Tailored Stories

Chiara Aina*

November 3, 2022

Job Marker Paper
Latest version available here

Abstract

I study the problem of a persuader who proposes a set of models to manipulate how a boundedly rational agent interprets observable signal realizations. The persuader does not control or know the signal the agent observes. 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 her expected utility under the adopted model. I characterize the extent of belief manipulability in this setting and show that the agent may hold inconsistent beliefs across signal realizations — posterior beliefs across realizations do not average to the prior — because each signal may trigger the adoption of a different model. While persuasion can mislead the agent, the extent to which she is vulnerable to it is driven by her initial beliefs. Polarization is inevitable if agents with sufficiently different priors are exposed to the same conflicting models. I apply this framework to political polarization, conflict of interests in finance, lobbying, and self-persuasion.

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, Tristan Gagnon-Bartsch, George Loewenstein, Fabio Maccheroni, Delong Meng, Matthew Rabin, Andrei Shleifer, Tomasz Strzalecki, Adi Sunderam, Heidi Thysen, 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 event, we might reach contrasting conclusions. Voters may disagree on the outcome of an election. Consumers often 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. One potential explanation for reaching divergent conclusions in such cases is that we adopt different narratives to interpret the same data. Narratives link what we observe to what we want to understand: they provide interpretations of events.[1] Thus, influencing the narratives people adopt can be a powerful tool to manipulate and persuade them. Indeed, when making sense of the observed data, one might rely on narratives provided by more knowledgeable sources, such as political figures, financial advisors, or experts considered trustworthy. This type of persuasion is powerful because it allows for changing people's beliefs without controlling or even knowing what they observe.

I study the problem of persuading a boundedly rational agent by strategically providing interpretations of possible events independent of what is actually observed. Consider an agent (the receiver, she) who after observing a new piece of information about the relevant payoff state takes an action that affects both her payoff and the persuader's (the sender, he). This additional information on the unknown state, a signal, is generated by a fixed stochastic process. The sender cannot manipulate the signal or the process generating it. Still, he can provide the receiver with one or multiple ways of interpreting the possible signal realizations, called *models*. Following Schwartzstein and Sunderam (2021), a model provides likelihood functions assigning to each state a distribution of signals conditional on that state.[2] Persuasion arises because the receiver adopts the most plausible model given what she observed. This is formalized by adopting the model that maximizes the likelihood of the realized signal given her prior, the *fit*. Without knowing the signal realization, the sender strategically communicates models to manipulate how the receiver interprets the different signal realizations.

Suppose a politician wishes to persuade a (representative) voter that he is the legitimate president 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 models about the election system. Assume that the politician communicates only the model according to which the voting system is fair. Since there is only one model available, the voter always adopts that. Then, once the election outcome is revealed, the voter would recognize the politician as president if the latter is the reported winner, while she would not otherwise. How can the politician be recognized as the legitimate president regardless of the election outcome? He cannot manipulate the reported election outcome or the voting system. Therefore, before the vote, the politician also promotes a conspiracy theory according to which elections are rigged.[3]

[1] The Cambridge Dictionary defines a narrative as "a particular way of explaining or understanding events." Despite 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 when reviewing the related literature. Barron and Fries (2022) provides a detailed discussion of the current conceptualization of narratives in economics in their appendix.

[2] Sometimes, while discussing examples, I informally refer to models as narratives or stories.

[3] For simplicity of exposition, I describe these inconsistent models as provided by a single agent. Alternatively,

Exposure to multiple models allows inconsistent reasoning to take root: each election outcome triggers the adoption of a different model. The voter's initial beliefs play a crucial role because they drive which model is adopted based on the reported outcome. Assume that the voter expects the politician to win the election fairly. If the politician is the reported winner, the most plausible model is the one about the just voting system; however, if the politician is not the reported winner, the conspiracy theory resonates best with the voter. This is equivalent to the voter holding an inconsistent interpretation across election outcomes: "if this politician is reported as winner, the election system is fair; otherwise, elections are rigged."[4] As a result, the voter is updating upwards her beliefs about the politician being the legitimate winner, recognizing him as president, 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 conflicting models.

To what extent can the sender manipulate the receiver's beliefs using models? To study this question, it is necessary to keep track of the beliefs the receiver holds conditional on every signal realization. Therefore, the main object of the analysis is an array of the receiver's posterior beliefs conditional on each signal, called *vector of posterior beliefs*. In the previous example, this means describing the voter's beliefs conditional on both election outcomes: when the politician is the reported winner of the election and when he is not. The main result of this paper characterizes the set of feasible vectors of posterior beliefs. It conveys two main insights.

First, the sender can always bias the receiver's beliefs in a given direction. If many models are provided, each signal might lead the receiver to adopt a different model. As a result, the receiver's beliefs may be inconsistent across realizations: the prior cannot be expressed as a convex combination of the posteriors across signals. Bayesian models do not allow for this type of inconsistency. In this setting, the sender can induce a feasible and inconsistent vector of posteriors by providing as many models as possible signal realizations. In the example, the voter's posteriors are both higher than her prior. The politician achieves this with two models: one tailored to the case of reported victory and one tailored to the case of reported loss.

Second, there are constraints on beliefs the sender is able to induce. Generally, not all vectors of posteriors are feasible. To induce a vector of posteriors, the sender should construct a set of tailored models so that each model is adopted conditional on the signal to which it has been tailored, inducing the desired posterior conditional on that signal. Because models compete with each other across signal realizations, such a set of models does not always exist. The intuition is the following. There is a trade-off between how well a model can explain a signal and how much it can move posteriors far from the prior given that signal. To ensure that each signal triggers the adoption of its tailored model, the posteriors across realizations should not be too distant from the prior overall: the theorem pins down the extent to which this is feasible. Also, the more a model fits a signal, the more freedom to move posteriors away from the prior

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.

[4]The following are examples of other domains in which agents might hold inconsistent interpretations, as a result of selecting different models conditional on different facts. While interpreting a grade at school, a student that believes she is competent in a subject might believe the following story: "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."

conditional on the other signals with other models. Therefore, the maximal belief manipulability is generated by *maximal overfitting*: each tailored model maximally fits the target signal given the desired posterior. To better convey intuitions behind these formal results, I introduce a graphical approach for the special case of binary signal and state (hereafter, binary case). This also yields a graphical construction of which vectors of posterior beliefs are feasible.

Having explored the limits of belief manipulability, I turn to the question of what makes the receiver more vulnerable to persuasion. Initial beliefs play a crucial role. 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. When her prior is 50-50, the receiver is fully persuadable: the sender can provide a set of models to make her hold any beliefs regardless of what she observes. More generally, 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 priors. If the signals are at least as many as the states, a receiver with a uniform prior can be persuaded to believe anything. If there are fewer signals than states, the receiver is not fully persuadable regardless of her prior.

What if the receiver does not only consider the models provided by the sender? In an extension, I allow the receiver to hold initially a model by default. She adopts other models only if these are better at explaining new information with respect to this default model. I characterize the set of feasible vectors of posteriors in this case and show how a default model constrains belief manipulation. Then, I prove how this result is connected with 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.

I present several stylized applications that fit this setting. First, I elaborate on the above example of the politician and the voter. Being exposed to multiple models 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 stories is one of the central features of the "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. Even if beneficial for the persuader, such a communication strategy can have severe consequences for the audience: conflicting models 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 models, their beliefs always diverge further. Regardless of what happens, voters adopt different models 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.[5]

Second, 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

[5]The 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 stylized facts in line with my predictions.

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 models that could be picked up depending on the private information of the investors, e.g., past financial experience. The advisor can manipulate investors regardless of their 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. 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. This does not work for investors with pessimistic priors; thus the advisor needs to adjust his communication to persuade them.

The third application explores a multiple-selves setting in which an agent can distort her own beliefs by manipulating the perceived informativeness of observable signals. This proposed mechanism can deliver the classic implications of the literature on motivated beliefs but also gives a bound on belief distortion. First, I comment on how leaving data open to interpretations allows inconsistent reasoning to take root and may be used to achieve self-serving beliefs. Second, I illustrate how models allow an agent to distort her confidence to offset her time-inconsistent preferences and commit to a costly action.

Last, I show how a strategic persuader could challenge a shared model 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 effects of smoking, where the so-called "merchants of doubt" (e.g., Michaels, 2008; Oreskes and Conway, 2011) provided alternative ways of interpreting scientific evidence. Holding a shared initial model does not deter polarization in a population of heterogeneous receivers.

This paper speaks directly to two strands of economic literature — narratives and persuasion — that have both flourished 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 (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 approach. Inspired by the interdisciplinary research on sense-making (Andreassen, 1990; Weick, 1995; DiFonzo and Bordia, 1997; Chater and Loewenstein, 2016), Schwartzstein and Sunderam (2021) formalize the concept of models as used in this paper and assume that individuals prefer the model that best fits the observed data and prior knowledge. They study the problem of manipulating a receiver endowed with a default model by strategically providing her with a model after a public signal is realized (*ex-post*). Instead, I investigate a setting in which the sender commits to his communication strategy without knowing the signal realization (*ex-ante*). The reason is two-fold. First, it is a sensible assumption. Shifting communication ex-ante may give more credibility to the sender. For example, a voter may be skeptical to hear the politician claiming elections to be rigged only after he lost the election. Also, the sender might be unable to learn the information available to the receiver in some cases. For example, an investor might prefer not to disclose to her financial advisor some relevant private information, such as previous experiences. Second, ex-ante commitment imposes a constraint on the sender. The main result of the present paper addresses this. Extending this result to the case in which the receiver has a

default model allows comparability with Schwartzstein and Sunderam (2021). I find that, with a default model, the sender can attain the same outcome with ex-ante or ex-post communication of models.[6] However, because models compete with each other, the set of ex-post optimal models might not be optimal if provided ex-ante.

The strategic provision of models implies significant differences from the previous literature on persuasion. The sender does not alter the signal the receiver observes, unlike the cheap talk literature (e.g., 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 on Bayesian persuasion, started by Kamenica and Gentzkow (2011) and continued by many generalizations of their framework (e.g., Alonso and Câmara, 2016; Ely, 2017; Galperti, 2019; Ball and Espín-Sánchez, 2021). In broad terms, these papers are about persuasion by generating information: the sender commits to an experiment that maps each state realization to a distribution of signals. In the political example, this translates into the politician manipulating the voting system and its accuracy. Because the chosen signal generating process induces a distribution over the receiver's posterior beliefs, such distribution must be Bayes-plausible: the expected posterior has to average to the prior. With this paper, I relax the Bayes plausibility constraint in a disciplined manner. By providing models ex-ante, the sender can induce posteriors across realizations unattainable with Bayesian persuasion. However, this communication strategy generally imposes restrictions on what the sender can achieve without the ability to modify how the signal is generated or which signal is observed.[7]

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 $\mu_0 \in int(\Delta(\Omega))$.[8] 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

[6] The sender can induce any posterior by proposing a model ex-post if the receiver has no default model. This is not always the case ex-ante, as shown in the main theorem. Without default model, it depends on the receiver's prior and sender's preferences if ex-ante commitment lowers is costly for the sender compared to ex-post communication. I discuss this in Section 3.3 and provide an example in Section 4.2.

[7] Ichihashi 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. 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. 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.

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

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.[9]

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 under the signal generating process. Let $\boldsymbol{\mu} = (\mu_s)_{s\in S} \in [\Delta(\Omega)]^S$ be a vector of posterior beliefs: it 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:

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

Given a set of models M, the receiver's resulting vector of posterior beliefs is such that for each signal the posterior is induced by the model with the highest fit, i.e., $\boldsymbol{\mu}^M = \left(\mu_s^{m_s^*}\right)_{s\in S}$. 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

Before presenting the results, 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 occurs via maximum likelihood: once exposed to a set of models, the receiver adopts the model that fits best the observed data given her prior.[10] The receiver does not come up with every model she is willing to entertain, but

[9] I further assume that the receiver does not update her prior if she observes a signal that occurs with zero probability under every model communicated by the sender.

[10] The literature of belief updating under ambiguity considers a maximum likelihood updating rule, introduced

she compares only the models she was exposed to. Importantly, only one model is used to make inference.[11] 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 of fit.[12] 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.[13] Second, I assume the receiver to be naïve. In particular, the receiver does not know the true model and does not form subjective beliefs about the possible models. As a result, she cannot anticipate the sender's value function even if she knows or can learn about the sender's preferences. This prevents the receiver from being strategic about the models she receives. If she could form beliefs about the true model, it would allow her to learn more about the 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 — and the sender cannot access it. Second, the sender can communicate as many models as he wishes. Because he does not know the signal realization, he has incentives to send multiple 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 — unless the sender wishes to persuade multiple receivers.

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.

[11] The "traditional" alternative would have been to assume the receiver to hold 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.

[12] I abstract from the reasons why this is the case. For example, it might be that the most plausible model is adopted because people believe what they are prepared to hear, or it might also be that communicated models are stored in the receiver's memory and the best-fitted one is the easiest to retrieve (e.g., Bordalo et al., 2017).

[13] Model selection via maximum likelihood is equivalent to selecting the model with the higher posterior probability given the signal, starting from a flat prior over proposed models. Indeed, there is some evidence that people choose the most probable hypothesis. Simple 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).

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 posterior beliefs. 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 posteriors can be induced, 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: the expected posterior must be equal to the prior.[14] By contrast, this paper assumes 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. Once the sender knows what vectors of posteriors he can induce, he optimizes his objective over this set of feasible vectors of posteriors.

As a first step, I show an equivalent representation between models (information structures with fixed signal space) and vectors of posteriors (the support of the distribution 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 posteriors that are Bayes-consistent. Also, let $\boldsymbol{\mu}^m$ be the vector of posteriors such that each posterior is induced by model m. Bayes-consistency is the only restriction that Bayesian updating imposes on vectors of posteriors.

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 a model can have when inducing posterior μ_s.

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

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 2: conditional on a signal, the maximal fit for a target posterior coincides with the reciprocal of the maximal

[14] There is a recent literature studying the set of feasible joint posterior belief distributions for multiple Bayesian agents (e.g., Arieli et al., 2020; Morris, 2020; Mathevet et al., 2020) or multiple signals (Levy et al., 2022).

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 only Bayes-consistent vectors of posterior beliefs are feasible when a single model is proposed.

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.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.

Theorem 1 (Many Models)**.** *The set of feasible vectors of posterior beliefs is*

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

Allowing for multiple models expands the feasibility set with respect to the case of one model only.[15] As a consequence, vectors of posteriors that are not Bayes-consistent are feasible. Therefore, 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. However, it is not always the case that 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 two conditions are satisfied. First, each model is tailored to a specific signal realization, inducing the desired posterior conditional on that signal.[16] Second, each model is adopted conditional on the signal it has been tailored to. The latter condition introduces 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 realizations 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

[15] When 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.

[16] Providing 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.

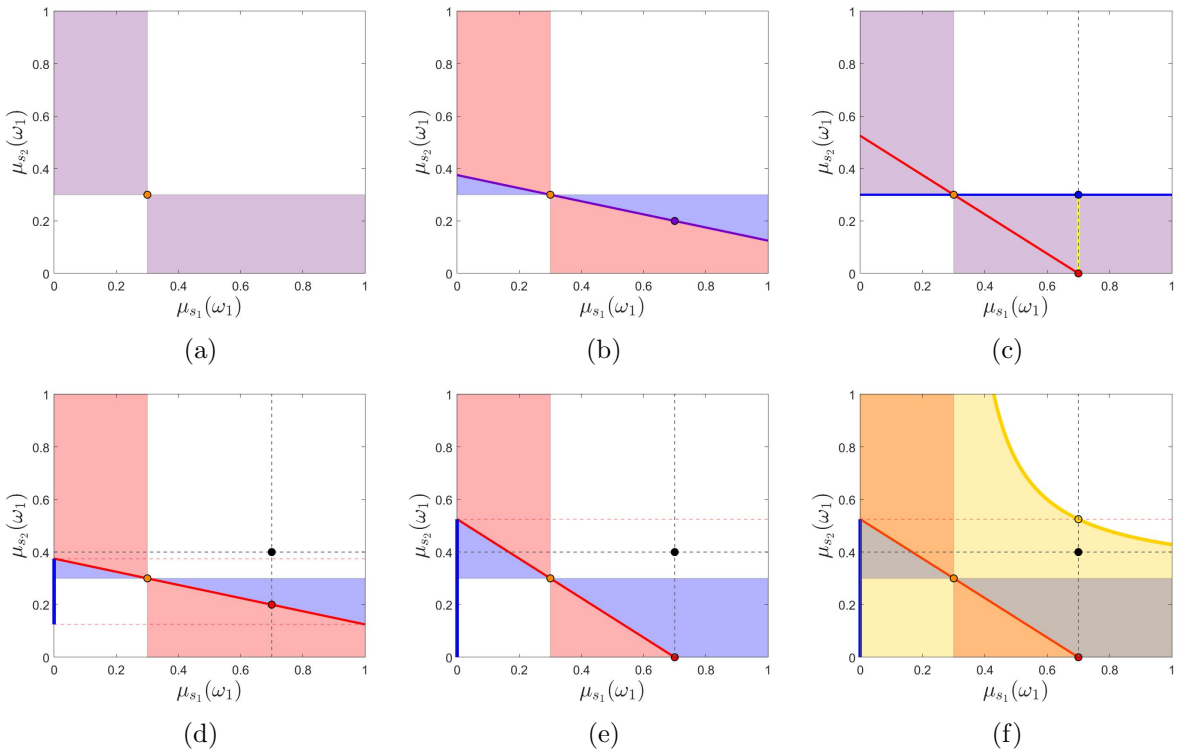


Figure 1: Graphical intuition for the binary case

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 point 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 condition implies that updating beliefs always in the same direction is impossible. In the binary case, there is a one-to-one map between Bayes-consistent vectors of posterior beliefs and models.[17] Hence, every point in the purple area corresponds to a model.

Focusing on a model as in Figure 1b, it is possible to observe some of its properties graphically. To do so, consider the purple line passing through the vector of posteriors induced by that

[17]The 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.

model (purple point) and the prior (orange point). This is the *isofit* line associated with the considered model: all the points on that line correspond to models with the same fit conditional on each signal.[18] The slope of the isofit line 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 $\mu_{s_1}(\omega_1) = 0.7$, 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 $\boldsymbol{\mu}$ is feasible if it is possible to find two models m_1 and m_2, respectively, inducing μ_{s_1} and μ_{s_2} and adopted conditional on s_1 and s_2. In Figure 1, the target vector is the black point. 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 imply that the target vector is unfeasible. The model m_1 I started with did not allow to induce the target vector of posteriors, but others could. As shown in Figure 1c, many models induce the target posterior conditional on s_1 with different fit levels. In particular, consider the model m_1 inducing μ_{s_1} with maximal fit in Figure 1e. I focus on this one because increasing $\Pr^{m_1}(s_1)$, $\Pr^{m_1}(s_2)$ decreases, and thus 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 point in Figure 1f — exemplifies how to construct the upper frontier of the set of feasible vectors of posteriors. All vectors below that line (yellow

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

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

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 $\mu^m_{s_2}(\omega_1) = \frac{\mu_0(\omega_1)}{\Pr^m(s_2)} - \frac{\Pr^m(s_1)}{\Pr^m(s_2)}\mu^m_{s_1}(\omega_1)$. 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.

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* $\min_{\omega\in\Omega}\mu_0(\omega) \geq \frac{1}{|S|}$*, 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 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. Moreover, with fewer signals than states, the receiver is never fully persuadable because the condition of Proposition 2 is never satisfied. Therefore, the 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. 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 a 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 $\mu_{0,\varepsilon} = (\mu_{0,\varepsilon}(\omega_1), \mu_{0,\varepsilon}(\omega_2)) = \left(\frac{1}{2}-\varepsilon, \frac{1}{2}+\varepsilon\right)$ and $\mathcal{F}_\varepsilon$ the set of the feasible vectors of posteriors with respect to this prior.

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

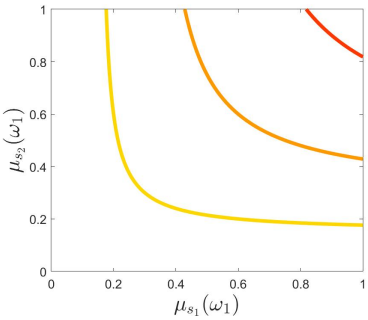


Figure 2: Frontier of the feasibility set, by prior

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 $\mu_0(\omega_1) = 45\%$.

3.3 Sender's Problem

Given these results, I turn to the sender's problem. Informed about the receiver's prior, the sender knows to what extent he can manipulate her beliefs. 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, she does not update her beliefs: she holds $\boldsymbol{\mu}^{\varnothing}$ 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 $\boldsymbol{\mu} \in \mathcal{F}$ such that its value is higher than the value of the prior: $V(\boldsymbol{\mu}) \geq V(\boldsymbol{\mu}^{\varnothing})$.

The sender faces two main restrictions in choosing his communication strategy. 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 $\boldsymbol{\mu} \in [\Delta(\Omega)]^S$ can be induced.*

Second, the sender provides models without knowing the signal realization. What is the impact of this assumption on the sender's expected utility? Knowing which signal the receiver observes allows the sender to communicate a tailored model inducing the desired posterior. Avoiding competition among models across signal realizations, he could induce any vector of posteriors.[19] The cost of committing ex-ante to models equals the gap between the unconstrained maximal sender's value over any vector of posteriors and the maximal sender's value over the feasible

[19]This 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.

vectors of posteriors:

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

When all vectors of posteriors are feasible, the sender cannot gain from communicating models ex-post. However, the commitment cost is positive if the unconstrained maximum is unfeasible. In particular, this measure captures how much the sender is willing to pay to learn the data available to the receiver. It can be used to comment on the value of microtargeting.[20] Knowing the private information available to the receiver allows the sender to tailor models perfectly. I discuss an example of this in Section 4.2.

4 Applications

This section discusses several applications. The first formalizes the political example outlined in the introduction and sheds light on the polarizing consequences of conflicting models. Then, I provide suggestive evidence of this mechanism. Second, a financial application illustrates the sender's optimization problem in 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.[21] The growing interest in understanding fake news revealed that people have a hard time in distinguishing true and false stories: fake news is widely shared and believed (Allcott and Gentzkow, 2017).

Building on the political example presented in the introduction, I illustrate why it is in the interest of the sender to communicate conflicting models.[22,23] 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.

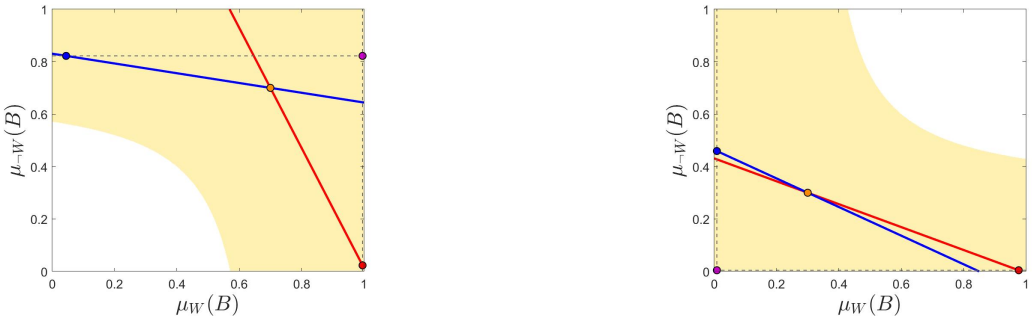
Politician Bob is running for the presidency. He wants to be recognized as president by part of the voters regardless of the reported election outcomes. To make sure of that, he is spreading two different models about the reliability of the election system. Let the state space be $\{B, \neg B\}$,

[20] This is a well-established practice in marketing: analyzing online information on potential customers to create and convey the most effective message. As a result, different ads are shown to different groups of consumers. 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`.

[21] The 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.

[22] For simplicity of exposition, I describe this strategy as adopted by a single politician. One can imagine different members of the same party implementing the same strategy in a coordinated manner.

[23] Interestingly, 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 (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.



(a) $\mu_0(B) = 70\%$ (b) $\mu_0(B) = 30\%$

Figure 3: Firehose of falsehood, by 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.

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 $\boldsymbol{\mu} = \left(\mu^f_W, \mu^c_{\neg W}\right)$ (purple point). This type of voter always recognize Bob as the legitimate president regardless the election outcome. Thus, Bob achieves his goal of being recognized as president by at least part of the voters. However, this has repercussions on other type of voters. Voters that expect Bob to win with low probability never recognize him as president (Figure 3b).

Interestingly, for different priors, the same pair of models 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 models persuade a voter with $\mu_0(B) < 33\%$ to never support Bob as president, always believing him to be an illegitimate president.

4.1.1 Inevitable Polarization

The previous example illustrates how the exposure to conflicting models might be a strong driver of inevitable polarization in a population of heterogeneous agents. It is possible to generalize this result for the binary setting.

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 each signal.

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 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 such that receivers with prior higher (lower) than the threshold would hold extreme high (low) posteriors regardless of the signal realization.[24] 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 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, e.g., Dixit and Weibull, 2007), 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: 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), 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.

Several mechanisms have been proposed in the literature to understand the determinants of this phenomenon. Often polarization is associated with confirmation bias, formalized for the first time by Rabin and Schrag (1999). They assume agents misinterpret new information as supportive 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

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

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 a uniform prior over states. Baliga et al. (2013) provides an explanation for polarization based on ambiguity aversion, in which agents hedge against uncertainty by 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 compared to state space (Andreoni and Mylovanov, 2012) or private signals on the interpretation of evidence (Benoît and Dubra, 2019). Recent papers discuss how mistakes in source credibility could amplify polarization. Cheng and Hsiaw (2022) investigate the belief distortion due to double-using the data to update beliefs on the states while also updating beliefs on source accuracy. Also, this mechanism can lead agents to disagree on how to interpret the same data. Gentzkow et al. (2021) shows how a small bias in data perception due to ideological preferences can cause divergent beliefs about both the state and the source precision, 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 models, both given different priors observing the same outcome and given the same prior observing different outcomes.

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 widespread voter fraud in the election, yet competing narratives on dysfunctional elections were broadly diffused. 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. 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 idea that the election system is fair fits well with the application discussed in the previous section. In what follows, I discuss some suggestive evidence of the mechanism I formalized above: conflicting models drive polarization in a population of heterogeneous voters. In particular, when exposed to conflicting models: (i) voters with different initial beliefs adopt different models on the election system once the outcome realizes, and (ii) voters with the same initial belief adopt different models if they observe different outcomes. I rely on insights on the 2020 US election provided by Persily and Stewart (2021) to discuss stylized facts in line with 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 expect Donald Trump to win and Democrats expect Joe Biden to win. On average, this assumption is verified: at the end of October 2020, the expected winner of the presidential election was Donald Trump for 85% of Republicans and Joe Biden for 73% of Democrats.[25,26]

[25] See Appendix B for more details about the distributions of priors.

[26] This pattern in priors is consistent with motivated beliefs or wishful thinking: voters wish their partisan candidate to win, influencing their expectations. I assume these motives might affect initial beliefs but not model selection or the updating procedure. A proper test of this paper should account for these confounding forces.

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, an extreme partisan polarization occurred: the gap between Democrats and Republicans went from 10.9% to 51.7%.[27] This suggests that voters with different priors adopt different models once the election outcome is observed: 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 (Persily and Stewart, 2021). A potential explanation is the disproportionate spread of distrustful narratives during the 2020 election campaign compared to previous elections.

Suggestive evidence about the second prediction can be found by looking at how voters' confidence in state elections changes depending on the state's reported election outcome. Figure 5 reports data on the confidence in state elections by the percentage of Trump share of votes. Republicans mostly distrust the accuracy of the state elections if residents 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 (Clark and Stewart, 2021). This gap supports the idea that voters with similar initial beliefs adopt different models if they observe different realizations. This pattern would be difficult to explain without invoking the idea that they are exposed to conflicting models. Indeed, the same graph for Democrats barely exhibits a discontinuity. Since most of these alternative narratives about the election accuracy were right-leaning, 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).

4.2 Financial Advice

Next, I illustrate the optimization problem for a financial advisor who wants to persuade investors to make a specific investment. It is well-known that commissions on investments could lead to a conflict of interest for the advisor. I consider the case in which the advisor knows that the investors' past financial experience influences their beliefs about the quality of new investments.[28] However, the advisor does not have access to this piece of private information. Nonetheless, he has the incentive to persuade the receiver to invest as much as possible.

To manipulate the investor, the advisor can propose different ways to predict future returns based on past returns. In finance, two important alternatives are that returns can exhibit pre-

[27] The 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.

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

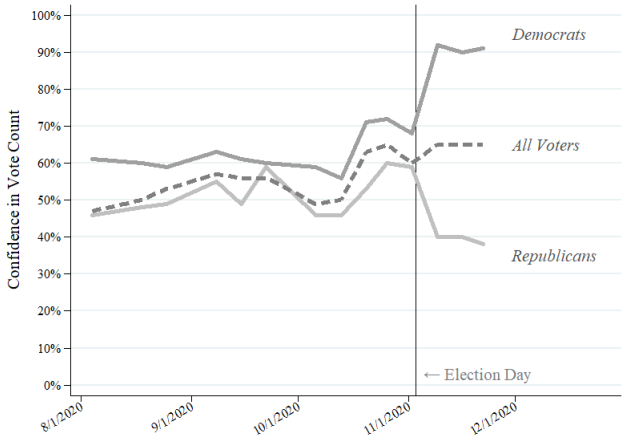
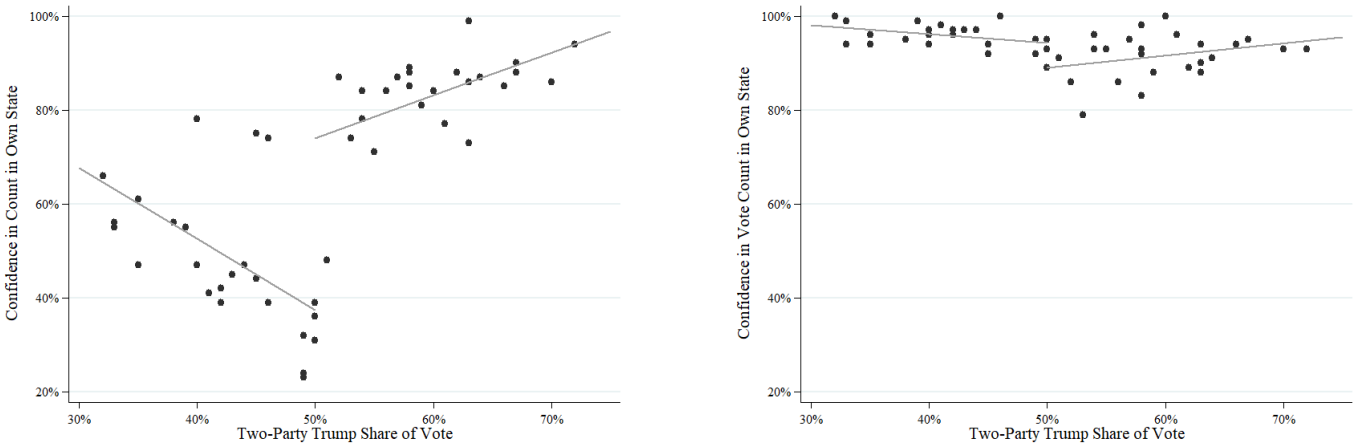


Figure 4: Accuracy of vote count (Persily and Stewart, 2021)

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.



(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 (SPAE), November 2020.

dictable mean reversion or predictable continuation. When returns exhibit predictable mean reversion, high past returns predict low future returns, and a contrarian strategy —selling following high returns — is profitable. When returns exhibit predictable continuation or momentum, high past returns predict high future returns, and a return chasing — buying following high returns is profitable. Both phenomena have been empirically well documented in finance, with Fama and French (1992) and Lakonishok et al. (1994) finding predictable mean reversion and Jegadeesh and Titman (1993) finding momentum in US stocks. Professionals rely on empirical measures to detect these patterns and choose the most effective trading strategy. Still, inexperienced investors might interpret past financial performance through the strategy that resonates best with their initial beliefs. The advisor can use simplified versions of these theories to his advantage. As shown formally in the following 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 model in terms of future performance. However, if the investor is pessimistic about the advisor-preferred asset, communicating these two models is counterproductive, and the advisor needs to adjust his communication.

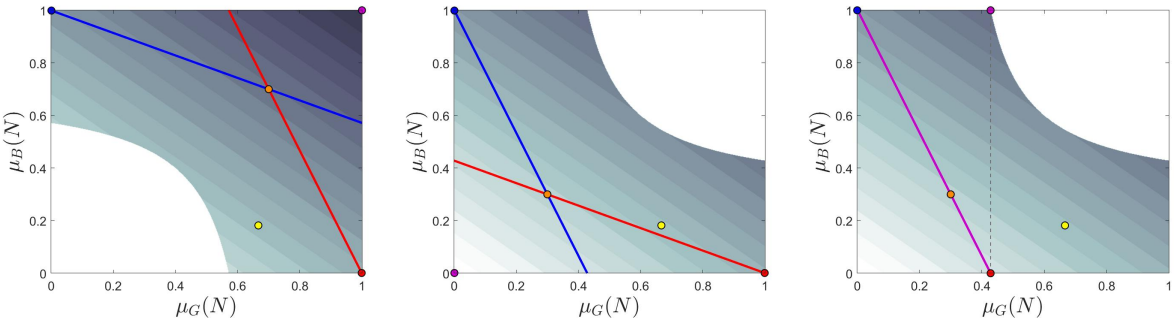
I discuss these insights in the context of choosing a hedging strategy. Hedging aims to limit the risk of uncertain events on financial assets. Usually, it involves diversification in offsetting or opposite positions. Formally, each investor has to allocate one unit of endowment over two possible outcomes, $\Omega = \{N, \neg N\}$, where N is the event of normal conditions and $\neg N$ is the event that extreme or catastrophic circumstances occur, e.g., financial crisis or fluctuations of foreign currency in the economic domain, or flood or drought as extreme weather events. 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 30% and *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.

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 $r_N > r_{\neg N} = 0$, his value function is $V(\boldsymbol{\mu}) = \sum_{s \in \{G,B\}} \Pr^t(s) \ r_N \ \mu_s(N)$. The sender expects normal conditions with probability of 40%. I assume the sender to know the investors' prior, but not their private information. The sender knows the true model where a positive past experience positively (negatively) correlates with the success (failure) of the investment: $\pi^t(G|N) = 75\%$ and $\pi^t(G|\neg N) = 25\%$. Given this, he expects investors with past good experience with probability 45% and with a past negative experience with probability 55%.

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. 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.

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 never hedges against extreme circumstances. Intuitively, this means that the advisor can leverage any past experience of the investor and move her beliefs always in the advantageous direction. He needs two models to achieve that. One option is to expose the investors to the following pair of models: (1) model m_1 suggesting a perfect positive correlation 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 read as simplified versions of the momentum ("early success predicts long-run success") and mean-reversion theories ("what goes down goes up"). Because of their optimistic initial beliefs, investors adopt the first given a good experience and the second given a bad experience. As a result, they never hedge against adverse events.



(a) $\mu_0(N) = 70\%$, optimized (b) $\mu_0(N) = 30\%$, not optimized (c) $\mu_0(N) = 30\%$, optimized

Figure 6: Financial advice, by investors' prior and advisor's communication

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 is 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.

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 model — 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 at $\boldsymbol{\mu}^* = ((0.43, 0.57), (0, 1))$ at the right-top kink (Figure 6c). 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 investors, inducing $(0, 1)$ (blue point). According to model m_1, a bad experience is a perfectly revealing signal of extreme conditions. In contrast, a past positive experience is an indefinite news. This model encourages only investors with a positive experience to have a positive α_N, and, indeed, it is tailored to those. Again, model m_2 is an oversimplified version of the mean-reverse theory, which pushes investors with a bad experience to invest the whole endowment in the advisor-preferred

outcome. To conclude, note that since the first-best of convincing all pessimistic investors to invest in outcome N fully is not attainable, the advisor shifts to the second-best: convincing the largest group of investors (the ones with bad experiences) to set $\alpha_N = 100\%$ while increasing α_N for the other group (the ones with good experiences) as much as possible.

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 to convince them to invest in his preferred outcome fully, inducing $\bar{\boldsymbol{\mu}} = ((1,0),(1,0))$ (right-upper corner in all figures). With pessimistic investors, $\bar{\boldsymbol{\mu}}$ is unfeasible. Therefore, the cost of commitment is $\Delta = V(\bar{\boldsymbol{\mu}}) - V(\boldsymbol{\mu}^*) = 74\% r_N$. In contrast, with optimistic investors $\Delta = 0$ because the sender can always achieve his maximal payoff.

4.3 Self-Persuasion

This paper 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.[29] I consider a multiple-selves setting, where the conscious mind (receiver) demands the unconscious one (sender) to supply models. This proposed mechanism to achieve self-serving beliefs can deliver the classic implications of this literature, but it also provides a bound on belief distortion.

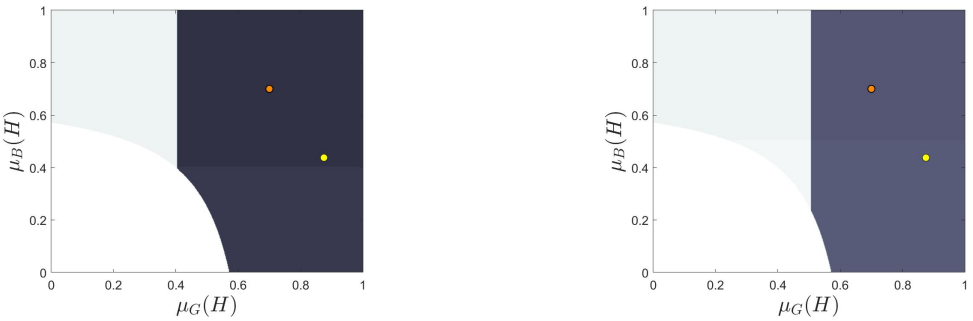
Confirmation bias can emerge because the agent initially keeps signals open to many favorable interpretations. This could be the case of a student who subconsciously likes to think to be good at school and thus leaves the informativeness of grades open to two interpretations: grades are a good measure of own ability, or grades are based on luck. Once she observes the grade, she wants to form accurate beliefs to decide how much to study for the next test. However, she always keeps high confidence in her abilities because she expects to be good at school. This occurs because she adopts the model according to which grades are informative after a good grade, but she believes grades do not convey much information after a bad grade.

Depending on the agent's subconscious preferences, beliefs might be distorted in any direction. The previous example assumes opposing goals for the sender and the receiver, such as higher self-confidence and accuracy. To better compare aligned and misaligned preferences, the next example focuses on a setting where there is a variable controlling the degree of misalignment of incentives among the parties. Building on the motivation problem of Bénabou and Tirole (2002), I explore a multiple-selves setting in which an agent distorts her own interpretations of signals to offset her time-inconsistent preferences and commit to a costly task.

The agent can have high (H) or low (L) abilities. She receives either a good (G) or a bad (B) signal. After observing the feedback at $t = 1$, she decides whether to take an action with disutility c that, with high abilities, would yield benefit v at $t = 2$. The agent at $t = 0$ (before the signal) acts as the sender, choosing potential interpretations of the future signals, while the receiver is the agent at $t = 1$ (after the signal). Because the agent has quasi-hyperbolic discounting (e.g., Laibson, 1997; O'Donoghue and Rabin, 1999), time inconsistency leads to

[29] Papers on motivated beliefs conjecture different sources of motivations or different channels through which beliefs are distorted, e.g., via direct utility (e.g., Köszegi, 2006; Brunnermeier and Parker, 2005) or via instrumental value associated with the beliefs (e.g., Bénabou and Tirole, 2002). For a survey, see Bénabou (2015).

misaligned incentives: at $t = 0$, both the cost and the benefit are in the future and present bias does not undermine the choice to act, unlike at $t = 1$. After the signal, the costly action if her beliefs given the signal are higher than $\frac{c}{\beta\delta v}$, where $\delta \leq 1$ is the discount factor and $\beta > 0$ is the present bias. Instead, before the signal, acting is optimal if the updated beliefs is higher than $\frac{c}{\delta v}$, which is lower than the relevant threshold at $t = 1$ if she suffers from present bias, e.g., $\beta < 1$.[30] The agent might have the incentives to distort her own interpretations of signals to avoid a future lack of willpower.



(a) No Present bias (b) Present bias

Figure 7: Motivated beliefs, by present bias

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 is the sender's vector of posteriors induced by the true model.

Consider an agent that initially believes to have high abilities with probability 70%. Signals are quite accurate, $\pi^t(G|H) = \pi^t(B|L) = 75\%$. Taking the costly action is always optimal at her initial beliefs and when she does not suffer from present bias. Indeed, updating her prior using the true model maximizes her ex-ante expected payoff (Figure 7a). She does not distort her beliefs in case of aligned incentives over time. However, self-deception could be beneficial in the case of sufficiently severe present bias (Figure 7b). Before the signal, she anticipates that the imminent cost of the action will be more salient than the future reward at the moment of the decision. Thus, conditional on the bad signal, confidence in her abilities will not be high enough to act. She overcomes this by distorting the perceived informativeness of upcoming signals — either discarding the signals as uninformative or believing only the good signal to be accurate enough. Belief manipulation allows her to stay motivated as in Bénabou and Tirole (2002), but through a different mechanism — manipulating how she interprets feedback rather than assuming 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 hold initially 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,

[30] To see how these thresholds were derived, let $a \in \{0, 1\}$ be the action. At $t = 1$, taking the action given signal s leads to $U^1(1) = -c + \mu_s(H)\ \beta\delta v$, while no action leads to $U^1(0) = 0$. At $t = 0$, the utility of taking the action is $U^0(1) = \beta\delta\big(-c + \mu_s^t(H)\ \delta v\big)$, while $U^0(0) = 0$.

as often individuals bring ways of interpreting data either generated on their own or provided by others in the past. I show how a default model restricts which beliefs the sender can induce.

5.1 Feasible Vectors of Posterior Beliefs with a Default Model

Assume that the receiver is endowed with a default model d known by the sender. The set-up is otherwise the same as in Section 2. The receiver adopts the model with the highest fit conditional on the observed signal s among the set of models M she has been 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)**.** *The set of feasible vectors of posterior beliefs is*

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

The proposed models compete not only 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.

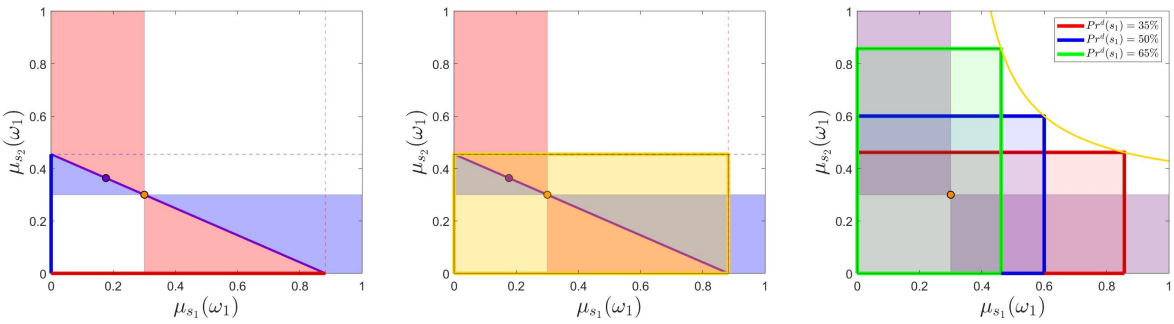
Interestingly, the presence of a default model eliminates the cost of ex-ante commitment for the sender. While, in the absence of a default model, every posterior is feasible when the sender can communicate a model knowing the signal, this is not the case if the receiver is endowed with a default model. 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. It shows that providing multiple models without knowing the signal (*ex-ante*) guarantees to induce any posterior achievable by communicating a model after observing the signal realization (*ex-post*). Ex-ante commitment does not prevent the sender from attaining the same expected utility he would get for each signal realization with ex-post communication. 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:

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

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

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

Note that this corollary does not imply that the sender should communicate ex-ante the same set of models that he would find optimal ex-post for each signal. Doing so could be self-defeating in some cases. With a 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, the sender might need extra care



(a) (b) (c)

Figure 8: Graphical intuition of Theorem 2 and Proposition 6

in choosing which models to communicate if there are more than two signals.

I conclude this section by discussing how the default model restricts the belief manipulability of the receiver. Indeed, 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: the set of feasible vectors of posteriors in the absence of a default model (Theorem 1) is the union of the sets of feasible vectors of posteriors with a default model for all default models.

Proposition 6.

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

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

5.2 Merchants of Doubt

"Doubt is our product, since it is the best means of competing with the 'body of fact' that exists

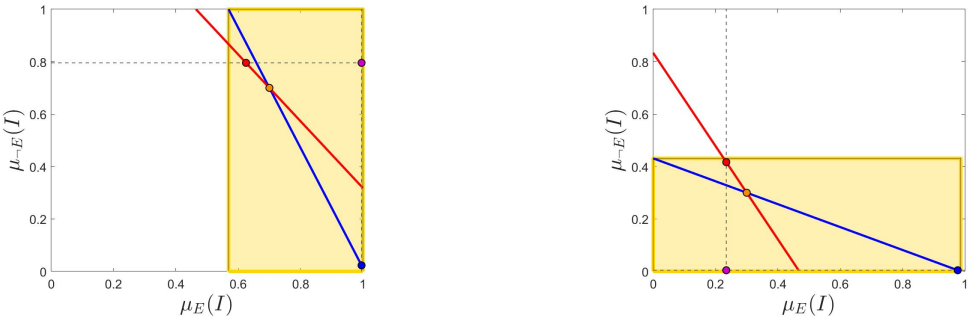
in the minds of the general public. It is also the means of establishing a controversy."

— 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 model 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 (e.g., blaming other factors, false positive results). Their aim was to delay regulations, defeat delegations, and insinuate doubt in the population. Ultimately, they diluted consensus, despite the scientific community having no doubt. This strategy guaranteed that even with new evidence emerging over time, the general public had already been exposed to competing ways of interpreting new facts and found them worthy of consideration. The following example illustrates how a competing model leads to polarization even when the population initially shared the same view.



(a) $\mu_0(I) = 70\%$ (b) $\mu_0(I) = 30\%$

Figure 9: Merchants of doubt, by agent's prior

Consider two states $\{I, \neg I\}$, where I is the event that the issue is real, e.g., smoking causes cancer. 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, $\pi^d(E|I) = \pi^d(\neg E|\neg I) = 99\%$ (blue point).

A lobby wants to strategically challenge this shared model 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 claim is that, if the issue is true, evidence emerges randomly ($\pi^m(E|I) = 50\%$), but there is a high 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 posteriors induced by this model corresponds to the red point. Once exposed to both models, the resulting posteriors vector corresponds to the purple point.

The default model induces a vector of posteriors almost identical for all agents. However, the sets of feasible vectors differ drastically. As a result, introducing an alternative model leads to

diverging beliefs regardless of the evidence. If initially doubtful about the issue, any piece of evidence makes agents more reluctant to believe the issue is real (Figure 9b). Instead, agents initially expecting the issue to be real become even more confident (Figure 9a). Strategically introducing a conflicting model promotes doubt among agents with different initial beliefs. Sharing the same default model does not prevent polarization.

6 Relationship to the Literature

This paper mostly contributes to three 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, starting with Shiller (2017, 2019). The present paper builds on the formalization of narratives as models 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.[31] I adopt this model selection rule and build on their framework, but investigating 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 assume that best-fitted models are adopted. Ichihashi and Meng (2021) sequentially combines Bayesian and model persuasion. They study a persuader who first designs and interprets information to persuade the receiver once the signal realizes. Their paper differs from mine in two ways: first, the sender chooses the signal generating process; second, they study ex-post communication of models. 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 after 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, recent papers take different approaches to what makes models persuasive. Ispano (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 model (the marginal distribution of news is undistorted).[32] He argues that coherence implies Bayes-consistent posteriors across signals and 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.

[31] Closely related, Levy and Razin (2020) study aggregation of forecasts over time. They assume the decision maker looks 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 an information structure with a binary signal. Similar to my results, the prior plays a crucial role in the evolution of beliefs.

[32] I assume models to be coherent by definition. However, one could argue that the receive 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.

Another influential way to formalize narratives is by describing them as 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 for estimation. The authors quantify the worst-case distortion when the proposed model is flexible in which variables enter the model under the constraint that the estimated model cannot distort the marginal distributions of individual variables. Increasing the number of variables in the model can lead to an almost perfect correlation.[33] Directed acyclical graphs are also used by Eliaz 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.[34]

Moreover, there are 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). 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, narrative that better fit the observed data are the more persuasive. Hence, narratives are a highly effective tool of persuasion and are hard to protect against. Moreover, in one treatment, the advisor does not have the opportunity to tailor the narrative to the data the investor observes. Unlike the present paper, the persuader is restricted to one narrative. Results show that advisors are less effective in moving beliefs to their target in this treatment. Overall, however, 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 focusing on narratives. Papers investigate their impact on beliefs and behavior (Graeber et al., 2022; Bursztyn et al., Forthcoming; Hillenbrand and Verrina, 2022; Harrs 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

[33] Also, according to (Olea et al., Forthcoming), including irrelevant covariates in models helps achieve 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.

[34] Also Izzo et al. (2021) assume that agents choose the model with the highest likelihood given the observed data. This translates into favoring the model with the smallest mean squared error in their setting.

main differences with respect to previous 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. A recent 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 in equilibrium. Eliaz et al. (2021a) builds on the classic cheap-talk game with multidimensional messages, relaxing the assumption that the receiver is capable of interpreting the equilibrium messages and allowing the sender to supply interpretations for them. 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 sometimes be attained. 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 which is then interpreted by Bayesian receivers. This restricts the sender to induce only Bayes-plausible distribution of posteriors, unlike this paper.

In a previous literature, ambiguity aversion has motivated ambiguous communication strategies by senders, proposing several explanations or messages. This was studied in cheap-talk games (Kellner and Le Quement, 2018, 2017) and in Bayesian persuasion (Beauchêne et al., 2019). The latter paper studies a sender who 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 and find that ambiguous persuasion could be more beneficial compared to Bayesian persuasion.

Biased Beliefs 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. The reason is two-fold. First, each signal generating process maps into a richer object, the vector of posterior beliefs. Second, the value of a posterior conditional of one signal is compatible with many possible signal generating processes. However, in the rich literature on biased beliefs (for a survey see Benjamin, 2019), most papers look only 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.[35]

Other papers suggest different criteria to form beliefs in uncertain settings that could lead to violation of the Bayes-consistency, such as the literature on belief updating with ambiguity-aversion. In the case of multiple priors, there are two main paradigms of Bayesian updating: full Bayesian (Jaffray, 1992; Pacheco Pires, 2002) and maximum likelihood (Dempster, 1967; Shafer, 1976; Gilboa and Schmeidler, 1993). A common consequence of both is that information increases the relevant ambiguity because the agent's set of beliefs dilates (Seidenfeld and Wasserman, 1993). As a result, ambiguity-averse agents should lower their valuations of bets for every possible piece of information they could receive ("all news is bad news"). Shishkin

[35] An exception is Esponda et al. (2020). Their focus is learning in presence of a initial misconception (focusing on the special case of base rate neglect). They use the vector of posteriors space to illustrated graphically part of their results. Their figures shows that participants' beliefs are sometimes Bayes-inconsistent. Other example can be found in the literature on polarization (see Section 4.1.1).

and Ortoleva (2021) experimentally test the dilation property and find that ambiguity-averse subjects do not lower their value of bets conditional on information. Epstein et al. (2021) define sensitivity to signal ambiguity as the attitude towards the uncertainty of the signal generating process and report that a fraction of subjects is averse to signal ambiguity using a lab experiment. According to their definition, being sensitive to signal ambiguity implies a violation of Bayes-consistency property extended to preferences.

This paper shows that inconsistent updating across signal realizations could result from adopting different models. This could be an explanation for some of the evidence on asymmetric updating in beliefs (e.g., Eil and Rao, 2011; Sharot, 2011; Ertac, 2011; Coutts, 2019; Möbius et al., 2022; Drobner and Goerg, 2021).[36]

7 Conclusion

Typically persuasion has been studied in settings where the persuader can control the information observed by the agent prior to making a decision, e.g., either by sending a persuasive message or providing new informative data. However, sometimes this is not possible. In such contexts, although the persuader cannot control or even know the information that the agent uses in making a decision, I demonstrate that the scope for persuasion using model is large, but generally bounded by initial beliefs.

Bayesian models assume that beliefs should be consistent across possible realizations: the receiver cannot update her beliefs in the same direction conditional on every signal. Exposure to multiple models can lead to the violation of this property. Because the receiver is boundedly rational in choosing the interpretation of outcomes she observes, each signal realization might trigger the adoption of a different model. Thus, the sender can leverage multiple models to induce beliefs across signals that the sender cannot attain by choosing the signal generating process as in Bayesian persuasion.

Several extensions could be explored in future research. First, this paper focuses on a problem with only a sender and a receiver. I discuss the consequences of conflicting models 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 effects models have. Moreover, I consider only one sender communicating multiple models. This can also be interpreted as a coordinated strategy by senders with the same incentives. The extension to the default model is the first step towards studying the competition among senders because the sender strategically responds to a model the receiver already holds. Much remains to be investigated in relation to multiple (uncoordinated) senders with possibly misaligned incentives. Second, I impose no restrictions on which models the sender is willing to supply and

[36]Results 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 monetary rewards. In contrast, 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 be crucial in switching on and off asymmetric updating, which is in line with the mechanism proposed in this paper. Indeed, the provision of models depends on whether it is possible to keep signals open to multiple interpretations (e.g., immediate vs. no resolution of uncertainty) or what incentives motivate the supply of interpretations (e.g., financial vs. positive beliefs).

the receiver is willing to accept. On the one hand, senders might be reluctant to communicate models too far from the true one. For example, belief distortion may bear 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 they also display truth-telling preferences to some extent. These frictions could be incorporated into a theoretical extension of this paper. On the other hand, receivers might consider only some types of models, e.g., models should never report a positive correlation between some signals and states, or models should explain a series of conditionally independent signals. Research along this line could shed light on how these restrictions harm or benefit the receiver.

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

Arieli, Itai, Yakov Babichenko, Fedor Sandomirskiy, and Omer Tamuz (2020) "Feasible joint posterior beliefs," *arXiv preprint arXiv:2002.11362*.

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.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2017) "Memory, attention, and choice," *The Quarterly journal of economics*.

Brunnermeier, Markus K and Jonathan A Parker (2005) "Optimal expectations," *American Economic Review*, 95 (4), 1092–1118.

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott (Forthcoming) "Opinions as Facts,"Technical report.

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.

Fama, Eugene F and Kenneth R French (1992) "The cross-section of expected stock returns," *the Journal of Finance*, 47 (2), 427–465.

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 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.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann (2022) "Stories, Statistics, and Memory."

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 Kőszegi (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," *SocArXiv. February*, 19.

Jaffray, Jean-Yves (1992) "Bayesian updating and belief functions," *IEEE transactions on systems, man, and cybernetics*, 22 (5), 1144–1152.

Jegadeesh, Narasimhan and Sheridan Titman (1993) "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of finance*, 48 (1), 65–91.

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.

Kőszegi, 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.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny (1994) "Contrarian investment, extrapolation, and risk," *The journal of finance*, 49 (5), 1541–1578.

Levendusky, Matthew S (2009) "The microfoundations of mass polarization," *Political Analysis*, 17 (2), 162–176.

Levy, Gilat, Inés Moreno de Barreda, and Ronny Razin (2022) "Persuasion with correlation neglect: a full manipulation result," *American Economic Review: Insights*, 4 (1), 123–38.

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.

Mathevet, Laurent, Jacopo Perego, and Ina Taneva (2020) "On information design in games," *Journal of Political Economy*, 128 (4), 1370–1404.

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*.

Morris, Stephen (2020) "No Trade and Feasible Joint Posterior Beliefs," *Working paper*.

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.

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 and Pietro Ortoleva (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.

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.

Weick, Karl E (1995) *Sensemaking in organizations*, 3: Sage.

Yang, Jeffrey (2022) "A Criterion of Model Decisiveness," *Mimeo*.

A Appendix: Proofs

Proof of Lemma 1. Consider the two statements separately.

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

Consider $\boldsymbol{\mu} \in \mathcal{B}$. 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 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 posteriors induced by such a model is $\boldsymbol{\mu}$.

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

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

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. □

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

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 beliefs 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 $\mu \in \mathcal{B} \backslash \{\mu^{\varnothing}\}$ for binary signal. □

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

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

Fix $p \in \left[0, \bar{\delta}(\mu_s)^{-1}\right]$. 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:

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

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,

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

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

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

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

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

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

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

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

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:

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

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

□

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

Proof of Theorem 1. Take a vector of posterior beliefs $\boldsymbol{\mu} \notin \mathcal{B}$, otherwise it would be trivially feasible by Proposition 1. Inducing $\boldsymbol{\mu}$ 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 $\boldsymbol{\mu} \in \mathcal{F}$. In what follows, I show that there exists a set of models inducing $\boldsymbol{\mu}$. 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 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

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

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 $\mu_0(\bar{\omega})\left(1 - \mu_{s_k}(\tilde{\omega})\right) > \mu_0(\bar{\omega})\left(1 - \mu_0(\tilde{\omega})\right) > \mu_{s_k}(\bar{\omega})\left(1 - \mu_0(\tilde{\omega})\right)$. This implies $\mu_0(\bar{\omega}) > \mu_{s_k}(\bar{\omega})$, which is a contradiction.

Also, posteriors sum up to one for every s:

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

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

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

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,

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

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 $\boldsymbol{\mu} \notin \mathcal{F}$. □

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 $s' \neq s, s' \in S$, $\delta(\mu_s(\omega)) \leq \frac{1}{1-\delta(\mu_{s'})^{-1}}$. 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':

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

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

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 $\mu_s(\omega_1) < \mu_0(\omega)$. 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 $\mu_s(\omega_1) \geq \mu_0(\omega)$. 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

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

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 $\mu_{0,\varepsilon} = (\mu_{0,\varepsilon}(\omega_1), \mu_{0,\varepsilon}(\omega_2)) = \left(\frac{1}{2}-\varepsilon, \frac{1}{2}+\varepsilon\right)$ and $\mathcal{F}_\varepsilon$ 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.

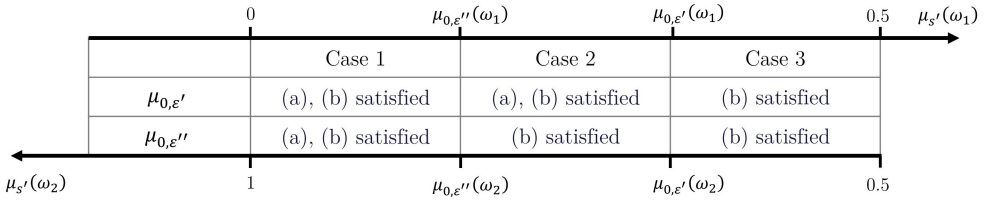


Figure 10

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, $\bar{R}_{\varepsilon'';s}(\omega_1) \leq \bar{R}_{\varepsilon'}(\omega_1; s)$. 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 $\mu_s \in \mathcal{F}_{\varepsilon'}$. □

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 $\boldsymbol{\mu} \in [\Delta(\Omega)]^S$ can be induced.

Take an arbitrary vector of posterior beliefs $\boldsymbol{\mu}$. To induce $\boldsymbol{\mu}$, 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:

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

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 models 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 model 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 $\boldsymbol{\mu}$. Inducing $\boldsymbol{\mu}$ 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) $\Pr^{m_k}(s_k) \geq \Pr^{m}(s_k)$ for each $m \in \{m_1, \ldots, m_K\} \cup \{d\}$.

Assume $\boldsymbol{\mu} \in \mathcal{F}^d$. I show that there exists a set of models inducing $\boldsymbol{\mu}$. Following the proof of Theorem 1, construct the vector of posteriors $\boldsymbol{\mu}^{m_k}$ and the fit levels $\left(\Pr^{m_k}(s)\right)_{s\in S}$ induced by each m_k. 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 $\boldsymbol{\mu} \in \mathcal{F}^d$. 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 $\boldsymbol{\mu} \notin \mathcal{F}^d$. 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 $\boldsymbol{\mu}$, each tailored model m_k inducing the posterior μ_{s_k} has to be adopted conditional on s_k. Thus, it must hold that $\Pr^{m_k}(s_k) \geq \Pr^{m}(s_k)$ for each $m \in \{m_1, \ldots, m_K\} \cup \{d\}$.

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 $\boldsymbol{\mu} \notin \mathcal{F}^d$. □

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

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

I show separately the two inclusions.

(i) Take $\boldsymbol{\mu} \in \mathcal{F}$. It is to be shown that for each $\boldsymbol{\mu} \in \mathcal{F}$ there exists a distribution over signals $p \in \Delta(S)$ such that $\bar{\delta}(\mu_s)^{-1} \geq p_s$ 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

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

as needed.

(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 $\sum_s \bar{\delta}(\mu_s) \geq 1$. Hence, $\boldsymbol{\mu} \in \mathcal{F}$

□

B Appendix: Additional Figures

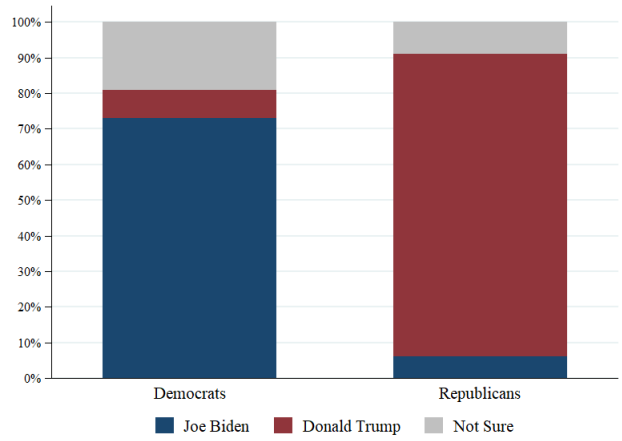


Figure 11: Priors on election winner, by party

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.

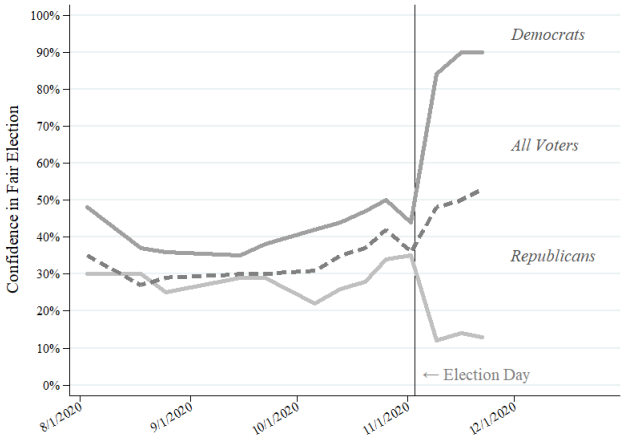


Figure 12: Confidence in fair election (Persily and Stewart, 2021)

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, 2020.