

Tailored Stories

Chiara Aina*

October 1, 2022

Job Marker Paper

Latest version available [here](#)

Abstract

Is it possible to persuade others only by providing interpretations of future events? I study the problem of persuading a boundedly rational agent without tampering with, or even knowing, the piece of information they observe. Persuasion is achieved by providing models, whereby the persuader can communicate ways of interpreting observable signals. The key assumption is that the agent adopts the model that best fits what is observed given her initial beliefs, and takes the action that maximizes their expected utility under the adopted model. Anticipating this, a persuader can strategically communicate models to manipulate how the receiver interprets the different realizations. I characterize the extent of belief manipulability in this setting and I show that the agent may hold inconsistent beliefs across signal realizations — posterior beliefs across signal realizations do not average to her prior — because different realizations might trigger the adoption of conflicting models. While persuasion can mislead the agent, its impact is bounded and depends on initial beliefs. I discuss this framework in the context of political polarization, conflict of interests in finance, lobbying, and self-persuasion. In particular, I show that polarization is inevitable if agents with sufficiently different initial beliefs are exposed to the same conflicting models.

Keywords: Persuasion, Narratives, Polarization.

JEL classification: D82, D83, D9.

*University of Zurich. Blümlisalpstrasse 10, 8006 Zürich, Switzerland; e-mail: chiara.aina@econ.uzh.ch. I am grateful to my advisors Nick Netzer and Jakub Steiner for their precious support, and to Joshua Schwartzstein for his invaluable guidance. For very helpful discussion and suggestions, I also thank Sandro Ambühl, Ian Ball, Kai Barron, Pierpaolo Battigalli, Roberto Corrao, Fabio Maccheroni, Delong Meng, Andrei Shleifer, Tomasz Strzalecki, Adi Sunderam, Roberto Weber, Jeffrey Yang, as well as seminar and conference participants at Ca' Foscari University, Carnegie Mellon University, CREED, Harvard, MIT, QMUL, UEA, ECBE, ESEM, IAREP/SABE, NASMES, and SITE.

1 Introduction

Beliefs are shaped by how we interpret the world. When we use different interpretations to make sense of the same fact, we might reach contrasting conclusions. Voters do not agree on the outcome of the elections. Consumers differ in how they evaluate companies based on the same public initiatives. Investors make different predictions based on the same past data. This occurs even when we share the same preferences and initial beliefs. An explanation is that we adopt different narratives to interpret the same event. Narratives link what we observe to what we want to understand: they provide interpretations of events.¹ Thus, controlling the narratives people adopt can be a powerful tool to manipulate and persuade them. Indeed, when making sense of the observed facts, one might rely on narratives provided by more knowledgeable sources, such as political figures, financial advisors, or experts considered trustworthy. This type of persuasion can be even more powerful than persuasive messages or new pieces of factual information because it gives a leeway to change people’s beliefs without tampering with, or even knowing, what they observe.

Is it possible to persuade others only by providing interpretations of future events? I study the general problem of persuading a boundedly rational agent by controlling her interpretation of signals she is about to receive. Consider an agent (the receiver, she) who, after observing an informative fact about the relevant payoff state, takes an action affecting both her payoff and also the one of the persuader (the sender, he). The additional information on the unknown state, a signal, is generated stochastically by a fixed process. Importantly, the sender cannot influence the signal or the process generating it, but he has the chance to provide the receiver with one or multiple ways of interpreting signals. Following Schwartzstein and Sunderam (2021), these interpretations — what I informally refer to as narratives or stories — are formalized as likelihood functions assigning to each state a distribution of signals conditional on that state, called *models*. Persuasion arises because the receiver adopts the interpretation that she finds more plausible given her prior beliefs. This is formalized by adopting the narrative that fits best the realized signal: the receiver selects the model that maximizes the likelihood of the observed data, given her prior. Anticipating this, the persuader strategically communicates models to manipulate how the receiver interprets the observed signal if this benefits him. Muddying the waters with multiple models, the persuader is able to corrupt the receiver’s reasoning, so that she would adopt different models depending on the observed fact, resulting in incoherent beliefs across signals.

Consider the example of a politician trying to get a (representative) voter’s support regardless of the reported election outcome. The voter will recognize the politician as president only if she strongly believes him to be the legitimate winner once she observes the reported election outcome. Before the election, the politician communicates to the voter narratives about the election system. Assume that the politician communicates only the narrative according to which the voting system is fair. Since there is only one story available the voter always adopts that. Then, once the election outcome is revealed, the voter would support the politician if the latter

¹Cambridge dictionary defines a narrative as “a particular way of explaining or understanding events”. In spite of the growing attention to this topic in economics, there is not yet a commonly shared definition of what a narrative is. Different ways of formalizing it have emerged in recent years and I discuss the main ones in the related literature. Barron and Fries (2022) provides a detailed discussion of the current conceptualization of narratives in economics in their appendix.

is the reported as winner, while she would not support him otherwise. In this case, the voter has consistent beliefs across election outcomes. However, this might not be the case if at the same time the politician also promotes a conspiracy theory according to which elections are rigged.² Being exposed to multiple stories allows inconsistent reasoning to take root: each election outcome triggers the adoption of a different narrative. The voter’s initial beliefs play a crucial role because they drive which narratives are adopted and following which outcome. Assume that the voter expects the politician to fairly win the election. If the politician is reported as winner, the most compelling story is the one about the just voting system; however, if the politician is not reported as winner, the conspiracy theory resonates best with the voter. This is equivalent to hold an inconsistent interpretation across election outcomes: “if reported as winner, the election system is fair; otherwise, elections are rigged.”³ As a result, the voter is updating upwards her beliefs about the politician being the legitimate winner, supporting him, regardless of the election outcome. The politician achieved this by leveraging how the voter makes sense of the reported election outcome, and he exploited it by providing narratives that benefit his own interests across all possible contingencies.

This paper shows the extent to which the sender can manipulate the receiver using narratives communicated before the release of new relevant information. To study this, it is necessary to keep track of which beliefs the receiver holds conditional on every observable signal. Therefore, the main object of the analysis is an array of the receiver’s posterior beliefs conditional on each signal realization, called the *vector of posterior beliefs*. In the context of the previous example, this means describing the voter’s beliefs conditional on both the election outcomes: when the politician is reported as the winner of the election, and when the politician is not reported as the winner of the election. The main result characterizes the maximal belief manipulability of the receiver, which the sender can attain by providing as many models as the signal realizations. This result conveys two main insights. First, if exposed to multiple models, the receiver may hold inconsistent beliefs across signal realizations, that is, the prior cannot be expressed as a convex combination of the posteriors across signals. This is appealing to a sender that wants to move the receiver’s beliefs always in the same direction. To do so, he should construct a set of tailored models so that each model is adopted conditional on the signal it has been tailored to, inducing the desired posterior conditional on that signal. Because models compete with each other across signal realizations, it is not always possible to construct a such set of models and induce all desired beliefs. The intuition is the following. Any model that leads beliefs to react a lot given a signal cannot maximize the likelihood of observing that signal very well, making it harder to be selected conditional on that signal. Therefore, the overall posteriors across signals cannot be too surprising with respect to the prior in all possible contingencies. Indeed, the second insight is that generally not all vectors of posteriors can be induced — making the result empirically testable. To better convey intuitions behind these formal results, I introduce

²For simplicity of exposition, I describe these inconsistent stories as provided by a single agent. Alternatively, one could think about this as a coordinated strategy implemented by different agents. The receiver might be less sensitive to this type of contradiction, and the credibility of the sources would be less likely to be questioned.

³The following are examples of other domains in which agents might hold inconsistent interpretations, as a result of selecting different interpretations conditional on different facts. While interpreting a grade at school, a student that believes to be competent in a subject might hold the following: “if it’s a good grade, it must be very informative about ability; if it’s a bad grade, it does not convey much information.” When learning about the new Covid-19 vaccine, somebody skeptical about vaccines might think: “if clinical trials report the vaccine as safe, tests were conducted in a hurry; if clinical trials report the vaccine as unsafe, tests were conducted properly.”

a graphical approach to this setting, restricting the attention to the special case of binary signal and state (hereafter, binary case). This includes also a graphical construction of which vectors of posteriors the receiver can hold.

Having explored the limits of belief manipulability, I turn to the question of what makes the receiver more vulnerable to persuasion. The answer points to the crucial role of initial beliefs. In the binary case, the sets of feasible vectors of posteriors can be ordered based on the prior: the closer the receiver’s prior is to the uniform distribution, the more she can be manipulated. In particular, the receiver is fully persuadable — she can hold any beliefs regardless of what she observes — when her initial beliefs are 50-50. More in general, I provide sufficient conditions for full manipulability. The sender has more leeway to manipulate if there are many signals to be interpreted and few states on which the receiver has dispersed prior on. Indeed, if the signals are at least as many as states, a receiver with uniform prior can believe anything.

All the information the sender needs to know what the receiver can be persuaded of is the receiver’s prior. With this, he communicates the set of models that induce the feasible vector of posteriors maximizing his expected utility. This setting imposes two main restrictions on the sender, which may partially deter manipulations. First, the sender can only provide interpretations of observable events. Consider the case in which the sender could convince the receiver that other signals could be observed, even if these signals can never occur. I show that it is enough to add one of these dummy signals to the list of realizations the receiver considers possible to achieve full manipulability. Second, the sender does not know the signal realization the receiver observes when he communicates models. Unless the receiver’s prior is such that all vectors of posteriors are feasible, this ex-ante commitment could be costly even for a sender communicating many models.

What if the receiver does not only consider the models provided by the sender? In an extension, I allow the receiver to initially hold a model as default. She adopts other models only if these are better at explaining new information with respect to this default model. I show that holding a default model deters belief manipulation and I characterize the set of feasible vectors of posteriors in this case. I show how this result is connected to the main result: the set of the feasible vectors of posteriors without a default model is the union of all the sets of the feasible vectors of posteriors with a default model for every default model. Last, I show how ex-ante commitment is not costly if the receiver has a default model and the sender communicates as many models as signal realizations: the same beliefs can be induced with ex-ante or ex-post model persuasion. However, even if the sender can attain the same outcome, the implementation should take into account the ex-ante commitment.

I present several stylized applications that fit this setting. First, I elaborate on the previously presented example of the politician and the voter. Being exposed to multiple narratives on the trustworthiness of the voting system might induce the voter to have a double standard in assessing candidates’ reported victory, always supporting one candidate regardless of the election outcome. I use this example to illustrate why it is in the best interest of the sender to provide conflicting models. Communicating a large number of possibly contradictory and untruthful messages is one of the central features of “firehose of falsehood”, a propaganda technique described by Paul and Matthews (2016) and usually associated with modern Russia. This type of disinformation campaign is reported to be generally effective in manipulating the

audience. However, even if beneficial for the persuader, such a communication strategy can have serious consequences: conflicting narratives lead to inevitable belief polarization in a population of heterogeneous voters. When voters with sufficiently different priors are exposed to the same pair of conflicting stories, their beliefs always diverge further. Regardless of what happens, voters adopt different stories to make sense of the election outcome and they update in opposite directions. I formalize this result for the binary case and I provide some suggestive evidence of this mechanism using the case of the 2020 US Presidential election.⁴

Second, I focus on a financial application to explain in detail how to construct a graphical solution for the sender’s problem. In particular, I study the misalignment of incentives between a financial advisor and investors with private information. Indeed, the setting I study in this paper is suitable not only for temporal interpretation (the sender communicates models before the signal realizes), but also for a private-information one: the receiver has access to the signal, but the sender does not. Thus, the advisor communicates different narratives which could be picked up depending on the private information of the investors, e.g., past financial experience. The advisor can manipulate investors regardless of her past experience, always moving her beliefs in an advantageous direction. If investors have favorable expectations towards the advisor-preferred asset, they always invest fully in that asset, making the ex-ante commitment costless. The advisor can achieve this by exposing investors to conflicting ways of looking at past data: there is either a perfectly positive or a perfectly negative correlation between past and future events. In contrast, this does not work for pessimistic investors. Also, communicating without knowing the pessimistic investors’ experience diminishes the sender’s utility in expectation.

The third application explores the phenomena of self-persuasion, whereby an agent could distort their beliefs by manipulating the perceived informativeness of the signals. Drawing from the literature on motivated beliefs, I investigate a multi-selves model in which an agent has incentives to distort her self-confidence in order to offset her time-inconsistent references. Building on the setting proposed by Bénabou and Tirole (2002), I show how overconfidence through distorting the feedback perceived precision can motivate in the sense of committing to a costly action. However, this happens only if one believes to suffer enough from present bias. More generally, I comment on how leaving facts open to interpretation allows inconsistent reasoning to take root and it may be one of the mechanisms people use to achieve self-serving beliefs. This proposed mechanism can deliver the classic implications of this literature, but it also gives a bound on belief distortion.

Last, I show how a strategic persuader could challenge a shared narrative to insinuate doubt and deepen polarization for agents differing in initial beliefs. I exemplify this in the context of the lost trust in science on issues like climate change and the health effect of smoking. Holding a shared initial model does not deter polarization.

This paper speaks directly to two strands of economic literature — narratives and persuasion — both flourishing in the last decade (see Section 6 for a more detailed discussion of the related literature). Starting from Shiller (2017, 2019), there has been an increasing formalization of narratives into economic literature using different notions: narratives as likelihood functions

⁴The 2020 US Presidential election provides an example of conflicting narratives communicated to voters before the release of the election outcome. Before the ballot, Donald Trump spread allegations on how elections could be rigged against him, especially through the vote-by-mail system. I use this case to illustrate some facts that are in line with my predictions.

(Schwartzstein and Sunderam, 2021), or directed acyclical graphs (Eliaz and Spiegler, 2020), or moral reasoning (Bénabou et al., 2018). This paper builds on the first. Inspired by the interdisciplinary research on sense-making (Andreassen, 1990; Weick, 1995; DiFonzo and Bordia, 1997; Chater and Loewenstein, 2016), Schwartzstein and Sunderam (2021) assume that individuals prefer narratives that make sense of observable data and prior knowledge, when uncertain about the right interpretation. Leveraging on this, persuaders may encourage the adoption of wrong interpretations to their advantage. The authors examine the strategic provision of models after a public signal has realized and observed by both the sender and the receiver. However, a receiver might be skeptical to accept narratives formulated only after the sender learns about the information available to the receiver, as the sender could formulate an ad-hoc story to fit the realized event. For example, a voter could be concerned to hear for the first time a politician claiming elections to be rigged only after he lost the election. Therefore, I investigate a setting in which the sender commits to his communication strategy without knowing the signal realization, while the receiver’s choices are conditional on the signal realization as in Schwartzstein and Sunderam (2021). The reason is two-fold. First, shifting to an ex-ante perspective may give more credibility to the sender’s stories from the receiver’s perspective. A second motive for ex-ante communication is that the sender needs to preemptively frame information because he might not be able to reach the receiver later on. The paper shows that ex-ante commitment to the sender’s communication is not a trivial problem because models compete with each other and the set of ex-post optimal models might not be optimal if provided ex-ante.⁵

Narratives as a tool for persuasion imply important differences from previous models of persuasion in the economic literature. In particular, the sender does not tamper with the signal the receiver observes, unlike the cheap talk literature (Milgrom, 1981; Crawford and Sobel, 1982). Moreover, there is a fixed signal generating process that cannot be manipulated. This is in stark contrast with the literature of Bayesian persuasion, started by Kamenica and Gentzkow (2011) and continued by many generalizations of their framework (e.g., Alonso and Cámara, 2016; Ely, 2017; Galperti, 2019; Ball and Espín-Sánchez, 2021). In broad terms, these papers are about persuasion by generating information: the sender chooses an experiment that maps each state realization to a distribution of signals, which in turn induces a distribution over the receiver’s posterior beliefs. The only restriction is that such distribution of posteriors must be Bayes-plausible — the expected posterior has to average to the prior. Thus, also in the Bayesian persuasion framework, the sender adopts an ex-ante perspective but by controlling the signal generating process. In the political example, this translates in the politician directly manipulating the voting system, and its accuracy, in order to get the voter’s support. This paper flips the problem by providing interpretations of signals, without the ability to modify how the signal is generated and which signals can be observed. It contributes to the literature on persuasion by showing how powerful is the strategic use of narratives, still imposing restrictions on what the sender can achieve. On one hand, this approach allows the sender to achieve a higher expected payoff: providing conflicting models, the receiver might update in the same desired direction for every possible signal. The main result of this paper relaxes the Bayes plausibility constraint,

⁵Unless the receiver initially considers another model by default, the sender can induce any posterior by proposing an interpretation of the publicly observed event. This is not the case for vectors of posteriors; thus, the main analysis in this paper studies the limits of persuasion in absence of a default model. Section 5 extends the main result to the case in which the receiver is endowed with a default model and allows comparability with Schwartzstein and Sunderam (2021).

but in a disciplined manner because not all beliefs are feasible across events. On the other hand, if the sender were to manipulate also the signals the receiver expects to observe, he could always achieve full manipulability.⁶

The rest of the paper is organized as follows: Section 2 sets up the framework. Section 3 addresses the question of what the receiver can be persuaded of, studies the comparative statics, and comments on the sender's problem. Section 4 illustrates applications. Section 5 extends the results to the case in which the receiver is endowed with a default model. Section 6 discusses the related literature. Section 7 concludes. All the proofs can be found in Appendix A.

2 Set-up

Two agents, sender and receiver, have utility functions $U^S(a, \omega)$ and $U^R(a, \omega)$ that depend on the receiver's action $a \in A$ and the state of the world $\omega \in \Omega$. They share a common prior $\mu_0 \in \text{int}(\Delta(\Omega))$.⁷ The receiver observes a signal $s \in S$. The state and signal spaces are finite and fixed. A *model* m is a map assigning to each state a distribution of signals conditional on that state: it specifies $\pi^m(s|\omega)$ for every $s \in S$ and $\omega \in \Omega$ with $\sum_s \pi^m(s|\omega) = 1$ for each $\omega \in \Omega$. Let $\mathcal{M} = [\Delta(S)]^\Omega$ be the set of all models. Conditional on signal s , a model m induces posterior belief μ_s^m via Bayes rule. I refer to the likelihood $\Pr^m(s) = \sum_{\omega \in \Omega} \mu_0(\omega) \pi^m(s|\omega)$ as to the *fit* of the model m given signal s .

Consider the following timing: without knowing the signal realization, the sender communicates a set of models to the receiver; given the observed signal, the receiver adopts a model to update her prior and chooses an action. In particular, I assume the receiver to act as follows. First, she adopts the model with the highest fit conditional on the observed signal s among the set of models M she has been exposed to:

$$m_s^* \in \arg \max_{m \in M} \Pr^m(s).$$

Then, she updates her prior using the adopted model and chooses the action that maximizes her expected utility:

$$a_s^* \in \arg \max_{a \in A} \mathbb{E}[U^R(a, \omega)],$$

where the expectation is taken with respect to the posterior $\mu_s^{m_s^*}$. When indifferent, the receiver adopts the model or the action maximizing the sender's expected utility.

In case of misaligned preferences, the sender has incentives to communicate a set of models with the purpose of influencing the receiver's action in order to maximize his expected utility rather than hers. The sender knows the receiver's preferences and the true model t , specifying the objective probabilities of signals. Let $\boldsymbol{\mu} = (\mu_s)_{s \in S} \in [\Delta(\Omega)]^S$ be a vector of posterior beliefs: it

⁶Ichihashi and Meng (2021) investigate the case in which the sender first designs the signal generating process, and then provides an interpretation of the observed signal. This is another approach to combine Schwartzstein and Sunderam (2021) and Kamenica and Gentzkow (2011). I do not study what happens if the sender can also manipulate how the signal is generated, but only the case in which he can convince the receiver that additional signals could be observed.

⁷This assumption is made for simplicity. The extension to heterogeneous priors is straightforward. See Section 4.2 for an example.

describes the posterior beliefs conditional on each signal realization. Thus, the value of a vector of posteriors $\boldsymbol{\mu}$ equals the sender’s expected utility given the receiver’s actions at those beliefs calculated using model t :

$$V(\boldsymbol{\mu}) = \sum_{s \in S} \Pr^t(s) \mathbb{E} [U^S(a_s^*, \omega)] .$$

Given a set of models M , the receiver’s resulting vector of posterior is such that for each signal the posterior is induced by the model with the highest fit, i.e., $\boldsymbol{\mu}^M = (\mu_s^{m_s^*})_{s \in S}$. Therefore, the sender chooses the set of models M^* that maximizes his value function at the resulting vector of posteriors:

$$M^* \in \arg \max_{M \subseteq \mathcal{M}} V(\boldsymbol{\mu}^M).$$

2.1 Discussion of Assumptions

Next, I discuss some of the assumptions behind this setting.

I start by focusing on the receiver. First, I relax the Bayes rationality of the receiver only partially: she updates her prior via Bayes rule once she has selected a model. Following Schwartzstein and Sunderam (2021), the model adoption is assumed to occur via maximum likelihood: once exposed to a set of models, the receiver adopts the model that fits best the observed data given her prior.⁸ It is equivalent to selecting the model with the higher posterior probability given the signal, starting from a flat prior over proposed models. This is because the receiver does not come up with all models she is willing to entertain, but she compare only the models she was exposed to. Importantly, only one model is used to make inference. This is in line with *Inference to the Best Explanation* (Harman, 1965; Lipton, 2003): only the best hypothesis is used to make inference. However, this theory is agnostic on what means the best. Here, I consider as a measure the goodness the fit. Indeed, there is some evidence that people choose the most probable hypothesis.⁹ A line of research in cognitive psychology argues that hypotheses are supported by the same observations they are supposed to explain and the more they explain, the more confidence we give to that hypothesis (Koehler, 1991; Pennington and Hastie, 1992; Lombrozo and Carey, 2006). Douven and Schupbach (2015a,b) provide evidence of the importance of explanatory power in updating and predicting estimates of posterior probabilities. Instead, the “traditional” alternative would have been to assume the receiver to hold fully Bayesian beliefs — not selecting any model but forming posteriors beliefs over models given the signal and priors over models as well. In this case, her posteriors are calculated as the average of the posteriors given each model weighted by the posterior of each model given the observed signal.¹⁰ This could be computationally intense and would exhaust our finite

⁸The literature of belief updating under ambiguity considers a maximum likelihood updating rule, introduced by Dempster (1967) and Shafer (1976), then axiomatized by Gilboa and Schmeidler (1993). When agents consider multiple prior over states, they only update the subset of priors that maximizes the probability of the realized event. Similar in spirit, I do not consider multiple priors over states but multiple models that could be used for updating. The other most common rule for updating in the case of multiple prior is full Bayesian updating: subjects update prior-by-prior and retain ambiguity in their posteriors. These updating rule might lead to similar inconsistent updating as the one discussed in this paper, but with different implications.

⁹Simple and more probable explanations are valued (Einhorn and Hogarth, 1986; Thagard, 1989), but in the absence of a simplicity difference people prefer more probable explanations (Lombrozo, 2007). Given the fixed structures of states and signals here, all models are equally simple/complex in this setting.

¹⁰Formally, let $\rho \in \Delta(M)$ with ρ^m is the prior over model m among the proposed ones. A fully Bayesian

cognitive resources in most situations, while maximum-likelihood selection circumvents these challenges.¹¹

Second, I assume the receiver to be naive, so she does not question the sender’s purpose. Indeed, the receiver does not have enough information to be strategic: she would need two pieces of information, the sender’s preferences, and his adopted model. This prevents the receiver from anticipating the sender’s value function and, hence, being strategic about the models she receives. In particular, without anticipating all possible models she is willing to entertain, there is nothing to guide her on how to form beliefs on the sender’s model. If she were able to do this, it would allow her to learn more about the true model and state directly, ignoring the sender’s proposed models. This full naïveté assumption is a common starting point in the literature (e.g., Heidhues and Kőszegi, 2018; Eyster, 2019) and it incorporates several motives leading the receiver to underreact to the sender’s private information and incentives.

The sender’s behavior differs in a two-fold manner from Schwartzstein and Sunderam (2021). First, the sender communicates models without knowing the signal realization. This allows for a *temporal* interpretation: a public signal will be observed by both agents, but the sender has to provide models before its realization. Also, this assumption can accommodate a *private-information* interpretation: the receiver might hold some private information on the state — the signal realization — and the sender cannot access it. Throughout the paper, I mostly focus on the first interpretation for simplicity; however, all results equally apply to both. Second, I do not restrict the sender’s capacity to send multiple narratives. Because the sender does not know the signal realization, he has incentives to send more models that could be picked up depending on the realization. Note that there is no need to have more models than the number of signal realizations. This explains why in the interim model by Schwartzstein and Sunderam (2021), the authors do not consider multiple models provided by the sender.

3 Ex-Ante Model Persuasion

In this section, I characterize the extent to which the receiver’s belief can be manipulated. I start with some preliminary results. First, I illustrate the connection between models and vectors of posteriors. Second, I pin down the trade-off between how well a model can fit the observed realization and how much a model can move the posterior away from the prior. Then, I state the main result of the paper and I provide a graphical intuition for the binary case. Last, I discuss some comparative statics and the sender’s problem.

3.1 Preliminaries

To solve the sender’s problem in the Bayesian persuasion literature, it is pivotal to characterize not only the posteriors the receiver might attain but also with which probabilities these

posterior is calculated as $\mu_s^B(\omega) = \sum_{m \in M} \rho_s^m \mu_s^m(\omega)$ where $\rho_s^m = \frac{\rho^m \Pr^m(s)}{\sum_{m' \in M} \rho^{m'} \Pr^{m'}(s)}$.

¹¹I abstract from the reasons why this could be the case. For example, it might be that the most compelling model is adopted because people believe what they are prepared to hear or it might also be that models communicated ex-ante are stored in the receiver’s memory and the best-fitted one is the easiest to retrieve.

posteriors can be attained, i.e., the distribution of the receiver's posteriors. The relevant constraint is that, given a prior, each information structure feasible for the sender corresponds to a Bayes-plausible distribution over posteriors. By contrast, this paper assumes that there is a fixed distribution over the signals induced by the true model t . Therefore, to solve the sender's problem, it is enough to study the vectors of posteriors that the receiver could hold. In what follows, I show that there is an equivalent representation between models (information structures with fixed signal space) and vectors of posteriors under a condition comparable to Bayes-plausibility. That is, I say that a vector of posterior beliefs $\boldsymbol{\mu}$ is *Bayes-consistent* if the prior μ_0 is a convex combination of the posteriors across signals $(\mu_s)_{s \in S}$. Let $\mathcal{B} \subset [\Delta(\Omega)]^S$ be the set of all vectors of posterior beliefs that are Bayes-consistent. Also, let $\boldsymbol{\mu}^m$ be the vector of posteriors such that each posterior is induced by model m .

Lemma 1. *For each Bayes-consistent vector of posterior beliefs $\boldsymbol{\mu} \in \mathcal{B}$ there exists a model that induces $\boldsymbol{\mu}$ and each model m induces a Bayes-consistent vector of posterior beliefs $\boldsymbol{\mu}^m \in \mathcal{B}$.*

Next, I focus on the trade-off between how well a model can fit data and how much a model can move beliefs. Fix a target posterior μ_s conditional on a signal s . Define the *movement* for posterior μ_s in state ω as $\delta(\mu_s(\omega)) = \frac{\mu_s(\omega)}{\mu_0(\omega)}$ and the *maximal movement* for μ_s as $\bar{\delta}(\mu_s) = \max_{\omega \in \Omega} \delta(\mu_s(\omega))$. With this, it is possible to characterize the set of fit levels that a model can have when inducing posterior μ_s .

Lemma 2. *Fix a posterior μ_s . For every $p \in [0, \bar{\delta}(\mu_s)^{-1}]$, there exists a model inducing μ_s with fit $\Pr^m(s) = p$, and every model inducing μ_s has fit $\Pr^m(s) \in [0, \bar{\delta}(\mu_s)^{-1}]$.*

Intuitively, there is less freedom in terms of fit levels to induce posteriors further from the prior. Schwartzstein and Sunderam (2021) characterize the upper bound in Lemma 1: conditional on a signal, the maximal fit for a target posterior coincides with the reciprocal of the maximal movement. Indeed, any model that leads beliefs to react a lot given a signal realization (higher movement) cannot fit the data well (lower fit).

3.2 Feasible Vectors of Posterior Beliefs

In this section, I characterize the set of feasible vectors of posteriors that the receiver could hold. The first result is straightforward and shows that using one model only Bayes-consistent vectors of posteriors are feasible.

Proposition 1 (One Model). *If $|M| = 1$, then the set of feasible vectors of posterior beliefs equals \mathcal{B} .*

Next, I consider the case in which the receiver is exposed to many models.¹² This is the main result of the paper and it shows how the set of feasible vectors of posteriors is characterized by a simple condition: the sum over signals of the maximal fit levels associated with the posterior beliefs exceeds one. Then, the characterization follows from Lemma 2 according to which the maximal fit of a posterior coincides with the reciprocal of the maximal movement for that belief.

¹²Exposing the receiver to a number of models equal to the number of signals allows maximal belief manipulability. More models would not enlarge the set because at most one model is adopted conditional on each signal realization, while fewer models would make some vectors unfeasible. As the sender can send as many models as he wishes, I focus on the maximal manipulability of the receiver.

Theorem 1 (Many Models). *If $|M| \geq |S|$, then the set of feasible vectors of posterior beliefs is*

$$\mathcal{F} = \left\{ \boldsymbol{\mu} \in [\Delta(\Omega)]^S : \sum_{s \in S} \bar{\delta}(\mu_s)^{-1} \geq 1 \right\}.$$

This theorem carries two key takeaways. First, allowing for multiple models expands the feasibility set with respect to the case of one model only.¹³ As a consequence, vectors of posteriors that are not Bayes-consistent are feasible. This means that, with many models, the sender can achieve a vector of posteriors that can be the support of a distribution of posteriors unattainable with Bayesian persuasion. Second, generally, not all vectors of posteriors are feasible. The theorem illustrates a trade-off in movement across signal realizations: moving a posterior away from the prior restricts how much movement is allowed for posteriors conditional on other signals. Thus, not “anything goes.”

A vector of posterior beliefs is feasible if there exists a set of models such that each model is tailored to a specific signal realization inducing the desired posterior conditional on that signal. This implies that each of such models must be adopted conditional on the signal it has been tailored to, introducing an analog of the incentive compatibility constraint for models depending on their fit levels across signal realizations. The proof shows that if a vector satisfies the condition of Theorem 1 such a set of models exists, otherwise it cannot. As models compete with each other across realizations, the higher fit a model has inducing the posterior, the more freedom there is to induce posteriors conditional on other signals with other models. Therefore, the maximal fit associated with each posterior pins down the extent to which each posterior contributes to the vector’s feasibility: if low, the other posteriors should compensate by being closer to the prior; if high, the other posteriors could be further away from the prior. In particular, the frontier of the feasibility set — the furthest vectors of posteriors from the prior that are still feasible — is generated by *maximal overfitting*: each tailored model induces the desired posterior with maximal fit conditional on the target signal. Closer vectors to the prior are always feasible.

3.2.1 Graphical Intuition

This section introduces a graphical approach to this setting in order to provide intuitions of these results. To do so, I focus on the binary case.

Given the binary state, let the posterior probability of state ω_1 identify the posterior beliefs in the following graphs. The axes represent the posterior attached to ω_1 conditional on each signal realization and each point in this graph represents a vector of posterior beliefs. I represent the prior $\mu_0(\omega_1) = 0.3$ as the vector of posteriors such that each posterior equals the prior, that is the orange dot in all figures. In Figure 1a, the purple area depicts the Bayes-consistent vectors of posteriors, \mathcal{B} : a vector of posteriors is Bayes-consistent if, for each ω , it holds that either (i) $\mu_{s_1}(\omega) > \mu_0(\omega) > \mu_{s_2}(\omega)$, or (ii) $\mu_{s_1}(\omega) < \mu_0(\omega) < \mu_{s_2}(\omega)$. This means that it is not possible to update beliefs always in the same direction. In the binary case, there is a one-to-one map

¹³When all posteriors are induced by the same model (equivalent to $|M| = 1$), the feasibility condition is satisfied. Notice that $\sum_s \bar{\delta}(\mu_s^m)^{-1} \geq \sum_s \Pr^m(s) = 1$, where for each signal s , $\bar{\delta}(\mu_s^m)^{-1} \geq \Pr^m(s)$ by Lemma 2.

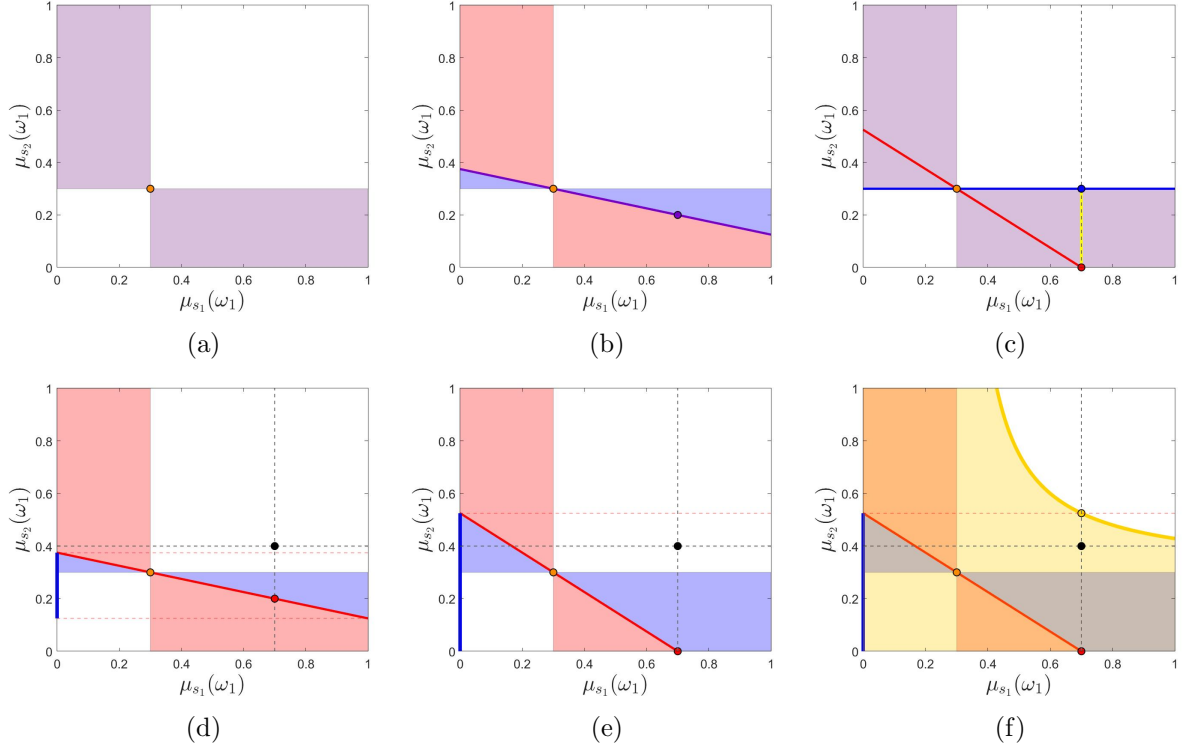


Figure 1: Graphical Intuition for the Binary Case

between Bayes-consistent vectors of posterior beliefs and models.¹⁴ Hence, every point in the purple area corresponds to a model.

Focusing on a model as in Figure 1b, it is possible to graphically observe some of its properties. To do so, consider the purple line passing through the vector of posteriors induced by that model (purple dot) and the prior (orange dot). This is the *isofit* associated with the considered model: all the points on that line correspond to models that have the same fit conditional on each signal.¹⁵ The slope of the isofit can be interpreted as follows: the steeper (flatter) the line, the higher the fit conditional on s_1 (s_2). For each level of fit, it is possible to partition \mathcal{B} into three subsets: vectors induced by models with the same fit (isofit line), vectors induced by models with higher fit conditional on s_1 (red area), and vectors induced by models with higher fit conditional on s_2 (blue area).

Given the prior, there is a multiplicity of models that induce the same posterior distribution conditional on a signal with different levels of fit. Consider the target posterior $\mu_{s_1}(\omega_1) =$

¹⁴The only exception is the vector for which the posterior conditional on every signal equals the prior. There are infinitely many *uninformative* models (they assign the same distribution of signals conditional on all states) inducing it. This statement is formalized in Appendix A.

¹⁵Formally, an isofit is the set of vectors of posterior beliefs that are induced by models that have the same fit conditional on every signal realization. For each $\varphi \in \Delta(S)$, formalize

$$I(\varphi) = \left\{ \mu \in [\Delta(\Omega)]^S : \forall \omega \in \Omega, \mu_0(\omega) = \sum_{s \in S} \varphi_s \mu_s(\omega) \right\}.$$

In the binary case, consider the Bayes-consistency constraint for ω_1 with weights given by the fit levels induced by model m and re-arrange to $\mu_{s_2}^m(\omega_1) = \frac{\mu_0(\omega_1)}{\Pr^m(s_2)} - \frac{\Pr^m(s_1)}{\Pr^m(s_2)} \mu_{s_1}^m(\omega_1)$. All models with the same fit ($\Pr^m(s_1), \Pr^m(s_2)$) correspond to points on this line. In this case, two models that have the same fit conditional on one signal have also the same fit conditional on the other signal. Thus, it is enough to look at the fit conditional on a signal only.

0.7, the dotted line in Figure 1c. The yellow line corresponds to all the models inducing the target $\mu_{s_1}(\omega_1)$: at each point, the fit varies. Among these models, the one with the highest fit conditional on s_2 is the model corresponding to the blue point, lying on the flattest isofit inducing the target. The model with the maximal fit conditional on s_1 is the one corresponding to the red point: a steeper line cannot induce the target. By Lemma 2, the fit of such model given s_1 is known: it is $\left(\max\left\{\frac{0.7}{0.3}, \frac{0.3}{0.7}\right\}\right)^{-1} = 43\%$.

In the binary case, a target vector μ is feasible if it is possible to find two models m_1 and m_2 , respectively, adopted conditional on s_1 and s_2 and inducing μ_{s_1} and μ_{s_2} . In Figure 1, the target vector is the black dot. Start with a model m_1 inducing μ_{s_1} (red dot) in Figure 1d. For any such model, I want to identify the compatible posteriors induced by other models conditional on s_2 if model m_1 is adopted conditional on s_1 . Therefore, focus on the blue area because this corresponds to models with the higher fit (thus, adopted) conditional on s_2 with respect to m_1 as in Figure 1b. The compatible posteriors conditional on s_2 with respect to m_1 are all the y-coordinates of points in the blue area: the blue line on the y-axis. However, the y-coordinate of the target vector does not lie in this set. This does not mean that the target vector is unfeasible because there are many models that induce the posterior conditional on s_1 changing fit level, as shown in Figure 1c. The model m_1 I started with did not allow me to induce the target but others could. In particular, consider the model m_1 inducing μ_{s_1} with maximal fit in Figure 1e. This is because increasing $\Pr^{m_1}(s_1)$, $\Pr^{m_1}(s_2)$ decreases and there are more models can be adopted conditional on s_2 . Thus, the range of compatible posterior distributions conditional on s_2 with respect to model m_1 expands. Maximally overfitting conditional on s_1 identifies the largest set of posteriors conditional on s_2 compatible with μ_{s_1} . With this m_1 , the y-coordinate of the target belongs to the set of compatible posteriors, thus there exists a model in the blue area that together with m_1 could induce the target. Furthermore, note that the dotted red line corresponds to the maximal posterior among the compatible ones given μ_{s_1} . The point for which this intersects μ_{s_1} — yellow dot in Figure 1f — exemplifies how to construct the upper frontier of the set of feasible vectors of posteriors. All vectors below that line (yellow area) are feasible.

3.2.2 Comparative Statics

In this section, I study what makes the receiver more vulnerable to persuasion. Generally, not all vectors of posterior beliefs are feasible. Interestingly, this is not the case when the receiver's minimal prior across states is sufficiently high, with respect to the reciprocal of the number of signals. In this case, the receiver is fully persuadable: any vector of posteriors can be induced.

Proposition 2. *If $\min_{\omega \in \Omega} \mu_0(\omega) \geq \frac{1}{|\mathcal{S}|}$, then all vectors of posterior beliefs are feasible.*

The proposition illustrates a simple test to check whether the receiver is fully persuadable. Two observations follow. First, the more signals, the more manipulability of the receiver's beliefs. Tailoring models to specific signals allows more feasible vectors of posteriors but also requires that models are compatible with each other across signals: the more signals, the less stringent this condition on the prior is. To exemplify this, I continue the example of Figure 1e. Consider a model inducing the target posterior 0.7 conditional on signal s_1 . To leave more freedom conditional on the other signal, set the fit to the maximum: $\Pr^m(s_1) = 43\%$. A model tailoring

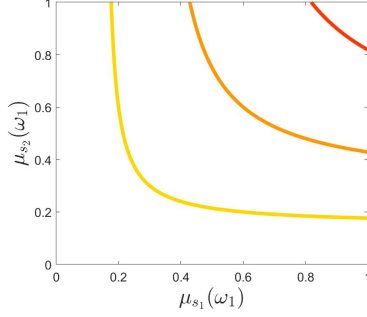


Figure 2: Frontier of the feasibility set, by prior

Notes: the lighter the color line, the further away from the uniform prior: yellow $\mu_0(\omega_1) = 15\%$, orange $\mu_0(\omega_1) = 30\%$, red $\mu_0(\omega_1) = 45\%$.

the other signal s_2 must have a fit higher than $\Pr^m(s_2) = 57\%$ to be adopted conditional on s_2 . If more signals were available, this constraint is less stringent because $\sum_{s \neq s_1} \Pr^m(s) = 57\%$. Hence, tailored models for the other signals would have more freedom overall to induce a posterior further away from the prior. Therefore, a receiver is more manipulable in a setting with many signals to be interpreted. Second, the minimal prior across states contains information regarding the set of feasible vectors of posteriors. To get an intuition for this, notice that the minimal prior across states is the lower bound for the maximal fit for any updated posteriors starting from a given prior, i.e., $\bar{\delta}(\mu_s)^{-1} \geq \min_{\omega \in \Omega} \mu_0(\omega)$ for any μ_s , pinning down the lower bound for the sum of maximal fit levels across signals. Also, note that by increasing the minimal prior over states the prior beliefs get closer to a uniform distribution. Hence, one can interpret the minimal prior across states as a measure of the concentration of beliefs with respect to uniformly distributed prior. Putting the pieces together, it holds that the more uniform the prior, the lower movement to induce further away posteriors, the more belief manipulability.

The next result follows from the last observation. If the signals are as many as states, any vector of posteriors is feasible if the receiver has uniformly distributed prior across states.

Corollary 1. *If $|S| \geq |\Omega|$ and $\mu_0(\omega) = \frac{1}{|\Omega|}$ for every $\omega \in \Omega$, all vectors of posterior beliefs are feasible.*

A stronger result holds focusing on the binary case. In this special case, the set of feasible vectors of posteriors can be ordered: the closer the receiver's prior is to 50-50, the more she can be manipulated (Figure 2). Without loss of generality, let $\mu_{0,\varepsilon} = (\mu_{0,\varepsilon}(\omega_1), \mu_{0,\varepsilon}(\omega_2)) = (\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon)$ and \mathcal{F}_ε the set of the feasible vectors of posteriors with respect to this prior.

Proposition 3 (Binary Case). *For $\varepsilon' < \varepsilon''$, it holds that $\mathcal{F}_{\varepsilon''} \subseteq \mathcal{F}_{\varepsilon'}$.*

3.3 Sender's Problem

Given these results, I turn to the sender's problem. Indeed, all the information the sender needs to know what the receiver can be persuaded of is her prior. Then, he maximizes his value on the set of feasible vectors of posteriors, knowing the receiver's preferences and anticipating the receiver's optimal action. Optimization is standard, except that the set of feasible vectors of posteriors could be non-convex as shown in Figure 1f for the binary case.

Since the receiver is not endowed with a model, she cannot interpret the realized signal without models provided by the sender. Therefore, I assume that, if the sender does not communicate any model to the receiver, she does not update her beliefs, discarding the realized signal: she holds μ^\varnothing such that $\mu_s^\varnothing = \mu_0$ for each signal s . The sender benefits from persuasion through models if there exists a feasible vector of posterior beliefs $\mu \in \mathcal{F}$ such that its value is higher than the value of the prior: $V(\mu) \geq V(\mu^\varnothing)$.

The sender faces two main restrictions in choosing his communication strategy. I investigate if these may deter persuasion. First, unlike the literature on Bayesian persuasion, the signal space is fixed and the sender cannot manipulate it. This is a crucial assumption and it restricts the sender's communications only to interpretations of observable events, $s \in S$. If this were not the case and the sender could add dummy signals, he could persuade the receiver to hold any beliefs in the original space. The intuition is that if the receiver believes that also other signals proposed by the sender were to be possible with $S' \supset S$, the sender could leverage those signals that cannot realize to manipulate beliefs further. Indeed, one dummy signal is enough to guarantee full manipulability.

Proposition 4. *Adding a dummy signal $s_0 \notin S$ to the signal space $S' = S \cup \{s_0\}$, any vector of posteriors on the original signal space $\mu \in [\Delta(\Omega)]^S$ can be induced.*

Second, I assume the sender provides models without knowing the signal realization. What is the impact of this assumption on the sender's expected utility? This coincides with his willingness to pay to avoid ex-ante commitment or to have the ability to perfectly correlate a model with each signal realization. Knowing the signal realization, he could send a tailored model to induce the desired posterior conditional on that signal. Thus, avoiding competition among provided models, he could induce any vector of posterior.¹⁶ Therefore, the cost of ex-ante commitment equals the gap between the unconstrained maximal sender's value over any vector of posteriors and the maximal sender's value over the feasible vectors of posteriors:

$$\Delta = \underbrace{\max_{\mu \in [\Delta(\Omega)]^S} V(\mu)}_{\text{no commitment}} - \underbrace{\max_{\mu \in \mathcal{F}} V(\mu)}_{\text{commitment}} \geq 0.$$

First, the commitment cost is strictly positive when the unconstrained maximum is unfeasible. For example, if the sender communicates only one model and the sender's value is maximized by a Bayes-inconsistent vector of posterior. In this case, the sender is always willing to pay to learn the signal realization in advance to better tailor the narrative to communicate. Second, the commitment cost is zero when all vectors of posteriors are feasible. Section 3.2.2 provides insightful results related to this as it discusses the cases in which all vectors of posterior are feasible. In this case, the sender cannot gain from perfectly tailored communication on signals.

This measure can help us understand in which circumstances commitment from ex-ante model persuasion is more or less costly. In the private-information interpretation, knowing the private information held would allow the sender to perfectly tailor the narrative to the receiver. This is a commonly used practice in marketing, known as *microtarget*: analyzing online information on the potential customers with the goal to create and convey the most effective message depending

¹⁶This is not the case if the receiver has a default model. In Section 5, I discuss and study the sender's cost of commitment if the receiver has a default model.

on preferences and personality.¹⁷ It is possible to use this framework to comment on the value of microtargeting in different contexts. I discuss an example of this in Section 4.2.

4 Applications

This section discusses several applications that could be read through the lenses of this framework. The first formalizes the political example outlined in the introduction and sheds light on the polarizing consequences of being exposed to conflicting stories. Then, I provide suggestive evidence of this mechanism. Second, a financial application illustrates the sender’s optimization problem in more detail. The third application discusses self-persuasion.

4.1 Firehose of Falsehood

Firehose of falsehood is a propaganda technique based on a large number of possibly contradictory and mutually inconsistent messages. It was defined by Paul and Matthews (2016) to describe the modern Russian propaganda.¹⁸ Usually, it is employed through coordinated operations led by official (e.g., Russia Today) or unofficial sources (e.g., the Internet Research Agency). This disinformation campaign is effective in entertaining, confusing, and ultimately manipulating the audience. In particular, inconsistent messages imply disregard for truthful reporting. The growing interest in understanding fake news revealed that people have a hard time in distinguishing true and false stories: not only fake news are widely shared and believed (Allcott and Gentzkow, 2017), but also falsehood spread faster, deeper, and more broadly than true news stories on Twitter (Vosoughi et al., 2018). Therefore, firehose of falsehood does not encounter the resistance we might hope for and a persuader willing to engage in political disinformation would find in it powerful tools to achieve his goals.

Building on the political example briefly presented in the introduction, I illustrate why it is in the interest of the sender to communicate conflicting stories. While this strategy is particularly successful for the persuader in manipulating a target receiver, it could lead to extreme polarization in beliefs and actions in the presence of receivers with sufficiently different priors. Thus, the next example illustrates the potential consequences of a fixed set of stories for receivers with different priors. For simplicity of exposition, I describe this strategy to be adopted by a single politician. However, one can imagine a party implementing such communication in a coordinated manner, with the help of different members or factions of the same political group. Indeed, narratives from different sources could come across as more persuasive than the ones coming from a single source contradicting oneself.¹⁹

¹⁷This is a well-established practice and there are several examples of companies running different ads to different groups of people. For an example, see <https://themarkup.org/news/2021/04/13/how-facebooks-ad-system-lets-companies-talk-out-of-both-sides-of-their-mouths>.

¹⁸The authors describe the distinct features of this phenomenon: (i) high number of channels and messages, (ii) lack of commitment to consistency or objective reality, and (iii) rapid, continuous, and repetitive communication. I focus on the first two dimensions.

¹⁹Interestingly, a study by Reich and Tormala (2013) argues that contradicting oneself — initially supporting something and then later switching to something else — might offer a persuasive advantage over both one-time opinions (supporting something once) and repeated consistent opinions (e.g., initially supporting something and then later supporting it again). The effect is moderated by trust in the source and it disappears if the conflicting opinions come from different sources.

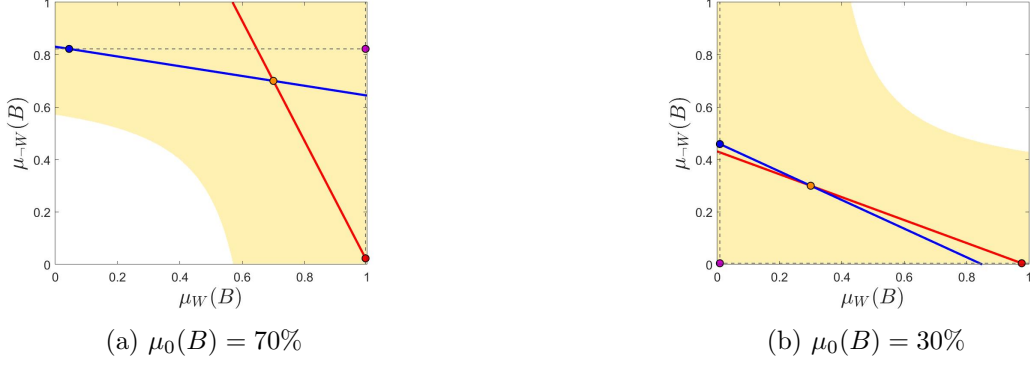


Figure 3: Firehose of Falsehood, for different voter's prior

Notes: the orange point corresponds to the voter's prior; the yellow area represents the set of feasible vectors of posteriors; the red and the blue points are respectively the vector of posteriors induced by models f and c , while the purple point is the resulting receiver's vector of posteriors.

Politician Bob is running for the presidency. He wants to be recognized as president by part of the voters regardless of the reported election outcomes. To make sure of that, he is spreading two different stories about the reliability of the election system. Let the state space be $\{B, \neg B\}$, where B is the event in which Bob is the legitimate winner of the election, and the signal state be $\{W, \neg W\}$, where W is the event in which Bob is reported as winner. Each voter recognizes Bob as president if, having observed the election outcome, she believes that Bob is the legitimate winner with a probability higher than 50%. The politician knows that mistakes in vote counting are very rare. To maintain credibility, this model is communicated: the system is fair with high precision, $\pi^f(W|B) = 99\%$ and $\pi^f(W|\neg B) = 1\%$. In addition, he is spreading a conspiracy theory, according to which if he were to win votes would not be truthfully reported: $\pi^c(W|B) = 1\%$. Otherwise, the votes are counted randomly: $\pi^c(W|\neg B) = 50\%$.

For simplicity, assume that voters expect Bob to be fairly elected with either high probability $\mu_0(B) = 70\%$ or low probability $\mu_0(B) = 30\%$. Figure 3 shows the vectors of posteriors induced by the fair model (red point) and conspiracy theory (blue point), by prior. It is enough to compare the slopes of isofit lines associated with the available models to understand which model is adopted conditional on each signal. For example, consider a voter that initially expects Bob to win fairly with high probability (Figure 3a). The red point lies on the steeper isofit line and the blue point lies on the flatter isofit line: the voter would adopt f conditional on W and c conditional on $\neg W$, resulting in $\boldsymbol{\mu} = (\mu_W^f, \mu_{\neg W}^c)$ (purple point). This type of voter always supports Bob regardless the election outcome. Thus, Bob achieves his goal of always being supported by at least part of the voters. However, this has repercussions on other type of voters. Voters that expects Bob to win with low probability never supports him (Figure 3b).

Interestingly, for different priors, the same pair of stories not only induces opposite actions conditional on both signals, but also polarizes beliefs. Indeed, whenever $\mu_0(B) \geq 33\%$, the voter is persuaded to support Bob regardless of the election outcome, holding a strong belief of his legitimacy; the same stories persuade a voter with $\mu_0(B) < 33\%$ to never support Bob as president, always believing him to be an illegitimate president. To see this, it is enough to calculate for which prior p , the fair narrative is adopted when Bob is reported winning, i.e., $\Pr^f(W) \geq \Pr^c(W)$. That is, $p \cdot 99\% + (1 - p) \cdot 1\% \geq p \cdot 1\% + (1 - p) \cdot 50\%$. Then, the conspiracy

theory is adopted when Bob is not reported as winner.

4.1.1 Inevitable Polarization

The previous example illustrates how the exposure to conflicting stories might be a strong driver of inevitable polarization in a population of heterogeneous agents, working around the interpretations of a publicly revealed fact, and sowing discord within the electorate regardless of the election outcome. It is possible to generalize this result for the binary setting.

To formalize this, two models m, m' are *conflicting* if one is such that $\pi^m(s_1|\omega_1) > \pi^m(s_1|\omega_2)$ and the other one is such that $\pi^{m'}(s_1|\omega_2) > \pi^{m'}(s_1|\omega_1)$. In words, to be conflicting each model must point to a different state given the signal; thus, a signal assigns a qualitatively different meaning depending on the adopted model. This definition is independent of the receiver's prior, but it guarantees that, for every prior, the resulting posteriors induced by m and m' have the same relation with respect to the prior.²⁰

The following result shows that conflicting models always lead to belief polarization.

Proposition 5 (Binary case, Polarization). *For each pair of conflicting models, there exists a threshold in prior p such that, for every signal s , it holds that (i) $\mu_s(\omega_1) < \mu_0(\omega_1)$ if $\mu_0(\omega_1) < p$, and (ii) $\mu_s(\omega_1) > \mu_0(\omega_1)$ if $\mu_0(\omega_1) > p$.*

The intuition is the following. Any pair of conflicting models induces a vector of posteriors that is not Bayes-consistent, with both posteriors higher or lower than the prior. This follows from the fact that each signal triggers the adoption of a different model. Because models are conflicting, the updating goes always in the same direction. Crucially, the prior drives in which direction the posteriors are stretched: there is a threshold in prior such that receivers with prior higher (lower) than the threshold would hold extreme high (low) posteriors regardless of the signal realization.²¹ In the presence of receivers with priors higher and lower than the threshold, there cannot be consensus on the interpretation of any event and posterior beliefs always diverge. This is in stark contrast with respect to models with Bayesian agents with heterogeneous priors. Baliga et al. (2013) shows that beliefs do not move in opposite directions even when agents have different priors if they agree on the likelihood functions and update using Bayes rule.

From a broader perspective, there are different ways to measure polarization: ideological polarization (the extent to which the electorate has divergent beliefs on ideological issues), partisan sorting (the extent to which voters identify with a party, e.g., Levendusky, 2009; Mason, 2015), and affective polarization (the extent to which party members dislike members of other parties, e.g., Iyengar et al., 2019). I focus on the first as discussed in Dixit and Weibull (2007): posteriors on states shift in different directions depending on the prior. This type of polarization has been documented for many decades. In the ground-breaking paper by Lord et al. (1979) and similar subsequent studies (e.g., Plous, 1991; Darley and Gross, 1983; Russo et al., 1998),

²⁰In particular, it holds that $\mu_{s_1}^m(\omega_1) > \mu_0(\omega_1) > \mu_{s_2}^m(\omega_1)$, and $\mu_{s_1}^{m'}(\omega_1) < \mu_0(\omega_1) < \mu_{s_2}^{m'}(\omega_1)$. Graphically, the prior defines four quadrants: μ^m and $\mu^{m'}$ would lie respectively in the South-East and North-West quadrants.

²¹The proposition is silent on the indifference case where the prior equals the threshold. In that particular case, it is the case that the two conflicting models are lying on the same isofit. Thus, it could be the case that posteriors are either Bayes-consistent or not, depending on the tie-breaking rule.

subjects were asked to read the same study relative to a controversial issue (e.g., capital punishment, nuclear technology), then judge whether it provides evidence for or against the issue, and finally report how the study change their beliefs. They all find that participants’ final attitudes were either more in favor if initially favorable to the issue, or less in favor if initially opposed to the issue. To understand the determinants of this phenomenon, several mechanisms have been proposed in the literature. Often polarization is associated with confirmation bias, which was formalized for the first time by Rabin and Schrag (1999). They assume agents misinterpret new information as supporting of current beliefs with an exogenous probability. A recent paper by Fryer et al. (2019) builds on this, assuming a similar distortion only to signals open to interpretation, and provides evidence of their predictions. They directly assume the prior to be driving the direction of polarization, while in Rabin and Schrag (1999) this role is assigned to early observed signals as agents start with uniform prior over states. Baliga et al. (2013) provides an explanation for polarization based on ambiguity aversion, in which agents hedge against uncertainty making predictions in different directions depending on the prior after intermediate signals. Other papers illustrate how polarization arises with Bayesian updating in the presence of additional relevant features, such as high dimensionality of signal space with respect to state space (Andreoni and Mylovanov, 2012) or private signal on the interpretation of evidence (Benoît and Dubra, 2019). Recently, polarization has been discussed as amplified by mistakes concerning trust in information sources. Cheng and Hsiaw (2022) investigate the belief distortion due to double using the data: first to update beliefs on source credibility, and second to update beliefs on the state mistakenly using the updated beliefs on credibility instead of priors. Also, this mechanism can lead agents to disagree on how to interpret the same data. Gentzkow et al. (2021) shows how small bias in data perception due to ideological preferences can cause divergent beliefs about both state and source accuracy, even with Bayesian updating. Unlike these papers, I contribute to this literature on polarization by highlighting why such divergence in beliefs could happen given a strategic supply of conflicting stories, both given different priors observing the same fact and given the same prior observing different facts.

4.1.2 The Case of 2020 US Presidential Election

The debate on the fairness of the 2020 US election fractured the American electorate. No evidence was found supporting the claims of a widespread voter fraud in the election, yet competing narratives on dysfunctional elections were broadly diffused. Among the reasons for this, it stands out the fact that these allegations were circulating for the entire election campaign. In particular, the incumbent president at the time Donald Trump cast doubts on the election system, especially on the mail-in ballots, well ahead of the election results — amplifying the spread of such theories. When ballots were tabulated, some voters adopted these narratives to interpret the election outcome, concluding the election to be rigged.

The preemptive provision of an alternative narrative with respect to the conventional belief that the election system is fair fits well the application discussed in the previous section. In what follows, I discuss some suggestive evidence of the mechanism I formalized above: conflicting stories drive polarization in a population of heterogeneous voters. In particular, when exposed to conflicting stories: (i) voters with different initial beliefs adopt different narratives on the election system once the signal realizes, and (ii) voters with the same initial belief adopt different

narratives if they observe different signals. Persily and Stewart (2021) provides some insights on the 2020 US election with attention to confidence in election fairness that speak to these two predictions. To allow comparability between the setting of this paper and the American bipartisan system, I assume that each voter expects his partisan candidate to win, i.e., before the election Republicans (Democrats) expect Donald Trump (Joe Biden) to win. On average, this assumption is verified: at the end of October 2020, the expected winner of the presidential election for Republicans (Democrats) was Biden with 6% (73%), and Trump with 85% (8%).²²

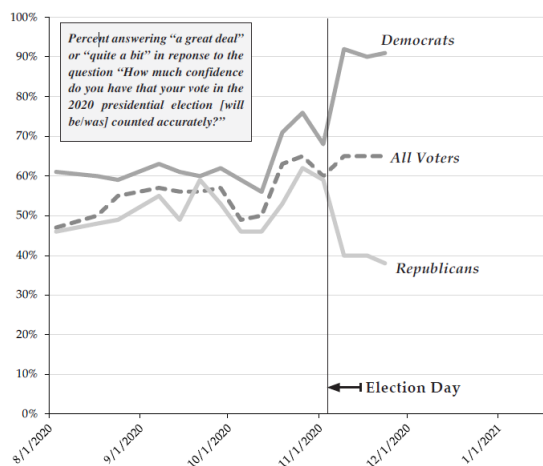


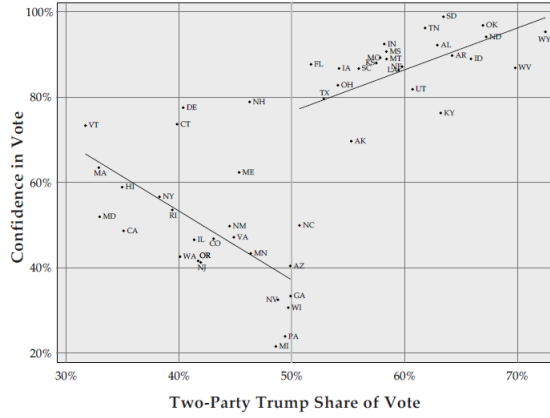
Figure 4: Accuracy of Vote Count (Persily and Stewart, 2021)

Notes: The y-axis shows the percentage answering “a great deal” or “quite a bit” in response to the question “How much confidence do you have that your vote in the 2020 presidential election [will be/was] counted accurately?”
Source: Economist/YouGov poll, 2020.

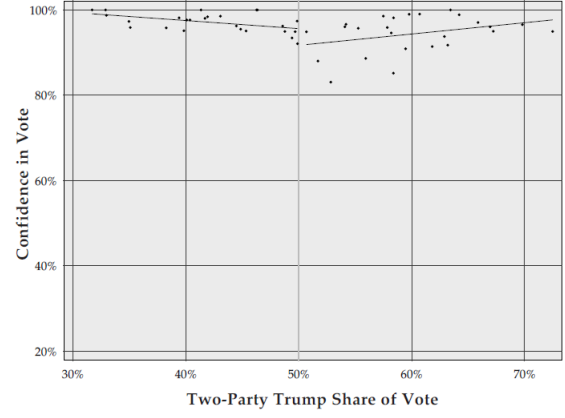
Figure 4 shows the confidence in accurate vote count over time. Persily and Stewart (2021) report that before the election around half of poll respondents expressed confidence that their own vote would be counted accurately, with Democrats slightly more confident than Republicans. After the release of the election outcome, while the aggregate measure remained unchanged, a drastic partisan polarization occurs: the gap between Democrats and Republicans went from 10.9 points to 51.7 points.²³ This suggests that voters with different priors adopt different narratives once the signal realizes: after the election, Democrats adopt the narrative claiming the election system to be fair, while Republicans adopt an alternative story questioning the integrity of the process. This effect is not unique to the 2020 election and it is also known as “winners-losers effect”: after the election supporters of the losing candidate tend to question the legitimacy of the election, while supporters of the winning candidate tend to gain confidence in the election system (Sances and Stewart, 2015; Sinclair et al., 2018). However, the 2020 gap is much wider than in previous elections. A potential explanation is the disproportionate spread

²²More details about the distribution of priors can be found in Appendix B. This pattern in priors is consistent with motivated beliefs or wishful thinking: voters wish their partisan candidate to win and this influences their expectations. I am assuming that these motives might affect prior beliefs, but not model selection or the updating procedure. A proper test of this paper should account for these confounding forces.

²³The same pattern can be observed regarding a similar question: “How much confidence do you have that the 2020 presidential election [will be held/was held] fairly?”. Along this measure, the pre-election gap was 15%, while the post-election one was 72.6%. The figure is reported in Appendix B.



(a) Republicans



(b) Democrats

Figure 5: Confidence in vote count in state elections (Persily and Stewart, 2021)

Notes: The y-axis shows the percentage answering “very confident” or “somewhat confident” in response to the question “How confident are you that votes in [state of residence] were counted as voters intended?”

Source: Survey of the Performance of American Elections (SPAEE), November 2020.

of distrustful narratives during the 2020 election campaign with respect to previous elections.²⁴

Suggestive evidence about the second prediction can be found by looking at how voters’ confidence in state election change depending on the reporter election outcome in the state. Figure 5 reports data on the confidence in state elections by the percentage of Trump share of votes, by party. Republicans mostly distrust the accuracy of the state elections if resident in states where Trump barely lost. The discontinuity in confidence vote between Republicans from states in which Trump barely lost and those from states in which Trump barely won is stark, and larger than in previous elections.²⁵ This gap supports the idea that voters with similar initial beliefs adopt different narratives once observed different outcomes when conflicting stories are proposed.²⁶ The same pattern is not found for Democrats, barely exhibiting a discontinuity. Since most of these alternative narratives about the election accuracy were right-leaning — supported by Donald Trump, and most spread in the Republican party —, it is reasonable to assume that Democrats discard them. Indeed, people tend to ignore messages inconsistent with their view or coming from sources perceived as untrustworthy (Graber, 1984). This could explain this difference among parties and it suggests an interesting extension of this paper.

4.2 Financial Advice

Next, I illustrate the optimization problem for a financial advisor that wants to persuade investors to make a specific investment. In this regard, it is well-known that commissions on

²⁴Persily and Stewart (2021) reports that the gap in voter confidence in one’s own vote being fairly counted (confidence in the overall fairness of the election) from before to after election was in 2004, 2008, and 2012 elections on average 22% (32%), in 2016 17% (18%), and in 2020 41% (58%).

²⁵Clark and Stewart (2021) reports this gap to be 15% in 2012, 20% in 2016, and 40% in 2020.

²⁶This claim concerns the voters at the threshold. Indeed, focusing on state elections, priors might also take in consideration the political tendency of the resident state, e.g., expectation of Trump’s victory in California compared to Idaho. At the threshold, these initial expectations should be comparable.

investments could lead to a conflict of interests for the advisor. I consider the case in which the advisor knows that the investors' past financial experience influence their beliefs about the quality of new investments.²⁷ However, the advisor does not have access to this piece of private information from investors. Nonetheless, he has the incentive to persuade the receiver to invest as much as possible.

Exposing the investor to multiple narratives on how to predict future returns based on past returns, the advisor can move the investor's beliefs always in an advantageous direction compared to the prior. Two important trading strategies in finance are mean-reversion and momentum. According to mean-reversion, an asset price moves close to its average over time; thus, given a deviation in one direction, one should expect the opposite direction. According to momentum, an asset price is expected to continue moving in the direction of its trend, making past deviations important predictors. Both strategies are effective in different situations. On one hand, professionals use both relying on empirical measures to determine the most effective. On the other hand, an inexperienced investor might interpret past financial performance through the strategy that resonates best with the available information and initial beliefs. The advisor can use simplified versions of these strategies to his advantage, proposing to investors both investment strategies, and letting these inconsistent ways of looking at the same information persuade investors to his advantage. As shown formally in the next example, an investor with favorable expectations toward the advisor-preferred asset will always fully invest in that asset, because any past data trigger the adoption of the most optimistic narrative in terms of future performance. However, if the investor is pessimistic about the advisor-preferred asset, communicating these two narratives is counterproductive and the advisor needs to adjust his communication.

I discuss these insights in the context of choosing a hedging strategy. Hedging has the purpose of limiting the risk of uncertain events on financial assets. Usually, it involves diversification in offsetting or opposite positions. To provide a graphical illustration of the financial advisor's optimization problem, I focus on the binary setting.

Investors need to decide how much to invest: each has to allocate his 1 unit of endowment over two possible outcomes, $\Omega = \{N, \neg N\}$, where N is the event under normal conditions and $\neg N$ is the event if extreme or catastrophic circumstances occur, e.g., flood or drought as extreme weather events, or financial crisis or fluctuations of foreign currency in the economic domain. This results in the choice of $\alpha = (\alpha_N, \alpha_{\neg N})$ with $\alpha_N + \alpha_{\neg N} = 1$. All investors have the same initial beliefs and I consider two cases: *optimistic* investors expecting normal conditions with probability 70% or *pessimistic* investors expecting extreme conditions with probability 70%. Also, each investor had a previous experience either with normal conditions (good experience, G) or with extreme conditions (bad experience, B) and they try to understand how this previous experience impacts the future one. Assuming a logarithmic utility over the outcomes, the investor's expected utility based on her posterior μ_s is $\mathbb{E}[U^R(\alpha)] = \sum_{\omega \in \{N, \neg N\}} \mu_s(\omega) \log(\alpha_\omega)$. The investor's optimal action α^* is to allocate a proportion of the endowment equal to the corresponding posterior $\alpha_\omega^* = \mu_s(\omega)$ for each outcome ω .

With misaligned incentives, the financial advisor receives a commission r_ω proportional to the receiver's allocation on outcome ω , i.e., $U^S(\alpha) = r_N \alpha_N + r_{\neg N} \alpha_{\neg N}$. Assuming $r_N > r_{\neg N} = 0$,

²⁷There is empirical evidence that personal experiences have a lasting impact on beliefs and behavior, such as the effect of having lived through a depression affects stock-market participation (Malmendier and Nagel, 2011).

his value function is $V(\boldsymbol{\mu}) = \sum_{s \in \{G, B\}} \Pr^t(s) r_N \mu_s(N)$. In this example, I relax the common prior assumption to better compare groups of investors with different priors, with the sender expecting normal conditions with only a probability of 40%. Moreover, the sender knows the true model where a positive past experience positively (negatively) correlates with the success (failure) of the investment with some noise: $\pi^t(G|N) = 75\%$ and $\pi^t(G|\neg N) = 25\%$. Given this, he expects that investors with past good experience with probability 45% and with a past negative experience with probability 55%.²⁸

Anticipating the optimal investor's allocation rule, the financial advisor wants to manipulate the investors' posteriors to maximize his profits. With a graphical approach, Figure 6 shows the financial advisor's indifference curves plotted on the feasible vectors of posteriors given the investors' prior. The darker the colored area, the higher the advisor's expected utility. To pin down the target vector of posteriors to maximize the advisor's value, it is enough to identify the point or area of the darkest color, that is, the vector of posteriors whose value is the highest among the feasible ones. Notice that the advisor's indifference curves are driven by the true model, his prior, and his incentives, on top of the investors' prior and incentives.

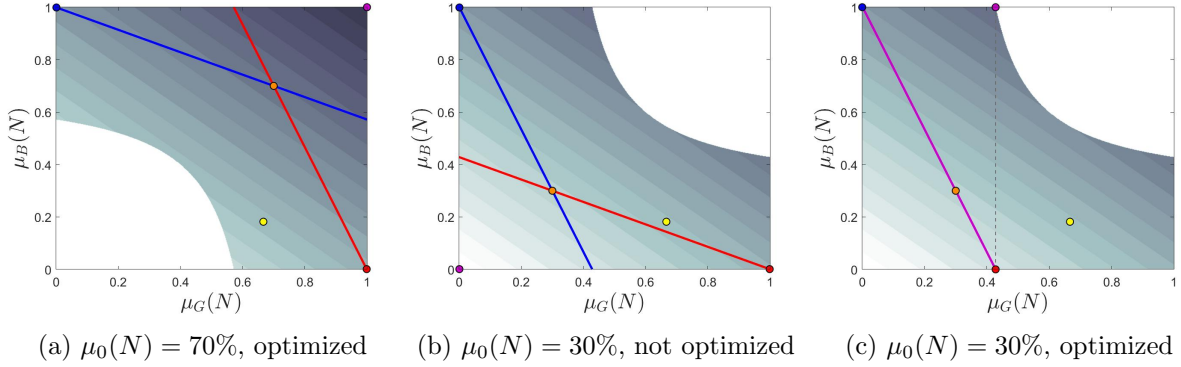


Figure 6: Financial Application

Notes: the financial advisor's value is illustrated for $r_N = 1$; the darker the colored area, the higher the sender's expected utility; the orange point is the investors' prior and the yellow point represents the advisor's vector of posteriors induced by his prior and true model; the red and the blue points are respectively the vector of posteriors induced by models m_1 and m_2 , while the purple point is the resulting vector of posteriors given these models.

Consider the optimistic investors (Figure 6a). The financial advisor does not want to discard the investor's experience as irrelevant, otherwise, the investors' beliefs would remain at the prior with an investment of $\alpha_N = 70\%$. Using multiple models, the advisor can attain a higher value. Indeed, the highest value for the advisor is achieved at the top-right corner where an optimistic investor always expects normal conditions and does not hedge against extreme circumstances — set always $\alpha_N = 100\%$. Intuitively, this means that the advisor can leverage any past experience of the investor and move her beliefs always in the advantageous direction. From the target vector, the advisor can construct the optimal set of models to communicate. Since the target is not Bayes-consistent, he needs two models to induce it. One option is to expose the investors to the following pair of models: (1) model m_1 suggesting a perfect positive correlation

²⁸It is possible to interpret this setting in a two-fold manner: (i) a financial advisor wants to manipulate an investor that had a good experience with probability 45% and negative with probability 55%; or (ii) a financial advisor wants to manipulate a population of investors with heterogeneous past experience, 45% had a positive experience and 55% had a negative experience. In this section, I adopt the second.

between past and future conditions, i.e., $\pi^{m_1}(G|N) = \pi^{m_1}(B|\neg N) = 1$ (red point), and (ii) model m_2 suggesting a perfect negative correlation between past and future conditions, i.e., $\pi^{m_2}(B|N) = \pi^{m_2}(G|\neg N) = 1$ (blue point). These can be interpreted as extremely simplified versions of the momentum (“early success predicts long-run success”) and mean-reverse theories (“what goes down goes up”). Because of their optimism, investors adopt the first given a good experience and the second given a bad experience. As a result, they fully invest regardless of past experience, never hedging against adverse events.

Instead, manipulating a pessimistic investor is not that easy. First, full investment is not attainable with pessimistic investors. The vector of posteriors in the top-right corner is not feasible given their prior (Figure 6b and 6c). Second, communicating the same pair of models tailored to the optimistic investors to the pessimistic ones is self-defeating. A pessimistic investor always adopts the most pessimistic story — model m_1 given a negative experience and model m_2 given a positive experience — and never invests in the advisor-preferred outcome (Figure 6b).

With a pessimistic investor, the maximal value the advisor can achieve is attained with the vector at the right-top kink (Figure 6c). Since this lies on the upper frontier $\boldsymbol{\mu}^* = ((0.43, 0.57), (0, 1))$, there exists only a combination of models to induce it. To see this, I draw the isofit corresponding to the maximal fit of $\mu_G(N) = 0.43$ (red line): only the point with $\mu_B(N) = 1$ lies on it. The optimal communication strategy is to entertain two models: (i) model m_1 such that $\pi^{m_1}(B|N) = 0$ and $\pi^{m_1}(G|\neg N) = 0.57$, inducing $(0, 0.43)$ (red point), and (ii) model m_2 as defined above for the optimistic investor, inducing $(0, 1)$ (blue point). According to model m_1 , a bad experience is a perfectly revealing signal of extreme conditions next, while a past positive experience is an ambiguous news because it is always the case given future normal conditions and it occurs more than randomly given future extreme conditions. This model encourages only investors with a positive experience to have a positive α_N and, indeed, it is tailored to those. Also in this case, model m_2 is an oversimplified version of the mean-reverse theory which pushes an investor with bad experience to fully invest in the advisor-preferred outcome. If investors are exposed to both these models, investors adopt the one that resonates best with them and the financial advisor will be able to maximize his profits. To conclude, note that since the first-best of convincing all pessimistic investors to fully invest in outcome N is not possible, the advisor shifts to the second-best: convincing the largest group of investors to fully invest in N (the ones with bad experience), while increasing α_N for the other group (the ones with good experience) as much as he can.

How much would the financial advisor be willing to pay to know the investors’ experience? This information would allow the advisor to perfectly target each group of investors with a tailored model able to convince them to fully invest in his preferred outcome. Thus, the advisor would not need to worry about how models compete with each other, communicating to each group of investors a model tailored to their past experience. This means inducing the vector of posteriors $\bar{\boldsymbol{\mu}} = ((1, 0), (1, 0))$ (right-upper corner in the figures). With pessimistic investors, $\bar{\boldsymbol{\mu}}$ is unfeasible with ex-ante provision of models. Therefore, the cost of ex-ante commitment is the gap between the maximal payoff to the sender $V(\bar{\boldsymbol{\mu}}) = r_S$ and the maximal value subject to the feasible set $V(\boldsymbol{\mu}^*) = 55\%r_N + 45\%(43\%r_N) = 74\%r_N$. In contrast, the cost of ex-ante commitment is zero with optimistic investors because the sender can always achieve his maximal payoff.

4.3 Self-Persuasion

Model persuasion can shed light on intra-personal phenomena as well. In this section, I contribute to the literature on motivated beliefs, discussing how an agent could distort her own beliefs by manipulating the perceived informativeness of observable signals. This proposed mechanism to achieve self-serving beliefs can deliver the classic implications of this literature, but it also gives a bound on belief distortion.

Papers on motivated beliefs conjecture different sources of motivations or different channels through which beliefs are distorted, e.g., via direct utility of beliefs/consumption of beliefs (e.g., Köszegi, 2006; Brunnermeier and Parker, 2005) or via instrumental value associated with the beliefs (e.g., Bénabou and Tirole, 2002). In general, these papers share the assumption that individuals search their mind for beliefs that are either positive or accurate, *demand side*, and at the same time their mind is able to distort beliefs from the truth subject to some reality constraints, *supply side* (for a survey see Bénabou, 2015). Analogously, I applied the ex-ante model persuasion in a multi-selves context, perhaps at an unconscious level, where the conscious mind (receiver) demands the unconscious one (sender) to supply beliefs in order to evaluate the decision. Depending on the agent's subconscious preferences, goals, or sophistication, the beliefs might be distorted in different directions. For instance, consider an individual having a subconscious supply of beliefs that boosts her beliefs in a desired direction regardless of the observed data. As a consequence, the individual would tend to pick the interpretation that best resonates with her initial beliefs, leading to *confirmation bias*. This could be the case of a student that likes to think to be good at school and thus leaves the informativeness of grades open to two interpretations: either grades are a good measure of own ability or grades are based on luck. If she expects to be good at math, once received a good grade on the math test she adopts the interpretation that grades are based mostly on ability. If she receives a bad grade, only bad luck could explain such a grade given her ability. Keeping the grade open to multiple interpretations allows the student to keep high self-esteem. However, her initial beliefs constrain to which extent she can manipulate her beliefs in the desired direction across possible contingencies.

These examples assume quite different goals for the sender and the receiver, such as higher self-confidence and accuracy. To highlight the mechanism, the next example focuses on an example where there is a variable controlling the degree of misalignment of incentives among the parties. This allows me to compare the consequences of aligned and misaligned preferences. In particular, I explore a multi-selves model in which an agent has incentives to distort her own interpretations of signals to offset her time-inconsistent preferences and commit to a costly task. As self-control problems are well-documented (Read and Van Leeuwen, 1998; Ariely and Wertenbroch, 2002; Augenblick et al., 2015), choosing a well-suited explanation for future events allows individuals to overcome these issues to maintain confidence.

I build on the motivation problem of Bénabou and Tirole (2002). They consider a risk-neutral individual with quasi-hyperbolic discounting (e.g., Laibson, 1997; O'Donoghue and Rabin, 1999). At $t = 0$, the individual can take an action $a \in \{0, 1\}$ that potentially affects her information at $t = 1$ with some utility flow. Based on this, at $t = 1$ she decides whether to take an action with disutility c that, if successful, would yield benefit v at $t = 2$. The probability of success,

discussed as self-confidence by the authors, may depend on either new information received or forgotten. Adopting this setting, I focus on the case in which the individual at $t = 0$ acts as the sender, choosing potential interpretations of the future signal to be received, while the receiver is the individual at $t = 1$. Even if their preferences are the same, their incentives are misaligned because of time inconsistency: at date 0, both the cost and the benefit are in the future and present bias would not undermine his choice to act, unlike in the next period. Formally, at date 1 no action a leads to $U^1(0) = 0$, while taking the action given s leads to

$$U^1(1) = u_1 + \beta\delta \mathbb{E}[u_2] = -c + \mu_s(\text{success}) \beta\delta v,$$

where $\delta \leq 1$ is the discount factor and $\beta > 0$ is the present bias. At date 1, the receiver would take the action if $\mu_s(\text{success}) \geq \frac{c}{\beta\delta v}$. Instead, the sender's utility when taking the action is

$$U^0(1) = u_0 + \beta \mathbb{E}[\delta u_1 + \delta^2 u_2] = u_0 + \beta\delta(-c + \mu_s^t(\text{success}) \delta v),$$

while $U^0(0) = u_0$. Hence, for the sender the action is optimal if $\mu_s^t(\text{success}) \geq \frac{c}{\delta v}$, which is lower than the relevant threshold of the receiver if she suffers from present bias, e.g., $\beta < 1$. Hence, the sender has the incentive to distort the receiver's interpretations of signals in order to avoid future lack of willpower.

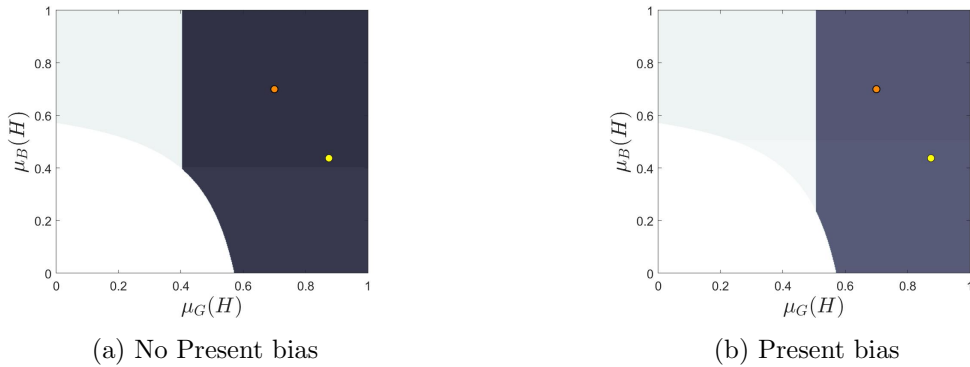


Figure 7: Motivated beliefs, time inconsistency

Notes: the example is parametrized as $c = 4$, $v = 10$, $\delta = 0.99$, and $\beta = 0.8$; the darker the colored area, the higher the sender's expected utility; the orange point is the prior and the yellow point represents the sender's vector of posteriors induced by the true model.

Building on this, I focus on a binary setting where the success depends on whether the individual's ability is high (H) or low (L). The signal she receives can be good (G) or bad (B). The agent initially believes to have high ability with probability 70%, while deep down she knows the signals are quite accurate, $\pi^t(G|H) = \pi^t(B|L) = 75\%$. Taking the action is always optimal at her initial beliefs and also when she does not suffer from present bias. Indeed, updating her prior using the true model maximizes her ex-ante expected payoff (Figure 1a). However, self-deception could be beneficial in the case of sufficiently severe present bias (Figure 1b). Assume that the agent is sophisticated in knowing her present bias. Before the signal, she knows that at the moment of the decision the imminent cost of the action, c , will be more salient than the future reward of a success, v . Thus, her confidence conditional on the bad signal will not be high enough to take the action. She can overcome this by distorting the perceived informativeness of the signal she is about to receive — either discarding the signals as uninformative

or believing only the good signal to be accurate enough. Belief manipulation would allow her to stay motivated through manufactured overconfidence. Unlike Bénabou and Tirole (2002), this is achieved by distorting the perceived signal's informativeness rather than assuming an memory loss or inattention.

5 Extension: Default Model

So far I assumed the receiver to only consider models proposed by the sender. In this section, I allow the receiver to initially hold a model, hereafter called *default model*: she considers also her default model on top of the models she is exposed to. This is a natural and realistic extension, as often individuals bring ways of interpreting facts either generated on their own or provided by others in past. As a result, the default model restricts the extent to which the receiver can be persuaded.

Formally, consider that the receiver is endowed with a default model d known by the sender. The set-up is the same as in Section 2, with the only exception that the receiver adopts the model with the highest fit conditional on the observed signal s among the set of models M she has exposed to and her default model: $m_s^* \in \arg \max_{m \in M \cup \{d\}} \Pr^m(s)$.

The following theorem characterizes the set of feasible vectors of posteriors in the presence of a default model.

Theorem 2 (Default, Many Models). *If $|M| \geq |S|$, then the set of feasible vectors of posterior beliefs is*

$$\mathcal{F}^d = \left\{ \mu \in [\Delta(\Omega)]^S : \forall s \in S, \bar{\delta}(\mu_s)^{-1} \geq \Pr^d(s) \right\}.$$

The proposed models do not only compete with each other but also with the receiver's default model given each signal realization: the better the fit of the default model given a signal, the less the sender can move beliefs given that signal. Note that the feasible set does not depend on the vector of posteriors induced by the default model, but only on the fit levels of the default model. As expected, the default model restricts the belief manipulability of the receiver: the constraint in Theorem 2 is tighter than in Theorem 1. To see this, take $\mu \in \mathcal{F}^d$. For every signal s it holds that

$$\bar{\delta}(\mu_s)^{-1} \geq \Pr^d(s) = 1 - \sum_{s' \neq s} \Pr^d(s) \geq 1 - \sum_{s' \neq s} \bar{\delta}(\mu_{s'})^{-1},$$

satisfying the condition of Theorem 1.

As a matter of fact, the two theorems are closely related as shown next: the set of the feasible vectors of posteriors in absence of a receiver's default model (Theorem 1) is the union of all the sets of the feasible vectors of posteriors with a default model for every default model.

Proposition 6. *The union of \mathcal{F}^d for all default models is \mathcal{F} , i.e.,*

$$\bigcup_{d \in \mathcal{M}} \mathcal{F}^d = \mathcal{F}.$$

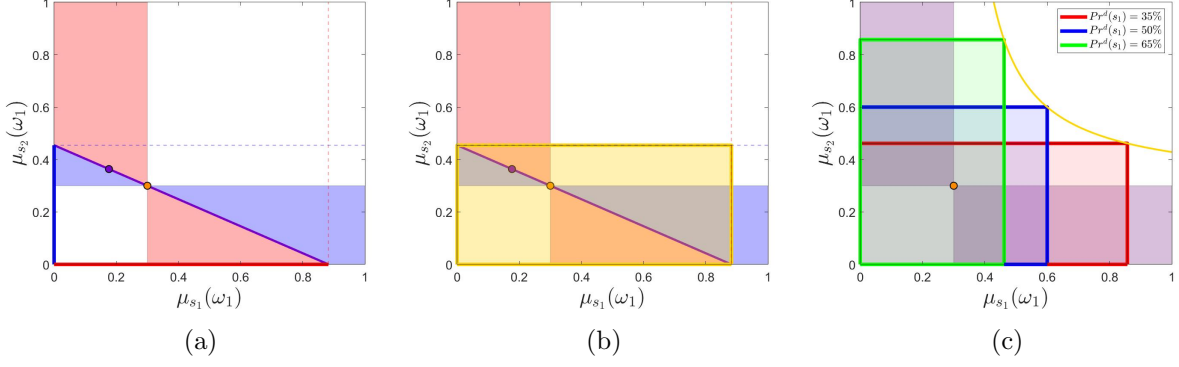


Figure 8: Graphical intuition of Theorem 2 & Proposition 6

Figure 8 provides a graphical intuition of these findings for the binary case. Consider a default model corresponding to the vector of posteriors μ^d depicted by the purple point in Figure 1b. Given this, the red area corresponds to models with higher fit conditional on s_1 and the blue area corresponds to models with higher fit conditional on s_2 with respect to d . Thus, the compatible posterior distributions conditional on s_1 and the compatible posterior distributions conditional on s_2 are respectively the ones on the red line and the blue line, in Figure 8a. Given the default model, the feasible vectors of posteriors are the ones in the yellow area in Figure 8b. The figure clarifies also why all the default models with the same fit levels — corresponding to that same isofit (purple line) — induce the same set of feasible vectors of posterior. Finally, Figure 8c helps building intuition for Proposition 6. The yellow line corresponds to the upper frontier of the feasible set of posterior beliefs for a receiver without a default model, while the colorful areas correspond to the feasible set of posteriors for a receiver with default models of fit given signal s_1 of 35% (red), 50% (blue), and 65% (green).

I conclude this section by discussing how the receiver's default model affects the cost of ex-ante commitment. In absence of the default model, every posterior is feasible. This is not the case if the receiver is endowed with a default model and Proposition 1 of Schwartzstein and Sunderam (2021) characterizes the feasible posterior beliefs in this setting. Given signal s , the posterior μ_s can be induced if $\bar{\delta}(\mu_s)^{-1} \geq \Pr^d(s)$. Theorem 2 generalizes this result for vectors of posteriors for the ex-ante stage. Providing multiple models ahead of time guarantees to induce any receiver's posterior achievable by communicating a model after observing the signal realization. Consequently, ex-ante commitment does not prevent the sender to attain the same expected utility he would get for each signal realization with ex-post communication, and thus has no cost for the sender. To formalize this, given a default model d , the sender's cost of ex-ante commitment is the gap between the maximal sender's value over ex-post feasible vectors of posteriors and the maximal sender's value over ex-ante feasible vectors of posteriors:

$$\Delta^d = \underbrace{\max_{\mu \in \text{post-}\mathcal{F}^d} V(\mu)}_{\text{no commitment}} - \underbrace{\max_{\mu \in \mathcal{F}^d} V(\mu)}_{\text{commitment}},$$

where $\text{post-}\mathcal{F}^d = \{\mu \in [\Delta(\Omega)]^S : \forall s \in S, \omega \in \Omega, \mu_s(\omega) \leq \frac{\mu_0(\omega)}{\Pr^d(s)}\}$.

Corollary 2. *With default model and providing multiple models, ex-ante commitment does not restrict the sender's value: $\Delta^d = 0$.*

Two remarks follow. First, the presence of a default model guarantees this result. Without a default model, the sender can induce any posterior beliefs ex-post as discussed in Section 3.3 and the commitment cost might or not be zero. Therefore, if the sender has the option to avoid the ex-ante commitment, the receiver’s default model plays a crucial role in restricting belief manipulation. Importantly, this abstracts welfare implications. If the receiver holds the wrong default model, the sender might never be able to induce truthful beliefs, either with ex-ante or ex-post communication. Second, this corollary does not imply that the sender should communicate ex-ante any set of models that he would find optimal ex-post for each signal. This could be self-defeating in some cases. With binary signal, the optimal ex-post models always work ex-ante. To see this, consider any two tailored models, m_1 for s_1 and m_2 for s_2 . It is enough to notice that $\Pr^{m_1}(s_1) \geq \Pr^{m_2}(s_1)$ implies $\Pr^{m_2}(s_2) \geq \Pr^{m_1}(s_2)$. However, with more than two signals, the sender might need extra care in choosing the models to communicate. Consider the following example with three signals and the three tailored models, whose fit levels are displayed in Table 1. While model m_2 is correctly adopted conditional on s_2 , the receiver would select model m_1 after observing both s_1 and s_3 . Even if these three models are ex-post optimal, there are not optimal ex-ante because they compete with each other in each signal realization.

	$\Pr^m(s_1)$	$\Pr^m(s_2)$	$\Pr^m(s_3)$
m_1	60%	10%	30%
m_2	20%	70%	10%
m_3	30%	50%	20%

Table 1: Fit levels for three tailored models

Notes: bold text highlights the higher fit level, corresponding to the selected model, for each signal realization.

5.1 Merchants of Doubt

“Doubt is our product, since it is the best means of competing with the ‘body of fact’ that exists in the minds of the general public. It is also the means of establishing a controversy.”

— Cigarette Executive (1969)

“Victory will be achieved when average citizens understand uncertainties in climate science.”

— Internal memo by The American Petroleum Institute (1998)

The strategic communication of an alternative narrative with respect to a commonly shared one might be used to deceive the public. This strategy was used by the tobacco industry and oil companies to challenge a well-established way of looking at the scientific evidence and to manufacture uncertainty on issues like the health effects of smoking and climate change (e.g., Michaels, 2008; Oreskes and Conway, 2011). These so-called “Merchants of Doubt” established a trustworthy presence in academia and media to discredit peer-reviewed articles providing alternative ways of interpreting the emerging scientific evidence (e.g., blaming other factors, false positive results).²⁹ Their aim was to delay regulations, defeat delegations, and insinuate doubt in the population (as emerged by quotes from internal memos reported above). Ultimately, they

²⁹On climate change, the most common stories range from “there is no climate change” and “if there is climate change, it is not man-made”; while on smoking the most heard is “other factor plays a role such as family history”.

diluted scientific consensus, despite the scientific community having no doubt. This strategy guaranteed that even with new evidence emerging over time, the general public has already been exposed to these competing ways of interpreting new facts and found both worthy of consideration. The following example simplifies this setting and illustrates how a competing narrative leads to polarization even when the population initially shared the same view.

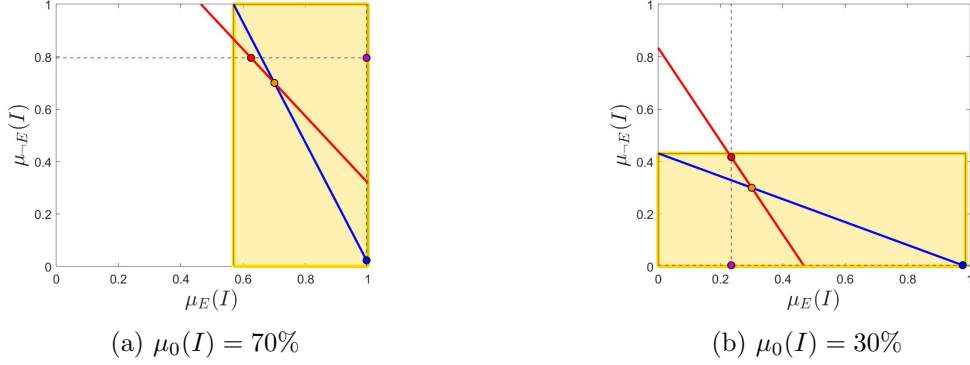


Figure 9: Merchants of Doubt, for different receiver's prior

Consider two states $\{I, \neg I\}$, where I is the event that the issue is real, e.g., smoking causes cancer or climate change. New evidence emerges, either in favor of the issues (E) or not ($\neg E$). By default, individuals trust science: favorable evidence means the issue is confirmed, and vice versa if unfavorable). Formally, assume $\pi^d(E|I) = \pi^d(\neg E|\neg I) = 99\%$ (blue point).

A lobby wants to strategically challenge this shared narrative in the population to induce disagreement on the issue. Even if agents all start with the same model, their prior plays an important role in determining the set of feasible vectors of posteriors. Figure 9 shows the feasible vectors for two different levels of prior, either supportive agents with $\Pr(I) = 70\%$ or skeptical agents with $\Pr(I) = 30\%$. Assume that the lobby's stand is that, if the issue is true, evidence emerges randomly ($\pi^m(E|I) = 50\%$), but there is a chance of false positive if the issue is not true because science searches for evidence in that direction ($\pi^m(E|\neg I) = 70\%$). The vector of posterior induced by this model corresponds to the red point and the resulting vector of posterior given both models corresponds to the purple point.

On one hand, the narrative supporting the scientific evidence induces a vector of posterior which is almost identical across initial beliefs; however, the feasible beliefs differ drastically even if the default model is shared. On the other hand, once the alternative narrative is introduced, agents with different priors end up with very different posterior regardless of the evidence — effectively prompting doubt and igniting a debate among agents with different priors. This illustrates how limited is the impact of a shared initial model in preventing polarization. On the contrary, this model can be used to deepen polarization by strategically providing a conflicting alternative.

6 Literature Review

This paper mostly contributes to two strands of literature. First, it contributes to the literature on narratives in economics. Second, it contributes to the rich literature on persuasion in economic theory. Also, it relates to the important literature on biased beliefs.

Narratives Recently, there has been an effort to incorporate narratives in economics, also prompted by Shiller (2017, 2019). This paper builds on the formalization of narratives as models as defined in Section 2, introduced by Schwartzstein and Sunderam (2021). They investigate what the receiver could be persuaded of when the sender can communicate an alternative model after the release of public data. A persuasive model is the one that fits better the observed data with respect to the receiver’s default model.³⁰ I adopt this model selection rule and build on their framework moving the strategic provisions on models before the signal realization. In Section 5, I provide a direct comparison and discussion of the effect of this ex-ante commitment for the sender. Recent papers build on this notion of narratives and the assumption that best-fitted models are adopted. Ichihashi and Meng (2021) sequentially combines model and Bayesian persuasion. They study a persuader that can design (choose the experiment that generates the signal) and then interpret information to persuade the receiver once the signal realizes. They characterize the sender’s optimal strategy using the concavification approach and show that can lead to a higher expected payoff than in Kamenica and Gentzkow (2011). Schwartzstein and Sunderam (2022) studies the social exchange of models in networks: agents start with a default shared model, but each comes up with a better interpretation given the release of data, shares it within their network, and then picks the best-fitting model among all shared ones. Social learning leads agents to have beliefs closer to the prior and feel better able to explain data than before. Moreover, two recent papers take different approaches to what makes models persuasive. Spano (2022) explores the following setting: before the signal realizes, the sender communicates a model and the receiver adopts the proposed model if it is coherent (conditional on a state, probabilities of each possible news sum to unity) and compatible with her default (true) model (the marginal distribution of news is undistorted).³¹ He argues that coherence implies Bayes-consistent posteriors across signals, but limits the scope of manipulation. Instead, Yang (2022) proposes a preference for “decisive models,” models that provide a strong recommendation regarding the best course of action.

Another influential way to formalize narratives is describing them as subjective causal models, expressed as directed acyclical graphs (Spiegler, 2016). Eliaz and Spiegler (2020) assumes agents prefer “hopeful narratives” that are empirically consistent, i.e., narratives that maximize anticipatory utility and correctly predict the empirical distribution of consequences. Their analysis focuses on the equilibrium as a long-run distribution over narrative-policy pairs. Instead, Eliaz et al. (2021b) studies to what extent a misspecified model can distort pairwise correlations between variables. In this paper, an analyst has incentives to show a strong correlation between two variables and he can propose a (mis)specification model to estimate it, fitting the data. The authors quantify the worst-case distortion when the proposed model is flexible in which variables enter the model and how, but it must be such that the estimated model cannot distort the marginal distributions of individual variables.³² Direct acyclical graphs are also used by

³⁰Closely related, Levy and Razin (2020) study aggregation of forecasts over time. They assume the decision maker to look for the most likely explanation for what she observes. Explanations are information structures consistent with previous forecasts and her prior. Thus, the signal space could vary across explanations — indeed, the analysis can be reduced to information structure with a binary signal. Similar to my results, the prior plays a crucial role in the evolution of beliefs.

³¹I assume models to be coherent by definition. However, one could argue that the receiver might hold an incoherent model ex-post. This follows from the sender proposing multiple models and how the receiver selects models across signals. To compare results, coherent and compatible models can only induce vectors of posteriors corresponding to the isofit line of the true model.

³²They find that increasing the number of variables in the model can lead to almost perfect correlation.

Eliasz et al. (2022) to study the proliferation of false narratives and their effect on political mobilization in a heterogeneous society of multiple social groups, and by Horz and Kocak (2022) to explore which conditions affect the effectiveness of authoritarian propaganda in reducing citizens’ protests. Other formal frameworks in which narratives have been defined are Bénabou et al. (2018), where narratives are described in terms of moral value, and Izzo et al. (2021), where narratives describe the linear relation between policies and their outcome.³³

Moreover, there are some recent experimental studies inspired by these formal notions of narratives (Barron and Fries, 2022; Charles and Kendall, 2022). In particular, Barron and Fries (2022) study narrative provision and adoption in a financial advice setting, building on an example discussed in Schwartzstein and Sunderam (2021).³⁴ They find that advisors with misaligned incentives communicate narratives biased from the truth, and are successful in manipulating investors’ beliefs in the desired direction. In particular, narratives that better fit the available data are more persuasive, in the sense that they move investors’ beliefs closer to the beliefs induced by the provided narrative. Hence, narratives are a highly effective tool of persuasion and hard to protect against. Moreover, in one treatment, the advisors do not have the opportunity to send tailored narratives because they have no access to the data the investors observe — shifting to a setting similar to the one studied in this paper. Results show that this treatment does not move beliefs closer to the truth, but persuaders are less effective in moving beliefs to their target. Unlike this paper, the persuader is restricted to one narrative and this might result in choosing a narrative that is not the one moving beliefs the most in the desired direction conditional on what it is actually observed. Moreover, in the experiment the information receivers have is a rich historical performance of 10 data points. The complexity of anticipating all possible observable dataset might reduce the ability of the sender to reason across possible contingencies observable by the receivers, leading to a suboptimal narrative. Overall, their paper is supportive of the framework introduced by Schwartzstein and Sunderam (2021) and adopted in this paper.

Finally, there is a noteworthy line of empirical research that studies narratives as pure stories: some papers investigate their impact on behavior (Hillenbrand and Verrina, 2022; Harris et al., 2021; Hagmann et al., 2020; Morag and Loewenstein, 2021), while others document how narratives about macroeconomic phenomena are spread (Andre et al., 2022, 2021).

Persuasion The impact of persuasion has been long studied in economics (see Little, 2022, for a comparison of approaches in a common framework). This paper contributes to this literature in exploring the consequences of providing interpretations of unknown events at the time of the communication. Thus, persuasion only occurs through narratives. This highlights two main differences with respect to previous well-known literature. First, the signal is undistorted, unlike leading papers such as Milgrom (1981) where the signal could be withheld, or Crawford and Sobel (1982) where the signal could be manipulated (for a review see Sobel, 2013). A recent

Also according to (Olea et al., Forthcoming), including irrelevant covariates in models helps in achieving better perceived predicted ability with a large dataset. Given the fixed state and signal space, all models have the same dimension in this paper and cannot exhibit this type of misspecification.

³³Also Izzo et al. (2021) assume that agents choose the model with the highest likelihood given the observed data. In their setting, this means choosing the model with the smallest mean squared error.

³⁴Similar to Schwartzstein and Sunderam (2021), persuaders propose a possibly biased interpretation of available data. In their design, this translates to communicating the parameters governing the data generating process — given the true causal relationship between parameters and observables.

paper by Gleyze and Pernoud (2022) investigates a cheap-talk game in which the receiver is not only uncertain about the state realization but also about the true model (which variables are payoff-relevant). They find that communication on models is impossible. Eliaz et al. (2021a) builds on the classic cheap-talk game with multidimensional messages, relaxing the assumption that the receiver is capable to interpret the equilibrium messages and allowing the sender to supply interpretations for them. The latter instructs which message components to consider to form beliefs on the sender’s state-message map, thus making inference on states. These strategic interpretations can be conditioned on both the state and the message, as opposed to the ex-ante commitment assumption presented in this paper. As a result, full persuasion can be attained depending on the prior and message dimensionality. Second, the persuader cannot influence the signal generating process. This is in stark contrast with the literature on Bayesian persuasion. Kamenica and Gentzkow (2011) and many generalizations of their framework (e.g., Alonso and Câmara, 2016; Ely, 2017; Galperti, 2019; Ball and Espín-Sánchez, 2021) are about persuasion by generating information with full commitment, then interpreted by Bayesian receivers. This restricts to only Bayes-plausible distribution of posteriors, unlike this paper.

Only by communicating interpretations, the sender is free in how many to provide. In previous literature, ambiguity aversions has motivated ambiguous communication strategies by senders, proposing several explanations or messages. This was studied both in cheap-talk games (Kellner and Le Quement, 2018, 2017) and in Bayesian persuasion (Beauchêne et al., 2019). Most closely related, the latter studies a sender chooses an ambiguous device — that is, multiple possible signal generating processes à la Kamenica and Gentzkow (2011), one of which will be implemented with unknown probabilities. Moreover, both the sender and the receiver are assumed to be ambiguity-averse. The authors characterize the sender’s optimal payoff using the concavification approach and find that ambiguous persuasion could be beneficial with respect to Bayesian persuasion. This is achieved through the construction of messages inducing the same posteriors, and thus the same action. Their paper differs from this, not only in terms of assumptions about the agents’ preferences but also in terms of the sender’s flexibility in interpretations. Unlike the Bayesian persuasion literature and thus also Beauchêne et al. (2019), this paper assumes the signal space to be fixed.

Biased Beliefs With this paper, I hope to contribute also to the literature on biases in belief updating by highlighting the importance of looking at beliefs updated conditional on all possible contingencies. It highlights how a signal generating process maps to a richer object — posteriors conditional on all possible signals — and how the value of a posterior conditional of one signal is compatible with many signal generating processes. In the rich literature on biased beliefs (for a survey see Benjamin, 2019), most papers look at deviations from Bayesian updating in belief formation along one dimension: the gap between the posteriors reported by the subject and the one calculated using Bayes rule, conditional the realized and observed signal realization. Some exceptions can be found in the literature on polarization (see Section 4.1.1). There are some exceptions. For example, Esponda et al. (2020) use the vector of posteriors space to illustrated graphically part of their results. Their focus is learning in presence of a initial misconception (focusing on the special case of base rate neglect). Their graphs shows that participants’ beliefs are sometimes Bayes-inconsistent.³⁵

³⁵There is substantial heterogeneity in the reported vectors of posteriors and how they violate Bayes-consistency. The most obvious case is the vector of posterior corresponding to the base-rate: fixing a state,

Other papers suggest different criteria to form beliefs in uncertain settings that could lead to violation of the Bayes-consistency constraint, such as the literature on belief updating with ambiguity-averse. In the case of multiple prior, there are two main paradigms of Bayesian updating: Full Bayesian (Jaffray, 1992; Pacheco Pires, 2002) and maximum likelihood updating (introduced by Dempster, 1967, and Shafer, 1976, then axiomatized by Gilboa and Schmeidler, 1993).³⁶ With the first, subjects update prior-by-prior and retain ambiguity in their posteriors. With the second, subjects only update the subset of priors that maximizes the probability of the realized event. A common consequence of both is that information increases the relevant ambiguity because the agent’s set of beliefs dilate (Seidenfeld and Wasserman, 1993). As a result, ambiguity-averse agents should lower their valuations for bets for every possible piece of information they could receive (“all news is bad news”). Shishkin et al. (2021) experimentally test the dilation property of updating models under ambiguity. They find that conditional on information ambiguity-averse subjects do not lower their value of bets. Epstein et al. (2021) interprets an extension of Bayes-consistent property to preferences as evidence of sensitivity to signal ambiguity, the attitude towards the uncertainty of the signal generating process.³⁷ Using a lab experiment, they report that a fraction of subjects is averse to signal ambiguity. Note that dilation of the prior is stronger than signal ambiguity (see Remark 2 of Epstein et al., 2021), but both imply violation of Bayes-consistency.

Also, this paper proposes that when inconsistent updating across signal realization could be the result of the adoption of different, and possibly distorted, models triggered by different signal realizations. This could be an explanation for the evidence on asymmetric updating in belief updating (e.g., Eil and Rao, 2011; Sharot, 2011; Ertac, 2011; Coutts, 2019; Möbius et al., 2022; Drobner and Goerg, 2021).^{38,39}

7 Conclusion

This paper explores whether it is possible to persuade others only by providing interpretations of events not yet known. My results show that not only it is possible, but also preemptively framing information can lead the receiver to hold inconsistent beliefs across observable facts. Allowing for

posteriors of 20% and 80% conditional on the two possible signals, both higher than the 15% prior probability.

³⁶For a comprehensive review, see Gilboa et al. (2013, Section 5).

³⁷Consider a setting with ambiguity on the informativeness of the signal generating process. Subjects are elicited their probability equivalents (PEs) of a draw from an Ellsberg urn, unconditionally and conditionally on two complementary signals: if the unconditional PE is higher than the average of the conditional PEs, it means that the observed signal is interpreted as providing weak support for both an event and its complement.

³⁸Results on asymmetric updating are mixed: some papers find more responsiveness to either good or bad news, while others find no difference. For example, Barron (2021) does not find evidence of asymmetric updating in a financial decision-making context where states differ in their monetary rewards, while Drobner (2022) shows that subjects update neutrally if they expect immediate resolution of the ego-relevant uncertainty, whereas they update optimistically if there is no resolution of uncertainty. This points to the idea that the underlying state and incentives might play a crucial role in switching on and off asymmetric updating. This is in line with the mechanism proposed in this paper as it is not always the case that different interpretations are adopted. It depends on whether it is possible to keep signals open to multiple interpretations (e.g., immediate vs. no resolution of uncertainty) and, in the case of self-persuasion, what the incentives that motivate the supply of interpretations are (e.g., financial vs. positive beliefs).

³⁹As mentioned in Section 4.3, several papers provide a theoretical explanation for asymmetric updating, explaining these motivates beliefs either via direct utility from self-serving beliefs (e.g., Köszegi, 2006; Brunnermeier and Parker, 2005) or via instrumental value associated with the beliefs (e.g., Bénabou and Tirole, 2002).

multiple stories to be communicated before the signal realizes, I provide a disciplined relaxation of the Bayes-plausibility constraint. This follows from the assumption that the receiver is boundedly rational in choosing the interpretation of facts she observes: she adopts the narrative that best explains the observed event given her priors. As a result, each signal realization might trigger the adoption of a different narrative. However, having observed only a signal realization, the receiver does not realize either which story she would have adopted conditional on a counterfactual signal realization or what beliefs she would have held. This gives the persuader more freedom to manipulate the receiver across signals. Thus, the sender can achieve a higher expected utility by sending interpretations of future realizations than by generating new information as in Bayesian persuasion. In this sense, the strategic use of narratives has the potential to exacerbate the misalignment of incentives between the two parties.

While communicating stories should be irrelevant according to the cheap-talk literature, it is plausible to assume that an individual does not know how to interpret a fact and adopts interpretations provided by others. I study the case in which such interpretations are communicated while the sender is not informed about the information the receiver has. The reason is twofold. First, receivers might be less open to entertaining stories formulated ex-post. From their perspective, it is a reasonable defensive mechanism that could be put in place easily. However, it is an open question whether people are equally receptive to narratives formulated before or after to the release of new data. This leads to the second reason to explore this setting: can ex-ante commitment limit the sender’s persuasiveness? As discussed in this paper, commitment can deter belief manipulability partially but it does not prevent the persuader from manipulating the receiver to his advantage if he is willing to use several models. Further work needs to be carried out to establish how the receiver can strategically deter model persuasion.

Many extensions remain to be explored. First, this paper focuses on a problem with only two agents, a sender and a receiver. I discuss the consequences of conflicting stories in a population of receivers with different priors. Future research should develop further insights on the sender’s optimization given a distribution of heterogeneous receivers, balancing the diverging effect models have on resulting vectors of posteriors for different receivers. Moreover, I consider only a sender communicating multiple stories. Alternatively, this can be interpreted as a coordinated strategy by senders with the same incentives. Moreover, the extension to the default model can be read as a first step to thinking about competition among senders. The sender strategically responds to a model the receiver already holds and the characterization of the feasible set of posteriors in presence of a default model informs the sender of the feasible outcomes. More can be studied in relation to multiple (uncoordinated) senders with possibly misaligned incentives. Second, I impose no restrictions on which models the sender is willing to send and the receiver is willing to accept. For example, whenever the sender communicates a story different from the one he believes in, he is untruthfully reporting how to interpret the signals. Senders might be reluctant to communicate stories too far from the true one, e.g., belief distortion bears some psychological costs for the sender, such as disappointment aversion, in line with the literature of psychological game theory (for a survey see Battigalli and Dufwenberg, 2022). In the experiment by Barron and Fries (2022), senders communicate biased narratives to their advantage but also display truth-telling preferences to some extent. These frictions could be incorporated into a theoretical extension of this paper. Moreover, receivers might consider only some types of models, e.g. never reporting a positive correlation between some signals and states, or with

more structure on the signal space assuming signals to be conditionally independent.

Furthermore, this paper discusses a wide range of applications, proposing a possible common mechanism that encompasses inter-personal (polarization, conflict of interest in financial markets, and lobbying) and intra-personal phenomena (overconfidence as motivation). These examples encourage research with the goal of testing the assumptions and implications of this research in these diverse settings.

References

- Allcott, Hunt and Matthew Gentzkow (2017) “Social media and fake news in the 2016 election,” *Journal of economic perspectives*, 31 (2), 211–36.
- Alonso, Ricardo and Odilon Câmara (2016) “Persuading voters,” *American Economic Review*, 106 (11), 3590–3605.
- Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart (2021) “Inflation narratives.”
- Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022) “Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples,” *The Review of Economic Studies*, 10.1093/restud/rdac008, rdac008.
- Andreassen, Paul B (1990) “Judgmental extrapolation and market overreaction: On the use and disuse of news,” *Journal of Behavioral Decision Making*, 3 (3), 153–174.
- Andreoni, James and Tymofiy Mylovanov (2012) “Diverging opinions,” *American Economic Journal: Microeconomics*, 4 (1), 209–32.
- Ariely, Dan and Klaus Wertenbroch (2002) “Procrastination, deadlines, and performance: Self-control by precommitment,” *Psychological science*, 13 (3), 219–224.
- Augenblick, Ned, Muriel Niederle, and Charles Sprenger (2015) “Working over time: Dynamic inconsistency in real effort tasks,” *The Quarterly Journal of Economics*, 130 (3), 1067–1115.
- Baliga, Sandeep, Eran Hanany, and Peter Klibanoff (2013) “Polarization and ambiguity,” *American Economic Review*, 103 (7), 3071–83.
- Ball, Ian and José-Antonio Espín-Sánchez (2021) “Experimental Persuasion.”
- Barron, Kai (2021) “Belief updating: does the ‘good-news, bad-news’ asymmetry extend to purely financial domains?” *Experimental Economics*, 24 (1), 31–58.
- Barron, Kai and Tilman Fries (2022) “Narrative Persuasion,” *Mimeo*.
- Battigalli, Pierpaolo and Martin Dufwenberg (2022) “Belief-dependent motivations and psychological game theory,” *Journal of Economic Literature*, 60 (3), 833–82.
- Beauchêne, Dorian, Jian Li, and Ming Li (2019) “Ambiguous persuasion,” *Journal of Economic Theory*, 179, 312–365.
- Bénabou, Roland (2015) “The economics of motivated beliefs,” *Revue d’économie politique*, 125 (5), 665–685.
- Bénabou, Roland, Armin Falk, and Jean Tirole (2018) “Narratives, imperatives, and moral reasoning.”

- Bénabou, Roland and Jean Tirole (2002) “Self-confidence and personal motivation,” *The quarterly journal of economics*, 117 (3), 871–915.
- Benjamin, Daniel J (2019) “Errors in probabilistic reasoning and judgment biases,” *Handbook of Behavioral Economics: Applications and Foundations* 1, 2, 69–186.
- Benoît, Jean-Pierre and Juan Dubra (2019) “Apparent bias: What does attitude polarization show?” *International Economic Review*, 60 (4), 1675–1703.
- Brunnermeier, Markus K and Jonathan A Parker (2005) “Optimal expectations,” *American Economic Review*, 95 (4), 1092–1118.
- Charles, Constantin and Chad Kendall (2022) “Causal Narratives.”
- Chater, Nick and George Loewenstein (2016) “The under-appreciated drive for sense-making,” *Journal of Economic Behavior & Organization*, 126, 137–154.
- Cheng, Haw and Alice Hsiaw (2022) “Distrust in experts and the origins of disagreement,” *Journal of economic theory*, 200, 105401.
- Clark, Jesse and Charles Stewart, III (2021) “The Confidence Earthquake: Seismic Shifts in Trust and Reform Sentiments in the 2020 Election,” *Available at SSRN*.
- Coutts, Alexander (2019) “Good news and bad news are still news: Experimental evidence on belief updating,” *Experimental Economics*, 22 (2), 369–395.
- Crawford, Vincent P and Joel Sobel (1982) “Strategic information transmission,” *Econometrica: Journal of the Econometric Society*, 1431–1451.
- Darley, John M and Paget H Gross (1983) “A hypothesis-confirming bias in labeling effects.,” *Journal of Personality and Social Psychology*, 44 (1), 20.
- Dempster, Arthur P (1967) “Upper and lower probability inferences based on a sample from a finite univariate population,” *Biometrika*, 54 (3-4), 515–528.
- DiFonzo, Nicholas and Prashant Bordia (1997) “Rumor and prediction: Making sense (but losing dollars) in the stock market,” *Organizational Behavior and Human Decision Processes*, 71 (3), 329–353.
- Dixit, Avinash K and Jörgen W Weibull (2007) “Political polarization,” *Proceedings of the National Academy of sciences*, 104 (18), 7351–7356.
- Douven, Igor and Jonah N Schupbach (2015a) “Probabilistic alternatives to Bayesianism: the case of explanationism,” *Frontiers in Psychology*, 6, 459.
- (2015b) “The role of explanatory considerations in updating,” *Cognition*, 142, 299–311.
- Drobner, Christoph (2022) “Motivated beliefs and anticipation of uncertainty resolution,” *American Economic Review: Insights*, 4 (1), 89–105.
- Drobner, Christoph and Sebastian J Goerg (2021) “Motivated belief updating and rationalization of information.”
- Eil, David and Justin M Rao (2011) “The good news-bad news effect: asymmetric processing of objective information about yourself,” *American Economic Journal: Microeconomics*, 3 (2), 114–38.
- Einhorn, Hillel J and Robin M Hogarth (1986) “Judging probable cause.,” *Psychological Bulletin*, 99 (1), 3.

- Eliaz, Kfir, Simone Galperti, and Ran Spiegler (2022) “False Narratives and Political Mobilization,” *arXiv preprint arXiv:2206.12621*.
- Eliaz, Kfir and Ran Spiegler (2020) “A model of competing narratives,” *American Economic Review*, 110 (12), 3786–3816.
- Eliaz, Kfir, Ran Spiegler, and Heidi C Thysen (2021a) “Strategic interpretations,” *Journal of Economic Theory*, 192, 105192.
- Eliaz, Kfir, Ran Spiegler, and Yair Weiss (2021b) “Cheating with models,” *American Economic Review: Insights*, 3 (4), 417–34.
- Ely, Jeffrey C (2017) “Beeps,” *American Economic Review*, 107 (1), 31–53.
- Epstein, Larry G, Yoram Halevy et al. (2021) “Hard-to-interpret signals.”
- Ertac, Seda (2011) “Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback,” *Journal of Economic Behavior & Organization*, 80 (3), 532–545.
- Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel (2020) “Mental Models and Learning: The Case of Base-Rate Neglect.”
- Eyster, Erik (2019) “Errors in strategic reasoning,” *Handbook of Behavioral Economics: Applications and Foundations* 1, 2, 187–259.
- Fryer, Roland G, Jr, Philipp Harms, and Matthew O Jackson (2019) “Updating beliefs when evidence is open to interpretation: Implications for bias and polarization,” *Journal of the European Economic Association*, 17 (5), 1470–1501.
- Galperti, Simone (2019) “Persuasion: The art of changing worldviews,” *American Economic Review*, 109 (3), 996–1031.
- Gentzkow, Matthew, Michael B Wong, and Allen T Zhang (2021) “Ideological bias and trust in information sources.”
- Gilboa, Itzhak, Massimo Marinacci, D Acemoglu, M Arellano, and E Dekel (2013) “Ambiguity and the bayesian paradigm. Advances in Economics and Econometrics: Theory and Applications,” in *Tenth World Congress of the Econometric Society*, D. Acemoglu, M. Arellano, and E. Dekel (Eds.).
- Gilboa, Itzhak and David Schmeidler (1993) “Updating ambiguous beliefs,” *Journal of economic theory*, 59 (1), 33–49.
- Gleyze, Simon and Agathe Pernoud (2022) “The Value of Model Misspecification in Communication.”
- Graber, Doris Appel (1984) *Processing the news: How people tame the information tide*: Longman Press.
- Hagmann, David, Julia Minson, and Catherine Tinsley (2020) “Personal narratives build trust across ideological divides.”
- Harman, Gilbert H (1965) “The inference to the best explanation,” *The philosophical review*, 74 (1), 88–95.
- Harrs, Sören, Lara Marie Müller, and Bettina Rockenbach (2021) “How Narratives Impact Financial Behavior.”
- Heidhues, Paul and Botond Köszegi (2018) “Behavioral industrial organization,” *Handbook of Behavioral Economics: Applications and Foundations* 1, 1, 517–612.

- Hillenbrand, Adrian and Eugenio Verrina (2022) “The asymmetric effect of narratives on prosocial behavior,” *Games and Economic Behavior*, 135, 241–270.
- Horz, Carlo and Korhan Kocak (2022) “How To Keep Citizens Disengaged: Propaganda and Causal Misperceptions.”
- Ichihashi, Shota and Delong Meng (2021) “The Design and Interpretation of Information,” *Available at SSRN 3966003*.
- Ispano, Alessandro (2022) “The perils of a coherent narrative.”
- Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood (2019) “The origins and consequences of affective polarization in the United States,” *Annual Review of Political Science*, 22 (1), 129–146.
- Izzo, Federica, Gregory J Martin, and Steven Callander (2021) “Ideological Competition,” *So-cArXiv. February*, 19.
- Jaffray, J-Y (1992) “Bayesian updating and belief functions,” *IEEE transactions on systems, man, and cybernetics*, 22 (5), 1144–1152.
- Kamenica, Emir and Matthew Gentzkow (2011) “Bayesian persuasion,” *American Economic Review*, 101 (6), 2590–2615.
- Kellner, Christian and Mark T Le Quement (2017) “Modes of ambiguous communication,” *Games and Economic Behavior*, 104, 271–292.
- (2018) “Endogenous ambiguity in cheap talk,” *Journal of Economic Theory*, 173, 1–17.
- Koehler, Derek J (1991) “Explanation, imagination, and confidence in judgment.,” *Psychological bulletin*, 110 (3), 499.
- Köszegi, Botond (2006) “Ego utility, overconfidence, and task choice,” *Journal of the European Economic Association*, 4 (4), 673–707.
- Laibson, David (1997) “Golden eggs and hyperbolic discounting,” *The Quarterly Journal of Economics*, 112 (2), 443–478.
- Levendusky, Matthew S (2009) “The microfoundations of mass polarization,” *Political Analysis*, 17 (2), 162–176.
- Levy, Gilat and Ronny Razin (2020) “Combining forecasts in the presence of ambiguity over correlation structures,” *Journal of Economic Theory*, 105075.
- Lipton, Peter (2003) *Inference to the best explanation*: Routledge.
- Little, Andrew T (2022) “Bayesian Explanations for Persuasion.”
- Lombrozo, Tania (2007) “Simplicity and probability in causal explanation,” *Cognitive psychology*, 55 (3), 232–257.
- Lombrozo, Tania and Susan Carey (2006) “Functional explanation and the function of explanation,” *Cognition*, 99 (2), 167–204.
- Lord, Charles G, Lee Ross, and Mark R Lepper (1979) “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence,” *Journal of Personality and Social Psychology*, 37 (11), 2098.
- Malmendier, Ulrike and Stefan Nagel (2011) “Depression babies: do macroeconomic experiences affect risk taking?” *The Quarterly Journal of Economics*, 126 (1), 373–416.

- Mason, Lilliana (2015) “‘I disrespectfully agree’: The differential effects of partisan sorting on social and issue polarization,” *American journal of political science*, 59 (1), 128–145.
- Michaels, David (2008) *Doubt is their product: how industry’s assault on science threatens your health*: Oxford University Press.
- Milgrom, Paul R (1981) “Good news and bad news: Representation theorems and applications,” *The Bell Journal of Economics*, 380–391.
- Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat (2022) “Managing self-confidence: Theory and experimental evidence,” *Management Science*.
- Morag, Dor and George Loewenstein (2021) “Narratives and Valuations,” *Available at SSRN 3919471*.
- O’Donoghue, Ted and Matthew Rabin (1999) “Doing it now or later,” *American economic review*, 89 (1), 103–124.
- Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (Forthcoming) “Competing models,” *Quarterly Journal of Economics*.
- Oreskes, Naomi and Erik M Conway (2011) *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*: Bloomsbury Publishing USA.
- Pacheco Pires, Cesaltina (2002) “A rule for updating ambiguous beliefs,” *Theory and Decision*, 53 (2), 137–152.
- Paul, Christopher and Miriam Matthews (2016) “The Russian ‘firehose of falsehood’ propaganda model,” *Rand Corporation*, 2 (7), 1–10.
- Pennington, Nancy and Reid Hastie (1992) “Explaining the evidence: Tests of the Story Model for juror decision making,” *Journal of personality and social psychology*, 62 (2), 189.
- Persily, Nathaniel and Charles Stewart, III (2021) “The Miracle and Tragedy of the 2020 US Election,” *Journal of Democracy*, 32 (2), 159–178.
- Plous, Scott (1991) “Biases in the assimilation of technological breakdowns: Do accidents make us safer?” *Journal of Applied Social Psychology*, 21 (13), 1058–1082.
- Rabin, Matthew and Joel L Schrag (1999) “First impressions matter: A model of confirmatory bias,” *The Quarterly Journal of Economics*, 114 (1), 37–82.
- Read, Daniel and Barbara Van Leeuwen (1998) “Predicting hunger: The effects of appetite and delay on choice,” *Organizational behavior and human decision processes*, 76 (2), 189–205.
- Reich, Taly and Zakary L Tormala (2013) “When contradictions foster persuasion: An attributional perspective,” *Journal of Experimental Social Psychology*, 49 (3), 426–439.
- Russo, J Edward, Margaret G Meloy, and Victoria Husted Medvec (1998) “Predecisional distortion of product information,” *Journal of Marketing Research*, 35 (4), 438–452.
- Sances, Michael W and Charles Stewart, III (2015) “Partisanship and confidence in the vote count: Evidence from US national elections since 2000,” *Electoral Studies*, 40, 176–188.
- Schwartzstein, Joshua and Adi Sunderam (2021) “Using models to persuade,” *American Economic Review*, 111 (1), 276–323.
- (2022) “Shared Models in Networks, Organizations, and Groups.”

- Seidenfeld, Teddy and Larry Wasserman (1993) “Dilation for sets of probabilities,” *The Annals of Statistics*, 21 (3), 1139–1154.
- Shafer, Glenn (1976) *A mathematical theory of evidence*, 42: Princeton university press.
- Sharot, Tali (2011) “The optimism bias,” *Current biology*, 21 (23), R941–R945.
- Shiller, Robert J (2017) “Narrative economics,” *American Economic Review*, 107 (4), 967–1004.
- (2019) *Narrative economics*: Princeton University Press Princeton.
- Shishkin, Denis, Pietro Ortoleva et al. (2021) “Ambiguous information and dilation: An experiment.”
- Sinclair, Betsy, Steven S Smith, and Patrick D Tucker (2018) ““It’s largely a rigged system”: voter confidence and the winner effect in 2016,” *Political Research Quarterly*, 71 (4), 854–868.
- Sobel, Joel (2013) “Giving and receiving advice,” *Advances in economics and econometrics*, 1, 305–341.
- Spiegler, Ran (2016) “Bayesian networks and boundedly rational expectations,” *The Quarterly Journal of Economics*, 131 (3), 1243–1290.
- Thagard, Paul (1989) “Explanatory coherence,” *Behavioral and brain sciences*, 12 (3), 435–467.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018) “The spread of true and false news online,” *Science*, 359 (6380), 1146–1151.
- Weick, Karl E (1995) *Sensemaking in organizations*, 3: Sage.
- Yang, Jeffrey (2022) “A Criterion of Model Decisiveness,” *Mimeo*.

A Appendix: Additional Proofs

Proof of Lemma 1. Consider the two statements separately.

(i) For each $\mu \in \mathcal{B}$, there exists a model that induces μ .

Consider $\mu \in \mathcal{B}$. As it is Bayes-consistent, there exists a distribution $\sigma \in \Delta(S)$ such that, for each state ω , the average of the posteriors $\mu_s(\omega)$ across signals weighted by $\sigma(s)$ equals the prior, i.e., $\sum_s \mu_s(\omega) \sigma(s) = \mu_0(\omega)$. For each σ , define a model such that, for each s and ω ,

$$\pi^\sigma(s|\omega) = \frac{\mu_s(\omega) \sigma(s)}{\sum_{s'} \mu_{s'}(\omega) \sigma(s')}.$$

This is a well-defined model because for each ω and s , $\pi^\sigma(s|\omega) \in [0, 1]$ and $\sum_s \pi^\sigma(s|\omega) = 1$.

Notice that the probability of signal s according to such a model is $\Pr^\sigma(s) = \sigma(s)$. To see this, calculate the probability of a signal according to the constructed model

$$\Pr^\sigma(s) = \sum_\omega \mu_0(\omega) \pi^\sigma(s|\omega) = \sum_\omega \left(\sum_{s'} \mu_{s'}(\omega) \sigma(s') \right) \left(\frac{\mu_s(\omega) \sigma(s)}{\sum_{s'} \mu_{s'}(\omega) \sigma(s')} \right) = \sigma(s) \sum_\omega \mu_s(\omega) = \sigma(s).$$

The posterior attached to state ω conditional on signal s induced by the model σ is

$$\mu_s^\sigma(\omega) = \frac{\mu_0(\omega) \pi^\sigma(s|\omega)}{\Pr^m(s)} = \frac{\mu_0(\omega)}{\sigma(s)} \left(\frac{\mu_s(\omega) \sigma(s)}{\sum_{s'} \mu_{s'}(\omega) \sigma(s')} \right) = \frac{\sum_{s'} \mu_{s'}(\omega) \sigma(s')}{\sigma(s)} \frac{\mu_s(\omega) \sigma(s)}{\sum_{s'} \mu_{s'}(\omega) \sigma(s')} = \mu_s(\omega).$$

Indeed, the vectors of posterior induced by such a model is μ .

In conclusion, for each $\sigma \in \Delta(S)$ such that $\sum_s \mu_s(\omega) \sigma(s) = \mu_0(\omega)$ (μ is Bayes-consistent), there exists a model that induces μ .

(ii) Each model m induces a vector of posterior beliefs that is Bayes-consistent $\mu^m \in \mathcal{B}$.

Consider as weights for the convex combination the distribution of the signals according to the model m : $(\Pr^m(s))_{s \in S}$. Given that $m \in [\Delta(S)]^\Omega$, it holds that it is a proper distribution with $\sum_s \Pr^m(s) = 1$. Then, for every $\omega \in \Omega$,

$$\sum_{s \in S} \Pr^m(s) \mu_s^m(\omega) = \sum_{s \in S} \Pr^m(s) \frac{\mu_0(\omega) \pi^m(s|\omega)}{\Pr^m(s)} = \sum_{s \in S} \mu_0(\omega) \pi^m(s|\omega) = \mu_0(\omega) \sum_{s \in S} \pi^m(s|\omega) = \mu_0(\omega).$$

Every vector of posterior beliefs induced by a model satisfies Bayes-consistency. \square

Corollary 3 (Binary Signal). Let $\mu^\varnothing = (\mu_0, \mu_0)$. For each vector of posterior beliefs $\mu \in \mathcal{B} \setminus \{\mu^\varnothing\}$ there exists a unique model that induces μ .

Proof of Corollary 3. Lemma 1 shows that each model induces a vectors of posteriors beliefs that is Bayes-consistent. To show the uniqueness of a model associated to a Bayes-consistent vector of posteriors in the binary signal, it is enough to show that there exists only one distribution over the signal space such that a vectors of posterior is Bayes-consistent.

Let $(\sigma_{s_1}, \sigma_{s_2}) = (\sigma, 1 - \sigma)$. For each state ω , the Bayes-consistency condition implies that

$$\mu_0(\omega) = \sigma \mu_{s_1}(\omega) + (1 - \sigma) \mu_{s_2}(\omega).$$

Then, it holds that $\sigma = \frac{\mu_0(\omega) - \mu_{s_2}(\omega)}{\mu_{s_1}(\omega) - \mu_{s_2}(\omega)}$.

Hence, $(\sigma_{s_1}, \sigma_{s_2})$ is a distribution over signals if either (i) $\mu_{s_1}(\omega) > \mu_0(\omega) > \mu_{s_2}(\omega)$, or (ii) $\mu_{s_1}(\omega) < \mu_0(\omega) < \mu_{s_2}(\omega)$ for every ω . These two conditions are equivalent to $\mu \in \mathcal{B} \setminus \{\mu^\emptyset\}$ for binary signal. \square

Proof of Lemma 2. Fix a posterior μ_s . Consider the two statements separately.

(i) For every $p \in [0, \bar{\delta}(\mu_s)^{-1}]$, there exists a model inducing μ_s with fit $\text{Pr}^m(s) = p$.

Fix $p \in [0, \bar{\delta}(\mu_s)^{-1}]$. To show that there exists a model with fit p inducing μ_s , I construct a vector of posteriors $\boldsymbol{\mu}$ such that (i) the target μ_s is induced conditional on s , and (ii) for each state ω , there exists a distribution over signals $\sigma \in \Delta(S)$ with the additional property $\sigma_s = p$ such that Bayes-consistency holds:

$$\sum_{s'} \mu_{s'}(\omega) \sigma_{s'} = \mu_s(\omega) \sigma_s + \sum_{s' \neq s} \mu_{s'}(\omega) \sigma_{s'} = \mu_0(\omega). \quad (\text{a})$$

Hence, it follows from Lemma 1 that there exists a model that induce this Bayes-consistent vector of posteriors and, thus so, with fit p .

Given the many degrees of freedom, there exists multiple vectors of posteriors that satisfy condition (a) as long as, for each ω ,

$$\mu_0(\omega) - \mu_s(\omega) p = \sum_{s' \neq s} \mu_{s'}(\omega) \sigma_{s'} \geq 0. \quad (\text{b})$$

For instance, fix a signal $s'' \neq s$ and, for each ω , let $\mu_{s''}(\omega) = \frac{\mu_0(\omega) - p \mu_s(\omega)}{1-p}$. Condition (a) is satisfied for the distribution $\sigma(s')$ such that $\sigma_s = p$, $\sigma_{s''} = 1 - p$, and $\sigma_{s'} = 0$ for all the other signals.

Condition (b) is implied by $p \in [0, \bar{\delta}(\mu_s)^{-1}]$. As the condition has to hold for every state, it holds that

$$p \leq \frac{\mu_0(\omega)}{\mu_s(\omega)} \leq \frac{1}{\max_{\omega} \frac{\mu_s(\omega)}{\mu_0(\omega)}} = \bar{\delta}(\mu_s)^{-1}.$$

(ii) Every model inducing μ_s has fit $\text{Pr}^m(s) \in [0, \bar{\delta}(\mu_s)^{-1}]$.

Consider an arbitrary model inducing μ_s conditional on s . It follows from Bayes rule that the fit of any m inducing the target $\mu_s^m = \mu_s$ conditional on s must be such that, for every ω

$$\text{Pr}^m(s) = \frac{\mu_0(\omega)}{\mu_s(\omega)} \pi^m(s|\omega).$$

Notice that if $\pi^m(s|\omega) = 0$ the fit equals 0 (minimal fit). Instead, if $\pi^m(s|\omega) = 1$, it follows that

$$\text{Pr}^m(s) \leq \frac{\mu_0(\omega)}{\mu_s(\omega)}.$$

Because this holds for every state, the maximal fit for μ_s is the minimum of the ratio across states, which equals the reciprocal of the maximal movement for μ_s :

$$\min_{\omega} \frac{\mu_0(\omega)}{\mu_s(\omega)} = \frac{1}{\max_{\omega} \frac{\mu_s(\omega)}{\mu_0(\omega)}} = \bar{\delta}(\mu_s)^{-1}.$$

The fit of a model that induces the target posterior can only take values in $[0, \bar{\delta}(\mu_s)^{-1}]$.

□

Proof of Proposition 1. It directly follows from Lemma 1.

□

Proof of Theorem 1. Take a vector of posterior beliefs $\mu \notin \mathcal{B}$, otherwise it would be trivially feasible by Proposition 1. Inducing μ requires a set of at most $K = |S|$ models $(m_k)_{k=1}^K$ such that each model m_k induce the distribution μ_{s_k} conditional on the signal s_k . This implies two conditions on each model m_k : (i) $\mu_{s_k}^{m_k} = \mu_{s_k}$, and (ii) $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each $j \neq k$.

Assume $\mu \in \mathcal{F}$. In what follows, I show that there exists a set of models inducing μ . Instead of directly constructing each model, I specify the vector of posteriors μ^{m_k} and the fit levels $(\Pr^{m_k}(s))_{s \in S}$ induced by each m_k . The corresponding distribution of posteriors is Bayes-plausible, thus corresponds to a model. Last, I show that each model m_k is chosen conditional on the signal s_k .

For each model m_k , specify the following posteriors and fit levels: for s_k , set $\mu_{s_k}^{m_k} = \mu_{s_k}$ and $\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1}$; otherwise, for $s \neq s_k$ and state ω , set

$$\mu_s^{m_k}(\omega) = \frac{\mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1} \mu_{s_k}(\omega)}{1 - \bar{\delta}(\mu_{s_k})^{-1}}, \quad \Pr^{m_k}(s) = \left(\frac{1 - \bar{\delta}(\mu_{s_k})^{-1}}{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}} \right) \bar{\delta}(\mu_s)^{-1}.$$

These probabilities are well-defined for every s and ω . First, each posterior is non-negative because $\bar{\delta}(\mu_{s_k})^{-1} \geq \delta(\mu_{s_k}(\omega))$ for every ω by definition of maximal movement, and is not higher than one because $\bar{\delta}(\mu_{s_k}) \geq \frac{1 - \mu_{s_k}(\omega)}{1 - \mu_0(\omega)}$ for every ω . To see the latter, let $\bar{\omega} = \arg \max_{\omega} \delta(\mu_{s_k}(\omega))$ and $\tilde{\omega} = \arg \max_{\omega} \frac{1 - \mu_{s_k}(\omega)}{1 - \mu_0(\omega)}$. Since $\bar{\delta}(\mu_{s_k}) \geq 1$, then $\mu_{s_k}(\bar{\omega}) \geq \mu_0(\bar{\omega})$ and $\mu_0(\tilde{\omega}) \geq \mu_{s_k}(\tilde{\omega})$ — otherwise the ratio $\frac{1 - \mu_{s_k}(\tilde{\omega})}{1 - \mu_0(\tilde{\omega})}$ would be less than one, thus trivially lower than the maximal movement. By contradiction, suppose that $\frac{\mu_{s_k}(\bar{\omega})}{\mu_0(\bar{\omega})} > \frac{1 - \mu_{s_k}(\tilde{\omega})}{1 - \mu_0(\tilde{\omega})}$. Then, it follows that $\mu_0(\bar{\omega})(1 - \mu_{s_k}(\tilde{\omega})) > \mu_0(\tilde{\omega})(1 - \mu_0(\tilde{\omega})) > \mu_{s_k}(\tilde{\omega})(1 - \mu_0(\tilde{\omega}))$. This implies $\mu_0(\bar{\omega}) > \mu_{s_k}(\tilde{\omega})$, which is a contradiction.

Also, posteriors sum up to one for every s :

$$\sum_{\omega} \mu_s(\omega) = \frac{(\sum_{\omega} \mu_0(\omega)) - \bar{\delta}(\mu_{s_k})^{-1} (\sum_{\omega} \mu_{s_k}(\omega))}{1 - \bar{\delta}(\mu_{s_k})^{-1}} = \frac{1 - \bar{\delta}(\mu_{s_k})^{-1}}{1 - \bar{\delta}(\mu_{s_k})^{-1}} = 1.$$

Second, each fit is non-negative because $\bar{\delta}(\mu_s)^{-1} \in (0, 1]$ for each s , and not bigger than one because $\mu \in \mathcal{F}$, implying for each s :

$$\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1} \geq 1 - \bar{\delta}(\mu_{s_k})^{-1} \geq \bar{\delta}(\mu_s)^{-1} (1 - \bar{\delta}(\mu_{s_k})^{-1}).$$

Notice that $\sum_{s \neq s_k} \Pr^{m_k}(s) = 1 - \bar{\delta}(\mu_{s_k})^{-1}$. Thus, $\sum_s \Pr^{m_k}(s) = 1$.

Finally, note that such constructed distribution of posterior is Bayes-plausible, i.e., $\mu_0(\omega) =$

$\sum_s \Pr^{m_k}(s) \mu_{s_k}^{m_k}(\omega)$ for each ω because for each state ω ,

$$\begin{aligned} \sum_{s \neq s_k} \Pr^{m_k}(s) \mu_{s_k}^{m_k}(\omega) &= \sum_{s \neq s_k} \frac{\mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1} \mu_{s_k}(\omega)}{1 - \bar{\delta}(\mu_{s_k})^{-1}} \left(\frac{1 - \bar{\delta}(\mu_{s_k})^{-1}}{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}} \right) \bar{\delta}(\mu_s)^{-1} \\ &= (\mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1} \mu_{s_k}(\omega)) \frac{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}}{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}} = \mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1} \mu_{s_k}(\omega). \end{aligned}$$

Therefore, it corresponds to a well-defined model.

Such constructed set of models induces $\boldsymbol{\mu}$ because each model inducing a posterior distribution conditional on a signal is chosen conditional on the signal is tailored to. Indeed, for every other model m_j with $j \neq k$:

$$\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq \underbrace{\left(\frac{1 - \bar{\delta}(\mu_{s_k})^{-1}}{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}} \right)}_{\leq 1} \bar{\delta}(\mu_{s_k})^{-1} = \Pr^{m_j}(s_k),$$

where the inequality follows from the fact that the sum of maximal fit levels for the target vector is greater or equal than one guaranteeing that the multiplying factor is less or equal than one.

Assume $\boldsymbol{\mu} \notin \mathcal{F}$. Then, it holds that $\sum_{s \in S} \bar{\delta}(\mu_s)^{-1} < 1$, equivalent to $\bar{\delta}(\mu_{s_k})^{-1} < 1 - \sum_{i \neq k} \bar{\delta}(\mu_{s_i})^{-1}$ for every k . If it were to exist a set of models inducing the target $\boldsymbol{\mu}$, each tailored model m_k inducing the posterior μ_{s_k} has to be adopted conditional on s_k . Thus, it must hold that $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each $j \neq k$. Notice that

$$\Pr^{m_j}(s_k) = 1 - \sum_{i \neq k} \Pr^{m_j}(s_i) \geq 1 - \sum_{i \neq k} \bar{\delta}(\mu_{s_i})^{-1},$$

since for every other signal the fit must be lower than the maximal fit associated to the target posterior conditional on that signal, i.e., $\Pr^{m_j}(s_i) \leq \Pr^{m_i}(s_i) \leq \bar{\delta}(\mu_{s_i})^{-1}$ for every i . This leads to a contradiction:

$$1 - \sum_{i \neq k} \bar{\delta}(\mu_{s_i})^{-1} > \bar{\delta}(\mu_{s_k})^{-1} \geq \Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k) \geq 1 - \sum_{i \neq k} \bar{\delta}(\mu_{s_i})^{-1}.$$

Therefore, it is not possible to construct a set of models to induce $\boldsymbol{\mu} \notin \mathcal{F}$. \square

Proof of Proposition 2. Let $K = |S|$ and $N = |\Omega|$. Assume that $\min_{\omega \in \Omega} \mu_0(\omega) \geq \frac{1}{K}$.

First, notice that the upper bound of the maximal movement is the reciprocal of the minimum prior across states: $\bar{\delta}(\mu_s) \leq \frac{1}{\min_{\omega \in \Omega} \mu_0(\omega)}$ for any posterior belief μ_s . To see this,

$$\bar{\delta}(\mu_s) = \max_{\omega \in \Omega} \frac{\mu_s(\omega)}{\mu_0(\omega)} \leq \max_{\omega \in \Omega} \frac{1}{\mu_0(\omega)} = \frac{1}{\min_{\omega \in \Omega} \mu_0(\omega)}.$$

By Lemma 2, the maximal fit has a lower bound: $\bar{\delta}(\mu_s)^{-1} \geq \min_{\omega \in \Omega} \mu_0(\omega)$.

Therefore, it holds that

$$\sum_{s \in S} \bar{\delta}(\mu_s)^{-1} \geq \sum_{s \in S} \min_{\omega \in \Omega} \mu_0(\omega) = K \min_{\omega \in \Omega} \mu_0(\omega) \geq 1,$$

where the last inequality follows from the assumption on the prior. □

Proof of Proposition 3 (Binary Case). First, I describe some properties of the characterizing condition of the set of feasible posterior beliefs, then I work by cases to show the inclusion.

Step 1. Fix a signal $s \in S$. I want to rewrite the characterizing condition in Theorem 1 in terms of posterior beliefs: for each state $\omega \in \{\omega_1, \omega_2\}$ and any other signal $s' \neq s, s' \in S$, $\delta(\mu_s(\omega)) \leq \frac{1}{1 - \bar{\delta}(\mu_{s'})^{-1}}$. For convenience, I rewrite the conditions for both states ω_1, ω_2 only in terms on one state. Then, the characterizing condition can be expressed as two constraints on $\mu_s(\omega_1)$ for each signal s' :

$$\mu_s(\omega_1) \leq \frac{\mu_0(\omega_1)}{1 - \bar{\delta}(\mu_{s'})^{-1}} = \bar{R}(\omega_1; s'), \quad (a)$$

$$\mu_s(\omega_1) \geq 1 - \frac{1 - \mu_0(\omega_1)}{1 - \bar{\delta}(\mu_{s'})^{-1}} = \underline{R}(\omega_1; s'). \quad (b)$$

Properties of Maximal Movement: it is useful to notice that

$$\bar{\delta}(\mu_s) = \max_{\omega} \left\{ \frac{\mu_s(\omega_1)}{\mu_0(\omega_1)}, \frac{\mu_s(\omega_2)}{\mu_0(\omega_2)} \right\} = \max_{\omega} \left\{ \frac{\mu_s(\omega_1)}{\mu_0(\omega_1)}, \frac{1 - \mu_s(\omega_1)}{1 - \mu_0(\omega_1)} \right\} = \begin{cases} \frac{\mu_s(\omega_1)}{\mu_0(\omega_1)}, & \text{if } \mu_s(\omega_1) \geq \mu_0(\omega_1) \\ \frac{\mu_s(\omega_2)}{\mu_0(\omega_2)}, & \text{if } \mu_s(\omega_1) < \mu_0(\omega_1). \end{cases}$$

Moreover, if $\mu_0(\omega_1) \geq 50\%$, $\bar{\delta}(\mu_s) \leq \frac{1}{1 - \mu_0(\omega_1)}$.

Claim: if $\mu_0(\omega_1) \geq 50\%$, condition (a) is not binding; otherwise, condition (b) is not binding.

Assume that $\mu_0(\omega_1) \geq 50\%$. Given the property of the maximal movement above, rewrite

$$\bar{R}(\omega_1; s') = \frac{\mu_0(\omega_1)}{1 - \bar{\delta}(\mu_{s'})^{-1}} \geq \frac{\mu_0(\omega_1)}{1 - 1 + \mu_0(\omega_1)} = 1,$$

which makes condition (a) always satisfied.

Instead, assume that $\mu_0(\omega_1) < 50\%$. Given the property of the maximal movement above, rewrite

$$\underline{R}(\omega_1; s') = 1 - \frac{1 - \mu_0(\omega_1)}{1 - \bar{\delta}(\mu_{s'})^{-1}} \leq 1 - \frac{1 - \mu_0(\omega_1)}{1 - \mu_0(\omega_1)} = 0,$$

which makes condition (b) always satisfied.

Claim: if $\mu_s(\omega_1) < \mu_0(\omega_1)$, condition (a) is not binding; otherwise, condition (b) is not binding.

Assume that $\mu_s(\omega_1) < \mu_0(\omega_1)$. Then, $\bar{R}(\omega_1; s') \geq 1$. To see this, rewrite

$$\bar{R}(\omega_1; s') = \frac{\mu_0(\omega_1)(1 - \mu_{s'}(\omega_1))}{1 - \mu_{s'}(\omega_1) - 1 + \mu_0(\omega_1)} = \frac{\mu_0(\omega_1)(1 - \mu_{s'}(\omega_1))}{\mu_0(\omega_1) - \mu_{s'}(\omega_1)},$$

where the denominator is positive. In this case, it simplifies to $\mu_{s'}(\omega_1)(1 - \mu_0(\omega_1)) \geq 0$, always satisfied.

Instead, assume that $\mu_s(\omega_1) \geq \mu_0(\omega_1)$. Then, $\underline{R}(\omega_1; s') \leq 0$. To see this, rewrite

$$\underline{R}(\omega_1; s') = 1 - \frac{(1 - \mu_0(\omega_1))\mu_{s'}(\omega_1)}{\mu_{s'}(\omega_1) - \mu_0(\omega_1)}.$$

Then it is enough to check that

$$\frac{(1 - \mu_0(\omega_1))\mu_{s'}(\omega_1)}{\mu_{s'}(\omega_1) - \mu_0(\omega_1)} \geq 1,$$

where the denominator is positive. In this case, it simplifies to $\mu_{s'}(\omega_1)(1 - \mu_0(\omega_1)) \geq 0$, always satisfied.

Step 2. Let $\mu_{0,\varepsilon} = (\mu_{0,\varepsilon}(\omega_1), \mu_{0,\varepsilon}(\omega_2)) = (\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon)$ and \mathcal{F}_ε the set of the feasible vectors of posteriors with respect to this prior with corresponding relevant thresholds $\bar{R}_\varepsilon(\cdot)$ and $\underline{R}_\varepsilon(\cdot)$.

To show that for $\varepsilon' < \varepsilon''$ $\mathcal{F}_{\varepsilon''} \subseteq \mathcal{F}_{\varepsilon'}$, I work by cases depending on the relation between the vector of posterior beliefs and the prior, summarized in Figure 10. Given the previous claims, some conditions are always satisfied in the first two cases.

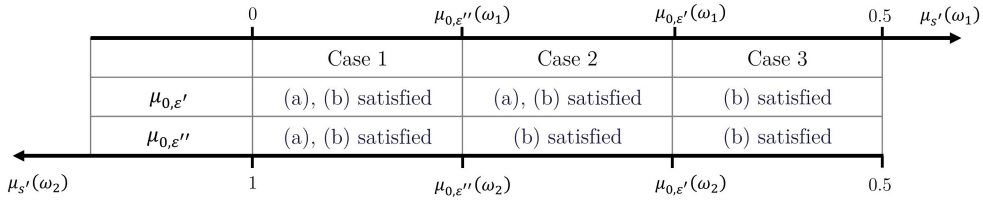


Figure 10

Case 1: $\mu_{s'}(\omega_1) \leq \mu_{0,\varepsilon'}(\omega_1) \leq \mu_{0,\varepsilon''}(\omega_1)$. Both conditions are always satisfied: all vector of posterior beliefs in this case are feasible for both priors.

Case 2: $\mu_{0,\varepsilon'}(\omega_1) \leq \mu_{s'}(\omega_1) \leq \mu_{0,\varepsilon''}(\omega_1)$. All vector of posterior beliefs in this case satisfies both conditions (a) and (b) for the prior $\mu_{0,\varepsilon'}$, but this is not true for $\mu_{0,\varepsilon''}$. Thus, if $\mu_{s'} \in \mathcal{F}_{\varepsilon''}$, then $\mu_{s'} \in \mathcal{F}_{\varepsilon'}$.

Case 3: $\mu_{s'}(\omega_1) \geq \mu_{0,\varepsilon'}(\omega_1) \geq \mu_{0,\varepsilon''}(\omega_1)$. Only condition (b) is guaranteed for either priors. Next, I show that condition (a) is more binding for $\mu_{0,\varepsilon''}$ with respect to $\mu_{0,\varepsilon'}$. To see this, it is enough to show that, for each s , $\bar{R}_{\varepsilon'';s}(\omega_1) \leq \bar{R}_{\varepsilon';s}(\omega_1)$. That is,

$$\bar{R}_{\varepsilon''}(\omega_1; s) = \frac{\mu_{0,\varepsilon''}(\omega_1)}{1 - \frac{\mu_{0,\varepsilon''}(\omega_1)}{\mu_s(\omega_1)}} \leq \frac{\mu_{0,\varepsilon'}(\omega_1)}{1 - \frac{\mu_{0,\varepsilon'}(\omega_1)}{\mu_s(\omega_1)}} = \bar{R}_{\varepsilon'}(\omega_1; s),$$

rearranged as

$$\frac{\mu_{0,\varepsilon''}(\omega_1)}{\mu_s(\omega_1) - \mu_{0,\varepsilon''}(\omega_1)} \leq \frac{\mu_{0,\varepsilon'}(\omega_1)}{\mu_s(\omega_1) - \mu_{0,\varepsilon'}(\omega_1)},$$

which is always verified since $\mu_{0,\varepsilon''}(\omega_1) \leq \mu_{0,\varepsilon'}(\omega_1)$. Thus, if $\mu_s \in \mathcal{F}_{\varepsilon''}$, then $\mu_s \in \mathcal{F}_{\varepsilon'}$. \square

Proof of Proposition 4. Add a dummy signal $s_0 \notin S$ to the signal space $S' = S \cup \{s_0\}$. I want to show that any vector of posterior on the original signal space $\mu \in [\Delta(\Omega)]^S$ can be induced.

Take an arbitrary vector of posterior beliefs μ . To induce μ , I construct a set of $K = |S|$ models $(m_k)_{k=1}^K \in [\Delta(S')]^\Omega$ such that each model m_k is tailored to induce the target distribution μ_{s_k} conditional on the signal s_k . This implies two conditions on each model m_k : (i) $\mu_{s_k}^{m_k} = \mu_{s_k}$, and

(ii) $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each $j \neq k$. The construction is similar to the proof of Theorem 1. Unlike that, the models are defined on the new signal space S' , but only $K = |S|$ models are necessary.

Instead of directly constructing each model, I specify the vector of posteriors $\boldsymbol{\mu}^{m_k}$ and the fit levels $(\Pr^{m_k}(s))_{s \in S}$ induced by each m_k . The corresponding distribution of posteriors corresponds to a unique model.

For each model m_k , specify the following posteriors and fit levels:

$$\mu_s^{m_k}(\omega) = \begin{cases} \mu_{s_k}(\omega) & \text{if } s = s_k, \\ \frac{\mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1} \mu_{s_k}(\omega)}{1 - \bar{\delta}(\mu_{s_k})^{-1}} & \text{if } s = s_0, \\ \text{any value} & \text{otherwise,} \end{cases} \quad \Pr^{m_k}(s) = \begin{cases} \bar{\delta}(\mu_{s_k})^{-1} & \text{if } s = s_k, \\ 1 - \bar{\delta}(\mu_{s_k})^{-1} & \text{if } s = s_0, \\ 0 & \text{otherwise.} \end{cases}$$

That is, for each signal realization, I am proposing the model inducing the target posterior with its maximal fit, while inducing the residual fit level conditional on the dummy signal s_0 . According to this information structure, all other signals are irrelevant as well as the posteriors conditional on them. This construction is well-defined, and it follows details from the proof of Theorem 1.

Such constructed set of models induces $\boldsymbol{\mu}$ because each model inducing a posterior distribution conditional on signal is chosen conditional on the signal is tailored to because the fit conditional on any other realizable signals $s \in S$ is zero:

$$\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq 0 = \Pr^{m_j}(s_k).$$

□

Proof of Proposition 5 (Binary Case, Polarization). Consider two conflicting stories m and m' such that $\pi^m(s_1|\omega_1) > \pi^m(s_1|\omega_2)$ and $\pi^{m'}(s_1|\omega_2) > \pi^{m'}(s_1|\omega_1)$. There exists a threshold in prior p such that, if $\mu_0(\omega_1) < p$, m' and m adopted respectively conditional on s_1 and s_2 ; otherwise, m and m' adopted respectively conditional on s_1 and s_2 . To see this, consider for which prior narrative m is adopted conditional on s_1 , that is $\Pr^m(s_1) > \Pr^{m'}(s_1)$ (and thus, m' is adopted conditional on s_2). The resulting condition is that

$$\mu_0(\omega_1) > p := \frac{1}{\frac{\pi^m(s_1|\omega_1) - \pi^{m'}(s_1|\omega_1)}{\pi^{m'}(s_1|\omega_2) - \pi^m(s_1|\omega_2)} + 1}.$$

Therefore, it follows that the resulting vector of posteriors is

$$\boldsymbol{\mu} = \begin{cases} \left(\mu_{s_1}^{m'}, \mu_{s_2}^m \right), & \text{if } \mu_0(\omega_1) < p \\ \left(\mu_{s_1}^m, \mu_{s_2}^{m'} \right), & \text{if } \mu_0(\omega_1) > p. \end{cases}$$

Note that, if $\pi(s_1|\omega_1) > \pi(s_1|\omega_2)$, then $\mu_{s_1}(\omega_1) > \mu_0(\omega_1) > \mu_{s_2}(\omega_1)$. Therefore, the resulting vector of posteriors is never Bayes-consistent: $\boldsymbol{\mu} \notin \mathcal{B}$. Furthermore, depending on the prior, Bayes-consistency is violated differently: for every signal s it holds that (i) $\mu_s(\omega_1) < \mu_0(\omega_1)$ if $\mu_0(\omega_1) < p$, and (ii) $\mu_s(\omega_1) > \mu_0(\omega_1)$ if $\mu_0(\omega_1) > p$. □

Proof of Theorem 2. Take an arbitrary vector of posterior beliefs μ . Inducing μ requires a set of at most $K = |S|$ models $(m_k)_{k=1}^K$ such that each model m_k is tailored to induce the distribution μ_{s_k} conditional on the signal s_k . This implies two conditions on each model m_k : (i) $\mu_{s_k}^{m_k} = \mu_{s_k}$, and (ii) $\Pr^{m_k}(s_k) \geq \Pr^d(s_k)$ for each $m \in \{m_1, \dots, m_K\} \cup \{d\}$.

Assume $\mu \in \mathcal{F}^d$. I show that there exists a set of models inducing μ . Construct the vector of posteriors μ^{m_k} and the fit levels $(\Pr^{m_k}(s))_{s \in S}$ induced by each m_k as in the proof of Theorem 1.

It only remains to show that each model m_k is chosen conditional on signal s_k . Model m_k is chosen with respect to the default model given s_k because $\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq \Pr^d(s_k)$ because $\mu \in \mathcal{F}^d$. Also, for every other model m_j with $j \neq k$:

$$\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq \underbrace{\left(\frac{1 - \bar{\delta}(\mu_{s_k})^{-1}}{\sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}} \right)}_{\leq 1} \bar{\delta}(\mu_{s_k})^{-1} = \Pr^{m_j}(s_k),$$

where the inequality follows from $\mu \in \mathcal{F}^d$ and, thus

$$\bar{\delta}(\mu_{s_k})^{-1} \geq \Pr^d(s_k) = 1 - \sum_{s \neq s_k} \Pr^d(s) \geq 1 - \sum_{s \neq s_k} \bar{\delta}(\mu_s)^{-1}.$$

Assume $\mu \notin \mathcal{F}^d$. Then, there must be a signal s_ℓ such that the characterizing condition does not hold: $\bar{\delta}(\mu_{s_\ell})^{-1} < \Pr^d(s_\ell)$. If it were to exist a set of models inducing the target μ , each tailored model m_k inducing the posterior μ_{s_k} has to be adopted conditional on s_k . Thus, it must hold that $\Pr^{m_k}(s_k) \geq \Pr^d(s_k)$ for each $m \in \{m_1, \dots, m_K\} \cup \{d\}$.

This leads to a contradiction because to have the model m_ℓ adopted conditional on s_ℓ it must be the case that $\Pr^{m_\ell}(s_\ell) \geq \Pr^d(s_\ell)$ but it also holds that $\Pr^d(s_\ell) > \bar{\delta}(\mu_{s_\ell})^{-1} \geq \Pr^{m_\ell}(s_\ell)$. Therefore, it is not possible to construct a set of models to induce $\mu \notin \mathcal{F}^d$. \square

Proof of Proposition 6. Note that \mathcal{F}^d depends only on the fit levels induced by the default model d — some distribution over signals $p \in \Delta(S)$ — we can rewrite the union of \mathcal{F}^d as

$$\begin{aligned} \bigcup_{d \in \mathcal{M}} \mathcal{F}^d &= \bigcup_{d \in \mathcal{M}} \left\{ \mu \in [\Delta(\Omega)]^S : \forall s \in S, \bar{\delta}(\mu_s)^{-1} \geq \Pr^d(s) \right\} \\ &= \left\{ \mu \in [\Delta(\omega)]^S : \exists p \in \Delta(S) \text{ such that } \forall s \in S, \bar{\delta}(\mu_s)^{-1} \geq p_s \right\}. \end{aligned}$$

I show separately the two inclusions.

(i) Take $\mu \in \mathcal{F}$. It is to be shown that for each $\mu \in \mathcal{F}$ there exists a distribution over signals $p \in \Delta(S)$ such that $\bar{\delta}(\mu_s) \geq p_s$ for every s . Since $\sum_{s \in S} \bar{\delta}(\mu_s)^{-1} \geq 1$, it is possible to construct p scaling down each reciprocal of the maximal movement so that the resulting components sum to one. For each s , set

$$p_s = \frac{\bar{\delta}(\mu_s)^{-1}}{\sum_{s' \in S} \bar{\delta}(\mu_{s'})^{-1}}.$$

This is a well-defined distribution over signals, because for each s it holds $p_s \in [0, 1]$ and

$\sum_s p_s = 1$. Moreover, for each signal s , we have

$$\bar{\delta}(\mu_s)^{-1} \geq \bar{\delta}(\mu_s)^{-1} \underbrace{\frac{1}{\sum_{s' \in S} \bar{\delta}(\mu_{s'})^{-1}}}_{\leq 1} = p_s.$$

Each $\mu \in \mathcal{F}$ also belongs to $\bigcup_{d \in \mathcal{M}} \mathcal{F}^d$. Hence, $\bigcup_{d \in \mathcal{M}} \mathcal{F}^d \supseteq \mathcal{F}$.

(ii) Take $\boldsymbol{\mu} \in \bigcup_{d \in \mathcal{M}} \mathcal{F}^d$. Then, for every signals s , there exists a distribution over signals $p \in \Delta(S)$ such that $\bar{\delta}(\mu_s)^{-1} \geq p_s$ for every s . Note that

$$\bar{\delta}(\mu_s)^{-1} \geq p_s = 1 - \sum_{s' \neq s} p_{s'} \geq 1 - \sum_{s' \neq s} \bar{\delta}(\mu_{s'}).$$

Thus, for each signal s , it holds that $\bar{\delta}(\mu_s)^{-1} \geq 1 - \sum_{s' \neq s} \bar{\delta}(\mu_{s'})$, equivalent to $\sum_s \bar{\delta}(\mu_s) \geq 1$. Each $\mu \in \bigcup_{d \in \mathcal{M}} \mathcal{F}^d$ also belongs to \mathcal{F} . Hence, $\mathcal{F} \supseteq \bigcup_{d \in \mathcal{M}} \mathcal{F}^d$.

□

B Appendix: Additional Figures

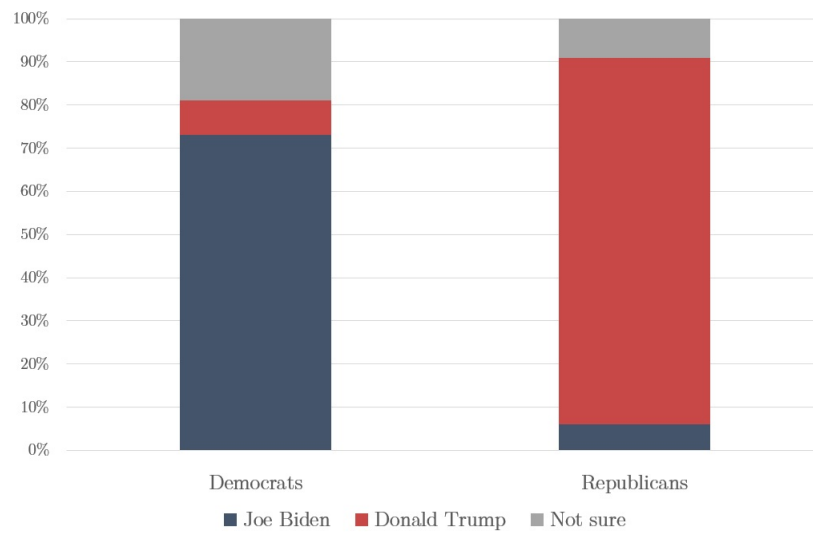


Figure 11: Priors on election winner by party affiliation

Notes: The y-axis shows the percentage of answers to the question “Who do you think will win the 2020 presidential election?” by reported party affiliation. *Source:* Economist/YouGov poll, October 25-27 2020.

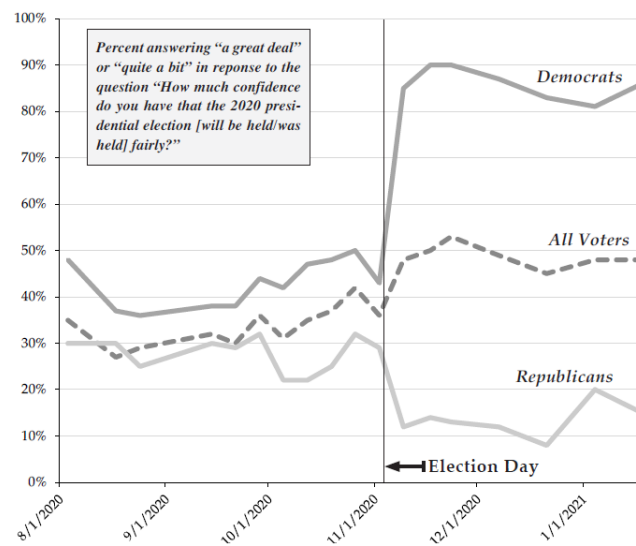


Figure 12: Confidence in fair election (Persily and Stewart, 2021)

Notes: The y-axis shows the percentage answering “a great deal” or “quite a bit” in response to the question “How much confidence do you have that the 2020 presidential election [will be held/was held] fairly?” *Source:* Economist/YouGov poll.