Tailored Stories

Chiara Aina*

September 24, 2025

Abstract

To what extent is it possible to manipulate beliefs by providing interpretations of unknown events? I characterize the limit of belief manipulability across signal realizations when the agent is exposed to a set of models to interpret observable signals and adopts the model that best fits what is observed. Because each signal could trigger the adoption of a different model, posteriors across signal realizations might not average to the prior. The scope of persuasion is large even for a persuader who does not control or know the signal the agent observes. I apply this framework to political polarization, finance, lobbying, and self-persuasion.

Keywords: Persuasion, Belief Manipulation, Polarization.

JEL classification: D82, D83, D9.

*Universitat Pompeu Fabra and Barcelona School of Economics. E-mail: chiara.aina@upf.edu. I am grateful to Nick Netzer, Joshua Schwartzstein, and Jakub Steiner for their guidance and support. For very helpful discussion and suggestions, I also thank Larbi Alaoui, Sandro Ambühl, Ian Ball, Kai Barron, Pierpaolo Battigalli, Roberto Corrao, Tristan Gagnon-Bartsch, George Loewenstein, Raquel Lorenzo, Fabio Maccheroni, Delong Meng, Marta Mojoli, Matthew Rabin, Andrei Shleifer, Tomasz Strzalecki, Adi Sunderam, Omer Tamuz, Heidi Thysen, Roberto Weber, Jeffrey Yang, and many others, including numerous seminar and conference participants at BU, CMU, Harvard, MIT, Northwestern, NYU, Princeton, Sciences Po, UCL, UPF, Warwick, Wharton, ECBE, NBER Political Economy, and SITE. I gratefully acknowledge the financial support from SNSF.

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 or initial beliefs. One potential explanation for reaching divergent conclusions in such cases is that we interpret the same data through different lenses. These interpretations shape how we connect what we observe to what we want to understand—guiding our reasoning like "stories" that help us make sense of the world. Often, when making sense of the observed data, individuals rely on interpretations provided by more knowledgeable sources, such as political figures, financial advisors, or experts considered trustworthy. Influencing these interpretations, therefore, becomes a powerful tool of persuasion, as it allows for manipulating their beliefs without controlling or even knowing what they observe.

I study the problem of persuading a boundedly rational agent by 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 realized signal or the process generating it. Still, he can provide the receiver with one or multiple ways of interpreting the possible signals, called *models*. Following Schwartzstein and Sunderam (2021), a model provides likelihood functions that assign a distribution of signals conditional on each state. 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 signals. The main result of this paper pins down the extent to which beliefs can be manipulated across signal realizations, thus providing clear bounds to what the sender can achieve using models to persuade.

Suppose a politician wishes to persuade a voter that he is the legitimate president regardless of the reported election outcome. The voter recognizes the politician as president only if she strongly believes him to be the legitimate winner after the election outcome is released.

Before the election, the politician communicates to the voter models about the election system. If he promotes only the model according to which the voting system is fair, the voter adopts it, recognizing the politician as president only if he is the reported winner. How can the politician ensure he is always recognized as the legitimate president? He cannot manipulate the reported election outcome or the voting system. However, before the vote, the politician or other members of his party could also spread a conspiracy theory according to which elections are rigged. 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 determine which model best fits the reported outcome. Consider a voter that expects the politician to win the election fairly. If the politician is the reported winner, the most plausible model is the one backing the fair voting system, but otherwise, the conspiracy theory resonates best with the voter. This is equivalent to an incoherent interpretation of the election outcomes: "if this politician is reported as the winner, the election system is fair; otherwise, elections are rigged." As a result, the voter becomes more confident that the politician is the legitimate winner, recognizing him as president, regardless of the election outcome.

To what extent can the sender manipulate the receiver's beliefs using models? Answering this question requires keeping track of the beliefs the receiver holds conditional on every signal realization. The main object of the analysis is a *vector of posterior beliefs*, representing the receiver's posteriors for every signal. 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 posteriors when the receiver lacks a default way of interpreting incoming data, that is, when she is most vulnerable to persuasion. This analysis is crucial for establishing the upper bound of belief manipulability and the limits of persuasion using models.[1] This characterization theorem conveys two main insights.

First, the sender can always bias the receiver's beliefs in any given direction. When many models are provided, the receiver selects the models that maximizes the likelihood of the

[1]This result also establishes an upper bound on belief manipulability under alternative assumptions on how the receiver updates her beliefs. Appendix A shows that the resulting vectors of posteriors for various information-based belief updating rules fall within the feasibility set. This includes cases where the receiver (i) updates beliefs using a model constructed as a convex combination of the models she was exposed to, (ii) is Bayesian with a prior over models, (iii) has a prior over models but biases her beliefs towards the best-fitting model, and (iv) updates her beliefs using the best-fitting model but underinfers.

observed realization such that each signal might lead the receiver to adopt a different model. As a result, even though the receiver updates beliefs using Bayes' rule within each model, all posteriors might be higher or lower than the prior and, hence, her beliefs may be inconsistent across realizations. This violates a core property of Bayesian models, expanding the scope of persuasion beyond what these models allow. 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 the beliefs the sender is able to induce, as generally not all vectors of posteriors are feasible. To induce a vector of posteriors, the sender should design 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 given 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 far it can move posteriors 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 maximal belief manipulability is generated by *maximal overfitting*. To see this, first, note that multiple models can induce the same posterior belief with different fit levels; then, when targeting a desired posterior given a signal, the tailored model should induce this belief with the highest possible fit given this target signal. This is because the more a model fits a signal, the more freedom to move posteriors away from the prior conditional on the other signals with other models.

The main theorem characterizes the vectors of posteriors that are feasible when the receiver is exposed to multiple models. This characterization provides a fundamental building block for analyzing belief manipulability across different contexts and strategic settings. The most natural application, which I focus on, is a sender who supplies a set of models to strategically persuade a receiver, but the same feasibility constraint applies when multiple senders compete to provide models or when models are not strategically supplied. Importantly, focusing on a single sender does not mean that he must personally provide every model or communicate them coherently. The sender could summarize the inconsistent models into

a single, incoherent model, as illustrated in the example,[2] or different agents could each communicate one of the models as part of a coordinated strategy dictated by the sender. In both cases, the receiver may overlook inconsistencies and perceive them as credible. For simplicity of exposition, I describe models as if supplied by a single sender.

Having explored the limits of belief manipulability, I turn to the question of what makes the receiver more vulnerable to persuasion. First, initial beliefs play a crucial role. In the binary case, the sets of feasible vectors of posteriors can be ordered: 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 and the sender can provide a set of models to make her hold any beliefs regardless of what she observes. More generally, I provide necessary and sufficient conditions for full manipulability. Second, I highlight the importance of a fixed signal space as the sender's communication is restricted only to interpretations of meaningful events. Third, I extend the main analysis by characterizing the set of feasible vectors of posteriors when the receiver holds a model by default and I show how adding one or more default models constraints belief manipulability.

I present several stylized applications to highlight how this framework can both deliver novel implications and replicate classic results of the literature with a different mechanism. These illustrations are also meant to emphasize the relevance of this setting in capturing persuasion in numerous real-world scenarios. Section 4.1 discusses the consequences of being exposed to conflicting models in a political setting, also known as "firehose of falsehood" (Paul and Matthews, 2016). I show that exposure to conflicting ways of interpreting new information has consequences both at the intra- and inter-personal levels: it reveals how confirmation bias can arise, not only by seeking confirming evidence, but also from selectively adopting models that align with prior beliefs, which in turns inevitably leads to polarization in a population of heterogeneous agents. Section 4.2 studies the misalignment of incentives between a financial advisor and investors with private information. I use this case to illustrate the optimal communication for the advisor and also how this framework is suitable to study a private-information setting. Communicating different models that could be picked up de-

[2]Other examples of inconsistent interpretations across contingencies include the following: While interpreting a grade, a student might find credible that "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 a new vaccine, somebody skeptical about vaccines might rely on the story "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." These interpretations cannot be represented by any well-defined model but rather an incoherent one reflecting the selection of different models across contingencies. See Ispano (2024) for a discussion on coherence in models.

pending on the private information of the investors, like past financial experience, allows the advisor to always move beliefs in the advantageous direction. Section 4.3 explores a multiple-selves setting in which an agent can distort her own beliefs by manipulating the perceived informativeness of observable signals or by leaving data open to interpretation. This mechanism can deliver the classic implications of the literature on motivated beliefs but also sets a bound on belief distortion. Section 5.2 shows how a strategic persuader could challenge a shared model to insinuate doubt and deepen polarization. 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" provided alternative ways of interpreting scientific evidence (e.g., Michaels, 2008; Oreskes and Conway, 2011). Holding a shared initial model does not deter polarization in a population of heterogeneous receivers.

This paper primarily contributes to the literature on persuasion, with a more detailed discussion deferred to Section 6. Persuasion is typically studied in settings where the sender controls either the signal observed by the receiver or the processes generating it. Instead, here the sender cannot alter the realized signal, unlike the cheap talk literature (e.g., Milgrom, 1981; Crawford and Sobel, 1982). Also, the signal generating process is exogenous and cannot be manipulated, in contrast to Bayesian persuasion (Kamenica and Gentzkow, 2011), where the sender endogenously designs and commits to a signal generating process. In the political example, this would translate into the politician manipulating the voting system and its accuracy. Committing to a known signal generating process requires the induced distribution of posteriors to satisfy Bayes plausibility—ensuring the expected posterior matches the prior. This paper relaxes Bayes plausibility in a disciplined manner: the sender can induce posteriors across signal realizations that are unattainable with Bayesian persuasion, though this communication strategy generally imposes restrictions on achievable outcomes. I discuss this comparison further in Section 3.3.

This paper builds on Schwartzstein and Sunderam (2021), which analyze persuasion when the sender supplies models and the receiver adopts the model that best fits the observed signal. While they focus on the problem of manipulating a receiver endowed with a default model after observing a public signal (*ex-post*), I analyze a setting where the sender commits to his communication strategy without knowing the signal realization (*ex-ante*). Shifting communication ex-ante is a sensible assumption, motivated by scenarios where ex-post communication is either infeasible or lacks credibility. For instance, the sender may need to

provide interpretations ahead of information release, such as in the political example, where voters might distrust a politician who claims election fraud only after losing; or the sender might lack access to the private information available to the receiver, as discussed in the financial application. Ex-ante commitment not only imposes a constraint on the sender—as formalized in the main result of this paper—but also illustrates the broad scope of persuasion across contingencies. Section 5 offers a direct comparison with Schwartzstein and Sunderam (2021), showing that ex-ante and ex-post communication lead to the same outcome in the presence of a default model. However, the set of models optimal ex-post might not be so ex-ante, highlighting the strategic implications of ex-ante commitment.

The paper proceeds as follows: Section 2 sets up the framework. Section 3 addresses the question of what the receiver can be persuaded of. Section 4 illustrates applications. Section 5 extends the results to the case in which the receiver has a default model. Section 6 discusses the related literature. Section 7 concludes. All proofs can be found in the appendix.

2 Set-up

Two agents, a sender and a 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 $\mu_0 \in int(\Delta(\Omega))$.[3] The receiver observes a signal $s \in S$. The state and signal spaces are finite and fixed. A *model* m is a map that assigns 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 $\sum_{s \in S} \pi^m(s|\omega) = 1$ for each $\omega \in \Omega$. Additionally, each model has to be such that there does not exist a signal $s \in S$ such that for each ω $\pi^m(s|\omega) = 0$.[4]

Let $\mathcal{M}$ be the set of all such models. Conditional on signal s, a model m induces posterior belief μ_s^m via Bayes rule. I refer to the likelihood $\Pr^m(s) = \sum_{\omega \in \Omega} \mu_0(\omega)\ \pi^m(s|\omega)$ as the *fit* of the model m given 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

[3] This assumption is made for simplicity. See Section 4.2 for an example with heterogeneous priors.

[4] This excludes cases where the agent has to update beliefs conditional on a zero-probability event, which only occurs if given signal s the agent adopts model m with $\Pr^m(s) = 0$. I opted for restricting these models rather than making an additional, and possibly more arbitrary, assumption on how agents update beliefs in such cases. This restriction does not affect the main result, but only the preliminary results of Section 3.1.

act as follows. First, she adopts the model with the highest fit conditional on the observed signal s among the set of models $M \subseteq \mathcal{M}$ 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 that maximizes the sender's expected utility.

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. 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 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^* with the purpose of influencing the receiver's action to maximize his value at the resulting vector of posteriors:

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

The key challenge in solving this problem lies in determining which vectors of posteriors the receiver could hold when exposed to multiple models. This contrasts with the Bayesian persuasion literature, where it is pivotal to characterize not only the posteriors the receiver might attain but also with which probabilities these posteriors can be induced, that is, the distribution of the receiver's posteriors. Instead, here, the probability of the signal realizations are exogenously determined by the true model. As a result, it is enough to identify the feasible vectors of posteriors, over which the sender optimizes his value.

Discussion of Assumptions. Before proceeding, I discuss some of the assumptions behind this setting, starting with the receiver. First, I relax 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 receiver adopts the model that maximizes the likelihood of observed data.[5] Appendix A shows that other information-based belief updating rules induce less biased beliefs compared to this. Importantly, the receiver does not come up with every model she is willing to entertain, but she compares only the models she was exposed to, and only one model is used to update beliefs. This is in line with *Inference to the Best Explanation* (Harman, 1965; Lipton, 2003): only the best hypothesis is used to make an inference. However, this theory is agnostic on what "the best" means. Here, I consider as a measure the goodness of fit.[6] This assumption is supported by empirical evidence. Aina and Schneider (2024) study how individuals update beliefs in the presence of multiple models and find that the most frequent and consistent updating rule is to select the best-fitting model. Also, 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 with multiple hypotheses.[7]

Second, I assume the receiver to be naïve. Because the receiver does not know the true model and does not form beliefs about the possible models, she cannot anticipate the sender's value even if she knows or can learn about the sender's preferences. This prevents the receiver from being strategic about the communicated models. If she could form beliefs about the true model, she would learn about the state directly, ignoring the sender's proposed models. Bauch and Foerster (2024) develop a cheap-talk game where the receiver takes into consideration the strategic incentives of the sender, as they study the communication of models ex-post where the receiver no longer needs to infer the true model.

[5]Similarly, in the literature on belief updating under ambiguity, when agents consider multiple priors over the state, they only update the subset of priors that maximizes the probability of the realized signal (e.g., Dempster, 1967; Shafer, 1976; Gilboa and Schmeidler, 1993).

[6]I abstract from the reasons 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 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).

[7]Model selection via maximum likelihood is equivalent to selecting the most likely model starting from a flat prior over the proposed models. There is some evidence that people choose the most probable hypothesis. Simpler 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).

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 Schwartzstein and Sunderam (2021) do not consider multiple models provided ex-post—unless the sender wishes to persuade multiple receivers.

3 Ex-Ante Model Persuasion

3.1 Preliminaries

As a first step, I show an equivalent representation between models and vectors of posteriors under a condition comparable, but weaker than Bayes-plausibility. A vector of posterior beliefs $\boldsymbol{\mu}$ is *Bayes-consistent* if the prior is a strict convex combination of the posteriors across signals: there exists $\varphi \in int(\Delta(S))$ such that $\mu_0 = \sum_{s \in S} \varphi_s \mu_s$. Let $\mathcal{B} \subset [\Delta(\Omega)]^S$ be the set of all Bayes-consistent vectors of posteriors. Let $\boldsymbol{\mu}^m$ be the vector of posteriors such that each posterior is induced by model m. The following result is a special case of Lemma 1 of Shmaya and Yariv (2016) and illustrates that Bayes-consistency is the only restriction that Bayesian updating imposes on vectors of posteriors.[8]

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

Next, I focus on the trade-off between how well a model can fit data and how much a model can move beliefs. Define the *movement* for μ_s in state ω as $\delta(\mu_s(\omega)) = \mu_s(\omega)/\mu_0(\omega)$ 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 a target posterior.

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*

[8]The appendix reports the proof for convenience. An analogous result for a more general signal space can be found in Bohren and Hauser (2024).

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 movement. In other words, models inducing larger belief shifts in response to new information exhibit a lower fit with the data. This result extends this insight by quantifying the constraints on fit when targeting a particular posterior, a critical step for understanding the trade-offs in designing a set of models.

3.2 Feasible Vectors of Posterior Beliefs

This section characterizes the set of feasible vectors of posteriors that the receiver could hold. The first result immediately follows from Lemma 1 and shows that only Bayes-consistent vectors of posterior beliefs are feasible when a single model is proposed.

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

Next, I consider the case in which the receiver is exposed to many models. The following theorem shows how a simple condition characterizes the set of feasible vectors of posteriors: the harmonic mean of the maximal movement across signals is not higher than the number of signal realizations. Let $x = (x_1, \ldots, x_N)$, then the harmonic mean is $H(x) = \left(\sum_{i=1}^{N} x_i^{-1}/N\right)^{-1}$.

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

$$\mathcal{F} = \left\{ \boldsymbol{\mu} \in [\Delta(\Omega)]^S : \; H\left(\bar{\delta}(\mu_s)\right) \leq |S| \right\}.$$

Allowing for multiple models expands the feasibility set. However, it is not the case that all vectors of posteriors are feasible because there is 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 and only if there exists a set of "tailored models."[9] A model is "tailored" to a specific signal realization if (i) it induces the target posterior conditional on that signal, and (ii) it is adopted conditional on that signal. The latter con-

[9]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.

dition 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 of posteriors satisfies the condition of Theorem 1 then a set of models satisfying these conditions exists, otherwise not. To see why, consider that models compete with each other across signal realizations. Then, a higher fit of the model tailored to induce a posterior given a signal allows greater flexibility for other models to induce posteriors conditional on the other realizations. 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. The frontier of the feasibility set—the furthest vectors of posteriors from the prior that are still feasible—is thus generated by *maximal overfitting*: each tailored model induces the desired posterior with maximal fit conditional on the target signal. Lemma 2 translates this intuition about competing fit levels into a condition on posterior movement, allowing Theorem 1 to be interpreted as a budget constraint on the total belief manipulation achievable across signal realizations.

3.2.1 Graphical Intuition

To provide intuitions for these results, I introduce a graphical approach for the binary case.

In the following graphs, the two axes represent the posteriors attached to state ω_1 for all two signal realizations; thus, each point represents a vector of posterior beliefs. I represent the prior $\mu_0(\omega_1) = 0.3$ as the vector for which all posteriors equal to the prior (orange point). The purple area in Figure 1a depicts all Bayes-consistent vectors of posteriors: either $\mu_{s_1}(\omega_1) > \mu_0(\omega_1) > \mu_{s_2}(\omega_1)$, or $\mu_{s_1}(\omega_1) < \mu_0(\omega_1) < \mu_{s_2}(\omega_1)$. Intuitively, updating beliefs always in the same direction is impossible. By Lemma 1, every point in the purple area corresponds to a model.[10]

Figure 1b focuses on a single model. The purple line passing through the induced vector of posteriors (purple point) and the prior is the *isofit* line associated with this model: all the

[10]With a binary signal, there is a one-to-one map between Bayes-consistent vectors of posteriors and models (Corollary 3, Appendix 7). The only exception is the vector for which all posteriors equal the prior for which there are infinitely many *uninformative* models inducing it.

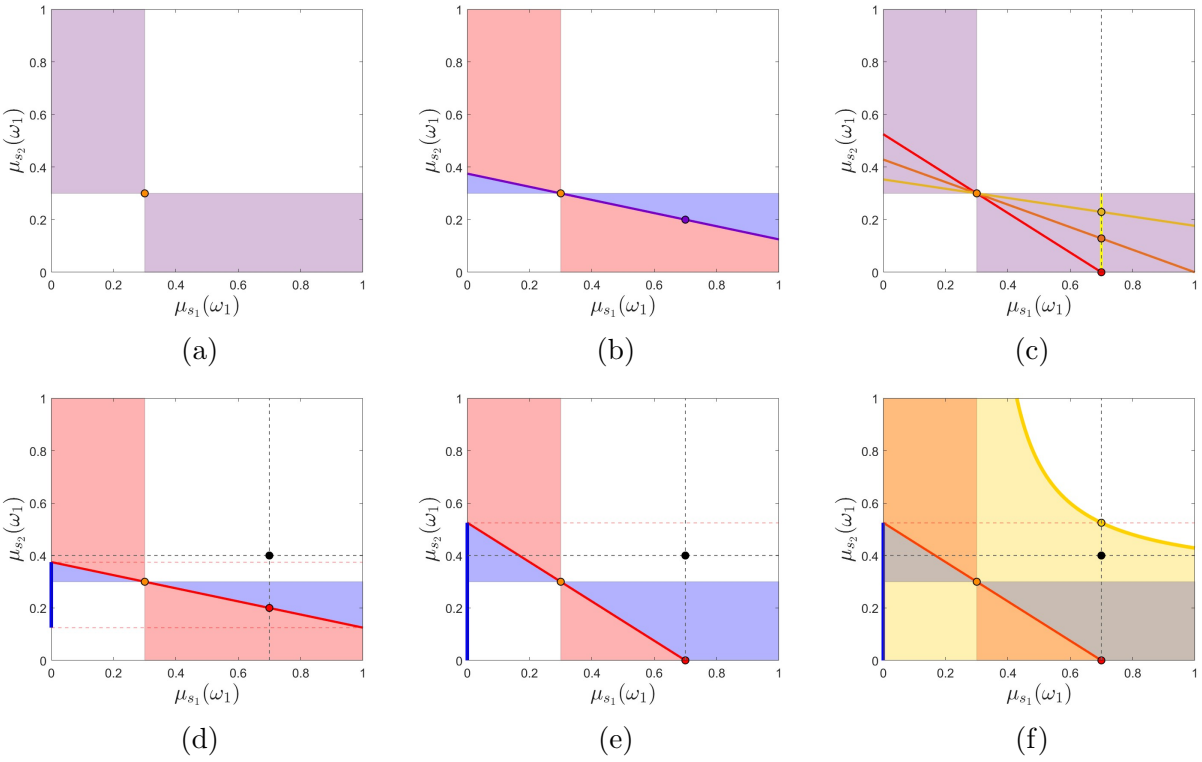


Figure 1: Graphical intuition for the binary case

points on the isofit line correspond to models with the same fit given each signal.[11] The slope of the isofit line can be interpreted as follows: the steeper (flatter) the line, the higher the fit given s_1 (s_2). For each fit level, 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 given s_1 (red area), and vectors induced by models with higher fit given s_2 (blue area).

Consider the target posterior $\mu_{s_1}(\omega_1) = 0.7$ (dotted line) in Figure 1c. There is a multiplicity of models (yellow line) that induce the same posterior conditional on the signal, but with different levels of fit. By Lemma 2, the maximal fit of a model inducing the target given s_1 is 43%. Such a model corresponds to the red point: a steeper line cannot induce the target. Points of lighter color represent models with lower fit levels (respectively, 30% and 15%).

Inducing a target vector of posteriors that is not Bayes-consistent (black point) requires two models: m_1 tailored to s_1 and m_2 tailored to s_2. I start by fixing model m_1 inducing

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

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

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-arranged to $\mu_{s_2}(\omega_1) = \mu_0(\omega_1)/\Pr^m(s_2) - \Pr^m(s_1)/\Pr^m(s_2)\mu_{s_1}(\omega_1)$. All models with the same fit $(\Pr^m(s_1), \Pr^m(s_2))$ correspond to points on this line.

the target μ_{s_1} and then identify the compatible posteriors given s_2 if m_1 is adopted given s_1. Consider model m_1 (red point) in Figure 1d. Because m_1 has to be adopted given s_1, a compatible model m_2 cannot lie in the red area (higher fit given s_1). The compatible posteriors given s_2 are all the y-coordinates of points in the blue area or on the isofit line (blue line on the y-axis). Even though the target μ_{s_2} does not lie in this set, it does not imply that the target is unfeasible. By Lemma 2, there are many models with different fit levels inducing μ_{s_1}. Figure 1e shows an alternative model m_1 that induce μ_{s_1} with maximal fit given s_1. As a result, the set of compatible posteriors given s_2 expands: by increasing $\Pr^{m_1}(s_1)$, $\Pr^{m_1}(s_2)$ decreases and thus more models can be adopted given s_2. This set includes μ_{s_2} and thus there exists a model m_2 that can induce the target together with m_1.

Maximal overfitting allows for maximal belief manipulability because it generates the largest set of posteriors given s_2 compatible with μ_{s_1}. The yellow point where the maximal compatible posterior given s_2 (dotted red line) intersects μ_{s_1} exemplifies how to construct the upper frontier of the feasibility set (yellow line) in Figure 1f. All vectors below this line (yellow area) are feasible.

3.2.2 Comparative Statics

Next, I study what makes the receiver more vulnerable to persuasion. Generally, not all vectors of posteriors are feasible. Interestingly, this is not the case when the receiver's minimal prior across the states is sufficiently high relative to the reciprocal number of signal realizations. In this case, the receiver is fully persuadable.

Proposition 2. *All vectors of posteriors are feasible if and only if* $\min_\omega \mu_0(\omega) \geq 1/|S|$.

The proposition illustrates a simple test to check whether the receiver is fully persuadable. Two observations follow. First, the minimal prior across the states contains information regarding the set of feasible vectors of posteriors. To get an intuition for this, notice that the reciprocal of the minimal prior across the states is the upper bound of the maximal movement, i.e., $\bar{\delta}(\mu_s) \leq 1/\min_\omega \mu_0(\omega)$ for any μ_s, pinning down the upper bound of the harmonic mean of the maximal movement across signals. Also, the minimal prior across the states can be interpreted as a measure of the concentration of beliefs because, by increasing the minimal prior, the prior beliefs get closer to a uniform distribution. Therefore, for priors closer to uniform, there is a lower movement on average to induce posteriors further away from the prior and thus more belief manipulability. Second, the receiver is more

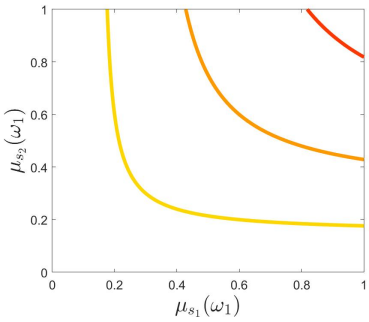


Figure 2: Frontier of the feasibility set, by prior

Notes: The lighter the color, the lower the prior: $\mu_0(\omega_1) = 15\%$ (yellow), 30% (orange), 45% (red).

manipulable in a setting with more signals to be interpreted. Tailoring models to specific signals allows more feasible vectors of posteriors but also requires models to be compatible with each other across signals: the more signal realizations, the less stringent this condition is. To exemplify this, continue the example of Figure 1e where a model m (red point) induces the target posterior given s_1 with maximal fit $\Pr^m(s_1) = 43\%$. As there are only two signals, any model must have a higher fit than $\Pr^m(s_2) = 57\%$ to be adopted given s_2. However, with more signals to be interpreted, this constraint would be less stringent because $\Pr^m(s_2) < \sum_{s \neq s_1} \Pr^m(s) = 57\%$. As a corollary of Proposition 2, we know that if the signals are at least as many as the states, a receiver with a uniform prior can be persuaded to believe anything; with fewer signals than states, the receiver is never fully persuadable.

A stronger result holds for the binary case where the feasibility sets 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 $\mathcal{F}_\varepsilon$ be the feasibility set with respect to prior $\mu_{0,\varepsilon} = (1/2 - \varepsilon, 1/2 + \varepsilon)$.

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

3.3 Sender's Problem

Given these results, I turn to the sender's problem. 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.

The sender faces three key restrictions. Unlike the literature on Bayesian persuasion in which the sender chooses the signal generating process, he cannot manipulate either the

true model or the signal space. On top of this, the sender communicates without knowing the realized signal. In what follows, I examine how these assumptions constrain persuasion.

First, the signal generating process—true model—cannot be manipulated. I show that proposing multiple models allows for a larger set of posteriors compared to what is attainable with Bayesian persuasion. Therefore, it is more beneficial to propose multiple models than to choose the true model.

Proposition 4. *Some feasible vectors of posteriors are not Bayes-consistent:* $\mathcal{B} \subset \mathcal{F}$.

If the sender could manipulate the true model on top of proposing models, he could increase his expected value by strategically manipulating the signal probabilities; nevertheless, the feasible vectors of posteriors would remain unchanged as the true model does not impact the constraint characterized in Theorem 1.

Second, the signal space is fixed. This restricts the sender's communication to interpretations of observable events in S. If this were not the case and the sender could add dummy signals—that is, realizations that can never occur, $s_0 \notin S$—, he could persuade the receiver to hold any beliefs in the original space. The intuition is that if the receiver believes that other signals proposed by the sender were also observable with $S' \supset S$, the sender could leverage those signals that cannot be realized to manipulate beliefs further. Indeed, one dummy signal is enough to guarantee full manipulability.[12]

Proposition 5. *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.*

As a third constraint, the sender provides models without knowing the signal realization. How does this impact 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, any vector of posteriors is feasible.[13] 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

[12] If a vector of posteriors is not feasible, the set of tailored models to induce the target have fit levels such that the compatibility constraint is not satisfied. However, by adding a dummy signal, these models can be modified by appropriately setting a positive fit given the dummy signal and decreasing the fit levels given the other signals to which the model is not tailored, so that the compatibility constraint is satisfied.

[13] This is not the case if the receiver has a default model, as discussed in Section 5.

over the feasible vectors of posteriors:

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

This measure captures the sender's willingness to pay to learn the data available to the receiver and can be used to comment on the value of micro-targeting.[14] I discuss an example of this in Section 4.2.

Often the sender's objective might be to maximize or minimize the receiver's belief in a particular state across all signal realizations. In such cases, providing models ex-ante is highly effective because $\Delta = 0$ as long as the relevant state ω^* is not too unlikely according to the receiver's prior. Even when it is, the sender can achieve his goal for all but one signal—and even for that signal, beliefs shift in the desired direction. The next result identifies the best outcome the sender can achieve assuming that his goal is to maximize the posterior belief for a target state for every signal realization.

Proposition 6. *Assume the sender wants to maximize $\mu_s(\omega^*)$ for every s. If $\mu_0(\omega^*) \geq 1/|S|$, the sender can achieve his first-best of $\mu_s(\omega^*) = 1$ for every s; otherwise, the sender can ensure $\mu_s(\omega^*) = 1$ for every $s \neq s'$, while conditional on the remaining signal s' the posterior can be at most $\mu_{s'}(\omega^*) = \mu_0(\omega^*)/\left(1 - \mu_0(\omega^*)(|S| - 1)\right)$.*

This result also provides valuable insights for situations where the sender's objective is to always induce a specific action from the receiver within a finite set of available ones. In the judge-prosecutor example from Kamenica and Gentzkow (2011), the prosecutor wishes the judge to convict the defendant, regardless the investigation's outcome. To trigger this action, the judge must hold beliefs that the defendant is guilty with a probability higher than 50% for every signal. According to Proposition 6, the prosecutor could successfully achieve this by providing models ex-ante to the judge whenever the judge's initial belief about the defendant being guilty exceeds 1/3.

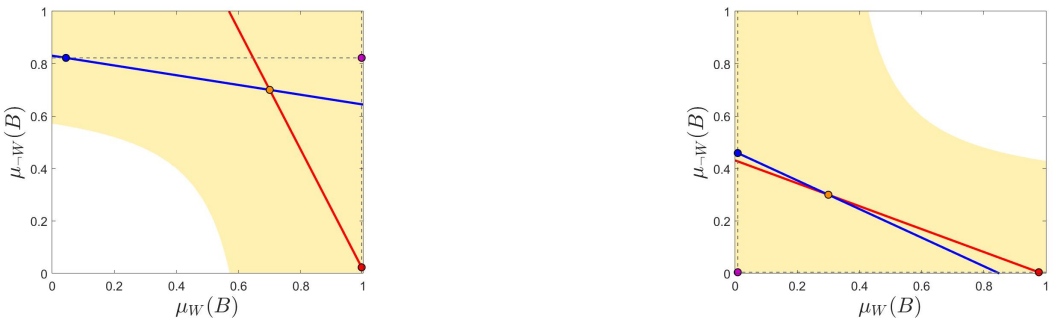
[14] This is a well-established practice in marketing: analyzing online information on potential customers to create and convey the most effective message. As a result, different ads are shown to different groups of consumers. For an example, see `https://themarkup.org/news/2021/04/13/how-facebooks-ad-system-lets-companies-talk-out-of-both-sides-of-their-mouths`.

4 Applications

This section discusses several applications. The first formalizes the political example outlined in the introduction, focusing 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. The third application discusses self-persuasion.

4.1 Polarization

Firehose of falsehood is a propaganda technique based on a large number of possibly contradictory and mutually inconsistent messages, defined by Paul and Matthews (2016) to describe Russian propaganda. Spreading conflicting models can be an effective strategy for a persuader in manipulating a target audience with destabilizing consequences for society.



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

Figure 3: Firehose of falsehood, by voter's prior

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

Politician Bob is running for president and the election outcome is soon to be revealed. Let the state space be $\{B, \neg B\}$, where B is the event in which Bob is the legitimate winner of the election, and the signal space be $\{W, \neg W\}$, where W is the event in which Bob is reported as the winner. For simplicity, assume that voters expect Bob to be fairly elected with either high probability $\mu_0(B) = 70\%$ or low probability $\mu_0(B) = 30\%$. Each voter recognizes Bob as president if, once observed the election outcome, she believes that Bob is the legitimate winner with a probability higher than 50%. Before the election outcome is released, voters are exposed to two different models about the election system. According to the official narrative, mistakes in vote counting are very rare: $\pi^f(W|B) = 99\%$ and $\pi^f(W|\neg B) = 1\%$. Meanwhile, Bob's party spreads a conspiracy theory according to which elections will be

rigged against him: if Bob were to win, votes would not be truthfully reported $\pi^c(W|B) = 1\%$; otherwise, the votes would be counted randomly $\pi^c(W|\neg B) = 50\%$.

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 with high prior (Figure 3a). The red point lies on the steeper isofit line and the blue point lies on the flatter one; therefore, the voter would adopt f conditional on W and c conditional on $\neg W$ (purple point). This voter always recognizes Bob as the legitimate president regardless of the election outcome. In contrast, voters with low prior never recognize him as president (Figure 3b) because they adopt c given W and f given $\neg W$. This example illustrates how, with heterogeneous receivers, the exposure to the same pair of conflicting models not only induces opposite actions given each signal, but also lead to inevitable polarization. It is possible to generalize this insight for the binary setting.

Two models m, m' are *conflicting* if $\pi^m(s_1|\omega_1) > \pi^m(s_1|\omega_2)$ and $\pi^{m'}(s_1|\omega_2) > \pi^{m'}(s_1|\omega_1)$. In other words, to be conflicting, each model must point to a different state given each signal. The following result highlights the consequences of being exposed to conflicting models.

Proposition 7 (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 extremely high (low) posteriors regardless of the signal realization.[15] In the previous example, voters with $\mu_0(B) > 33\%$ are persuaded to support Bob regardless of the election outcome, holding a strong belief in his legitimacy; the same models lead voters with $\mu_0(B) < 33\%$ to never recognize Bob as president, always believing him to be an illegitimate president.

[15]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.

This result highlights how the exposure to conflicting models generates two important phenomena. First, it leads to confirmation bias through the selective adoption of the model confirming the prior. Second, 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, leading to inevitable polarization. This contrasts with Bayesian models, where heterogeneous priors do not generate opposite shifts in beliefs (Baliga et al., 2013).

Several mechanisms have been proposed to understand the determinants of polarization. Often, polarization is associated with confirmation bias, first formalized by Rabin and Schrag (1999), which assumes agents misinterpret new information as supportive of early evidence with an exogenous probability. Fryer et al. (2019) builds on this, assuming a similar distortion for signals open to interpretation in the direction of the prior, and provides evidence for their predictions. In contrast, in this paper, confirmation bias is not assumed but arises endogenously from adopting the best-fitting model in the presence of conflicting models. Other explanations include ambiguity aversion (Baliga et al., 2013), private signals about the interpretation of evidence (Benoît and Dubra, 2019), and a high dimensionality of the signal space compared to the state space (Andreoni and Mylovanov, 2012). Recent papers also highlight how mistakes in source credibility can amplify disagreement (Cheng and Hsiaw, 2022; Gentzkow et al., 2024).

Unlike these approaches, I show that polarization can emerge from exposure to conflicting models. This generates a distinct channel of polarization: agents differ in how they interpret data. This form of model polarization occurs in two forms. First, agents with sufficiently different priors adopt different models to explain the same observed outcome, consistent with classic experiments (Lord et al., 1979; Darley and Gross, 1983; Plous, 1991; Russo et al., 1998). Second, when agents with similar priors observe different outcomes, they also adopt different models to make sense of the different information. Appendix B provides suggestive evidence of this mechanism using the case of the 2020 US presidential election.

4.2 Financial Advice

Next, I illustrate the optimization problem of 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 investors' information about past financial performance influences their beliefs about future investments.

However, the advisor does not have access to this piece of private information. Nonetheless, he has the incentive to persuade the investor to invest as much as possible.

To manipulate the investor, the advisor can propose different ways to predict future returns based on past returns.[16] In finance, two important alternatives are that returns can exhibit predictable mean reversion or predictable continuation (Barberis et al., 1998). 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, but inexperienced investors might interpret past performances through the strategy that resonates best with their initial beliefs and advisors could use simplified versions of these theories to their advantage. As shown in the following example, an investor with favorable expectations toward the advisor-preferred asset would always fully invest in that asset because any past data trigger the adoption of the most optimistic model in terms of future performance. Instead, communicating these models to a pessimistic investor can be counterproductive, and the advisor needs to adjust his strategy.

Formally, each investor has to allocate one unit of endowment over two possible outcomes: the stock market going up (U) or going down (D). This results in the choice of a hedging strategy $\alpha = (\alpha_U, \alpha_D)$ with $\alpha_U + \alpha_D = 1$. All investors have the same initial beliefs and I consider two cases: *optimistic* investors expecting the market going up with a probability of 70%, and *pessimistic* investors expecting the same outcome with a probability of 30%. Each investor holds selective information regarding the past dynamic of the stock market, indicating either a positive (G) or a negative past performance (B), and tries to understand how this can predict the future one. For example, she only relies on her previous experience on the stock market, only samples information for a restricted period, or reads some newspapers reporting limited information.[17] Assuming a logarithmic utility over the outcomes, the

[16] According to Reich and Tormala (2013), contradicting oneself—initially supporting something and then later switching to something else—might offer a persuasive advantage over both one-time and repeated opinions. This effect, moderated by trust, disappears if the conflicting opinions come from different sources.

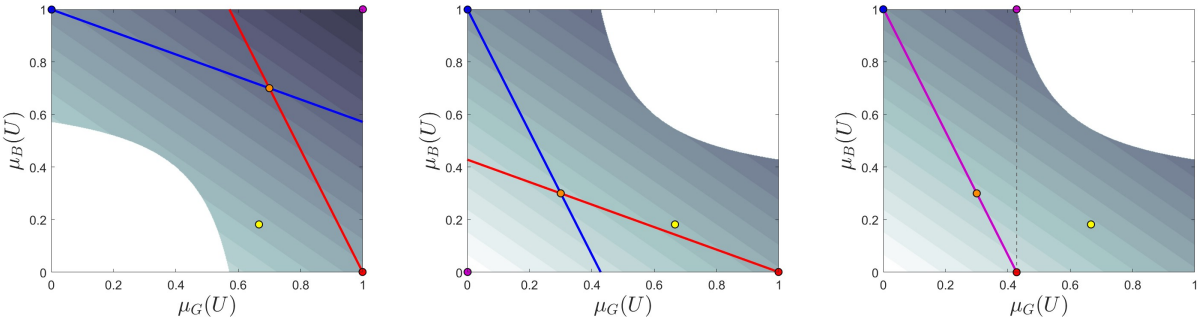
[17] Empirical evidence shows that personal experiences have a lasting impact on beliefs and behavior, such as how having lived through a depression affects stock market participation (Malmendier and Nagel, 2011).

investor's expected utility based on her posterior μ_s is $\mathbb{E}[U^R(\alpha)] = \sum_{\omega\in\{U,D\}} \mu_s(\omega)\log(\alpha_\omega)$. The investor's optimal action is to allocate a proportion of the endowment equal to the corresponding posterior, $\alpha^* = \mu_s$.

The financial advisor receives a commission proportional to the receiver's allocation on outcome: $U^S(\alpha) = r_U\ \alpha_U + r_D\ \alpha_D$. Assuming $r_U > r_D = 0$, $V(\boldsymbol{\mu}) = \sum_{s\in\{G,B\}} \Pr^t(s)\ r_U\ \mu_s(U)$. The advisor expects the stock market going up with probability of 40% and knows the true model, where a positive past information positively (negatively) correlates with an upward (downward) trend in the stock market: $\pi^t(G|U) = \pi^t(B|D) = 75\%$. The advisor does not know exactly what type of information the investors have looked at, but he expects 45% of investors have had a positive impression in the past, while 55% have had a negative one. Figure 4 shows the financial advisor's indifference curves plotted on the feasible vectors of posteriors given the investors' prior; they are driven by the true model, his prior, and his incentives, on top of the investors' prior and incentives.

Consider the optimistic investors (Figure 4a). The financial advisor does not want to discard the investors' information as irrelevant, otherwise the investors' beliefs would remain at the prior with an investment of $\alpha_U = 70\%$. By Proposition 6, the advisor can attain full manipulability by proposing multiple models. The highest value for the advisor is achieved at the top-right corner, where an optimistic investor always expects the stock market to go up and never hedges against the opposite outcome. Intuitively, this means that the advisor can leverage any past experience of the investor and always move her beliefs 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 performance, i.e., $\pi^{m_1}(G|U) = \pi^{m_1}(B|D) = 1$ (red point), and (ii) model m_2 suggesting a perfect negative correlation between past and future performance, i.e., $\pi^{m_2}(B|U) = \pi^{m_2}(G|D) = 1$ (blue point). These can be read as simplified versions of the momentum ("early success predicts long-run success") and mean reversion ("what goes down goes up"). Because of their optimistic initial beliefs, investors adopt the first given G and the second given B and never hedge against the stock market going down.

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. Second, communicating the same pair of models tailored to the optimistic investors to the pessimistic ones is self-defeating. A pessimistic investor would always adopt



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

Figure 4: Financial advice, by investors' prior and advisor's communication

Notes: 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 the vector of posteriors induced by m_1 and m_2, respectively, while the purple point is the resulting vector of posteriors given these models; the darker the colored area, the higher the advisor's value for $r_U = 1$.

the most pessimistic model and never invest in the advisor-preferred outcome (Figure 4b).

With a pessimistic investor, $\boldsymbol{\mu}^* = ((0.43, 0.57), (0, 1))$ is the maximal value the advisor can achieve by Proposition 6 (right-top kink in Figure 4c). The optimal communication strategy is to entertain two models: (i) model m_1 such that $\pi^{m_1}(B|U) = 0$ and $\pi^{m_1}(G|D) = 0.57$ (red point), and (ii) model m_2 as defined above for the optimistic investors (blue point). According to m_1, B is a perfectly revealing signal of the stock market going down. In contrast, G is indefinite news. This model encourages only investors with positive information to have $\alpha_U > 0$, and, indeed, it is tailored to those. Again, m_2 is a version of mean reversion, which pushes investors with negative information to invest the whole endowment in the advisor-preferred outcome. Note that since the first-best outcome of convincing all pessimistic investors to fully invest in outcome U is not attainable, the advisor shifts to the second-best: convincing the largest group of investors (the ones with negative information) to set $\alpha_U = 100\%$ while increasing α_U for the other group (the ones with positive information) 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, inducing $\bar{\boldsymbol{\mu}} = ((1, 0), (1, 0))$. With pessimistic investors, $\bar{\boldsymbol{\mu}}$ is unfeasible, thus 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.[18] 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.

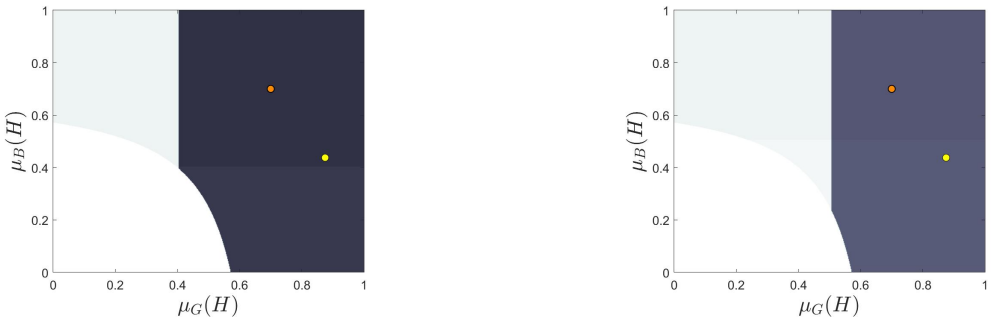
Confirmation bias can emerge because the agent keeps signals open to different interpretations. This could be the case of a student who thinks, and subconsciously likes, to be intelligent and thus leaves the informativeness of grades open to two interpretations: grades are a good measure of own ability, or grades are based on luck. She always keeps high confidence in her abilities because she believes grades to be informative after a good grade but not to convey much information after a bad grade. In such a manner, inconsistent updating across signals could result from selectively adopting models, which could be an explanation for some of the evidence on asymmetric updating (e.g., Eil and Rao, 2011; Sharot, 2011; Ertac, 2011; Coutts, 2019; Möbius et al., 2022; Drobner and Goerg, 2022).[19]

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 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 misaligned

[18] Papers on motivated beliefs conjecture different sources of motivations or 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).

[19] 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) finds no 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 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. 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).

incentives. After the signal, it is optimal to take the costly action if the beliefs given the signal are higher than $c/(\beta\delta v)$, 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 belief is higher than $c/(\delta v)$, which is lower than the relevant threshold at $t = 1$ if $\beta < 1$. 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 5: Motivated beliefs, by present bias

Notes: The orange point is the prior and the yellow point is the vector of posteriors induced by t; the darker the colored area, the higher the sender's value for $c = 4$, $v = 10$, $\delta = 0.99$, and $\beta = 0.8$.

Consider an agent that initially believes to have high abilities with a probability of 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. Updating her prior using the true model maximizes her ex-ante expected payoff (Figure 5a). 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 5b). 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, the receiver only considers models proposed by the sender. In this section, I allow the receiver to initially hold a model, hereafter called the *default model*, that she considers on top of the models she is exposed to, and I show how this restricts belief manipulability.

5.1 Feasible Vectors of Posterior Beliefs with a Default Model

Assume the receiver to be endowed with a default model. The set-up is otherwise the same as in Section 2. The receiver adopts the model with the highest fit given the observed signal s among the set of models M and her default model d: $m_s^* \in \arg\max_{m \in M \cup \{d\}} \Pr^m(s)$. The following theorem characterizes the feasible vectors of posteriors in this setting.

Theorem 2 (Default, Many Models)**.** *The set of feasible vectors of posteriors given d is*

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

Moreover, $\mathcal{F}^d \subseteq \mathcal{F}$.

The default model restricts belief manipulability. The proposed models compete not only with each other but also with the default model in each signal realization: the higher the fit of the default model given a signal, the less the sender can move beliefs conditional on that signal with additional models.[20]

Interestingly, the presence of a default model eliminates the cost of ex-ante commitment. While, in the absence of a default model, every posterior is feasible when the sender communicates 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 posteriors in this setting. Given a default model d, the sender's cost of ex-ante commitment is the gap between the maximal sender's value over the ex-post feasible vectors

[20] A receiver endowed with a default model might be more skeptical to adopt other models. It is straightforward to extend this model adoption rule, along with the associated result, to the case where the receiver only switches to one of the proposed models if its fit exceeds that of the default model by a specified threshold, in the spirit of Ortoleva (2012). This would restrict the feasibility further around the prior depending on the threshold, but the intuition would be qualitatively unchanged.

of posteriors and the maximal sender's value over the 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 post-$\mathcal{F}^d = \{\boldsymbol{\mu} \in [\Delta(\Omega)]^S : \ \forall s \in S, \ \omega \in \Omega, \ \ \mu_0(\omega)/\mu_s(\omega) \leq \Pr^d(s)\}$, by following Schwartzstein and Sunderam (2021). Because post-$\mathcal{F}^d = \mathcal{F}^d$, providing multiple models without knowing the signal realization (*ex-ante*) allows for the same posteriors achievable by communicating a model given the realized signal (*ex-post*).

Corollary 1. *With a default model, ex-ante commitment does not restrict the sender's value:* $\Delta^d = 0$.

However, this result does not imply that the sender should communicate ex-ante the set of ex-post optimal models. Doing so could be self-defeating, especially for signal spaces with more than two realizations.[21]

Theorem 2 can be extended to the case where the receiver considers a set of default models, $D \subseteq \mathcal{M}$. In this case, the scope for manipulation narrows further, as the feasible set of posteriors is the intersection of the feasibility sets for each individual model in D. While each additional default model can shrink the feasible set, this set is never empty, as all feasibility sets contain the vector for which all posteriors equal the prior.

Corollary 2. *Let* $\boldsymbol{\mu}^{\varnothing} = (\mu_0)_{s \in S}$. *The set of feasible vectors of posteriors given* D *is*

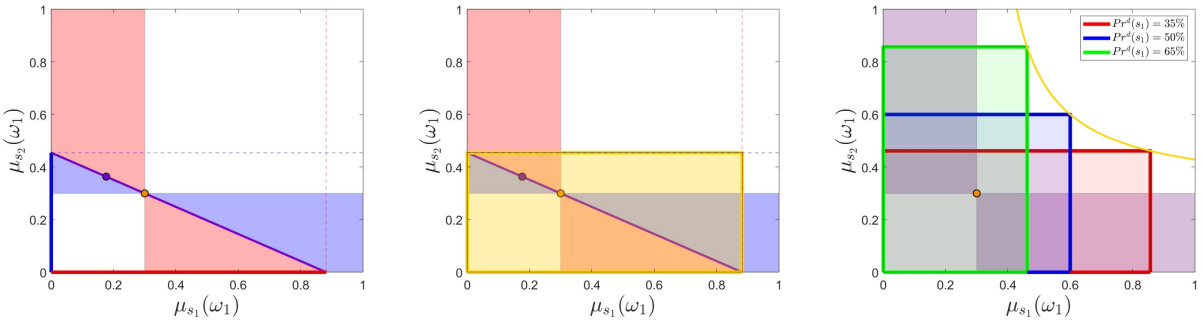
$$\mathcal{F}^D = \bigcap_{d \in D} \mathcal{F}^d \supseteq \bigcap_{d \in \mathcal{M}} \mathcal{F}^d = \boldsymbol{\mu}^{\varnothing}.$$

I conclude this section with a result that links Theorem 1 and Theorem 2. The two theorems are closely related: the set of feasible vectors of posteriors in the absence of a default model is the union of the feasibility sets with a default model for all default models.

Proposition 8.

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

[21] With a binary signal, the ex-post optimal 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 $\Pr^{m_2}(s_2) \geq \Pr^{m_1}(s_2)$. The same is not guaranteed for a larger signal space.



(a) (b) (c)

Figure 6: Graphical intuition of Theorem 2 and Proposition 8

Figure 6 provides a graphical intuition of these results for the binary case. Consider a default model (purple point) in Figure 6a. Given its isofit line (purple line), the red area corresponds to models with a higher fit given s_1, and the blue area corresponds to models with a higher fit given 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 on the axes, which together generate the feasible vectors of posteriors (yellow area) in Figure 6b. These figures also clarify why all the default models with the same fit levels—corresponding to that same isofit line—induce the same feasible vectors of posteriors. Figure 6c helps building intuition for Proposition 8. The yellow line corresponds to the upper frontier of the feasibility set without a default model, while the colored areas correspond to the feasibility sets in the presence of 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."

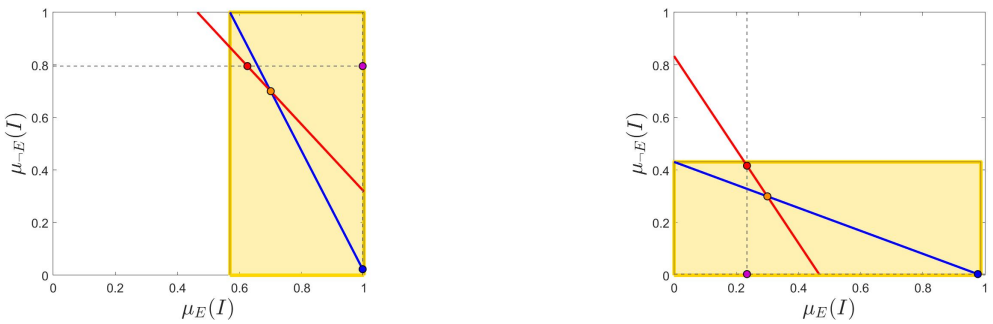
— Cigarette Executive (1969)

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

— Internal memo by The American Petroleum Institute (1998)

Even when agents share a common initial model, strategic communication of an alternative model can be dilute consensus. Tobacco and oil companies used this strategy to challenge scientific consensus and promote competing interpretations of evidence on the health effects of smoking and climate change—the so-called "merchants of doubt" (e.g., Michaels, 2008;

Oreskes and Conway, 2011). The example below illustrates that a shared initial model is insufficient to prevent polarization in a heterogeneous population.



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

Figure 7: Merchants of doubt, by agent's prior

Notes: The orange point is the prior; the red and the blue points are the vector of posteriors induced by m and d, respectively, while the purple point is the resulting vector of posteriors given these models; the yellow area represents the set of feasible vectors of posteriors given d.

Consider a binary state $\{I, \neg I\}$, 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, $\pi^d(E|I) = \pi^d(\neg E|\neg I) = 99\%$. Even if all agents start with the same model, the sets of feasible vectors differ drastically depending on their prior. Figure 7 shows these sets for supportive agents with $\mu_0(I) = 70\%$ and skeptical agents with $\mu_0(I) = 30\%$.

A lobby wants to strategically challenge this shared model in the population to induce disagreement on the issue. 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 default model induces a vector of posteriors almost identical for all agents (blue point). However, introducing an alternative model (red point) 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 7b). By contrast, agents initially expecting the issue to be real become even more confident (Figure 7a). Strategically introducing a conflicting model promotes doubt among agents with different initial beliefs and sharing the same default model does not deter polarization.

6 Relationship to the Literature

The impact of persuasion has long been studied in economics (see Little, 2023, for a comparison of approaches in a common framework). This paper contributes to this literature by exploring the consequences of providing interpretations of events that are unknown at the time of the communication. This type of persuasion differs in two key ways from the existing 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. Close to this paper, Eliaz et al. (2021a) build 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 assumed this paper, and as a result, full persuasion can sometimes be attained. Second, the persuader cannot influence the signal generating process, which 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, 2022) are about persuasion by generating information which is then interpreted by Bayesian receivers (for an extension to non-Bayesian receivers see de Clippel and Zhang, 2022). This restricts the sender to inducing only Bayes-plausible distributions of posteriors, unlike this paper. Moreover, a strand of previous literature studies senders who engage in ambiguous communication—by proposing several explanations or messages—to persuade receivers who are ambiguity averse. This was studied in cheap-talk games (Kellner and Le Quement, 2017, 2018) and in Bayesian persuasion (Beauchêne et al., 2019).

This paper builds on the framework of model persuasion introduced by Schwartzstein and Sunderam (2021), where agents are assumed to adopt the best-fitting model.[22] I adopt the same model selection rule but investigate the ex-ante strategic provision of models, which allows a direct comparison with the literature of Bayesian persuasion. Section 5 offers a direct comparison and a discussion of the effect this ex-ante commitment has on belief manipulability compared to the ex-post perspective adopted in Schwartzstein and

[22]Levy and Razin (2021) study the aggregation of forecasts over time. They assume the agent to look for the most likely explanation—information structures consistent with previous forecasts and the prior. Thus, the signal space could vary across explanations, but 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.

Sunderam (2021). Related contributions include Ichihashi and Meng (2021), who combine Bayesian persuasion with ex-post model persuasion in a sequential framework; Bauch and Foerster (2024), who relax the assumption of naïve receivers in ex-post persuasion; Jain (2024), who analyze optimal information design anticipating strategic misinterpretation; and Schwartzstein and Sunderam (2024), who study the exchange of models in networks. Other papers take different approaches to what makes models persuasive (e.g., Ispano, 2024; Yang, 2024; Wojtowicz, 2024).[23]

A complementary approach to model persuasion represents narratives as causal models, represented by directed acyclic graphs (Spiegler, 2016). In this framework, Eliaz and Spiegler (2020) assume 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. Spiegler et al. (2025) and Eliaz and Spiegler (2024) use this framework to study political competition and news media, respectively. Eliaz et al. (2021b) study to what extent a misspecified model can distort pairwise correlations between variables. Even if the misspecified model cannot distort the marginal distributions of individual variables, increasing the number of variables in the model can lead to an almost perfect correlation.[24] Alternative frameworks in which narratives have been formally investigated are Bénabou et al. (2018), Izzo et al. (2023),[25] Bilotta and Manferdini (2024), Szeidl and Szucs (2025), Montiel Olea and Prat (2025); see Barron and Fries (2024b) for a discussion of this concept.

A growing body of experimental work finds strong support for the idea that agents' beliefs are susceptible to the strategic provision of interpretations (Barron and Fries, 2024a; Charles and Kendall, 2024; Ambuehl and Thysen, 2023). In particular, Barron and Fries (2024a) study narrative provision and adoption in a financial advice setting, building on an example of Schwartzstein and Sunderam (2021), and overall their evidence supports this framework. They find that advisors with misaligned incentives promote biased narratives

[23] In Ispano (2024), the sender communicates a model before the signal is realized and the proposed model is adopted if it is coherent (conditional on a state, probabilities of each possible news sum to one) and compatible with her default model (the marginal distribution of news is undistorted). I assume models to be coherent by definition. However, one could argue that the receiver might hold an incoherent model ex-post. This follows from how the receiver selects models across signals when exposed to many. To compare results, coherent and compatible models can only induce vectors of posteriors on the isofit line of the true model.

[24] According to Montiel Olea et al. (2022), including irrelevant covariates in models can also improve the perceived predicted ability with large datasets. In this paper, given the fixed state and signal space, all models have the same dimension and cannot exhibit this type of misspecification.

[25] Also in Izzo et al. (2023), the model (described as linear relations between policies and their outcome) with the highest likelihood given the observed data (thus, the smallest mean squared error) is adopted.

and successfully manipulate investors' beliefs, especially when the narratives fit well with the observed data. Building on the theoretical framework of this paper, Aina and Schneider (2024) investigate how individuals update their beliefs when confronted with multiple models, distinguishing between different updating rules. Selecting the best-fitting model is the most frequently applied rule in their data, providing strong empirical support for the receiver's model selection rule assumed in this paper. Moreover, in line with this paper's insights, they document the systematic emergence of Bayes-inconsistent vectors of posterior beliefs in the presence of competing models, skewed in the direction of selecting the best-fitting model for every signal realization. This finding underscores the potential impact of persuasion using models and the possibility of relaxing the Bayes-plausibility constraint due to behavioral biases. Finally, a noteworthy empirical literature investigates the impact of narratives on beliefs and behavior (e.g., Morag and Loewenstein, 2024; Graeber et al., 2023; Bursztyn et al., 2023; Andre et al., 2024, 2022).

7 Conclusion

Persuasion has typically been studied in settings where the persuader controls the information observed before making a decision, such as by sending a persuasive message or providing new informative facts. However, sometimes this is not possible. In such cases, although the persuader cannot control or even know the information used in the decision, I demonstrate that the scope for persuasion using models is large, but generally bounded.

Bayesian models assume that beliefs should be consistent across possible realizations: the receiver cannot update her beliefs in the same direction given every signal. Exposure to multiple models can lead to the violation of this property. Because of receiver's bounded rationality, each signal might trigger the adoption of a different model. Thus, the sender can leverage multiple models to induce beliefs that cannot be attained by choosing the signal generating process as in Bayesian persuasion. Aina and Schneider (2024) provide strong empirical support for this insight, showing that when individuals are exposed to conflicting models, they often hold inconsistent beliefs across contingencies, highlighting the potential impact of persuasion using models.

Several extensions are left to be explored in future research. First, this paper focuses on a problem with only one sender and one 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. My extension to the default model is the first step towards studying 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 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 on psychological game theory (for a survey see Battigalli and Dufwenberg, 2022). In the experiment by Barron and Fries (2024a), senders communicate biased narratives to their advantage but they also display truth-telling preferences to some extent. Incorporating these motives might lead to insightful predictions. On the other hand, receivers might consider only some types of models depending on the context. Research along this line could shed light on how these restrictions impact welfare.

This paper discusses a wide range of applications, proposing a possible common mechanism encompassing 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.

Appendix

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 $\boldsymbol{\mu} \in \mathcal{B}$. Hence, there exists $\varphi \in int(\Delta(S))$ such that $\mu_0 = \sum_s \varphi_s \mu_s$. For each φ, define a model such that $\pi^{\varphi}(s|\omega) = (\mu_s(\omega)\ \varphi_s)/\mu_0(\omega), \forall s, \omega$. This model is a well-defined because $\pi^{\varphi}(s|\omega) \in [0,1], \forall s, \omega$, and $\sum_s \pi^{\varphi}(s|\omega) = 1$. Notice that $\Pr^{\varphi}(s) = \varphi_s$. Therefore, this model belongs to $\mathcal{M}$, since $\varphi \in int(\Delta(S))$ and $\mu_0 \in int(\Delta(\Omega))$, and it induces $\boldsymbol{\mu}$:

$$\mu_s^{\varphi}(\omega) = \frac{\mu_0(\omega)\pi^{\varphi}(s|\omega)}{\Pr^m(s)} = \frac{\mu_0(\omega)}{\varphi_s}\left(\frac{\mu_s(\omega)\ \varphi_s}{\mu_0(\omega)}\right) = \mu_s(\omega), \quad \forall \omega, s.$$

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

Consider as weights for the convex combination the fit levels of model m: $(\Pr^m(s))_{s\in S}$. As $m \in \mathcal{M}$ and $\mu_0 \in int(\Delta(\Omega))$, this is a well-defined distribution in $int(\Delta(S))$. Then, $\boldsymbol{\mu} \in \mathcal{B}$:

$$\sum_s \Pr^m(s)\ \mu_s^m(\omega) = \sum_s \Pr^m(s)\ \frac{\mu_0(\omega)\pi^m(s|\omega)}{\Pr^m(s)} = \mu_0(\omega)\sum_s \pi^m(s|\omega) = \mu_0(\omega), \quad \forall\omega. \quad \square$$

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

Proof of Corollary 3. Given Lemma 1, it is only left to show uniqueness in the case of binary signal. Let $(\varphi_{s_1}, \varphi_{s_2}) = (\varphi, 1-\varphi)$. Bayes-consistency implies $\mu_0(\omega) = \varphi\mu_{s_1}(\omega) + (1-\varphi)\mu_{s_2}(\omega)$; then, $\varphi = (\mu_0(\omega) - \mu_{s_2}(\omega))/(\mu_{s_1}(\omega) - \mu_{s_2}(\omega)), \forall\omega$. Therefore, $(\varphi_{s_1}, \varphi_{s_2}) \in int(\Delta(S))$ if (i) $\mu_{s_1}(\omega) > \mu_0(\omega) > \mu_{s_2}(\omega)$ or (ii) $\mu_{s_1}(\omega) < \mu_0(\omega) < \mu_{s_2}(\omega)$, $\forall\omega$. $\square$

Claim 1. *For every ω, $\bar{\delta}(\mu_s)^{-1} \leq (1-\mu_0(\omega))/(1-\mu_s(\omega))$.*

Proof of Claim 1. Let $\bar{\omega} = \arg\max_\omega \delta(\mu_s(\omega))$. Rewrite the condition as

$$\bar{\delta}(\mu_s) = \frac{\mu_s(\bar{\omega})}{\mu_0(\bar{\omega})} \geq \frac{1-\mu_s(\omega)}{1-\mu_0(\omega)} = \frac{\sum_{\omega'\neq\omega}\mu_s(\omega')}{\sum_{\omega'\neq\omega}\mu_0(\omega')}, \quad \forall\omega.$$

This is equivalent to $\sum_{\omega'\neq\omega}\mu_s(\bar{\omega})\mu_0(\omega') \geq \sum_{\omega'\neq\omega}\mu_0(\bar{\omega})\mu_s(\omega')$, $\forall\omega$, which is satisfied because $\mu_s(\bar{\omega})\mu_0(\omega') \geq \mu_s(\omega')\mu_0(\bar{\omega})$, $\forall\omega'$, by definition of maximal movement. $\square$

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 $\boldsymbol{\mu}$ such that (i) the target μ_s is induced conditional on s, and (ii) there exists $\varphi \in int(\Delta(S))$ such that Bayes-consistency holds with the additional property $\varphi_s = p$:

$$\sum_{s'} \mu_{s'}(\omega)\ \varphi_{s'} = \mu_s(\omega)\ \varphi_s + \sum_{s'\neq s} \mu_{s'}(\omega)\ \varphi_{s'} = \mu_0(\omega), \quad \forall\omega. \tag{a}$$

Such $\boldsymbol{\mu}$ is well-constructed if $\mu_0(\omega) - \mu_s(\omega)\ p = \sum_{s'\neq s}\mu_{s'}(\omega)\ \varphi_{s'} \geq 0, \quad \forall\omega$, which is always verified for $p \leq \mu_0(\omega)/\mu_s(\omega) = \bar{\delta}(\mu_s)^{-1}$. Then, Lemma 1 guarantees that there exists a model inducing this Bayes-consistent vector of posteriors with fit p given s.

Given the many degrees of freedom, there are multiple $\boldsymbol{\mu}$ satisfying conditions (a). For instance, consider $\boldsymbol{\mu}$ with $\mu_{s'}(\omega) = (\mu_0(\omega) - p\ \mu_s(\omega))/(1-p), \forall \omega$ and $s' \neq s$. These are well-defined posteriors by Claim 1 and $p \leq \bar{\delta}(\mu_s)^{-1}$. Condition (a) is satisfied for $\varphi_s = p$ and $\varphi_{s'} = (1-p)/(|S|-1), \forall s' \neq s$.

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

Consider an arbitrary model m with $\mu_s^m = \mu_s$. First, $\Pr^m(s) > 0, \forall s$ because $m \in \mathcal{M}$ and $\mu_0 \in int(\Delta(\Omega))$. Second, $\Pr^m(s) = \frac{\mu_0(\omega)}{\mu_s(\omega)} \pi^m(s|\omega) \leq \frac{\mu_0(\omega)}{\mu_s(\omega)}, \forall \omega$ by Bayes rule. Because this holds for every state, the maximal fit for μ_s is the minimum of this ratio across states:

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

Proof of Proposition 1. It directly follows from Lemma 1. $\square$

Proof of Theorem 1. Inducing an arbitrary $\boldsymbol{\mu}$ requires a set of at most $K = |S|$ models $(m_k)_{k=1}^K$ such that each $m_k \in \mathcal{M}$ is tailored to s_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 $j = 1, \ldots, K$.

Assume $\boldsymbol{\mu} \in \mathcal{F}$. Note that the condition of Theorem 1 can be rewritten as $\sum_s \bar{\delta}(\mu_s)^{-1} \geq 1$. I show that there exists a set of tailored models inducing $\boldsymbol{\mu}$. Instead of constructing each model, I specify $\boldsymbol{\mu}^{m_k}$ and the fit levels $\left(\Pr^{m_k}(s)\right)_{s \in S}$ and show that $\boldsymbol{\mu}^{m_k} \in \mathcal{B}$. Thus, the corresponding model exists by Lemma 1. Last, I show that each m_k is adopted given s_k.

For each m_k, specify: for s_k, $\mu_{s_k}^{m_k} = \mu_{s_k}$ and $\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1}$; for $s \neq s_k$ and every ω,

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

Posteriors are well-defined by definition of maximal movement and Claim 1. Fit levels are non-negative because $\bar{\delta}(\mu_s)^{-1} \in (0, 1]$ for every s, less than one because $\boldsymbol{\mu} \in \mathcal{F}$, and sum to one because $\sum_{s \neq s_k} \Pr^{m_k}(s) = 1 - \bar{\delta}(\mu_{s_k})^{-1}$. Such $\boldsymbol{\mu}^{m_k}$ is Bayes-consistent for $\left(\Pr^{m_k}(s)\right)_{s \in S}$:

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

Each m_k is adopted conditional on s_k because for every model model m_j it holds

$$\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq \underbrace{\left(\frac{1-\bar{\delta}(\mu_{s_k})^{-1}}{\sum_{s\neq s_k}\bar{\delta}(\mu_s)^{-1}}\right)}_{\leq 1 \text{ for } \boldsymbol{\mu}\in\mathcal{F}} \bar{\delta}(\mu_{s_k})^{-1} = \Pr^{m_j}(s_k).$$

Assume $\boldsymbol{\mu} \notin \mathcal{F}$. Then, it holds that $\sum_s \bar{\delta}(\mu_s)^{-1} < 1$, equivalent to $\bar{\delta}(\mu_{s_k})^{-1} < 1 - \sum_{i\neq k}\bar{\delta}(\mu_{s_i})^{-1}, \forall k$. If it were to exist a set of models inducing $\boldsymbol{\mu}$, each tailored model m_k inducing μ_{s_k} would be adopted given s_k, that is, $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k), \forall j \neq k$. Notice that for each m_j:

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

since $\Pr^{m_j}(s_i) \leq \Pr^{m_i}(s_i) \leq \bar{\delta}(\mu_{s_i})^{-1}$ for every i by Lemma 2. This leads to a contradiction:

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

Proof of Proposition 2. Let $K = |S|$.

Assume that $\min_\omega \mu_0(\omega) \geq 1/K$. Notice that $\bar{\delta}(\mu_s)^{-1} \geq \min_\omega \mu_0(\omega)$ because

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

Then, for every $\boldsymbol{\mu}$ it holds that $\sum_s \bar{\delta}(\mu_s)^{-1} \geq \sum_s \min_\omega \mu_0(\omega) = K \min_\omega \mu_0(\omega) \geq 1$.

Assume that $\min_\omega \mu_0(\omega) < 1/K$. I show that there exists at least one $\boldsymbol{\mu} \notin \mathcal{F}$. Let $\underline{\omega} = \arg\min_\omega \mu_0(\omega)$. Consider $\boldsymbol{\mu}$ such that for every s $\mu_s(\underline{\omega}) = 1$ and $\mu_s(\omega) = 0$, $\forall \omega \neq \underline{\omega}$. Then, $\bar{\delta}(\mu_s)^{-1} = \min_\omega \mu_0(\omega), \forall \mu_s$. Hence, $\boldsymbol{\mu} \notin \mathcal{F}$ because $\sum_s \bar{\delta}(\mu_s)^{-1} = K \min_\omega \mu_0(\omega) < 1$. $\square$

Claim 2 (Binary Case)**.** *If $\mu_0(\omega_1) \leq 50\%$ and $\mu_s(\omega_1) \leq \mu_0(\omega_1)$ for some s, $\boldsymbol{\mu} \in \mathcal{F}$.*

Proof of Claim 2. Take $\boldsymbol{\mu} = (\mu_s, \mu_{s'})$ with $\mu_s(\omega_1) \leq \mu_0(\omega_1) \leq 50\%$. Note that $\bar{\delta}(\mu_s)^{-1} = \mu_0(\omega_2)/\mu_s(\omega_2)$ and $\bar{\delta}(\mu_{s'})^{-1} \geq \min_w \mu_0(\omega) = \mu_0(\omega_1)$. Hence, $\bar{\delta}(\mu_s)^{-1} + \bar{\delta}(\mu_{s'}) \geq 1$. $\square$

Proof of Proposition 3 (Binary Case). Take $\varepsilon' < \varepsilon''$. I show that it is never the case that $\boldsymbol{\mu} \in \mathcal{F}_{\varepsilon''}$ and $\boldsymbol{\mu} \notin \mathcal{F}_{\varepsilon'}$. By Claim 2, any $\boldsymbol{\mu}$ such that $\exists s : \mu_s(\omega_1) \leq 1/2 - \varepsilon'$, then $\boldsymbol{\mu} \in \mathcal{F}_{\varepsilon'}$. Thus, consider $\boldsymbol{\mu}$ such that $\mu_s(\omega_1) > 1/2 - \varepsilon' > 1/2 - \varepsilon'', \forall s$. Let $\bar{\delta}_\varepsilon(\mu_s) =$

$\max_\omega \mu_s(\omega)/\mu_{0,\varepsilon}(\omega)$. Then,

$$\bar{\delta}_{\varepsilon''}(\mu_s) = \frac{\mu_s(\omega_1)}{1/2 - \varepsilon''} \geq \frac{\mu_s(\omega_1)}{1/2 - \varepsilon'} = \bar{\delta}_{\varepsilon'}(\mu_s).$$

It follows that if $\sum_s \bar{\delta}_{\varepsilon''}(\mu_s)^{-1} \geq 1$, $\sum_s \bar{\delta}_{\varepsilon'}(\mu_s)^{-1} \geq 1$. That is, if $\mu_s \in \mathcal{F}_{\varepsilon''}$, then $\mu_s \in \mathcal{F}_{\varepsilon'}$. □

Proof of Proposition 4. Let $K = |S|$, $N = |\Omega|$, $\underline{\omega} = \arg\min_\omega \mu_0(\omega)$ and $p = \mu_0(\underline{\omega})$.

Take any $\boldsymbol{\mu} \in \mathcal{B}$. By Lemma 1, there exists a model m such that $\boldsymbol{\mu}^m = \boldsymbol{\mu}$. Then, $\boldsymbol{\mu}^m \in \mathcal{F}$ because $\sum_s \bar{\delta}(\mu_s^m)^{-1} \geq \sum_s \Pr^m(s) = 1$, since $\bar{\delta}(\mu_s^m)^{-1} \geq \Pr^m(s), \forall s$ by Lemma 2.

It is left to show that there exists $\boldsymbol{\mu} \in \mathcal{F}$ such that $\boldsymbol{\mu} \notin \mathcal{B}$. If $\min_\omega \mu_0(\omega) \geq 1/K$, all vectors of posteriors are feasible by Proposition 2. If $\min_\omega \mu_0(\omega) < 1/K$, consider $\boldsymbol{\mu}$ such that, $\forall s$, $\mu_s(\underline{\omega}) = Kp$ and $\mu_s(\omega) = ((1 - Kp)\mu_0(\omega))/(1 - p)$, $\forall \omega \neq \underline{\omega}$. These are well-defined for $p < 1/K$. Then, $\bar{\delta}(\mu_s) = \delta(\mu_s(\underline{\omega})) = K \geq (1 - Kp)/(1 - p) = \delta(\mu_s(\omega))$, $\forall \omega$. Hence, $\boldsymbol{\mu} \in \mathcal{F}$ because $\sum_s \bar{\delta}(\mu_s)^{-1} = \sum_s 1/K = 1$, and $\boldsymbol{\mu} \notin \mathcal{B}$ because it induces the same $\mu_s \neq \mu_0$, $\forall s$. □

Proof of Proposition 5. Consider a dummy signal $s_0 \notin S$ and the enlarged signal space: $S' = S \cup \{s_0\}$. It can be shown that any $\boldsymbol{\mu} \in [\Delta(\Omega)]^S$ can be induced.

Take $\boldsymbol{\mu} \notin \mathcal{F}$, otherwise the statement would be trivially true. To induce $\boldsymbol{\mu}$ on S, I show that there exists a set of $K = |S|$ models $(m_k)_{k=1}^K$ such that each $m_k \in \mathcal{M} \subset [\Delta(S')]^\Omega$ is tailored to induce μ_{s_k} given s_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. Instead of constructing each m_k, I specify $\boldsymbol{\mu}^{m_k}$ and $(\Pr^{m_k}(s))_{s \in S}$ and show that $\boldsymbol{\mu}^{m_k} \in \mathcal{B}$. Thus, the corresponding model exists by Lemma 1.

For each m_k, specify: for s_k, $\mu_{s_k}^{m_k} = \mu_{s_k}$, and for $s \in S'$ with $s \neq s_k$, for every ω

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

Posteriors are well-defined by definition of $\bar{\delta}(\mu_s)$ and Claim 1. Also, set $\Pr^{m_k}(s_0) = 1 - \sum_{k=1}^K \bar{\delta}(\mu_{s_k})^{-1}$, and for $s \neq s_0$ $\Pr^{m_k}(s) = \bar{\delta}(\mu_s)^{-1}$. The fit levels are well-defined because $\boldsymbol{\mu} \notin \mathcal{F}$ and $\bar{\delta}(\mu_s)^{-1} \in (0, 1]$. Each $\boldsymbol{\mu}^{m_k}$ is Bayes-consistent for $(\Pr^{m_k}(s))_{s \in S}$ because

$$\sum_{s \in S'} \mu_s^{m_k}(\omega)\Pr^{m_k}(s) = \mu_{s_k}(\omega)\bar{\delta}(\mu_{s_k})^{-1} + \frac{\mu_0(\omega) - \bar{\delta}(\mu_{s_k})^{-1}\mu_{s_k}(\omega)}{1 - \bar{\delta}(\mu_{s_k})^{-1}}(1 - \bar{\delta}(\mu_{s_k})^{-1}) = \mu_0(\omega), \ \forall \omega.$$

Each m_k is adopted given s_k since $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each j. □

Proof of Proposition 6. Fix a state $\omega^* \in \Omega$. Assume the sender wants to maximize $\mu_s(\omega^*), \forall s$. Let $K = |S|$ and $p = \mu_0(\omega^*)$. If $p \geq 1/K$, there exists $\boldsymbol{\mu} \in \mathcal{F}$ such that $\mu_s(\omega^*) = 1, \forall s$ by Proposition 2.

Assume $p < 1/K$ and fix a signal $s' \in S$. Consider a $\boldsymbol{\mu}$ such that $\mu_s(\omega^*) = 1, \forall s \neq s'$. To be feasible $\bar{\delta}(\mu_{s'}) \leq 1/(1-p(K-1))$ because $\sum_{s \neq s'} \bar{\delta}(\mu_s)^{-1} = p(K-1)$. At most, the remaining posterior could be set to $\mu_{s'}(\omega^*) = p/(1-p(K-1))$.

It is left to show that such there exists $\boldsymbol{\mu} \in \mathcal{F}$ with these features. Consider $\boldsymbol{\mu}$ with $\mu_s(\omega^*) = 1, \forall s \neq s'$, $\mu_{s'}(\omega^*) = p/(1-p(K-1))$, and $\mu_{s'}(\omega) = \mu_0(\omega)(1-\mu_{s'}(\omega^*))/(1-p), \forall \omega \neq \omega^*$; this vector of posteriors is well-defined for $p < 1/K$. Note that $\bar{\delta}(\mu_{s'}) = \delta(\mu_{s'}(\omega^*))$ because

$$\delta(\mu_{s'}(\omega^*)) = \frac{1}{1-p(K-1)} \geq \frac{1}{1-p}\left(1 - \frac{p}{1-p(K-1)}\right) = \delta(\mu_{s'}(\omega)), \quad \forall \omega \neq \omega^*,$$

which is verified for $p < 1/K$. Thus, $\boldsymbol{\mu} \in \mathcal{F}$. □

Proof of Proposition 7 (Binary Case, Polarization). Consider m and m' with $\pi^m(s_1|\omega_1) > \pi^m(s_1|\omega_2)$ and $\pi^{m'}(s_1|\omega_2) > \pi^{m'}(s_1|\omega_1)$. $\Pr^m(s_1) > \Pr^{m'}(s_1)$ is equivalent to:

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

If $\mu_0(\omega_1) < p$, $\boldsymbol{\mu}^M = \left(\mu_{s_1}^{m'}, \mu_{s_2}^{m}\right)$, otherwise $\boldsymbol{\mu}^M = \left(\mu_{s_1}^{m}, \mu_{s_2}^{m'}\right)$. As m and m' are conflicting, $\boldsymbol{\mu}^M \notin \mathcal{B}$ with, $\forall s$, (i) $\mu_s(\omega_1) < \mu_0(\omega_1)$ if $\mu_0(\omega_1) < p$, or (ii) $\mu_s(\omega_1) > \mu_0(\omega_1)$ if $\mu_0(\omega_1) > p$. □

Proof of Theorem 2. Inducing an arbitrary $\boldsymbol{\mu}$ requires at most $K = |S|$ models $(m_k)_{k=1}^K$ such that each $m_k \in \mathcal{M}$ is tailored to s_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 $\boldsymbol{\mu} \in \mathcal{F}^d$. First, note that $\boldsymbol{\mu} \in \mathcal{F}$ because the condition of Theorem 1 is satisfied:

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

To induce $\boldsymbol{\mu}$, for each m_k construct $\boldsymbol{\mu}^{m_k}$ and $\left(\Pr^{m_k}(s)\right)_{s \in S}$ following the proof of Theorem 1. Then, because $\boldsymbol{\mu}^{m_k} \in \mathcal{B}$, the corresponding model exists by Lemma 1, and each m_k is adopted with respect to $m \in \{m_1, \ldots, m_K\}$ given s_k. It is only left to show that each m_k is adopted given s_k with respect to d. Because $\boldsymbol{\mu} \in \mathcal{F}^d$: $\Pr^{m_k}(s_k) = \bar{\delta}(\mu_{s_k})^{-1} \geq \Pr^d(s_k)$.

Assume $\boldsymbol{\mu} \notin \mathcal{F}^d$. Then, there must be a signal s_ℓ such that $\bar{\delta}(\mu_{s_\ell})^{-1} < \Pr^d(s_\ell)$. If it were to exist a set of models inducing $\boldsymbol{\mu}$, each tailored model m_k inducing μ_{s_k} would be adopted given s_k, that is, $\Pr^{m_k}(s_k) \geq \Pr^m(s_k)$ for each $m \in \{m_1, \ldots, m_K\} \cup \{d\}$. This leads to a contradiction because $\Pr^d(s_\ell) > \bar{\delta}(\mu_{s_\ell})$ by assumption. □

Proof of Proposition 8. Because $\mathcal{F}^d$ depends only on $\left(\Pr^d(s)\right)_{s\in S}$, rewrite:

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

Take $\boldsymbol{\mu} \in \mathcal{F}$. It is to be shown that for each $\boldsymbol{\mu} \in \mathcal{F}$ there exists $p \in int(\Delta(S))$ such that $\bar{\delta}(\mu_s)^{-1} \geq p_s$, $\forall s$. Set $p_s = \bar{\delta}(\mu_s)^{-1}/\sum_{s'} \bar{\delta}(\mu_{s'})^{-1}, \forall s$. This is a well-defined distribution in $int(\Delta(S))$ since $\bar{\delta}(\mu_s)^{-1} \in (0, 1]$. As needed, $\bar{\delta}(\mu_s)^{-1} \geq p_s$ since $\sum_s \bar{\delta}(\mu_s)^{-1} \geq 1$.

Take $\boldsymbol{\mu} \in \bigcup_{d\in\mathcal{M}} \mathcal{F}^d$. Then, there exists $p \in int(\Delta(S))$ such that $\bar{\delta}(\mu_s)^{-1} \geq p_s, \forall s$. 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, $\bar{\delta}(\mu_s)^{-1} \geq 1 - \sum_{s'\neq s} \bar{\delta}(\mu_{s'}), \forall s$. □

References

Aina, Chiara and Florian H. Schneider (2024) "Weighting Competing Models."

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2020) "The Polarization of Reality," 110, 324–328.

Alonso, Ricardo and Odilon Câmara (2016) "Persuading voters," *American Economic Review*, 106 (11), 3590–3605.

Ambuehl, Sandro and Heidi C Thysen (2023) "Competing Causal Interpretations: A Choice Experiment."

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart (2024) "Narratives about the Macroeconomy."

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*.

Andreoni, James and Tymofiy Mylovanov (2012) "Diverging opinions," *American Economic Journal: Microeconomics*, 4 (1), 209–32.

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 (2022) "Experimental Persuasion."

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny (1998) "A model of investor sentiment," *Journal of financial economics*, 49 (3), 307–343.

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 (2024a) "Narrative Persuasion."

——— (2024b) "Narrative Persuasion: A Brief Introduction."

Battigalli, Pierpaolo and Martin Dufwenberg (2022) "Belief-dependent motivations and psychological game theory," *Journal of Economic Literature*, 60 (3), 833–82.

Bauch, Gerrit and Manuel Foerster (2024) "Strategic communication of narratives."

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," *Quarterly Journal of Economics*, 117 (3), 871–915.

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

Bilotta, Francesco and Giacomo Manferdini (2024) "Coarse Memory and Plausible Narratives."

Bohren, J Aislinn and Daniel N Hauser (2024) "Behavioral Foundations of Model Misspecification."

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2017) "Memory, attention, and choice," *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 (2023) "Opinions as Facts," *Review of Economic Studies*, 90 (4), 1832–1864.

Charles, Constantin and Chad Kendall (2024) "Causal Narratives."

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."

de Clippel, Geoffroy and Xu Zhang (2022) "Non-Bayesian Persuasion," *Journal of Political Economy*, 130 (10), 2594–2642.

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.

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 (2022) "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 and Ran Spiegler (2020) "A model of competing narratives," *American Economic Review*, 110 (12), 3786–3816.

——— (2024) "News media as suppliers of narratives (and information)," *arXiv preprint arXiv:2403.09155*.

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.

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.

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 (2024) "Ideological bias and trust in information sources."

Gilboa, Itzhak and David Schmeidler (1993) "Updating ambiguous beliefs," *Journal of Economic Theory*, 59 (1), 33–49.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann (2023) "Stories, Statistics, and Memory."

Harman, Gilbert H (1965) "The inference to the best explanation," *The Philosophical Review*, 74 (1), 88–95.

Ichihashi, Shota and Delong Meng (2021) "The Design and Interpretation of Information."

Ispano, Alessandro (2024) "The perils of a coherent narrative."

Izzo, Federica, Gregory J Martin, and Steven Callander (2023) "Ideological Competition," *American Journal of Political Science*, 67 (3), 687–700.

Jain, Atulya (2024) "Informing agents amidst biased narratives."

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.

Köszegi, Botond (2006) "Ego utility, overconfidence, and task choice," *Journal of the European Economic Association*, 4 (4), 673–707.

Laibson, David (1997) "Golden eggs and hyperbolic discounting," *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.

Levy, Gilat and Ronny Razin (2021) "A maximum likelihood approach to combining forecasts," *Theoretical Economics*, 16 (1), 49–71.

Lipton, Peter (2003) *Inference to the best explanation*: Routledge.

Little, Andrew T (2023) "Bayesian explanations for persuasion," *Journal of Theoretical Politics*, 35 (3), 147–181.

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?" *Quarterly Journal of Economics*, 126 (1), 373–416.

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*.

Montiel Olea, José Luis, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (2022) "Competing models," *Quarterly Journal of Economics*, 137 (4), 2419–2457.

Montiel Olea, José Luis and Andrea Prat (2025) "Competing Ideologies: Fit, Simplicity, and Fear."

Morag, Dor and George Loewenstein (2024) "Narratives and valuations," *Management Science*.

O'Donoghue, Ted and Matthew Rabin (1999) "Doing it now or later," *American Economic Review*, 89 (1), 103–124.

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.

Ortoleva, Pietro (2012) "Modeling the change of paradigm: Non-Bayesian reactions to unexpected news," *American Economic Review*, 102 (6), 2410–2436.

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," *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.

——— (2024) "Shared Models in Networks, Organizations, and Groups."

Shafer, Glenn (1976) *A mathematical theory of evidence*, 42: Princeton university press.

Sharot, Tali (2011) "The optimism bias," *Current biology*, 21 (23), R941–R945.

Shmaya, Eran and Leeat Yariv (2016) "Experiments on decisions under uncertainty: A theoretical framework," *American Economic Review*, 106 (7), 1775–1801.

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," *Quarterly Journal of Economics*, 131 (3), 1243–1290.

Spiegler, Ran, Kfir Eliaz, and Simone Galperti (2025) "False Narratives and Political Mobilization," *Journal of the European Economic Association*.

Szeidl, Adam and Ferenc Szucs (2025) "A Model of Populism as a Conspiracy Theory," *American Economic Review*, 115 (9), 3214–3247.

Thagard, Paul (1989) "Explanatory coherence," *Behavioral and Brain Sciences*, 12 (3), 435–467.

Wojtowicz, Zachary (2024) "Model Diversity and Dynamic Belief Formation."

Yang, Jeffrey (2024) "A Criterion of Model Decisiveness."

A Online Appendix: Other Belief Updating Rules

This appendix provides additional results to illustrate why the set characterized in Theorem 1 can be interpreted as an upper-bound on belief manipulability for a class of assumptions on how the receiver forms beliefs given the models she is exposed to. I consider the information-based belief updating rules in Table A1, varying either how the model adoption or how the inference occurs, and show that each of these rules induces a feasible vector of posteriors.

Rule	**Model Adoption**	**Inference**	**Statement**
ML Selection	Select best-fitting model	Bayes rule	
Mixed Model	Convex combination of models	Bayes rule	Proposition A1
Bayesian	Bayesian weights given priors over models	Bayes rule	Proposition A2
Biased-ML Bayesian	Bayesian weights biased towards best-fitting model	Bayes rule	Proposition A3
ML Underinference	Select only best-fitting model	Underinference	Proposition A4

Table A1: Belief updating rules discussed in Appendix A

A.1 Formal Statements

I start by considering a receiver that updates beliefs using a new model constructed by mixing the models she has been exposed.

Proposition A1 (Mixed Model)**.** *Assume the receiver to update beliefs using a mixed model αM constructed as a convex combination of models with weights $\alpha^m \in [0,1]$ for every $m \in M$ with $\sum_m \alpha^m = 1$. The resulting vector of posteriors $\boldsymbol{\mu}^{\alpha M}$ is Bayes-consistent, thus feasible.*

Proof of Proposition A1. The mixed model αM is defined for each ω and s as

$$\pi^{\alpha M}(s|\omega) = \sum_{m \in M} \alpha^m \pi^m(s|\omega).$$

This model is always well-defined because (1) for each ω and s, $\pi^{\alpha M}(s|\omega) \in [0,1]$, (2) for each ω, $\sum_s \pi^{\alpha M}(s|\omega) = \sum_s \sum_m \alpha^m \pi^{m_k}(s|\omega) = \sum_m \alpha^m \sum_s \pi^{m_k}(s|\omega) = 1$, and (3) it belongs to $\mathcal{M}$ because $M \subseteq \mathcal{M}$. Hence, $\boldsymbol{\mu}^{\alpha M} \in \mathcal{B} \subset \mathcal{F}$ by Lemma 1 and Proposition 4. □

Next, I look at the traditional case in which the receive is Bayesian with priors over models.

Proposition A2 (Bayesian)**.** *Assume the receiver to be Bayesian with prior over models $\rho \in \Delta(M)$. The resulting vector of posteriors $\boldsymbol{\mu}^{(\rho,M)}$ is Bayes-consistent, thus feasible.*

Proof of Proposition A2. A Bayesian agent with prior ρ forms posterior for ω and s as

$$\mu_s^{(\rho,M)}(\omega) = \sum_{m\in M} \rho_s^m \ \mu_s^m(\omega), \quad \text{with} \ \ \rho_s^m = \frac{\rho^m \ \Pr^m(s)}{\sum\limits_{m'\in M} \rho^{m'} \ \Pr^{m'}(s)}.$$

This is equivalent to update beliefs using a mixed model with the weights equals to the priors over models: $\boldsymbol{\mu}^{(\rho,M)} = \boldsymbol{\mu}^{\rho M}$. To see this, calculate the posterior for every ω and s:

$$\begin{aligned}\mu_s^{(\rho,M)}(\omega) &= \sum_{m\in M} \underbrace{\frac{\rho^m \ \Pr^m(s)}{\sum\limits_{m'\in M} \rho^{m'} \ \Pr^{m'}(s)}}_{\rho_s^m} \underbrace{\frac{\mu_0(\omega)\pi^m(s|\omega)}{\Pr^m(s)}}_{\mu_s^m(\omega)} \\ &= \frac{\mu_0(\omega) \sum\limits_{m\in M} \rho^m \pi^m(s|\omega)}{\sum\limits_{m\in M} \rho^m \ \Pr^m(s)} = \frac{\mu_0(\omega)\pi^{\rho M}(s|\omega)}{\Pr^{\rho M}(s)} = \mu^{\rho M}(\omega).\end{aligned}$$

Thus, $\boldsymbol{\mu}_s^{(\rho,M)} \in \mathcal{B} \subset \mathcal{F}$ by Proposition A1. □

Then, I study the case in which the receiver has priors over the models, but biases her Bayesian beliefs towards the best-fitting model's prediction.[26] If the bias was maximal, the receiver updates beliefs as in the main text; if the bias is minimal, the receiver is Bayesian.

Proposition A3 (Biased-ML Bayesian)**.** *Assume the receiver to form beliefs as a convex combination between the Bayesian posterior with prior over models $\rho \in \Delta(M)$ and best-fitting model's posterior: for every s and $\beta \in [0,1]$, $\mu_s^{\beta(\rho,M)} = \beta\mu_s^{m_s^*} + (1-\beta)\mu_s^{(\rho,M)}$. The resulting vector of posteriors $\boldsymbol{\mu}^{\beta(\rho,M)}$ is feasible.*

Proof of Proposition A3. To show that $\boldsymbol{\mu}^{\beta(\rho,M)} \in \mathcal{F}$, I show that there exists a model for which $\boldsymbol{\mu}^{\beta(\rho,M)}$ belongs to the set of feasible vectors of beliefs given this model as default. Then, by Theorem 2, $\boldsymbol{\mu}^{\beta(\rho,M)}$ is feasible. In particular, I show that $\boldsymbol{\mu}^{\beta(\rho,M)} \in \mathcal{F}^{\rho M}$ where ρM is the mixed model with weights given by the prior over the model ρ.

Notice that

$$\begin{aligned}\delta\left(\mu_s^{\beta(\rho,M)}(\omega)\right) &= \frac{\beta\mu_s^{m_s^*}(\omega) + (1-\beta)\mu_s^{(\rho,M)}(\omega)}{\mu_0(\omega)} = \beta\delta\left(\mu_s^{(\rho,M)}(\omega)\right) + (1-\beta)\delta\left(\mu_s^{m_s^*}(\omega)\right) \\ &\leq \beta\bar{\delta}\left(\mu_s^{(\rho,M)}\right) + (1-\beta)\bar{\delta}\left(\mu_s^{m_s^*}\right) \leq \max\left\{\bar{\delta}\left(\mu_s^{(\rho,M)}\right), \bar{\delta}\left(\mu_s^{m_s^*}\right)\right\}, \ \forall s,\omega.\end{aligned}$$

[26] While interesting, I do not study the characterizing condition to generalize Theorem 1 for this belief updating rule. To do so, I would have to make arbitrary assumptions on how the receiver forms prior beliefs on the models she has been exposed to, which would ultimately drive the result. For example, assuming the receiver to form uniform beliefs on the proposed models might create incentives to communicate more models than signals to dilute her prior.

Because this holds for every ω, then it follows that

$$\bar{\delta}\left(\mu_s^{\beta(\rho,M)}\right)^{-1} \geq \left(\max\left\{\bar{\delta}\left(\mu_s^{(\rho,M)}\right), \bar{\delta}\left(\mu_s^{m_s^*}\right)\right\}\right)^{-1}.$$

To show that $\boldsymbol{\mu}^{\beta(\rho,M)} \in \mathcal{F}^{\rho M}$, it has to hold that $\bar{\delta}\left(\mu_s^{\beta(\rho,M)}\right)^{-1} \geq \Pr^{\rho M}(s), \forall s$. This condition is verified because $\left(\max\left\{\bar{\delta}\left(\mu_s^{(\rho,M)}\right), \bar{\delta}\left(\mu_s^{m_s^*}\right)\right\}\right)^{-1} \geq \Pr^{\rho M}(s), \forall s$. To see this, consider the two cases separately. First, $\bar{\delta}\left(\mu_s^{(\rho,M)}\right)^{-1} \geq \Pr^{\rho M}(s), \forall s$ by Lemma 2 and $\boldsymbol{\mu}^{\rho M} = \boldsymbol{\mu}^{(\rho,M)}$ (see the proof of Proposition A2). Second, by Lemma 2 and $\Pr^{m_s^*}(s) \geq \Pr^m(s) \forall m$,

$$\bar{\delta}\left(\mu_s^{m_s^*}\right)^{-1} \geq \Pr^{m_s^*}(s) \geq \sum_m \rho^m \Pr^m(s) = \Pr^{\rho M}(s), \forall s. \qquad \square$$

Last, I consider the case in which, once adopted the best-fitting model, the receiver underinfers compared to the Bayesian prediction as there is ample evidence that individuals mostly underinfer from signals (Benjamin, 2019).

Proposition A4 (ML Underinference)**.** *Assume the receiver to select the best-fitting model but underinfer when applying Bayes rule to update beliefs by a factor of $\theta \in [0,1]$. The resulting vector of posteriors $\boldsymbol{\mu}^{M_\theta}$ is feasible.*

Proof of Proposition A4. Once adopted model m given s, the receiver stays closer to the prior by a factor $1-\theta$: $\mu^{m_\theta}(\omega) = \theta\mu_s^m(\omega) + (1-\theta)\mu_0(\omega)$. If $\theta = 1$, the receiver uses Bayes rule (as in the main text); otherwise, she does not update.

Notice that $\bar{\delta}(\mu_s^{m_\theta}) \leq \bar{\delta}(\mu_s^m)$ for every m. Since $\theta \in [0,1]$ and $\bar{\delta}(\mu_s) \geq 1$ for every μ_s, it holds

$$\bar{\delta}(\mu_s^{m_\theta}) = \max_\omega \frac{\theta\mu_s^m(\omega) + (1-\theta)\mu_0(\omega)}{\mu_0(\omega)} = \theta \max_\omega \frac{\mu_s^m(\omega)}{\mu_0(\omega)} + (1-\theta) = \theta\bar{\delta}(\mu_s^m) + (1-\theta) \leq \bar{\delta}(\mu_s^m).$$

Recall that $m_s^* \in \arg\max_{m \in M} \Pr^m(s)$. Theorem 1 implies that $\boldsymbol{\mu}^M \in \mathcal{F}$ with $\sum_s \bar{\delta}\left(\mu_s^{m_s^*}\right)^{-1} \geq 1$. Since $\bar{\delta}\left(\mu_s^{m_\theta}\right)^{-1} \geq \bar{\delta}(\mu_s^m)^{-1}$, it holds that $\sum_s \bar{\delta}\left(\mu_s^{m_{s,\theta}^*}\right)^{-1} \geq 1$. Therefore, $\boldsymbol{\mu}^{M_\theta} \in \mathcal{F}$. $\square$

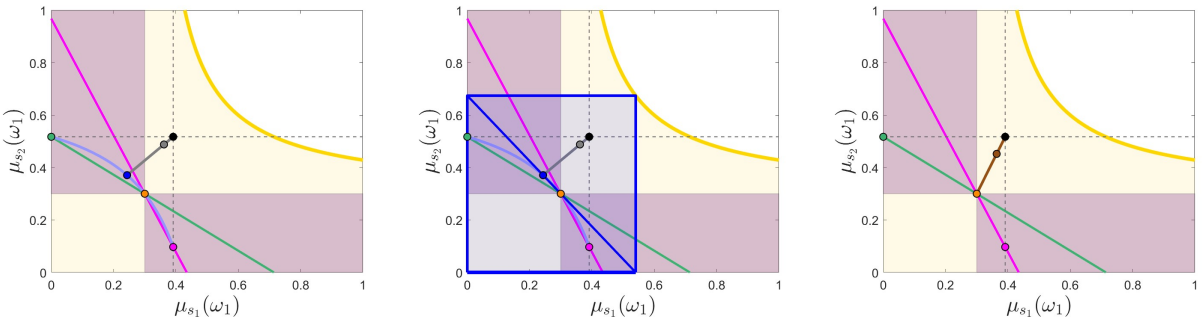
A.2 Graphical Intuition

Consider the receiver to be exposed to two models, $M = \{m_1, m_2\}$. In Figure A1, the pink and green points respectively corresponds to the induced vectors of posteriors of m_1 and m_2. The black point illustrated the resulting vector of posterior from selecting the best-fitting model $\boldsymbol{\mu}^M$.

If the receiver is Bayesian or uses a mixed model, her beliefs lie on the light blue line in Figure A1a. This line always lies in the Bayes-consistency area. The blue point represents

the resulting vectors of posteriors $\boldsymbol{\mu}^{(\rho,M)}$ for equal prior over models $\rho = (0.5, 0.5)$ and $\boldsymbol{\mu}^{\rho M}$ calculated using mixed model ρM.

To look at the case in which the receiver forms Bayesian beliefs biased towards the best-fitting model, consider the gray line: this illustrates $\boldsymbol{\mu}^{\beta(\rho,M)}$ for every β. If $\beta = 0$, this coincides with $\boldsymbol{\mu}^{(\rho,M)}$ and if $\beta = 1$ $\boldsymbol{\mu}^{M}$. The gray point shows the case in which $\beta = 0.8$. The proof of Proposition A3 shows that such beliefs are always feasible using the construction of Figure A1b. I show that $\boldsymbol{\mu}^{\beta(\rho,M)}$ is feasible by showing that it always belongs to the set of feasible vector of posterior given the mixed model ρM as default.



(a) (b) (c)

Figure A1: Graphical intuition for the binary case

Finally, if the receiver underinfers given the observed signals, the resulting vectors of posterior for every θ connecting $\boldsymbol{\mu}^{M}$ and $\boldsymbol{\mu}_0$, that is, the brown line of Figure A1c. The brown point illustrates this for $\theta = 0.7$. Because these vectors of posteriors are closer to the prior, they are always feasible.

B Online Appendix: Suggestive Evidence on Model Polarization in the 2020 US Presidential Election

This appendix provides some suggestive evidence of the polarization mechanism discussed in Section 4.1, using the 2020 US presidential election as a case study. Unlike in previous election campaigns, competing narratives about the fairness of the electoral process circulated well before the vote: one portraying the system as reliable, the other casting doubt on the fairness of the election system. The preemptive diffusion of these conflicting narratives marked a departure from earlier elections and fits closely with the example discussed.

When exposed to conflicting models, we should observe: (i) voters with different initial beliefs adopt different models on the election system once the outcome is observed, 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; that is, 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.[27] Figure B1 shows more details about the distributions of priors by reported party affiliation.

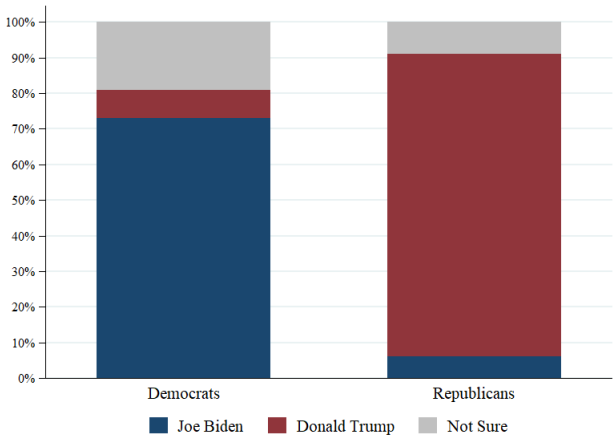


Figure B1: Priors on election winner, by party affiliation

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

[27]This pattern in priors can be 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. With this assumption, this framework can rationalize the findings of Alesina et al. (2020), who document that objective facts are perceived differently depending on partisan affiliation.

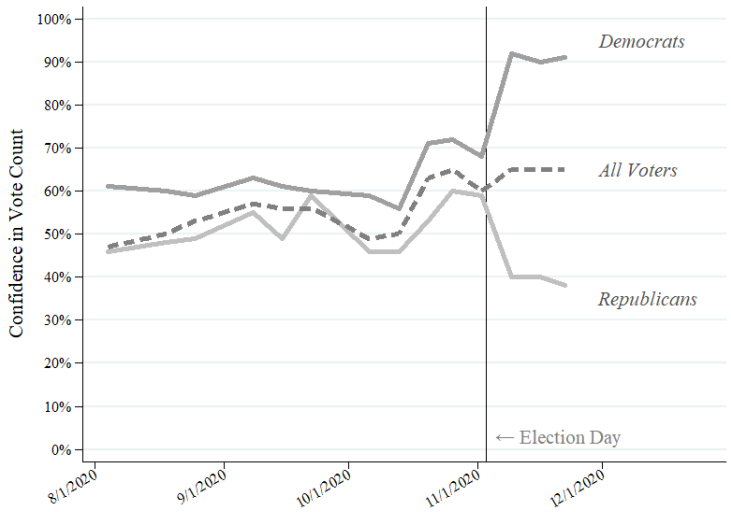


Figure B2: Accuracy of vote count (Persily and Stewart, 2021)

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

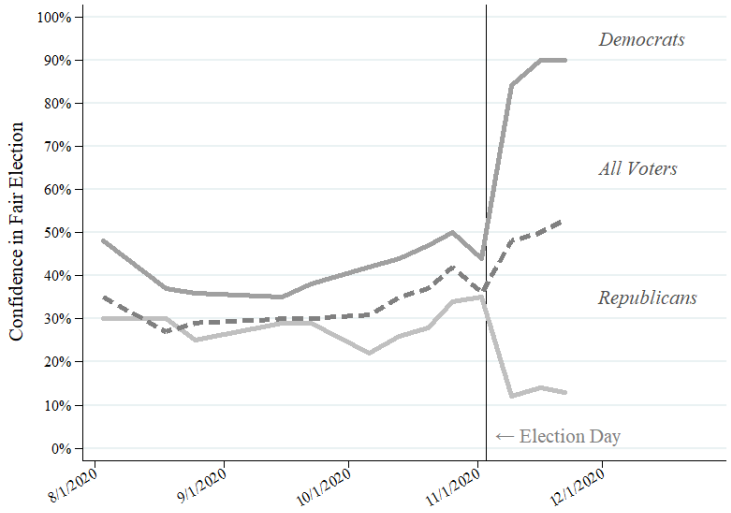
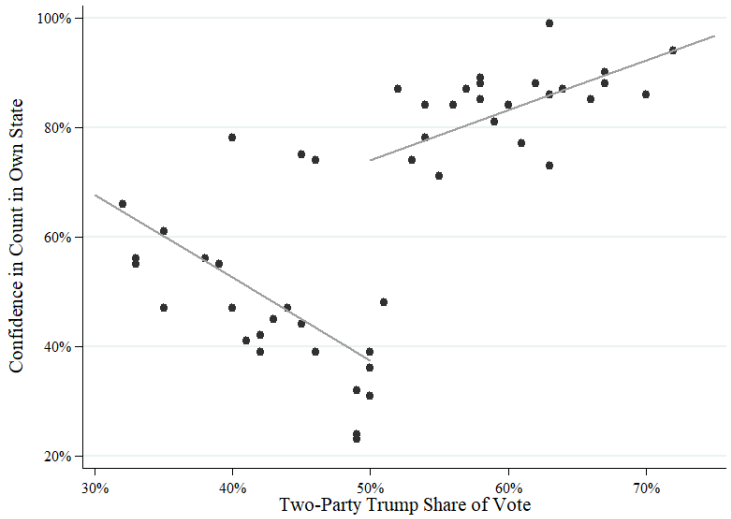


Figure B3: Confidence in fair election (Persily and Stewart, 2021)

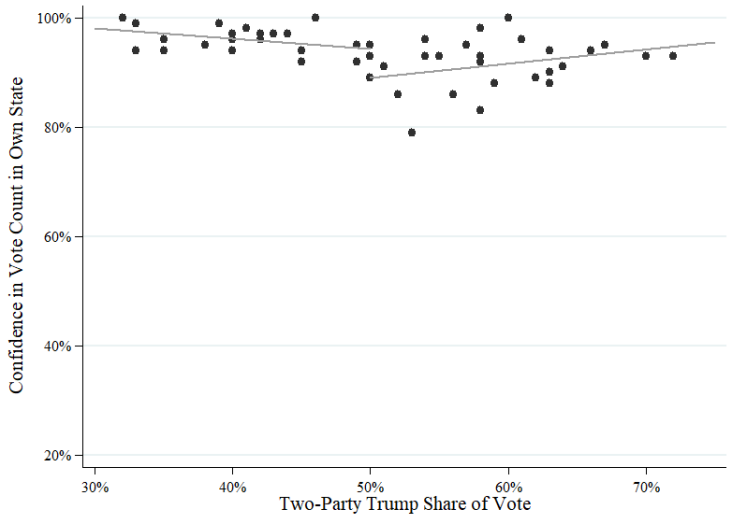
Notes: The y-axis shows the percentage answering "a great deal" or "quite a bit" in response to the question "How much confidence do you have that the 2020 presidential election [will be held/was held] fairly?" *Source:* Economist/YouGov poll, 2020.

Figure B2 shows the confidence in accurate vote count over time. tpersily2021miracle 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, the aggregate measure remained unchanged but an extreme partisan polarization occurred: the gap between Democrats and Republicans went from 10.9% to 51.7%. A similar pattern appears in Figure B3, which examines responses to a question on whether the presidential election was held fairly. Along this measure, the pre-election gap was 15%, while the post-election one was 72.6%. This suggests that voters with different priors adopt different models once the election outcome is observed: after the election, Democrats adopted the narrative claiming the election system to be fair, while Republicans adopted an alternative story questioning the integrity of the process. This effect is not unique to the 2020 election, and it is also known as the "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 B4 reports data on the confidence in state elections by the percentage of Trump's share of votes. Republicans mostly distrust the accuracy of the state elections if they live 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 rigged elections were right-leaning, it is not unreasonable to assume that Democrats discarded them.



Republicans



Democrats

Figure B4: 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.