how we ensure that subjects are aware of the use of narratives. We then move on to describe the four sets of treatments with variations that occur in stage II, where the subsequent information is balanced arguments to be acquired endogenously (main treatment *Endo*), to be given exogenously (main treatment *Exo*), to be given exogenously after two weeks (mechanism treatment *Long*), or where the subsequent information is the exact counter-narrative (mechanism treatment *Oppo*). Information processing is always incentivized using a novel accuracy-based estimation task, which we will explain in detail. At the end of stage I and at the end of stage II, we elicit subjects' attitudes towards the issue to evaluate the impact of slanted narratives and the extent to which subjects counteract them with additional information.

In stage I, subjects are randomly exposed to the narratives constructed as in Section 2.2. Subjects

[12] See Barron and Fries (2024) for an overview of different conceptualizations of narratives. See Ochs et al. (2009) and Herman (2011) for understanding narratives not as a binary category but as a matter of degree. Bursztyn et al. (2023) provide evidence that people generally confuse, and thus disagree, about what are facts and what are opinions.

[13] Results from our mechanism treatments provide support for this claim, which we discuss later. Our approach is thus similar to the baseline in the context of framing effects proposed by Druckman (2001), where those who are exposed to competing frames are considered as "de-framed".

are randomly exposed to the pro-GMO narrative in *Pro* treatment, the anti-GMO narrative in *Anti* treatment, or both narratives in a randomized order in *Both* treatment. A key feature of stage I is that all subjects are explicitly informed upfront about the construction and the random assignment of the narratives. Specifically, they are told that the narratives are constructed based on the same facts published in academic articles or news reports, but slanted towards one side of the issue and that the assignment of narratives is completely random. Furthermore, each narrative is explicitly labeled as either *pro-GMO* or *anti-GMO* so that subjects know which version of the narrative they are assigned to read. When subjects finish reading, we ask them a few comprehension questions to make sure that they have actually read the article and more importantly that they have grasped the critical facts in the article. Subjects can try as many times as they want or go back to read again, and can proceed to stage II only if they answer all comprehension questions correctly.

2.3.1 Main Treatment *Endo*: Endogenous Argument Acquisition in Stage II

In treatment *Pro.Endo*, *Anti.Endo*, and *Both.Endo*, we offer the opportunity for subjects to acquire further arguments from both sides of the issue in stage II. To achieve this, we provide a total of eight additional arguments reasoning either for or against the use of GMO mosquitoes to eliminate disease-carrying mosquitoes or the use of GMO technology in general. Four of these arguments are leaning toward the pro-GMO side and the other four arguments are leaning toward the anti-GMO side. We construct these arguments based on discussions on an online platform, Quora, and edit these arguments so that they are comparable in terms of length and clarity. Appendix B contains the list of these arguments.

When offered the opportunity, each subject in *Pro.Endo*, *Anti.Endo*, and *Both.Endo* can indicate the number of arguments she would like to read from either side. Arguments are randomly drawn from the corresponding side and presented to the subject in random order. Upon reading each argument, subjects are asked to rate how convincing each argument is on a seven-point scale, ranging from "extremely unconvincing" to "extremely convincing".

The setting of *Endo* represents typical everyday situations where individuals form initial impressions based on media coverage that is more descriptive and comprehensive, and subsequently exchange arguments or acquire additional information in a more fragmented manner.[14] In the second wave of the experiment, we include a question about different types of information that people would normally choose to get a more balanced view, including balanced arguments, comprehensive articles from the opposing side, comprehensive articles from the preferred side, etc., to get a sense of whether what is captured by

[14] Other situations can often exist, for example, where people are first exposed to fragmented arguments before they have access to full media exposure. These situations are also interesting for future research.

our setting reflects people's preference for information sources when it comes to counteracting.

Incentivizing information acquisition and processing: prediction of *Benchmark* Importantly, we incentivize information acquisition and processing by setting up novel accuracy-based incentives: subjects are asked to estimate the attitude of a separate *Benchmark* group that has read all the narratives and all the arguments. Subjects can earn up to one dollar if their estimates are correct. As such, these arguments have instrumental value. In order to make a sensible estimation, subjects in treatment *Pro.Endo*, *Anti.Endo*, and *Both.Endo* should acquire *all* the arguments *Benchmark* group has seen. This is especially true when the information is not monetarily costly to acquire or process (e.g., each argument takes a native English speaker around 20 seconds to read).[15]

Assuming that a more balanced intake of information can lead to convergence or a less biased judgment, in the sense of Mullainathan and Shleifer (2005), subjects are expected to counteract or attenuate the influence of slanted narratives on their attitudes and judgments by acquiring additional information. Next, we explain how we measure their attitudes.

Outcomes of interest: attitudes toward the issue After randomly exposing subjects to narratives in stage I, we elicit their attitudes toward the issue using three outcome measures. First, we ask subjects to report on a nine-point scale whether they think the practice of releasing GMO mosquitoes to prevent the spread of disease is an appropriate use of technology or is taking technology too far.[16] Second, we ask subjects to allocate a donation worth 100 dollars between a pro-GMO mosquito organization and an anti-GMO mosquito organization. The total amount of donation is fixed to rule out other motives that drive donation decisions, such as social preference or image concern. We inform subjects that we make the donation according to the average of all responses.

Recall that to incentivize information acquisition, we ask subjects (except those in *Benchmark*, the target group) to estimate the attitude of a separate *Benchmark* group. Specifically, subjects are asked to estimate the proportion of the total donation *Benchmark* allocates to the pro-GMO organization on average. Subjects will earn a bonus worth up to one dollar if their estimates are close enough to the actual average allocation of *Benchmark*. This second-order belief measure about the *Benchmark* is our third outcome measure.

The purpose of including this additional second-order belief measure is two-fold. First, we include the estimate with real incentives, intending to replicate early evidence in psychology based on

[15]We will evaluate the role of information processing cost in more detail and, in particular, explore how our results depend on additional variables that approximate the opportunity cost and cognitive capacity of the subjects.

[16]This scale was used by PEW in its survey on Americans' views on the genetic engineering of animals.

non-incentivized measures (Brenner et al., 1996). Compared to the donation decision, which is also incentivized, this accuracy-based measure is free from other motives such as the perceived quality of the organizations. Second and most importantly, including this measure allows us to incentivize and encourage information acquisition in stage II, which we have discussed earlier.

After stage II, we repeat all three measures of attitudes that we use after stage I for all subjects. The main focus is on the subjects' attitudes toward the issue *before and after* access to additional information. This design choice allows for a within-subject comparison of our outcome variables, which we can correlate with other variables to investigate potential channels of why counteracting slanted narratives is successful or unsuccessful.

There are two potential concerns of measuring our main outcome variables twice using a within-subject design. The first is that within-subject variations can potentially lead to an experimenter demand effect. We deem this not concerning as it is not *ex ante* obvious in which stage the experimenters are looking for an effect or in which direction the effect is expected. Nevertheless, we explicitly address this concern by measuring subjects' sensitivity to the experimenter's demand and their beliefs about the purpose of the experiment in additional treatments, which we discuss in Section 6. The second concern is that subjects may want to appear consistent in their responses (Falk and Zimmermann, 2013), leading to a smaller effect of information acquisition. Although there is little evidence supporting this potential concern in recent research with similar design features,[17] we use multiple outcome measures, including an incentivized one, to mitigate this concern, as suggested in Haaland et al. (2023).

2.3.2 Main Treatment *Exo*: Exogenous Argument Provision in Stage II

Recall that in stage II of the main treatment *Endo*, subjects endogenously acquire additional arguments to counteract slanted narratives. To evaluate the impact of all the additional arguments per se, we add a second set of main treatments *Exo*, where all the additional arguments are provided exogenously to subjects in random order in stage II. These treatments create a "counterfactual" case that allows us to isolate the extent to which subjects counteract slanted narratives had they read all the available information we provide. In this treatment set, subjects go through stage I in the same way as in the main treatment *Endo*, but are asked to read all the arguments instead of making their own decisions about information acquisition in stage II.

We will refer to the additional treatments as *Pro.Exo* and *Anti.Exo*, depending on the narrative(s) randomly assigned in stage I. Note that *Both.Exo* is effectively identical to the *Benchmark* explained

[17] For instance, Roth and Wohlfart (2020) do not find any significant effect of eliciting priors on the posteriors in an information provision experiment on macroeconomic risk.

earlier. In what follows, we may use *Benchmark* and *Both.Exo* interchangeably.

2.3.3 Mechanism Treatment *Long*: Exogenous Argument Provision after Two Weeks

In the second wave of data collection, we include an additional set of mechanism treatments, *Long*, to examine whether the impact of slanted narratives persists two weeks after the initial exposure. Specifically, the *Pro.Long* and *Anti.Long* treatments replicate the *Pro.Exo* and *Anti.Exo* treatments with a two-week interval between stage I and stage II.

After the main results from the main treatment, we are most interested in the difference between the *Pro.Long* and *Anti.Long* treatments, which renders a *Both.Long* group redundant. However, since the implementation of the second wave data collection slightly differs from the first wave,[18] we nevertheless collect additional data for *Both.Long* to establish a valid benchmark group as the target of estimation. *Both.Long* is therefore not a proper treatment group but an instrumental group used to incentivize reading and evaluating arguments, ensuring the consistency across treatments as much as possible. *Both.Long* simply replicates the *Benchmark* treatment without a two-week interval with a much smaller sample and its data is not used for analysis.

2.3.4 Mechanism Treatment *Oppo*: Exogenous Opposing Narrative Provision

Although we will examine several different channels for whether subjects can counteract the influence of narratives in the main experiment, we cannot completely rule out alternative channels related to the differences between subsequent arguments and initial narratives due to our very design choice. These alternative channels range from arguments not addressing the concerns of the narratives, to arguments complementing the initial narratives thus making them stronger, to arguments being less persuasive than narratives, to arguments differing from narratives in terms of length and format, and so on.[19] Modifying the exact textual details to address each of these channels seems infeasible, since some of them are contradictory to others. Instead, we add another set of mechanism treatments *Oppo* in the second wave of data collection, in which we replace the stage II arguments with the exact counter-narrative that subjects do not initially see. This set of treatments allows us to explicitly test the role of these channels as a whole, in addition to those already found in the original experiment.

[18] All differences and the reasons for them are documented in the pre-registration of the mechanism treatments. We will additionally highlight some of these differences in Section 2.4, especially when they are introduced to address a specific concern.

[19] While we acknowledge that narratives and arguments are not directly comparable, in daily information consumption, it is common for people to encounter a mixture of information of different formats and quality that provides coherent or disjoint content. We will discuss the empirical frequency with which people seek different sources of information in order to gain an unbiased perspective later on, and compare the convincingness of narratives and arguments in Appendix F.

Subjects in *Oppo* treatments go through stage I just like those in other treatments. But instead of reading eight balanced arguments in stage II, subjects in *Pro.Oppo* and *Anti.Oppo* are provided with the other narrative that they do not initially read in stage I. After reading the second narrative, subjects are asked to rate how persuasive this second narrative is.

Similar to treatment *Both.Long*, we implement *Both.Oppo* as a benchmark group to incentivize reading and evaluating subsequent information. In *Both.Oppo*, subjects read two narratives in random order and evaluate the convincingness of two narratives. Subjects in *Pro.Oppo* and *Anti.Oppo* are incentivized to estimate the donation allocation made by subjects in treatment *Both.Oppo*. But unlike *Both.Long* which functions merely as the target of estimation, *Both.Oppo* is implemented as a proper treatment group with the aim of obtaining subjects' evaluations of both narratives. These evaluations enable us to examine how two groups of different sides process identical information, which is otherwise impossible as *Pro.Oppo* only evaluates the anti-GMO narratives and vice versa.

This set of treatments is particularly important and informative for a number of reasons. First, *Oppo* is in a sense a more delicate *Both* group in stage I of the main experiment, with an active elicitation of attitudes between two narratives. Thus, it allows us to explicitly test whether *Both* group we construct as in Mullainathan and Shleifer (2005) can stand as an appropriate "narrative-free" benchmark. Second, *Oppo* represents almost the best possible theoretical condition for counteracting the influence of narratives, since subjects are forced to directly juxtapose opposing narratives. Real-life situations that resemble such a point-to-point opposition scenario, where the counter-narrative is constructed as the exact opposite, are extremely rare, if not non-existent.

2.4 Overview of the Experiment, Adjustments in Wave 2, and Procedure

Table 1 provides an overview of the experiment, summarizing all four sets of treatments described in the previous section. In the second wave of data collection, where we run additional mechanism treatments, we introduce some implementation changes in addition to the treatment variation described above. These adjustments are either due to implementation constraints or to address a specific aspect of robustness. Major implementation differences with conceptual reasons between the two waves of data collection are highlighted and discussed below.[20]

Screening As discussed in Section 2.1, in the first wave of data collection, our main treatments, we set up the experiment to minimize the role of prior opinions. This is intended to represent a situation in

[20]An outline of minor modifications due to implementation constraints, such as wording, donation organization, etc., can be found in the pre-analysis plan.

Table 1: Overview of the Experiment

	Wave 1: Main Treatments		Wave 2: Mechanism Treatments	
	Endo	*Exo*	*Long*	*Oppo*
Panel A: design				
	Elicitation of prior and knowledge			
Pre-screening	Screening out		*No screening out*	
Stage I	Exposure to narratives & Stage I elicitation of attitudes			
$ Incentives	Original level		*Doubled*	
Stage II Timing	Immediately	Immediately	*In two weeks*	Immediately
Stage II Info Acq	Endogenous	Exogenous	Exogenous	Exogenous
Stage II Info	Arguments	Arguments	Arguments	*Second narrative*
	Stage II elicitation of attitudes & *Additional measures*			
Panel B: procedure				
Platform	Mturkdata		Prolific	
Date	May, 2020		May, 2024	
N	362	354	$300 \rightarrow 219$	448

Notes: This table summarizes the experimental design and implementation of the four treatment sets. Treatments *Endo* and *Exo* were pre-registered at AEA RCT Registry (AEARCTR-0005595). Treatments *Nrt* and *Long* were pre-registered at AEA RCT Registry (AEARCTR-0013582). In both waves, we are targeting a sample from the US population with a track record of approval rate above 95%, and this sample from both platforms is similar in terms of observable characteristics.

which people are not yet polarized, and in which counteracting should have the best chance. Therefore, we intentionally pre-screened subjects based on their stated level of knowledge about the issue and their prior attitudes to ensure the best chance of counteracting, and screened out those who self-reported being very knowledgeable about the issue or having a strong prior on the issue during the recruitment stage.

In the second wave, we again elicit knowledge and prior attitudes before the experiment. To gain additional insight into the role of priors, we do not exclude any subjects during recruitment. While we still perform the main analysis on the same "screened" subsample to ensure that the results in the second wave are comparable to those in the first wave, we separately examine how including those with strong priors or knowledge might affect our results.

Doubled Incentives and Stakes In the first wave of data collection, the incentives for information acquisition and processing are half a dollar, while the donation level we use to measure attitudes in an

incentivized way is one hundred dollars. A valid concern is that the incentives are too weak for people to care. In the second wave of data collection, we double the monetary incentives and the stake size to explicitly address robustness to this concern.

Additional Variables In addition to the variables already included in the first wave, we include some additional variables in the second waves to explicitly investigate the role of alternative channels, including the experimenter demand effect (De Quidt et al., 2018), rational inattention (Maćkowiak et al., 2023), and induced group identity (Chen and Li, 2009).

First, to evaluate the role of experimenter demand effect, we include an open-ended question asking subjects to guess the purpose of the study, as recommended by De Quidt et al. (2018), as well as a novel measure of demand sensitivity inspired by De Quidt et al. (2018) and Zizzo (2010). Specifically, subjects are presented with a price list containing two options as a bonus payment. One option is paired with the message "it would be nice if some of you would choose this option" with a decreasing monetary value from \$1.1 to \$0.5, and the other option is fixed at \$1. For simplicity, we do not impose the single-switch rule but instead use the total number of times the former option is selected as a measure of subjects' willingness to sacrifice financially to please the experimenter.

Second, to evaluate the role of rational inattention, we include questions about self-reported memory, degree of cognitive reflection (Frederick, 2005), and opportunity costs measured by subjects' reported hourly wage. In addition, we measure subjects' intrinsic stake in counteracting narratives by explicitly asking them the extent to which they strive to maintain an impartial/unbiased position toward the issue.

Third, to evaluate the role of induced group identity (Chen and Li, 2009), we include a measure of individual-level groupiness (Kranton et al., 2020; Kranton and Sanders, 2017), the tendency to treat in-group and out-group differently.

Sample and Procedures The first wave of the experiment, including treatments *Endo* and *Exo*, were administered through Mturkdata[21] in May 2020. The pre-screening survey took less than 1 minute and was completed by 969 respondents. On average, respondents earned 5 cents for participating. Of those, 802 (82.7%) respondents passed the pre-screening according to the exclusion criterion explained earlier, that is, they did not see themselves as very knowledgeable about GMO mosquitoes and did not have extreme attitudes about GMO mosquitoes. Qualified respondents were immediately invited to proceed to the main experiment, which took around 15 more minutes. Approximately 95% of the subjects

[21]Mturkdata manages its own high quality samples for academic research based on Mturk (http://www.mturkdata.com/). We targeted a sample among the US population with a track record and an approval rate of over 95%.

who passed the pre-screening survey decided to continue with our experiment.[22] We obtained a total of 716 complete observations, with more than 100 observations in each treatment group. Subjects who completed the main experiment earned a minimum of 2 dollars and an average of 2.3 dollars.

The second wave of the experiment, including treatments *Oppo* and *Long*, were administered through Prolific in May 2024. We include the pre-screening questions to control for prior and knowledge without excluding any respondent as we did for the main treatments. The distribution of prior and knowledge of the subjects is very similar in the two waves of our experiment; see Figure 10 in Appendix D. A total of 448 respondents completed the *Oppo* treatments and earned 2.97 dollars on average. 300 respondents completed stage I of the *Long* treatments.[23] Two weeks later, they were invited to the second part of the experiment, of which 219 respondents completed. Subjects earned a fixed amount of 1 dollar for completing stage I and an average of 3 dollars for completing stage II of the experiment.

2.5 Hypotheses

In this section, we present our main hypotheses for the initial main treatments and additional hypotheses for the mechanism treatments, all of which are included in our pre-analysis plans.

Stage I: the impact of slanted narratives Our first hypothesis concerns the immediate effect of slanted narratives on attitudes after initial exposure, when participants are explicitly made aware of the random assignment of narratives. This hypothesis serves as a first stage throughout the main treatments and the mechanism treatments.

As described earlier, we design the experiment to give subjects the best possible conditions to counteract the effects of slanted narratives. We also use comprehension questions to ensure that the subjects have all understood the key underlying facts on which we have built the slanted narratives. Despite all the efforts, people may still be persuaded by information known to be partial, potentially due to heuristics as "what you see is all there is" (Enke, 2020; Kahneman, 2011). Our first hypothesis is thus that subjects can nevertheless be swayed by slanted narratives despite full awareness, setting the stage for our stage II. We also hypothesized a similar effect on subjects' incentivized estimates of the benchmark's attitude, aiming to replicate the effect found in Brenner et al. (1996) in the context of hypothetical legal disputes when estimates are not incentivized.

[22] Approximately another 6% of the subjects who continued later dropped out of the study without completing it. Dropout rates are not correlated with treatment assignment.

[23] This does not include the 51 respondents in *Both.Long*, which is implemented merely as an instrumental group and excluded from the analysis, as explained in Section 2.3.3.

Hypothesis 1. *After being exposed to slanted narratives, subjects in Pro are more positive toward the use of GMO mosquitoes than those in Both. Subjects in Both are more positive toward the use of GMO mosquitoes than those in Anti.*

Stage II: the effect of slanted narratives after information acquisition Our second hypothesis relates to the effect of slanted narratives after information acquisition, especially its magnitude compared to that of stage I. According to the standard Bayesian benchmark, given the weak priors in our setting, the acquisition of additional information would dilute the effect of initial exposure to slanted narratives, leading to a convergence of attitudes.[24] Conditional on finding an impact of slanted narratives on attitudes in stage I, we expect the impact to be smaller after people use the opportunity to acquire further information from both sides, given that we have intentionally designed the most favorable conditions for counteracting those biases.

Hypothesis 2. *slanted narratives have a smaller impact on attitudes after information acquisition than before information acquisition.*

Stage II: pattern of information acquisition and evaluation Our third hypothesis concerns the specific patterns of information acquisition (at the extensive margin) and evaluation (at the intensive margin). By examining the relationship between the pattern of argument acquisition and evaluation and the effectiveness of counteracting at the individual level, we hope to shed light on potential channels of how counteracting works.

Hypothesis 3. *Subjects exposed to slanted narratives acquire a higher proportion of confirming arguments and evaluate them more positively compared to those who have access to both narratives.*

Additional hypotheses regarding long-run effects and mechanisms Our mechanism treatments *Long* and *Oppo* allow us to explore the long-term effects and the boundary condition of narratives. In treatment *Long*, we expect more counteracting effects two weeks after the initial exposure, as participants may forget over time. In treatment *Oppo*, if one of the channels associated with the differences between arguments and narratives is at play, we expect participants in *Anti.Oppo* and *Pro.Oppo* to be able to counteract the first-stage effect to a greater extent than in the main treatment. In both cases, however, the extent of counteracting is an open question, and the results are informative in either case.

[24] When the belief space is multi-dimensional, the opposite prediction can also be rationalized by Bayesian updating, i.e., posterior beliefs may appear polarized after information acquisition (e.g., Benoît and Dubra, 2019; Roel and Staab, 2023). This is not the case in our setting. More generally, see Fryer et al. (2019) for a discussion of why Bayesian updating is non-falsifiable in experiments once the experimental setting is embedded in a richer world-view.

Hypothesis 4. *The effect of narratives in Anti.Long and Pro.Long will decay after two weeks.*

Hypothesis 5. *Subjects in Anti.Oppo and Pro.Oppo are able to counteract the first-stage effect of narratives to a greater extent than in the main treatment.*

In addition, in these mechanism treatments, we are still interested in understanding participants' information processing. However, it is difficult to predict whether we shall observe the same level of distortion in information processing as in the main treatments for two reasons. First, as previously hypothesized, biases in information processing in stage II may co-occur with the extent of counteracting. If we expect more counteracting or a decaying effect of narratives on attitudes, then the correlational outcome, bias in information processing, is likely to reduce as well. Second, once we replace eight arguments with one narrative, the number of observations we can use to detect the effect becomes 1/8 of the original. We decided to keep the number of subjects per treatment the same in the second wave of data collection in order to make the results comparable to those in the first wave. This means that we would most likely not have enough power to detect the effect.

3 Counteracting with Arguments

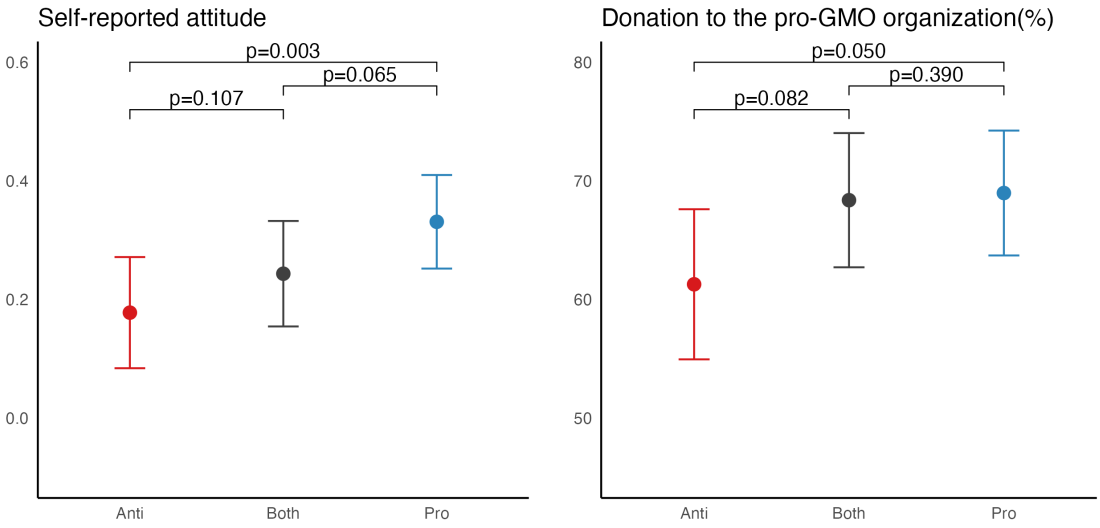
In this section, we start by demonstrating the immediate impact of slanted narratives on subjects' attitudes and whether subjects are indeed aware thereof. Next, we investigate the extent to which subjects are able to counteract such impact by acquiring additional arguments from both sides, focusing on treatment *Endo*. We then explore mechanisms related to information processing that may account for our findings, combining results from treatment *Exo*.

3.1 Do Slanted Narratives Affect Attitudes?

We start by showing the average treatment effect of narratives in stage I. Recall that we implemented two measures of attitudes: (1) the self-reported attitude towards the practice of using GMO mosquitoes to eliminate mosquito-carrying diseases and (2) the allocation of donations between a pro-GMO organization and an anti-GMO organization. Figure 2 presents the average self-reported attitude towards GMO mosquitoes and the average donation to the pro-GMO organization across narratives.

Consistent with Hypothesis 1, attitudes are swayed in the direction of the narrative slant. The left panel of Figure 2 shows that, according to their self-reported attitudes, subjects who read a pro-GMO article are the most positive while those who read an anti-GMO narrative are the most negative toward

Figure 2: Stage I Attitude across Narratives



Notes: This figure presents the self-reported attitude (in the left panel) and donation to the pro-GMO organization (in the right panel) after reading the assigned narratives (95% CI). Self-reported attitude ranges from "taking technology too far" to "appropriate use of technology" and is normalized to [-1,1]. Donation is the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects in treatments *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. The p-values are obtained from one-sided Mann-Whitney U test.

the use of GMO mosquitoes (one-sided Mann-Whitney U test, *Pro* vs. *Anti*: $p = 0.003$; *Pro* vs. *Both*: $p = 0.065$; *Both* vs. *Anti*: $p = 0.107$). The right panel of Figure 2 shows that narratives also affect the allocation of donations in a similar manner. Subjects in *Pro* groups allocate a significantly larger fraction of donation to the pro-GMO organization than those in the *Anti* groups (one-sided Mann-Whitney U test, *Pro* vs. *Anti*: $p = 0.05$; *Pro* vs. *Both*: $p = 0.39$; *Both* vs. *Anti*: $p = 0.082$).

We then estimate the impact of slanted narratives on attitudes using OLS regressions that control for other variables that may also play a role. Table 2 presents the results of regressions with different model specifications. The dependent variables are the two measures of attitudes: self-reported attitude (Columns 1-3) and donation allocation (Columns 4-6). The independent variable $Narrative$ indicates to which narrative(s) individuals are exposed: 1 if the subject reads a pro-GMO article, -1 if the subject reads an anti-GMO article, and 0 if the subject reads both articles. Control variables include prior attitudes, knowledge, and interaction terms to capture heterogeneous effects.

Regression results in Table 2 demonstrate a statistically significant persuasion effect of the slanted narratives on both self-reported attitudes and donation allocations. As a sanity check, prior attitudes were found to be significant predictors of both outcomes, but the impact of narratives was not found to be heterogeneous among people with different prior attitudes, as shown in Column 3 and Column 6. Meanwhile, the impact of narratives appears to vary between groups with different levels of knowledge.

Table 2: Effect of Narratives on Attitudes Before Information Acquisition

	Self-reported attitude			Donation		
	(1)	(2)	(3)	(4)	(5)	(6)
Narrative	0.077^{**}	0.083^{***}	0.096^{***}	3.761^{*}	4.113^{**}	5.933^{***}
	(0.030)	(0.026)	(0.031)	(2.016)	(1.739)	(2.122)
Prior		0.376^{***}	0.385^{***}		23.719^{***}	24.212^{***}
		(0.031)	(0.032)		(2.137)	(2.163)
Knowledge		0.041	0.036		0.431	0.671
		(0.028)	(0.028)		(1.911)	(1.933)
Narrative*Prior			0.060			-1.085
			(0.037)			(2.553)
Narrative*Knowledge			-0.064^{*}			-3.379
			(0.035)			(2.407)
R^2	0.017	0.312	0.322	0.010	0.268	0.273
Adj. R^2	0.015	0.306	0.312	0.007	0.262	0.262
Num. obs.	362	362	362	362	362	362

Notes: This table presents OLS estimates of the effect of narratives on attitudes. The dependent variables are the two measures of attitudes toward GMO mosquitoes: (1) Self-reported attitudes ranging from "taking technology too far" to "appropriate use of technology", standardized to [-1,1], and (2) proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes treatments *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. Standard errors are given in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Those who have a greater understanding of the issue are less susceptible to the influence of narratives, as evidenced by the negative coefficient of the interaction term between knowledge and narrative.

Result 1. *After exposure to slanted narratives, subjects in Pro are more positive toward the use of GMO mosquitoes than those in Both, and than those in Anti.*

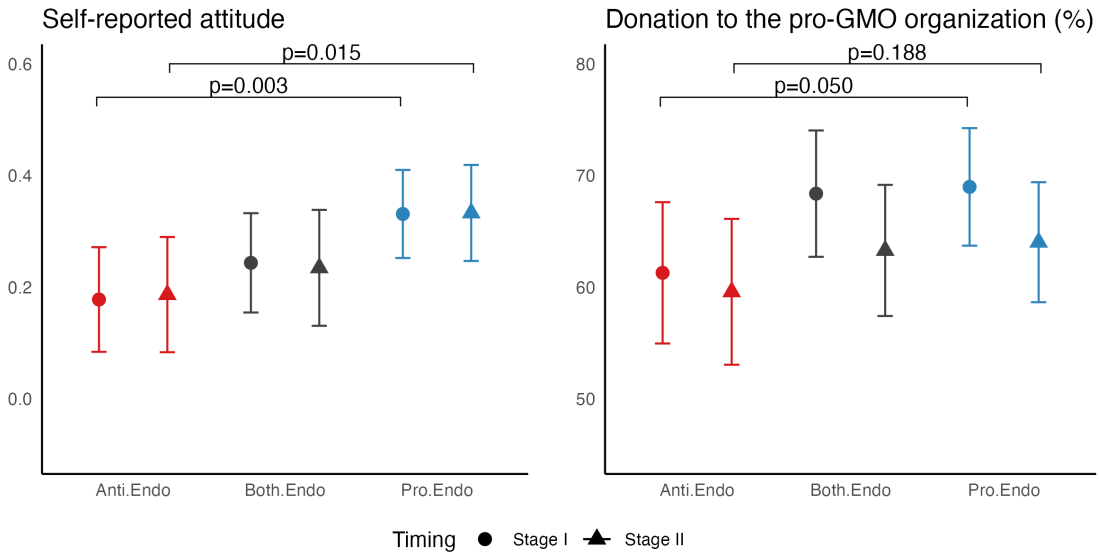
Note that according to our pre-registration (AEARCTR-0005595), we had planned to use the estimate of the benchmark's donation as another outcome of interest, with the aim of replicating previous findings in the literature (Brenner et al., 1996; Lord et al., 1979) that suggest estimates are biased towards the prior. However, we were unable to replicate these early results as our findings showed that narratives do not affect estimates in the first place (see Figure 11 in Appendix E). The discrepancy in results between our study and previous studies could possibly be attributed to different incentive schemes, as we awarded accuracy while previous studies did not. Given that we did not find any first-stage impact of biased narrative on estimates, we deviated from our pre-analysis plan and focused on the other two measures of attitude in the remainder of the paper, reporting results on the estimates only in Appendix E.

3.2 Counteracting by Acquiring Arguments

To see whether information acquisition helps subjects counteract slanted narratives, we now proceed to compare the impact of slanted narratives before and after information acquisition.

First, we present the persistence of the average treatment effect after information acquisition, as shown in Figure 3. The treatment effect after information acquisition is comparable to that before information acquisition. Specifically, subjects who read a pro-GMO narrative still self-report significantly more positive attitudes toward the issue than subjects who read an anti-GMO narrative (one-sided Mann-Whitney U test, $p = 0.015$). Regarding donation allocations, while *Pro.Endo* still donates more to the pro-GMO organization than *Anti.Endo*, the difference is no longer statistically distinguishable (one-sided Mann-Whitney U test, $p = 0.188$).

Figure 3: Stage I and II Attitude across Narratives in *Endo* treatments



Notes: This figure compares the self-reported attitude (in the left panel) and donation to the pro-GMO organization (in the right panel) after reading the assigned narrative(s) and after information acquisition (95% CI). Self-reported attitude ranges from "taking technology too far" to "appropriate use of technology" and is normalized to [-1,1]. Donation is the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects in treatments *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. The p-values are obtained from one-sided Mann-Whitney U test.

Next, we estimate the impact of slanted narratives on attitudes after information acquisition using OLS regressions on stage II attitudes with control variables, as shown in Table 3. Along with the model specifications used in stage I, we add the number of acquired arguments and its interaction terms as additional control variables. Regression results suggest that exposure to slanted narratives remains a significant predictor of self-reported attitudes, although not of donation allocations anymore.

To test Hypothesis 2, we compare the magnitude of the coefficient of $Narrative$ in stage I and stage

Table 3: Effect of narrative on attitude After Information Acquisition

	Self-reported attitude			Donation		
	(1)	(2)	(3)	(4)	(5)	(6)
Narrative	0.105	0.113**	0.128**	1.589	1.988	2.823
	(0.066)	(0.057)	(0.061)	(4.032)	(3.619)	(3.838)
Prior		0.396***	0.403***		21.102***	21.261***
		(0.036)	(0.037)		(2.290)	(2.321)
Knowledge		0.029	0.027		1.571	1.844
		(0.032)	(0.033)		(2.040)	(2.067)
Narrative*Prior			0.038			−2.110
			(0.043)			(2.739)
Narrative*Knowledge			−0.047			−0.948
			(0.041)			(2.574)
Nr. Arguments	0.020*	0.010	0.010	1.445**	0.921	0.927
	(0.011)	(0.010)	(0.010)	(0.678)	(0.611)	(0.612)
Narrative*Nr. Arguments	−0.008	−0.008	−0.009	0.123	0.118	0.161
	(0.014)	(0.012)	(0.012)	(0.828)	(0.743)	(0.748)
R^2	0.022	0.278	0.282	0.016	0.213	0.215
Adj. R^2	0.014	0.268	0.268	0.007	0.202	0.199
Num. obs.	362	362	362	362	362	362

Notes: This table presents OLS estimates of the effect of narratives on attitudes after information acquisition. The dependent variables are the two measures of attitudes toward GMO mosquitoes: (1) Self-reported attitudes ranging from "taking technology too far" to "appropriate use of technology", standardized to [-1,1], and (2) proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes treatments *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. Standard errors are given in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

II regressions (Column 2 and 5 in Table 2 v.s. Column 2 and 5 in Table 3). Our analysis shows that the impact of slanted narratives on attitudes *does not* differ before and after information acquisition. The coefficients of $Narrative$ are similar in both regressions (two-sided Z test, $p = 0.632$ for self-reported attitude and $p = 0.597$ for donations), indicating that the opportunity of acquiring further arguments from both sides only has limited success in counteracting the impact of slanted narratives on attitudes.

Alternatively, we examine Hypothesis 2 by testing whether information acquisition leads to a smaller dispersion of attitudes, which is another indicator of convergence. Using the two-sided variance ratio test, we find that the variances of both self-reported attitude and donation do not decrease, but instead *increase* after information acquisition ($p = 0.025$ for self-reported attitude and $p = 0.314$ for donations). This provides further evidence that, contrary to Hypothesis 2, information acquisition does not dilute the

effect of slanted narratives.

Result 2. *Slanted narratives do not have a smaller effect on attitudes after information acquisition than before information acquisition.*

Consistency An alternative explanation is that subjects may want to appear consistent with their initial responses (Falk and Zimmermann, 2013), for example, as a means of reducing cognitive dissonance (Acharya et al., 2018; Festinger, 1962). This explanation is inconsistent with our finding in Section 3.3 that the variance of attitudes actually increases after information acquisition. In the design section, we have discussed that there appears to be limited evidence to support the concern about consistency in studies with similar design features (e.g., Roth and Wohlfart, 2020) and that our use of multiple measures should mitigate this concern. Nevertheless, the desire to appear consistent may share a common microfoundation with the "induced" confirmation bias channel we identified earlier. Further unpacking of the underlying mechanisms would be interesting for future research.

In summary, our results provide evidence that slanted narratives have a significant and robust effect on subjects' attitudes and donation allocations and that this effect can persist even after information acquisition. These results do not support our Hypothesis 2 about counteracting, and thus raise the question of why the opportunity to acquire additional arguments only has limited success in counteracting the impact of slanted narratives, which we will explore in the next section.

3.3 Why Information Acquisition Doesn't Help Much

In this section, we aim to understand why the opportunity to acquire additional arguments only has limited success in counteracting the impact of biased narratives. We explore four potential explanations in detail. The first possibility is that subjects find slanted narratives unpersuasive and thus do not feel the need to counteract. The second is that subjects may not acquire enough information to challenge their preconceptions. The third is that subjects may evaluate subsequent information in a biased manner, giving more weight to information that confirms their existing beliefs induced by narratives than to information that contradicts them. Finally, we also discuss the potential role of argument quality in explaining our results.

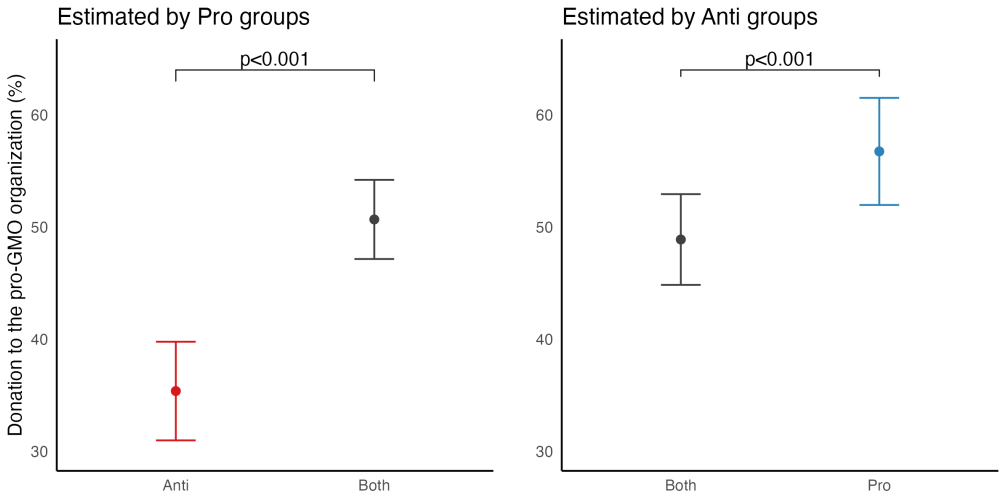
3.3.1 Are slanted narratives perceived to be unpersuasive?

The first possible explanation for the persistent effect of slanted narratives despite the explicit warning flag is that people may find narratives unpersuasive, so there is no need to do anything about

them. We start by examining how subjects perceive the consequences of being exposed to different narratives, i.e., whether and to what extent they anticipate the impact of slanted narratives.

Recall that we ask subjects to estimate the attitudes of their corresponding benchmark group in each set of the treatments. In addition to that, we also elicited subjects' second-order beliefs about the stage I attitude of their opposing side in an incentivized manner. Specifically, subjects in *Pro.Endo* group are asked to estimate the average stage I attitude of those in *Anti.Endo* and vice versa. These two estimates allow us to infer their beliefs about the extent to which slanted narratives affect attitudes. For instance, a subject who reads a pro-GMO narrative is asked to estimate not only the attitude of the benchmark group, that is, those who read both narratives, but also the attitude of those who read a narrative of the opposing side (anti-GMO). Thus, the difference between the two estimates reflects the subject's perception of the gap in attitude caused by the pro-GMO narrative, which equivalently is the impact of the narrative she herself reads. We compare the two estimates for *Pro.Endo* and for *Anti.Endo* respectively in Figure 4.

Figure 4: Evidence on the Awareness of the Impact of Narratives



Notes: This figure presents subjects' estimates of the attitude of their corresponding benchmark group and of group on opposite side (95% CI). The left panel compares the two estimates made by the group that reads a pro-GMO narrative. The right panel compares the two estimates made by the group that reads an anti-GMO narrative. Estimates range from 0 to 100, indicating the proportion of donation allocated to the pro-GMO organization. The sample includes treatments *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. The p-values are obtained from one-sided Wilcoxon signed-rank test.

Figure 4 presents evidence that subjects indeed anticipate an impact of narratives. The left panel of Figure 4 shows that subjects in *Pro.Endo* believe that *Anti.Endo* donates less to the pro-GMO organization than *Both.Endo* does. Reversely, subjects in *Anti.Endo* believe that those in *Pro.Endo* donate more to the pro-GMO organization than *Both.Endo*. The difference between an individual's two estimates is significant at the 0.01 level (Wilcoxon signed-rank test) for both groups.

Taken at face value, these results show that subjects can at least predict the direction of narrative slant effects on attitudes. If anything, subjects *overestimate* how persuasive the narratives are, since the actual treatment effect is much smaller in magnitude.[25] This result thus rules out the possibility that people do not counteract narratives because they find the narratives unpersuasive in the first place.

3.3.2 Information acquisition pattern

The second possible account for the persistent effect of biased narratives is related to the way subjects acquire information. Recall that to encourage subjects to acquire all eight arguments, we reward them for accurately estimating the stage II attitude of the *Benchmark* group, that is, of subjects who have read both narratives and all eight arguments.[26] Despite the financial incentives, subjects may acquire arguments sparingly or selectively because reading arguments costs attention, time, and cognitive efforts.[27] Not acquiring all the available information or acquiring only confirming information could potentially leave intact the impact of slanted narratives on attitudes. To investigate this possibility, we first examine the overall information acquisition pattern. Then we answer the counterfactual question of "what if people read all the arguments" with two additional treatments, *Pro.Exo* and *Anti.Exo*, in which reading all eight arguments is compulsory.

Figure 5 above presents the cumulative distribution of the total number of arguments acquired. Only 25.1% of the subjects choose to read all eight arguments. 68% of them read half or less than half. That is, indeed most subjects do not acquire all the available arguments even with the incentive.

We do not find any evidence that subjects acquire more arguments aligned with the initial narrative—known as confirmation-seeking on the extensive margin (Charness et al., 2021; Frey, 1986; Iyengar and Hahn, 2009; Klayman and Ha, 1987; Pariser, 2011; Prior, 2007; Sears and Freedman, 1967). In treatments *Pro.Endo* and *Anti.Endo*, about 85.6% of the subjects choose equal numbers of arguments from both sides. Only 6% of the subjects are confirmation-seeking in the sense that they acquire more arguments from the side aligned with the assigned narrative, whereas around 8% of the subjects are the opposite. This information acquisition pattern is similar across different sides.

Result 3. *Subjects exposed to slanted narratives do not acquire more confirming arguments than disconfirming arguments.*

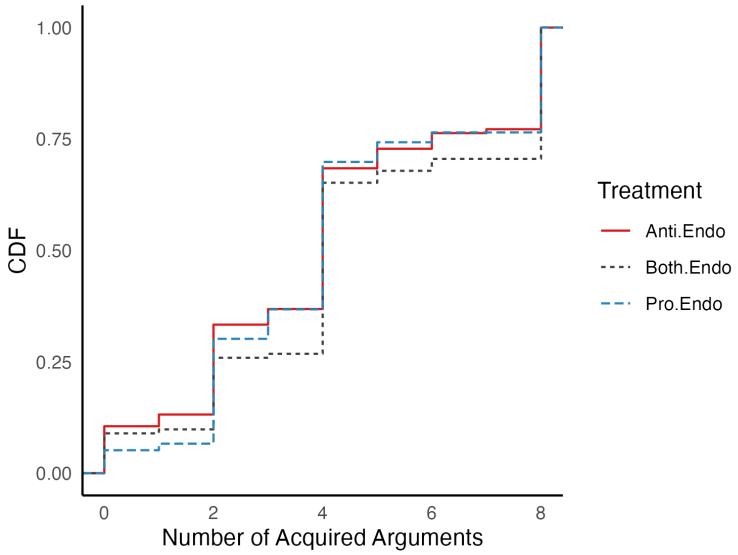
Thus, given the general pattern of information acquisition, it is possible that the failure to exploit the

[25] For example, the difference in donation between treatment *Pro.Endo* and treatment *Both.Endo* is negligible.

[26] As shown in Appendix F, reading all eight arguments improves estimation accuracy (one-sided t-test: $p = 0.064$), suggesting that acquiring all available arguments indeed has instrumental value for subjects to form more accurate beliefs about the benchmark's attitude.

[27] We will discuss the role of rational inattention along other potential channels in Section 6.

Figure 5: Total Number of Arguments Acquired across Narratives



Notes: This figure presents the cumulative distribution of the total number of arguments acquired across treatment groups: *Pro.Endo*, *Both.Endo*, and *Anti.Endo*. Subjects can acquire at least zero and at most eight arguments.

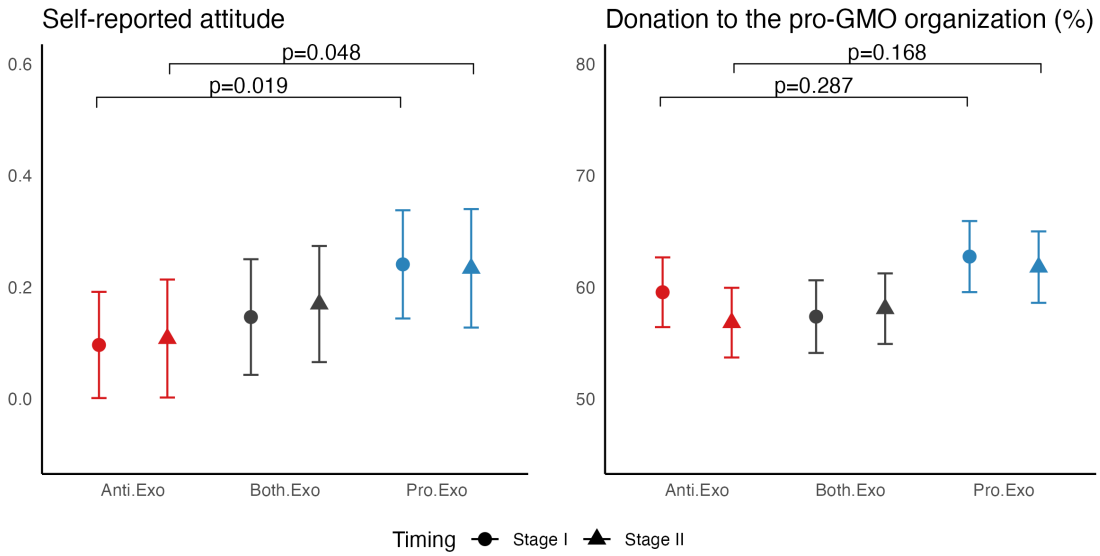
availability of arguments may explain why the effects of slanted narratives persist. We further explore this possibility by examining the additional treatments *Pro.Exo* and *Anti.Exo*, in which subjects are required to read all eight arguments presented in random order.

Effect of exogenously imposed information Similar to what we find in the *Endo* treatments, the impact of slanted narratives in the *Pro.Exo* and *Anti.Exo* treatments do not differ before and after subjects read the arguments.[28] Figure 6 illustrates the before and after attitudes (self-reported attitude and donation allocation) among subjects in *Exo* by the side of the initial narrative(s). Subjects in *Pro.Exo* have a significantly more positive self-reported attitude than those in *Anti.Exo*, and the difference persists after they read all the arguments (one-sided Mann-Whitney U test, before: $p = 0.019$; after: $p = 0.048$). Regarding donation allocation, subjects in *Pro.Exo* donate more to the pro-GMO organization than *Anti.Exo* before and also after reading all the arguments, although the difference is insignificant in both cases (one-sided Mann-Whitney U test, before: $p = 0.287$; after: $p = 0.168$).

To compare the impact of slanted narratives in Stage I with that of Stage II, we again use OLS regressions with controls and find no evidence of a difference. The regression results are reported in Table 4. The exposure to narratives is a significant predictor of subjects' attitudes, both before and after subjects read all the arguments. The magnitude of its impact in Stage I is indistinguishable from that

[28] On average, subjects spend a similar amount of time reading the arguments in the *Exo* treatments as in the *Endo* treatments.

Figure 6: Stage I and II Attitude across Narratives in *Exo* treatments



Notes: This figure compares the self-reported attitude (in the left panel) and donation to the pro-GMO organization (in the right panel) after reading the assigned narrative(s) and after information acquisition (95% CI). Self-reported attitude ranges from "taking technology too far" to "appropriate use of technology" and is normalized to [-1,1]. Donation is the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects in treatments *Pro.Exo*, *Both.Exo*, and *Anti.Exo*. *Both.Exo* is the *Benchmark* group. The p-values are obtained from one-sided Mann-Whitney U test.

in Stage II (two-sized Z test, $p = 0.82$ for self-reported attitude and $p = 0.78$ for donations). Thus, we conclude that the impact of slanted narratives persists even when subjects are required to read all the arguments provided. Using the variance ratio test, the variance in attitudes appears to increase rather than decrease ($p = 0.09$ for self-reported attitude and $p = 0.427$ for donations). These results suggest that attitudes do not converge even after all subjects have read the same set of additional arguments.

In summary, while the incentives did not lead the subjects to acquire all available information, this alone does not fully explain the continued impact of the initial narratives on attitudes. Even if subjects had read all the information provided, we did not observe any evidence that they would have counteracted the impact of slanted narratives. These findings suggest that factors other than insufficient information acquisition may be responsible for the persistent effect of initial narratives on attitudes.

Result 4. *Inadequate information acquisition cannot account for the persistent effect of initial narratives.*

3.3.3 Biased information evaluation

We now turn to information processing as a third potential account for the persistent effect of slanted narratives. Individuals who are initially exposed to a pro-GMO narrative and hence are more positive toward the issue may evaluate pro-GMO arguments more favorably than those initially exposed to an

Table 4: Effect of Narratives on Attitudes Before and After Reading Arguments in *Exo*

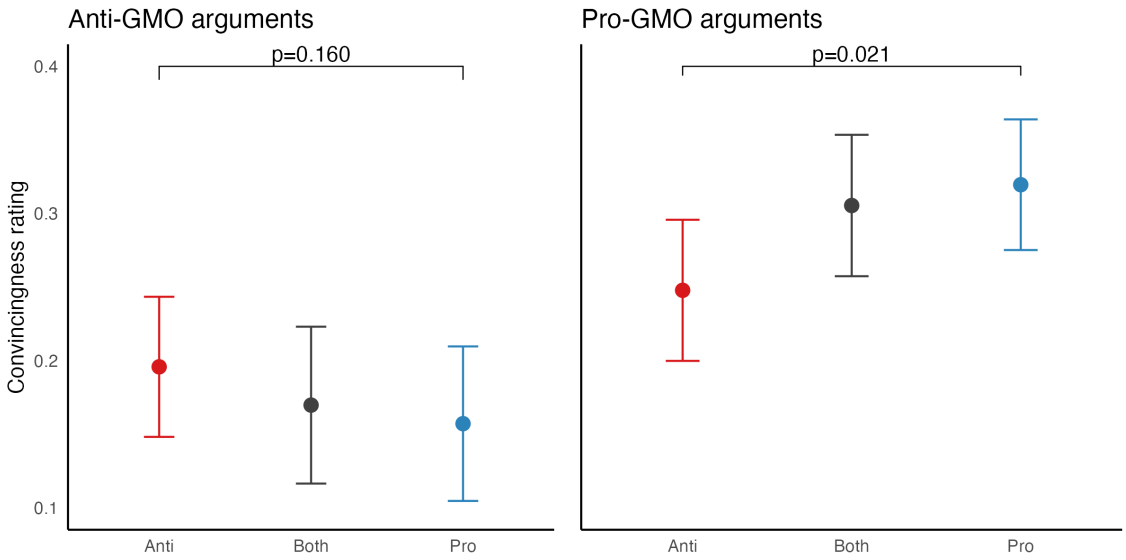
Stage I:	Self-reported attitude			Donation		
	(1)	(2)	(3)	(4)	(5)	(6)
Narrative	0.071**	0.107***	0.163***	1.469	3.638*	6.632***
	(0.035)	(0.029)	(0.036)	(2.267)	(2.013)	(2.469)
Prior		0.406***	0.399***		22.388***	21.872***
		(0.032)	(0.032)		(2.212)	(2.212)
Knowledge		0.048	0.044		−2.273	−2.473
		(0.033)	(0.033)		(2.271)	(2.262)
Narrative*Prior			−0.072*			−5.577**
			(0.041)			(2.800)
Narrative*Knowledge			−0.089**			−4.091
			(0.043)			(2.972)
R^2	0.011	0.330	0.345	0.001	0.227	0.242
Adj. R^2	0.009	0.324	0.336	−0.002	0.221	0.231
Num. obs.	354	354	354	354	354	354
Stage II:	Self-reported attitude			Donation		
	(7)	(8)	(9)	(10)	(11)	(12)
Narrative	0.063*	0.097***	0.133***	2.445	4.432**	6.567***
	(0.038)	(0.033)	(0.040)	(2.244)	(1.965)	(2.413)
Prior		0.393***	0.385***		22.615***	22.047***
		(0.036)	(0.036)		(2.160)	(2.161)
Knowledge		0.052	0.050		2.694	2.608
		(0.037)	(0.037)		(2.217)	(2.210)
Narrative*Prior			−0.088*			−6.280**
			(0.046)			(2.736)
Narrative*Knowledge			−0.040			−2.001
			(0.048)			(2.903)
R^2	0.008	0.271	0.281	0.003	0.250	0.263
Adj. R^2	0.005	0.264	0.270	0.001	0.244	0.253
Num. obs.	354	354	354	354	354	354

Notes: This table presents OLS estimates of the effect of narratives on attitudes before and after reading arguments. The dependent variables are the two measures of attitudes toward GMO mosquitoes: (1) Self-reported attitudes ranging from "taking technology too far" to "appropriate use of technology", standardized to [-1,1], and (2) proportion ([0,100]) of donations allocated to the pro-GMO organization. The first block of the table presents the results before reading arguments and the second block of the table presents the results after reading arguments. The sample includes treatments *Pro.Exo*, *Both.Exo*, and *Anti.Exo*. Standard errors are given in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

anti-GMO narrative. That is, they may exhibit a variant of confirmation bias toward the prior that is *induced* by the randomly assigned narratives. This may lead individuals to reinforce their attitudes induced by slanted narratives rather than challenge them. To test this hypothesis, we first investigate whether subjects display confirmation bias in their ratings of argument convincingness, as captured in Figure 7, and then examine whether this contributes to attitude change after argument acquisition.

We find support for this channel at the aggregate level. Despite the agreement that pro-GMO

Figure 7: Convincingness Rating of Arguments across Narratives



Notes: This figure presents the convincingness rating of anti-GMO arguments (the left panel) and of pro-GMO arguments (the right panel) across treatment groups exposed to different narrative(s) (95% CI). Ratings range from "Extremely unconvincing" to "Extremely convincing" and are normalized to [-1,1]. The sample includes treatment sets *Endo* and *Exo*. The p-values are obtained from one-sided Mann-Whitney U test.

arguments are in general more convincing than anti-GMO arguments, subjects in different treatments disagree on how much more convincing pro-GMO arguments are. They tend to rate arguments aligned with the assigned narrative more favorably than their counterparts. As illustrated in the right panel of Figure 7, *Pro* groups consider pro-GMO arguments more convincing than *Anti* groups do (one-sided Mann-Whitney U test, $p = 0.021$). In contrast, the left panel shows that *Anti* groups judge anti-GMO arguments as more convincing than *Pro* groups do, although the difference is not statistically significant (one-sided Mann-Whitney U test, $p = 0.16$).

To investigate the effect of slanted narratives on evaluation bias at the individual level, we construct a measure of each subject's bias in argument evaluation. This measure, denoted as d_i, captures how much she favors pro-GMO arguments over anti-GMO arguments relative to the average subject. Specifically, for each argument, subject i's rating on it is standardized to the mean rating of all subjects who have read it. Then we take the difference between her average standardized ratings on all pro-GMO arguments that she reads and on all anti-GMO arguments she reads as d_i. The larger d_i, the more subject i is biased toward the pro-GMO arguments compared with the average subject. We regress this evaluation bias on narratives and other control variables using OLS regressions. The results are reported in Table 5.

Regression results provide individual-level evidence of the *induced* confirmation bias. Consistent with the confirmation bias hypothesis, the coefficient of *Narrative* is significantly positive in columns

Table 5: Effect of Narratives on Evaluation Bias

	d_i: evaluation bias			
	(1)	(2)	(3)	(4)
Narrative	0.050^{*}	0.060	0.088^{**}	0.103^{**}
	(0.028)	(0.039)	(0.036)	(0.040)
Endogenous Info		0.067	0.025	0.022
		(0.046)	(0.043)	(0.043)
Prior			0.330^{***}	0.333^{***}
			(0.030)	(0.030)
Knowledge			0.071^{**}	0.069^{**}
			(0.029)	(0.029)
Narrative*Endogenous Info		−0.029	−0.050	−0.054
		(0.056)	(0.051)	(0.051)
Narrative*Prior				0.022
				(0.037)
Narrative*Knowledge				−0.039
				(0.036)
R^2	0.005	0.008	0.177	0.178
Adj. R^2	0.003	0.004	0.171	0.170
Num. obs.	677	677	677	677

Notes: This table presents OLS estimates of the effect of narratives on the bias in evaluating arguments. The dependent variables is a constructed measure of evaluation bias denoted as d_i. The sample includes subjects who have read at least one pro-GMO argument and one anti-GMO argument in treatment sets ***Endo*** and ***Exo***. Standard errors are given in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

1, 3 and 4, indicating that people tend to evaluate pro-GMO arguments more favorably if they were initially exposed to a pro-GMO narrative. This bias is robust when controlling for subjects' home-grown confirmation bias which can be captured by prior attitude and knowledge, as in columns 3 and 4.

Result 5. *Subjects evaluate arguments aligned with the initial narrative more positively compared to those initially exposed to the opposite narrative.*

The next question is whether this *induced* confirmation bias affects attitude after subjects read arguments. If so, naturally, we would expect a positive relationship between evaluation bias and the change in attitude. We estimate

$$\Delta att_i = \gamma_0 + \theta d_i + \sigma Narrative_i + X_i\beta + \epsilon_i \tag{1}$$

where $\Delta att_i = att_i^{post} - att_i^{pre}$ is the change of individual i's self-reported attitude or donation after information acquisition. The control variables include the number of arguments that the subject i reads, her prior attitude, and her knowledge about this issue. The results are presented in Table 6.

Table 6: Effect of Evaluation Bias on Attitudes Change

	Change in self-reported attitude			Change in donation		
	(1)	(2)	(3)	(4)	(5)	(6)
d_i: Evaluation Bias	0.145^{***}	0.147^{***}	0.178^{***}	4.276^{***}	4.298^{***}	5.262^{***}
	(0.020)	(0.020)	(0.022)	(1.298)	(1.296)	(1.416)
Narrative		0.009	-0.006		-3.216	-3.367
		(0.043)	(0.044)		(2.777)	(2.850)
Prior			-0.058^{***}			-2.932^{**}
			(0.018)			(1.199)
Knowledge			-0.015			2.666^{**}
			(0.016)			(1.062)
Nr. Arguments		0.004	0.003		0.879^{**}	0.766^{**}
		(0.005)	(0.005)		(0.347)	(0.346)
Narrative*Prior			-0.022			-1.313
			(0.021)			(1.351)
Narrative*Knowledge			0.043^{**}			2.188
			(0.020)			(1.335)
R^2	0.074	0.077	0.098	0.016	0.028	0.047
Adj. R^2	0.072	0.071	0.087	0.014	0.022	0.036
Num. obs.	677	677	677	677	677	677

Notes: This table presents OLS estimates of the effect of evaluation bias on the change in attitudes. The dependent variables are: (1) Change in self-reported attitudes ranging from "taking technology too far" to "appropriate use of technology", standardized to [-1,1], and (2) change in the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects who have read at least one pro-GMO argument and one anti-GMO argument in treatment sets *Endo* and *Exo*. Standard errors are given in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Regressions show that evaluation bias induced by slanted narratives is a significant predictor of attitude change. The more an individual favors pro-GMO arguments over anti-GMO arguments relative to an average subject, the more likely she shifts her attitude toward the pro-GMO stance and reallocates donations to the pro-GMO organization. Of course, causality could go both ways in our experiment and we do not intend to make causal claims.

In summary, we find that slanted narratives affect not only people's attitudes, but also the way people evaluate subsequent information. Slanted narratives can induce confirmation bias that potentially

reinforces the impact of narratives on attitudes rather than counteracts the impact. This behavioral mechanism provides a plausible explanation for why the impact of narratives persists despite the opportunity to acquire additional information from both sides.

3.3.4 Are arguments unpersuasive?

Finally, another alternative explanation is that the subsequent arguments presented in our experiment may be unpersuasive compared to the initial narratives. We present some suggestive results that at least some arguments are able to influence attitudes to a similar extent as initial narratives and that some arguments are perceived as even more convincing than narratives.

First, we assess the persuasion effect of these arguments on participants' attitudes by exploiting the random variation in argument presentation among subjects who read less than eight arguments. The regression results in Table 9 in Appendix F show that at least one *Pro* and one *Anti* argument significantly shift attitudes in the intended direction with a magnitude comparable to the persuasion effect of the initial narratives (two-sided z-test, *Pro* argument 3 vs. initial narratives: $p = 0.78$; *Anti* argument 2 vs. initial narratives: $p = 0.33$).

Second, we utilize the mechanism treatments *Oppo*, where subjects read and evaluate the narratives instead of arguments in the second stage as described in Section 2.3, to directly compare the perceived convincingness of arguments and narratives. As Figure 12 and Figure 13 in Appendix F show, while perceived convincingness varies between arguments, five of the eight arguments are rated as convincing. In particular, four arguments are perceived to be even more convincing than the initial narratives. Taken together, while arguments vary in quality, it is unlikely that the failure to counteract narratives can be entirely attributed to unpersuasive arguments.

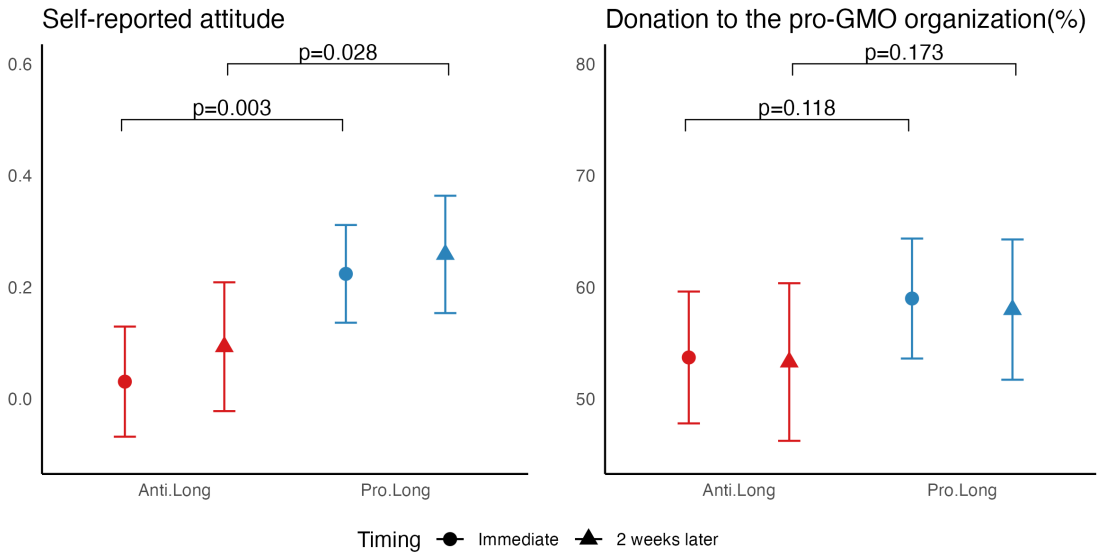
4 Counteracting with Time

In this section, we investigate whether people can counteract the effect of slanted narratives with the help of time, thereby informing the long-run dynamics of slanted narratives. To this end, we focus on the treatment set *Long* and examine the impact of slanted narratives two weeks after the initial exposure.

Figure 8 demonstrates the effect of slanted narratives on attitudes immediately after exposure and two weeks after exposure. Specifically, subjects who read a pro-GMO narrative still self-report significantly more positive attitudes toward the issue than subjects who read an anti-GMO narrative (one-sided Mann-Whitney U test, immediate: $p = 0.003$; two weeks later: $p = 0.028$). *Pro.Long* group also

donates more to the pro-GMO organization than the *Anti.Long* group, although neither the immediate difference nor the long-term difference are statistically significant (one-sided Mann-Whitney U test, immediate: $p = 0.118$; two weeks later: $p = 0.173$). The corresponding regression results can be found in Table 10 in Appendix G.[29]

Figure 8: Stage I and II Attitude across Narratives in *Long* Treatments



Notes: This figure compares the self-reported attitude (in the left panel) and donation to the pro-GMO organization (in the right panel) immediately after reading the assigned narrative(s) and two weeks after reading the narrative(s) (95% CI). Self-reported attitude ranges from "taking technology too far" to "appropriate use of technology" and is normalized to [-1,1]. Donation is the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects in treatments *Pro.Long* and *Anti.Long* screened based on prior and knowledge. The p-values are obtained from one-sided Mann-Whitney U test.

We observe a qualitatively similar but non-significant "induced" confirmation bias in the subjects' evaluation of the arguments two weeks after the initial exposure. *Pro.Long* rates pro-GMO arguments as more convincing than *Anti.Long* does (one-sided Mann-Whitney U test, $p = 0.145$). Conversely, *Anti.Long* considers anti-GMO arguments as more convincing than *Pro.Long* does (one-sided Mann-Whitney U test, $p = 0.200$). These patterns are also reflected in the regressions reported in Table 11 in Appendix G. We also find suggestive evidence that the evaluation bias induced by slanted narratives predicts change in self-reported attitude, as reported in Table 12 in Appendix G.

Our results are consistent with those of other studies that conduct follow-up surveys a few weeks after the initial study to examine persistent effects of information on beliefs and behaviors. For example, in one of these pioneering works by Kuziemko et al. (2015), 60% of the initial effect of information about

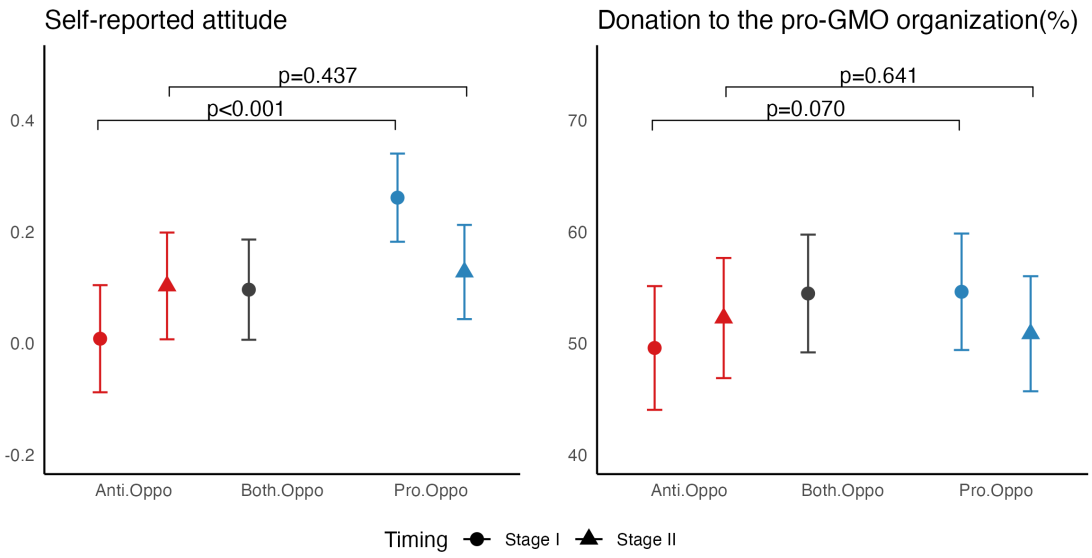
[29] It is worth noting that we conduct our analysis on the same "as-if" screened subsample to ensure comparability with the results in the main treatments. The inclusion of extreme priors does not weaken the first stage, but reduces the effect of narrative persuasion after argument acquisition. This finding reinforces our design objective of minimizing the role of prior opinions and adds important quantifiers to our results, that is, our results are most applicable in situations where subjects have modest priors and are uncertain—effectively, when people are not yet polarized.

income distribution on attitudes toward redistribution persists in the follow-up survey, while this number in our case is 52%. Nevertheless, we would like to emphasize that these results should be treated with caution, as the reduced effective sample size (due to the removal of the screening process), combined with the requirement for recontact in long-term treatments, inevitably limits the statistical power of the tests.

5 Counteracting with Opposing Narratives

In this section, we investigate whether people can counteract the effect of slanted narratives with exact opposing narratives, focusing on the treatment set *Oppo*. The results in this section allow us to explicitly investigate other potential channels related to the differences between arguments and narratives inherent to our design as a whole. Note that real-life situations in which the opposing narrative is constructed as the exact opposite and forcefully juxtaposed to individuals, resembling a point-to-point opposition scenario, are extremely rare. Therefore, this set of treatments also represents almost the best possible theoretical condition for counteracting the influence of narratives.

Figure 9: Stage I and II Attitude across Narratives in *Oppo* treatments



Notes: This figure compares the self-reported attitude (in the left panel) and donation to the pro-GMO organization (in the right panel) in stage I and in stage II (95% CI). Self-reported attitude ranges from "taking technology too far" to "appropriate use of technology" and is normalized to [-1,1]. Donation is the proportion ([0,100]) of donations allocated to the pro-GMO organization. The sample includes subjects in treatments *Pro.Oppo*, *Both.Oppo*, and *Anti.Oppo* screened based on prior and knowledge. The p-values are obtained from one-sided Mann-Whitney U test.

Figure 9 presents the stage I and stage II attitudes of the subjects in treatment *Oppo* by the side of the initial narrative(s). We again replicate the very strong first stage: subjects in *Pro.Oppo* self-report to be significantly more positive than those in *Anti.Oppo* (one-sided Mann-Whitney U test, $p < 0.001$). However,

the difference is attenuated after reading the other opposing narrative (one-sided Mann-Whitney U test, $p = 0.437$). We find a marginally significant difference between *Pro.Oppo* and *Anti.Oppo* for donations in stage I (one-sided Mann-Whitney U test, $p = 0.07$) and again the difference is mitigated in stage II: one-sided Mann-Whitney U test, $p = 0.641$).[30] Regressions are reported in Table 13 in Appendix G.[31]

The above results show that once subjects are presented with the direct counterpart of a slanted narrative, they can fully counteract its effects. There are two pieces of good news here. First, as explained earlier, this result provides empirical support for our design choice of using *Both* as the narrative-free benchmark group, since people appear to be unaffected by the order of presentation or the active elicitation of attitudes as long as they are exposed to both narratives. Second, under the best possible theoretical condition, where opposing narratives are by construction point-to-point opposition of each other, counteracting seems possible. This may suggest that precise and targeted confrontation is essential to success, echoing recent findings by Costello et al. (2024) that personalized AI-designed dialogues are lastingly effective in reducing conspiracy beliefs.

The bad news is that most people are *unlikely* to actively seek out the effective opposing narratives. Recall that we asked subjects what kind of information they would seek to maintain an unbiased perspective: only 26% of subjects choose comprehensive articles from the opposing side, which mimics the *Oppo* treatment. In contrast, about 40% prefer comprehensive articles from their own side, and 27% prefer balanced arguments. The likelihood that people will be exposed to the exact information needed for counteracting becomes even smaller once we factor in how algorithms control people's news feeds.

6 Discussion

In this section, we draw connections to a broader literature and discuss exploratory findings using additional variables added in the second wave of data collection, to shed light on other possible channels that could potentially contribute to the persistent effect of slanted narratives. Finally, we put together evidence collected across treatments and discuss what may or may not help counteract the effects of slanted narratives based on our findings, as well as the generalizability of our results.

[30] Both the stage I and stage II results are robust to using the full sample that includes those with strong prior and knowledge.

[31] We find no difference in the evaluation of the convincingness of the opposing narratives depending on the initial exposure. This is consistent with the possibility that biases in information processing in stage II may co-occur with the extent of counteracting. Also, due to the power issue discussed earlier, where the number of observations we have for this result is 1/8 of the original, we cannot rule out the possibility that the tests here are underpowered.

6.1 Alternative Explanations Connected with Other Literature

Below we assess the role of the experimenter demand effect, group identity, stakes, and rational inattention by examining heterogeneous treatment effects across two subgroups divided by a median split of the corresponding variables introduced in Section 2.4. All regression results mentioned in this section are reported in Table 14 and Table 15 in Appendix G. In addition, we briefly discuss how our results are related to the "order effect", such as primacy effect, in the literature.

Experimenter demand effect Since we explicitly inform subjects of the narrative slant as well as the random assignment, one concern is that our results may be driven by the experiment demand effect (EDE). Although it is not obvious *ex ante* in which stage the experimenters are looking for an effect or in which direction the effect is expected, we explicitly address this possibility in three steps and indeed find a limited empirical role of EDE.

First, we include an open-ended question about the purpose of the study at the end of the experiment in the second wave of the experiment to get a general idea of subjects' beliefs about the experimenters' demand. We manually code the subjects' responses to the open-ended question about the purpose of our study and find that very few subjects (2.6% in *Long* and 1.1% in *Oppo*) correctly guess our focus on the change in attitude after Stage II. Only about 36.4% of the responses in *Oppo* and 30% of the responses in *Long* are more or less related to the *stage I effect*, that is, how narratives affect people's opinions. Other answers are way off target. Thus, the experimenter demand effect, if it plays a role at all, likely contributes only to the stage I result, not to our main results on counteracting and mechanisms.

Second, when we classify subjects using a binary indicator of whether their guess about the stage I effect is close enough, we find no heterogeneous effects among the two groups of subjects. This shows that even the narrative persuasion result in stage I is unlikely to be entirely driven by EDE. In addition, this heterogeneity does not show up in the extent to which they counteract the effect in stage II either, suggesting a limited role for EDE in driving our main results regarding the counteracting effect of narratives and its mechanisms in stage II.

Third, we also measure the extent to which subjects are willing to change their behavior just to accommodate or please the experimenters, as inspired by Tsutsui and Zizzo (2014). As described in Section 2.4, subjects are presented with a price list comparing two bonus options, one of which is increasingly dominated by the other, and the dominated options are linked by the statement "it would be nice if some of you would choose this option". We use the number of dominated options chosen as a measure of sensitivity to the experimenter's demand. We find no heterogeneous treatment effects with

respect to sensitivity to EDE in either stage I or stage II.

Induced group identity A second explanation of our results is that subjects are not impacted by narratives per se but by the fact that they are assigned to the *pro* and *anti* groups, as our design is reminiscent of a minimal group paradigm. Such minimal grouping potentially creates a group identity, and in turn in-group bias in donation allocation and information processing. [32]

To see if our result is an artifact of in-group bias, we include a measure of minimal groupiness using the standard other-other allocation task and categorize subjects into "groupy" and "non-groupy" types (Chen and Li, 2009; Kranton et al., 2020; Kranton and Sanders, 2017). We find no heterogeneous treatment effects with regard to groupiness in stage I or stage II. In addition, no one mentions the group nature when guessing the purpose of the study in the open-ended question. Thus, our results are likely not driven by the group identity.

Rational inattention Another potential explanation for the failure to counteract is rational inattention (Gabaix, 2019; Maćkowiak et al., 2023). Properly processing new information to counteract the initial impact of narratives requires attention, cognitive effort, and time. A central prediction of rational inattention is that attentive information processing is negatively correlated with processing costs.[33]

In the post-survey of wave 2, we include a standard measure of cognitive ability (Frederick, 2005) and a self-reported measure of memory as a proxy of cognitive capacity. In addition, we asked subjects to report their hourly wage as a measure of the opportunity cost of participating in the experiment. All of these measures approximate the concept of information processing costs: those with higher cognitive ability or self-reported memory, as well as those with lower opportunity costs, may have lower information processing costs and thus be better able to process Stage II information and thus better counteract the influence of narratives.

We find no significant heterogeneous treatment effects with respect to cognitive ability or memory in stage II, although the sign of the coefficient is consistent with the prediction. We find a marginally significant heterogeneous treatment effect with respect to opportunity cost in the opposite direction. Note that our design is not tailored to test the predictions of rational inattention, and information processing costs may not be best captured by any of our measures; these results should only be taken as suggestive evidence that rational inattention plays a limited role.

[32] Recently, Bauer et al. (2023) find that in a belief updating context, people tend to place greater weight on ingroup information and that this observation is driven by those who are more "groupy".

[33] For instance, Fuster et al. (2022) find that low-numeracy individuals are less responsive to the information provided.

Stakes and Incentives Related to the rational inattention account, failure to counteract may also occur because the stakes are perceived to be too low. The stakes could be intrinsic, for example, if subjects care about being unbiased on the issue of GMOs. The stakes may be extrinsic, for example, if we reward accurate estimates of the "unbiased" benchmark's attitudes.

We ask subjects to what extent they strive to maintain an unbiased perspective on the issue of GMO technology in the post-survey. Most subjects (79.1%) are at least moderately committed to an unbiased perspective. Restricting our attention to these subjects in treatments *Long* who intrinsically care to be unbiased, we still find little evidence of counteracting.

Recall that in wave 2 of the experiment, we increase the extrinsic stake size and incentives for robustness. Specifically, we double the total amount of donation from 100 dollars to 200 dollars and double the reward for correct estimate from 0.5 dollar to 1 dollar. As shown in Section 5, the effect of narratives on donation allocation in *Long* is similar in magnitude to the main experiment, suggesting that an increase in stake size and incentives does not automatically prompt more counteracting, consistent with the stylized findings in the literature (e.g., Enke et al., 2023).

Order effect What seems relevant to our findings is the "order effect". Literature commonly refers to the order effect as the phenomenon that behavior is affected by the order of the tasks or information presented. The order effect may stem from the primacy effect (e.g., Peterson and DuCharme, 1967), which refers to "the tendency for facts, impressions, or items that are presented first to be better learned or remembered than material presented later in the sequence" (APA Dictionary).

We could explicitly test the presence of an order effect using treatment groups *Both*, in which subjects receive identical information, i.e. two narratives, in different orders in stage I before attitude elicitation. We find no evidence for the order effect with regard to either self-reported attitude or donation, in either wave 1 or wave 2 of the experiment.

Note that in our setting where only two pieces of information are presented, we cannot rule out the possibility that the primacy effect still plays a role, but is offset by a "recency effect," the opposite of the primacy effect.[34] In this regard, our results from the *Oppo* treatment suggest that the way subsequent information connects with previous information may be relevant in explaining the presence or the absence of the order effect (e.g., DuCharme, 1970), which is also related to the recent work where signals presented as retractions are underweighted (Gonçalves et al., 2024).

[34] In a very different setting with belief updating tasks, Coutts (2019) find that people remember the most recent signals better.

6.2 What helps counteracting and what doesn't?

First, simply attaching a warning flag or raising awareness is certainly not enough to safeguard people from being swayed by slanted narratives. Despite that all subjects are explicitly informed of the random assignment of narratives and that they fully anticipate the impact thereof, these narratives still create a visible, robust, and persistent divide between groups that initially hold similar views. Such a divide does not stem from the need to cater to the experimenter's demand or a randomly induced group identity, nor does it diminish with higher stakes, higher cognitive ability, or lower opportunity cost.

Second, providing cross-cutting information may help, depending on the exact information. More balanced arguments from both sides narrow the gap in attitude to a limited extent, while the exact counter-narrative constructed around the exact same set of facts closes the divide. While the latter seems promising, only 26% of subjects would choose comprehensive articles from the opposing side when they seek to maintain an unbiased perspectives. An open question is what features of the subsequent information make counteracting successful, which is both interesting and challenging given the rapid evolution of AI-powered information dissemination tools.

Finally, pre-existing priors and knowledge seem to "help" in the longer term. For those who do not start out with strong priors or knowledge, slanted narratives induce an opinion that persists for at least two weeks. Once we include people who already take sides, the effect of initial narratives tends to diminish over time. This is consistent with the stylized findings that it is difficult to change people's existing opinions in general. Taking this set of results at face value, it is perhaps more difficult to make normative judgments about the desirability of the persistence of narrative persuasion. On the one hand, strong priors make people immune to narrative persuasion or even manipulation; on the other hand, counteracting effort is considered important precisely because we typically regard rigid opinions or polarization as potentially harmful. We leave this challenging question for future work.

7 Conclusion

The optimists would like to believe that awareness prepares people for biased information in this information-rich world. Across two waves of data collection with eleven treatments, we find significant effects of narrative persuasion despite the fact that participants are explicitly made aware of the use of narratives and their random assignment.

In settings where careful design efforts are made to ensure that counteracting has the best chance, we find that such effects of narrative persuasion are difficult to counteract with additional balanced

arguments. This is potentially because the initial random exposure to narratives distorts the way people process subsequent information, which we refer to as *induced* confirmation bias. These effects qualitatively persist for at least two weeks after the initial exposure. Only when people are contrasted with the exact opposing narratives are they able to fully counteract the impact. These results highlight the importance of balancing and complete information provision in the first place.

Our results thus have several important implications for the design of debiasing interventions in other settings where counteracting is even harder, such as mass media or political campaigns. A potential lesson for regulating the information market in the public sphere is that preventing polarization is perhaps more effective than combating polarization that has already formed. Given that our experiment is conducted in a setting free from motivated beliefs, another implication is that susceptibility to slanted narratives as well as the induced confirmation bias could arise from fundamental errors in cognition or reasoning, independent of motivation.

While it seems straightforward to construct two versions of narratives that counteract each other in a controlled experiment, juxtaposing point-to-point information slant may be uncommon in the real world. Our data also suggest that less than one-third of people who strive to be "unbiased" would actively seek out this type of information. Thus, more work is needed to develop effective strategies for counteracting slanted narratives when complete exposure is not feasible. This could include further research into the qualitative features of arguments that make them less effective than the exact counter-narrative[35], and exploration of a combination of approaches, such as fact-checking, media literacy education, or other initiatives aimed at reducing cognitive errors.

[35]For example, a related qualitative difference also appears in recent work comparing whether people remember narratives or statistics better (Graeber et al., 2024).

References

Acharya, A., Blackwell, M., and Sen, M. (2018). Explaining preferences from behavior: A cognitive dissonance approach. *The Journal of Politics*, 80(2):400–411.

Alesina, A., Miano, A., and Stantcheva, S. (2023). Immigration and redistribution. *The Review of Economic Studies*, 90(1):1–39.

Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2024). Narratives about the macroeconomy. *SAFE Working Paper*.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Barron, K. and Fries, T. (2024). Narrative persuasion. *WZB Discussion Paper*.

Barron, K., Harmgart, H., Huck, S., Schneider, S. O., and Sutter, M. (2021). Discrimination, narratives and family history: An experiment with jordanian host and syrian refugee children. *Review of Economics and Statistics*, pages 1–34.

Bauer, K., Chen, Y., Hett, F., and Kosfeld, M. (2023). Group identity and belief formation: a decomposition of political polarization. *SAFE Working Paper*.

Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, imperatives, and moral reasoning. Technical report, National Bureau of Economic Research.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186.

Benoît, J.-P. and Dubra, J. (2019). Apparent bias: What does attitude polarization show? *International Economic Review*, 60(4):1675–1703.

Brenner, L. A., Koehler, D. J., and Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making*, 9(1):59–70.

Broockman, D. and Kalla, J. (2022). The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with fox news viewers. *OSF Preprints*, 1:1–42.

Bursztyn, L., Rao, A., Roth, C., and Yanagizawa-Drott, D. (2023). Opinions as facts. *The Review of Economic Studies*, 90(4):1832–1864.

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., and Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546.

Charness, G. and Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23.

Charness, G., Oprea, R., and Yuksel, S. (2021). How do people choose between biased information sources? evidence from a laboratory experiment. *Journal of the European Economic Association*, 19(3):1656–1691.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review*, 109(6):2294–2332.

Chiang, C.-F. and Knight, B. (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9).

Coppock, A., Ekins, E., Kirby, D., et al. (2018). The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science*, 13(1):59–87.

Costello, T. H., Pennycook, G., and Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814.

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

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: empirical evidence. *Annual Review of Economics*, 2(1):643–669.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Di Tella, R., Gálvez, R. H., and Schargrodsky, E. (2021). Does social media cause polarization? evidence from access to twitter echo chambers during the 2019 argentine presidential debate. Technical report, National Bureau of Economic Research.

Djourelova, M. (2023). Persuasion through slanted language: Evidence from the media coverage of immigration. *American Economic Review*, 113(3):800–835.

Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology*, 22(1):91–101.

DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, 85(1):66.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Eliaz, K. and Spiegler, R. (2020). A model of competing narratives. *American Economic Review*, 110(12):3786–3816.

Enke, B. (2020). What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., and Van De Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics and Statistics*, 105(4):818–832.

Falk, A. and Zimmermann, F. (2013). A taste for consistency and survey response behavior. *CESifo Economic Studies*, 59(1):181–193.

Festinger, L. (1962). *A theory of cognitive dissonance*, volume 2. Stanford university press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19:41–80.

Fryer, R. G., Harms, P., and Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 17(5):1470–1501.

Fuster, A., Perez-Truglia, R., Wiederholt, M., and Zafar, B. (2022). Expectations with endogenous information acquisition: An experimental investigation. *Review of Economics and Statistics*, 104(5):1059–1078.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1*, volume 2, pages 261–343. Elsevier.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gonçalves, D., Libgober, J., and Willis, J. (2024). Retractions: Updating from complex information.

Graeber, T., Roth, C., and Zimmermann, F. (2024). Stories, statistics, and memory. *The Quarterly Journal of Economics*, pages 2181–2225.

Haaland, I., Roth, C., and Wohlfart, J. (2023). Designing information provision experiments. *Journal of Economic Literature*, 61(1):3–40.

Harari, Y. N. (2015). Sapiens: A brief history of humankind. *New York: Harper Collins*.

Harrs, S., Müller, L. M., and Rockenbach, B. (2021). How optimistic and pessimistic narratives about covid-19 impact economic behavior. *ECONtribute Discussion Paper*.

Herman, D. (2011). *Basic elements of narrative*. John Wiley & Sons.

Hillenbrand, A. and Verrina, E. (2022). The asymmetric effect of narratives on prosocial behavior. *Games and Economic Behavior*, 135:241–270.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39.

Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Jehiel, P. and Steiner, J. (2020). Selective sampling with information-storage constraints. *The Economic Journal*, 130(630):1753–1781.

Jones, M. and Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1):59–99.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kendall, C. W. and Charles, C. (2024). Causal narratives. Technical report, National Bureau of Economic Research.

Klayman, J. and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211.

Kranton, R., Pease, M., Sanders, S., and Huettel, S. (2020). Deconstructing bias in social preferences reveals groupy and not-groupy behavior. *Proceedings of the National Academy of Sciences*, 117(35):21185–21193.

Kranton, R. E. and Sanders, S. G. (2017). Groupy versus non-groupy social preferences: Personality, region, and political party. *American Economic Review*, 107(5):65–69.

Kuziemko, I., Norton, M. I., Saez, E., and Stantcheva, S. (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098.

Maćkowiak, B., Matějka, F., and Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273.

Martin, G. J. and Yurukoglu, A. (2017). Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–99.

Morag, D. and Loewenstein, G. (2024). Narratives and valuations. *Management Science*.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *American Economic Review*, 95(4):1031–1053.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Ochs, E., Capps, L., et al. (2009). *Living narrative: Creating lives in everyday storytelling*. Harvard University Press.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Pennycook, G. and Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

Peterson, C. R. and DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, 73(1):61.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Roel, M. and Staab, M. (2023). The benefits of being misinformed: Information moderation under misperception.

Roth, C. and Wohlfart, J. (2020). How do expectations about the macroeconomy affect personal expectations and behavior? *Review of Economics and Statistics*, 102(4):731–748.

Schwardmann, P., Tripodi, E., and Van der Weele, J. J. (2022). Self-persuasion: Evidence from field experiments at international debating competitions. *American Economic Review*, 112(4):1118–1146.

Schwartzstein, J. and Sunderam, A. (2021). Using models to persuade. *American Economic Review*, 111(1):276–323.

Sears, D. O. and Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31(2):194–213.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

Shiller, R. J. (2020). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press.

Sunstein, C. R. (2017). # *republic*. Princeton University Press.

Thaler, M. (2023). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *American Economic Journal: Microeconomics*.

Tsutsui, K. and Zizzo, D. J. (2014). Group status, minorities and trust. *Experimental Economics*, 17:215–244.

Walter, N. and Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication research*, 47(2):155–177.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13:75–98.