Tailored Stories

Chiara Aina*

December 20, 2023

Abstract

To what extent is it possible to manipulate beliefs by providing interpretations of unknown events? I characterize the feasible posteriors across signals 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, Narratives, Polarization.
JEL classification: D82, D83, D9.

*Department of Economics, Harvard University. E-mail: chiaraaina@fas.harvard.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 Sandro Ambühl, Ian Ball, Kai Barron, Pierpaolo Battigalli, Roberto Corrao, Tristan Gagnon-Bartsch, George Loewenstein, 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 Boston University, CMU, CREED, Harvard, MIT, NYU, UPF, Wharton, ECBE, NBER Political Economy, and SITE.

1 Introduction

Beliefs are shaped by how we interpret the world. When we use different interpretations to make sense of the same event, we might reach contrasting conclusions. Voters may disagree on the outcome of an election. Consumers often differ in how they evaluate companies based on the same public initiatives. Investors make different predictions based on the same past data. This occurs even when we share the same preferences and initial beliefs. One potential explanation for reaching divergent conclusions in such cases is that we adopt different narratives to interpret the same data. Narratives link what we observe to what we want to understand: they provide interpretations of events.[1] Thus, influencing the narratives people adopt can be a powerful tool to manipulate and persuade them. Indeed, when making sense of the observed data, one might rely on narratives provided by more knowledgeable sources, such as political figures, financial advisors, or experts considered trustworthy. This type of persuasion is powerful because it allows for changing people's beliefs without controlling or even knowing what they observe.

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

[1] The Cambridge Dictionary defines a narrative as "a particular way of explaining or understanding events." Despite the growing attention to this topic in economics, there is not yet a commonly shared definition of what a narrative is. Different ways of formalizing it have emerged in recent years, and I discuss the main ones when reviewing the related literature. Barron and Fries (2022) provide a detailed discussion of the current conceptualization of narratives in economics in their appendix.

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

outcome. Before the election, the politician communicates to the voter models about the election system. Assume that the politician communicates only the model according to which the voting system is fair. Since there is only one model available, the voter always adopts that. Then, once the election outcome is revealed, the voter would recognize the politician as president if the latter is the reported winner, while she would not otherwise. How can the politician be recognized as the legitimate president regardless of the election outcome? He cannot manipulate the reported election outcome or the voting system. However, before the vote, the politician could also promote a conspiracy theory according to which elections are rigged.[3] Exposure to multiple models allows inconsistent reasoning to take root: each election outcome triggers the adoption of a different model. The voter's initial beliefs play a crucial role because they drive which model is adopted based on the reported outcome. Assume that the voter expects the politician to win the election fairly. If the politician is the reported winner, the most plausible model is the one about the just voting system; however, if the politician is not the reported winner, the conspiracy theory resonates best with the voter. This is equivalent to the voter holding an inconsistent interpretation across election outcomes: "if this politician is reported as the winner, the election system is fair; otherwise, elections are rigged."[4] As a result, the voter updates upwards her beliefs about the politician being the legitimate winner, recognizing him as president regardless of the election outcome. The politician achieved this by leveraging how the voter makes sense of the reported election outcome, and he exploited it by providing conflicting models.

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

First, the sender can always bias the receiver's beliefs in a given direction. If many models are provided, each signal might lead the receiver to adopt a different model. As a result, the receiver's beliefs may be inconsistent across realizations: all posteriors might be higher or lower than the

[3]For simplicity of exposition, I describe these inconsistent models as provided by a single agent. Alternatively, one could think about this as a coordinated strategy implemented by different agents, e.g., different members of the same party implementing the same strategy in a coordinated manner. The receiver might be less sensitive to this type of contradiction, and the credibility of the sources would be less likely to be questioned.

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

prior. Bayesian models do not allow for this type of inconsistency. In this setting, the sender can induce an inconsistent vector of posteriors by providing as many models as there are possible signal realizations. In the example, each voter's posterior is higher than her prior. The politician achieves this with two models: one tailored to the case of reported victory and one tailored to the case of reported loss.

Second, there are constraints on beliefs the sender is able to induce. Generally, not all vectors of posteriors are feasible. To induce a vector of posteriors, the sender should construct a set of tailored models so that each model is adopted conditional on the signal to which it has been tailored, inducing the desired posterior 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*: each tailored model maximally fits the target signal given the desired posterior. 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. To better convey the intuitions behind these formal results, I introduce a graphical approach for the special case of binary signal and state (hereafter, binary case). This also yields a graphical construction of which vectors of posterior beliefs are feasible.

The main theorem provides a feasibility result that goes beyond persuasion. The same constraint on belief manipulability holds regardless of how the receiver is exposed to models, e.g., multiple senders are supplying models or models are "in the air" (not strategically supplied). Such assumptions do not vary the feasible vectors of posteriors, but rather results on optimality. Moreover, the assumption on adopting the best-fitting model allows me to investigate the upper-bound on belief manipulability for a large class of assumptions on the receiver.[5] With this result, this paper aims to provide a building block to study belief manipulability using models. Persuasion with a monopolistic sender is the most natural application of this setting which I use through the paper to highlight the relevance of the findings and to benchmark it with the literature.

Having explored the limits of belief manipulability, I turn to the question of what makes the receiver more vulnerable to persuasion. Initial beliefs play a crucial role. In the binary case, the

[5]Appendix B shows that that the resulting vectors of posteriors with other information-based belief updating rules belong to the feasibility set. All the following cases satisfies this: (i) the receiver updates beliefs using a model constructed as a convex combination of the models she was exposed to, (ii) the receiver is Bayesian with prior over models, (iii) the receiver has priors over models but biases her Bayesian beliefs towards the best-fitting model, and (iv) the receiver updates her beliefs using the best-fitting model but underinfers.

sets of feasible vectors of posteriors can be ordered based on the prior: the closer the receiver's prior is to the uniform distribution, the more she can be manipulated. When her prior is 50-50, the receiver is fully persuadable: the sender can provide a set of models to make her hold any beliefs regardless of what she observes. More generally, I provide necessary and sufficient conditions for full manipulability. The sender has more leeway to manipulate if there are many signals to be interpreted and few states on which the receiver has dispersed priors. Even if the sender could manipulate the signal-generating process, the feasible vectors of posteriors would not be affected. The only way in which the sender could enlarge further this set would be to convince the receiver that other signals could be observed, even if these dummy signals can never occur.

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

I present several stylized applications that fit this setting. First, I discuss the consequences of being exposed to conflicting models in a political setting. Communicating a large number of possibly contradictory and untruthful stories is one of the central features of the "firehose of falsehood," a propaganda usually associated with modern Russia (Paul and Matthews, 2016). The exposure to conflicting ways of interpreting information can lead to both confirmation bias and inevitable polarization. I formalize this result for the binary case and I provide some suggestive evidence of this mechanism using the case of the 2020 US presidential election.[6] Second, I study the misalignment of incentives between a financial advisor and investors with private information. The framework of this paper is suitable to study a private-information setting: the receiver has access to the signal, but the sender does not. Communicating different models that could be picked up depending on the private information of the investors, like past financial experience, allows the advisor to always move beliefs in an advantageous direction. I illustrates the optimal communication strategy for the advisor for both optimistic and pessimistic investors. The third application explores a multiple-selves setting in which an agent can distort her own beliefs by manipulating the perceived informativeness of observable signals or leaving data open to interpretation. This proposed mechanism can deliver the classic implications of the literature on motivated beliefs but also sets a bound on belief distortion. I provide an example of how models allow an agent to distort her confidence to offset her time-inconsistent preferences and commit to a costly action. Last, I show how a strategic persuader

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

could challenge a shared model to insinuate doubt and deepen polarization for agents differing in initial beliefs. I exemplify this in the context of the lost trust in science on issues like climate change and the health effects of smoking, where the so-called "merchants of doubt" (e.g., Michaels, 2008; Oreskes and Conway, 2011) provided alternative ways of interpreting scientific evidence. Holding a shared initial model does not deter polarization in a population of heterogeneous receivers.

This paper speaks directly to two strands of economic literature — narratives and persuasion — that have flourished in the last decade (see Section 6 for a more detailed discussion of the related literature). Starting from Shiller (2017, 2019), there has been an increasing formalization of narratives in the economic literature using different notions: narratives as likelihood functions (Schwartzstein and Sunderam, 2021), or directed acyclical graphs (Eliaz and Spiegler, 2020). This paper builds on the first approach. Inspired by the interdisciplinary research on sense-making (Andreassen, 1990; Weick, 1995; DiFonzo and Bordia, 1997; Chater and Loewenstein, 2016), Schwartzstein and Sunderam (2021) formalize the concept of models as used in this paper and assume that individuals prefer the model that best fits the observed data and prior knowledge. They study the problem of manipulating a receiver endowed with a default model by strategically providing her with a model after a public signal is realized (*ex-post*). Instead, I investigate a setting in which the sender commits to his communication strategy without knowing the signal realization (*ex-ante*). The reason is two-fold. First, it is a sensible assumption. Shifting communication ex-ante may give more credibility to the sender. For example, a voter may be skeptical if the politician claims elections to be rigged only after he lost the election. Also, the sender might be unable to learn the information available to the receiver in some cases. For instance, an investor might prefer not to disclose to her financial advisor some relevant private information, such as previous experiences. Second, ex-ante commitment imposes a constraint on the sender, pinned down in the main result of this paper. Extending this result to the case in which the receiver has a default model allows comparability with Schwartzstein and Sunderam (2021): I find that the sender can attain the same outcome with ex-ante or ex-post communication of models.[7] However, because models compete with each other, the set of ex-post optimal models might not be optimal if provided ex-ante.

The strategic provision of models implies significant differences from the previous literature on persuasion. The sender does not alter the signal the receiver observes, unlike the cheap talk literature (e.g., Milgrom, 1981; Crawford and Sobel, 1982). Moreover, there is a fixed signal generating process that cannot be manipulated. This is in stark contrast with the literature on Bayesian persuasion, started by Kamenica and Gentzkow (2011) and continued by many generalizations of

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

their framework (e.g., Alonso and Câmara, 2016; Ely, 2017; Galperti, 2019; Ball and Espín-Sánchez, 2021). In broad terms, these papers are about persuasion by generating information: the sender commits to an experiment that maps each state to a distribution of signals. In the political example, this translates into the politician manipulating the voting system and its accuracy. Because the chosen signal generating process induces a distribution over the receiver's posterior beliefs, such a distribution must be Bayes-plausible: the expected posterior has to average to the prior. With this paper, I relax the Bayes plausibility constraint in a disciplined manner. By providing models ex-ante, the sender can induce posteriors across realizations unattainable with Bayesian persuasion. However, this communication strategy generally imposes restrictions on what the sender can achieve without the ability to modify how the signal is generated or which signal is observed.

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

2 Set-up

Two agents, 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 $\mu_0 \in int(\Delta(\Omega))$.[8] 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 it does not exists a signal $s \in S$ such that $\pi^m(s|\omega) = 0$ for each ω. 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 signal s.

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

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

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

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.

In case of misaligned preferences, the sender has incentives to communicate a set of models with the purpose of influencing the receiver's action in order to maximize his expected utility rather than hers. The sender knows the receiver's preferences and the true model t, specifying the objective probabilities of signals under the signal generating process. Let $\boldsymbol{\mu} = (\mu_s)_{s \in S} \in [\Delta(\Omega)]^S$ be a *vector of posterior beliefs*: it describes the posterior beliefs conditional on each signal realization. Thus, the value of a vector of posteriors $\boldsymbol{\mu}$ equals the sender's expected utility given the receiver's actions at those beliefs calculated using model t:

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

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^* that maximizes his value at the resulting vector of posteriors:

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

Discussion of Assumptions Before presenting the results, I discuss some of the assumptions behind this setting. I start by focusing on the receiver. First, I relax the Bayes rationality of the receiver only partially: she updates her prior via Bayes rule once she has selected a model. Following Schwartzstein and Sunderam (2021), the model that maximizes the likelihood of observed data is adopted.[9] Appendix B shows how 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.[10] This is in line with *Inference to the Best Explanation* (Harman,

[9]The literature on belief updating under ambiguity considers a maximum likelihood updating rule, introduced by Dempster (1967) and Shafer (1976), then axiomatized by Gilboa and Schmeidler (1993). When agents consider multiple priors over states, they only update the subset of priors that maximizes the probability of the realized event. The other most common rule for updating in the case of multiple priors is full Bayesian updating (Jaffray, 1992; Pacheco Pires, 2002): subjects update prior-by-prior and retain ambiguity in their posteriors. I do not consider multiple priors over states but multiple models that could be used for updating. In this setting, a recent paper (Frick et al., 2022) shows that maximum likelihood updating is maximally efficient in learning under ambiguity aversion.

[10]For example, a Bayesian receiver would average the prediction of each model weighted by the posterior of each model given the observed signal and her prior over models.

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.[11] A line of research in cognitive psychology argues that hypotheses are supported by the same observations they are supposed to explain, and the more they explain, the more confidence we give to that hypothesis (Koehler, 1991; Pennington and Hastie, 1992; Lombrozo and Carey, 2006). Douven and Schupbach (2015a,b) provide evidence of the importance of explanatory power in updating and predicting estimates of posterior probabilities.[12] Second, I assume the receiver to be naïve. The receiver does not know the true model and does not form beliefs about the possible models. As a result, 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, it would allow her to learn more about the state directly, ignoring the sender's proposed models. This full naïveté assumption is a common starting point in the literature (e.g., Heidhues and Kőszegi, 2018; Eyster, 2019) and it incorporates several motives leading the receiver to underreact to the sender's private information and incentives.

The sender's behavior differs in a two-fold manner from Schwartzstein and Sunderam (2021). First, the sender communicates models without knowing the signal realization. This allows for a *temporal* interpretation: a public signal will be observed by both agents, but the sender has to provide models before its realization. Also, this assumption can accommodate a *private-information* interpretation: the receiver might hold some private information on the state — the signal — and the sender cannot access it. Second, the sender can communicate as many models as he wishes. Because he does not know the signal realization, he has incentives to send multiple models that could be picked up depending on the realization. Note that there is no need to have more models than the number of signal realizations. This explains why, in the interim model by Schwartzstein and Sunderam (2021), the authors do not consider multiple models provided by the sender — unless the sender wishes to persuade multiple receivers. In addition to the aforementioned differences, I further refine the definition of models to exclude cases where the agent has to update beliefs conditional on a zero-probability signal.[13] This restriction only affects the preliminary results of Section 3.1, and does not impact the main results. Importantly, this restriction has bite only if the receiver is

[11] 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 it might also be that communicated models are stored in the receiver's memory and the best-fitted one is the easiest to retrieve (e.g., Bordalo et al., 2017).

[12] Model selection via maximum likelihood is equivalent to selecting the model with the higher posterior probability given the signal, starting from a flat prior over proposed models. There is some evidence that people choose the most probable hypothesis. Simple and more probable explanations are valued (Einhorn and Hogarth, 1986; Thagard, 1989), but in the absence of a simplicity difference, people prefer more probable explanations (Lombrozo, 2007).

[13] This only occurs if conditional on signal s the agent adopt model m with $\Pr^m(s) = 0$. Given interior priors, $\Pr^m(s) = 0$ if and only if $\pi^m(s|\omega) = 0$ for each ω. I opted for restricting models to the case in which this never occurs for any signal rather than making an additional, and possibly more arbitrary, assumption on how agents update beliefs after a zero-probability signal.

exposed to a single model that differs from the true one, or a set of models, such that all models have zero fit given some signal. In all other cases, the receiver would not adopt a model which deem the observed signal as unforeseeable.[14] The sender has no incentive to communicate this way because, as the following section shows, he has more leeway to manipulate beliefs by providing more models each tailored to a specific signal and when there are more signals to be interpreted.

3 Ex-Ante Model Persuasion

This section starts with two preliminary results. First, I illustrate the connection between models and vectors of posterior beliefs. Second, I pin down the trade-off between how well a model can fit the observed realization and how much a model can move the posterior away from the prior. Then, I state the main result of the paper and provide a graphical intuition for the binary case. Last, I discuss some comparative statics and the sender's problem.

3.1 Preliminaries

To solve the sender's problem in the Bayesian persuasion literature, it is pivotal to characterize not only the posteriors the receiver might attain but also with which probabilities these posteriors can be induced, i.e., the distribution of the receiver's posteriors. The relevant constraint is that each information structure feasible for the sender corresponds to a Bayes-plausible distribution over posteriors: the expected posterior must be equal to the prior.[15] By contrast, this paper assumes a fixed distribution over the signals induced by the true model. Therefore, to solve the sender's problem, it is enough to study the vectors of posteriors that the receiver could hold. Given this, the sender optimizes his value over this feasibility set.

As a first step, I show an equivalent representation between models (information structures with fixed signal space) and vectors of posteriors (the support of the distribution of posteriors) under a condition comparable to Bayes-plausibility. 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. Bayes-consistency is the only restriction that Bayesian updating imposes on vectors of posteriors.

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

[14] This issue is not relevant in Schwartzstein and Sunderam (2021), where the receiver is endowed with a default model and the sender strategically provides an alternative model with higher fit, or in Bayesian persuasion, where the sender chooses the true model and thus the receiver cannot encounter unforeseen realizations.

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

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 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. Indeed, any model that leads beliefs to react a lot given a signal realization (higher movement) cannot fit the data well (lower fit).

3.2 Feasible Vectors of Posterior Beliefs

In this section, I characterize the set of feasible vectors of posteriors that the receiver could hold. The first result is straightforward and shows that only Bayes-consistent vectors of posterior beliefs are feasible when a single model is proposed.

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

Next, I consider the case in which the receiver is exposed to many models. This 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.[16]

Theorem 1 (Many Models)**.** *The set of feasible vectors of posterior beliefs 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 always 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.[17] A model is tailored to a specific signal realization if (i) it induces the target posterior conditional

[16] The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the given set of values. Let $x = (x_1, \ldots, x_N)$, then the harmonic mean is $H(x) = \left(\sum_{i=1}^{N} x_i^{-1}/N\right)^{-1}$.

[17] Providing a number of models equal to the number of signals allows maximal belief manipulability. More models would not enlarge the set because, at most, one model is adopted conditional on each signal realization.

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

3.2.1 Graphical Intuition

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

In the following graphs, the axes represent the posteriors attached to state ω_1 conditional on each signal realization; thus, each point represents a vector of posterior beliefs. I represent the prior $\mu_0(\omega_1) = 0.3$ as the vector of posteriors with all posteriors equal to the prior (orange point). The purple area in Figure 1a depicts all Bayes-consistent vectors of posteriors: either $\mu_{s_1}(\omega) > \mu_0(\omega) > \mu_{s_2}(\omega)$, or $\mu_{s_1}(\omega) < \mu_0(\omega) < \mu_{s_2}(\omega)$. Intuitively, updating beliefs always in the same direction is impossible. By Lemma 1, every point in the purple area corresponds to a model.[18]

Figure 1b focuses on a model to observe some of its properties. 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 points on the isofit line correspond to models with the same fit conditional on each signal.[19] The slope of the isofit line can be interpreted as follows: the steeper (flatter) the line, the higher the fit conditional on s_1 (s_2). For each level of fit, it is possible to partition $\mathcal{B}$ into three subsets: vectors induced by models with the same fit (isofit line), vectors induced by models with higher fit given s_1 (red area), and vectors induced by models with higher fit given s_2 (blue area).

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

[19] 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

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

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.

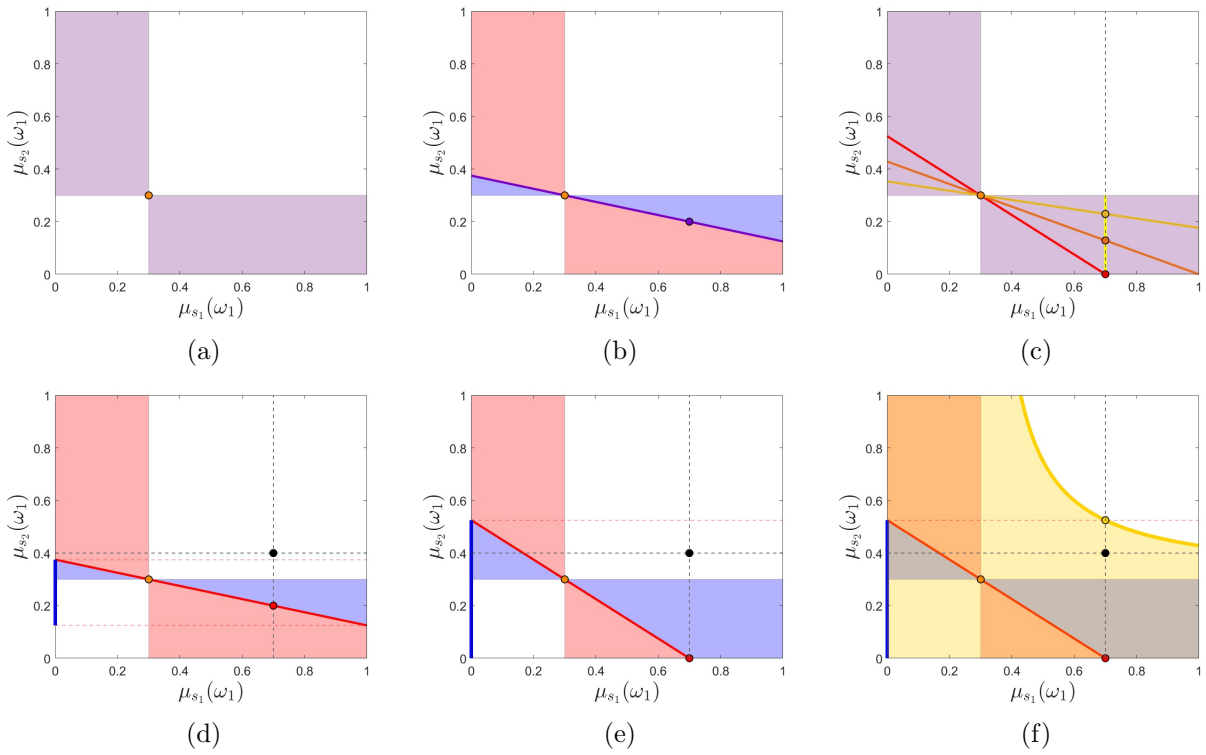


Figure 1: Graphical intuition for the binary case

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 distribution conditional on a signal 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 (black point) that is not Bayes-consistent requires two models: m_1 tailored to s_1 and m_2 tailored to s_2. I start by fixing a model m_1 inducing the target μ_{s_1} and then identify the compatible posteriors given s_2 if model 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 conditional on s_2 with respect to m_1 are all the y-coordinates of points in the blue area or on isofit line (blue line on the y-axis). Even if 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 that can induce 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 with respect to m_1 expands: by increasing $\Pr^{m_1}(s_1)$, $\Pr^{m_1}(s_2)$ decreases, and thus, more models can be adopted conditional on 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. While not all vectors of posterior beliefs are generally feasible, interestingly, this is not the case when the receiver's minimal prior across states is sufficiently high with respect to the reciprocal number of signal realizations. In this case, the receiver is fully persuadable: any vector of posteriors can be induced.

Proposition 2. *All vectors of posterior beliefs 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 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 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 states can be interpreted as a measure of the concentration of beliefs because, by increasing the minimal prior over states, 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 manipulable in a setting with many 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 (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 $\Pr^m(s_2) < \sum_{s \neq s_1} \Pr^m(s) = 57\%$. Moreover, the number of signals should be larger compared to the number of states; otherwise, with fewer signals than states, the receiver is never fully persuadable because the condition of Proposition 2 is never satisfied. The next result formalizes these key insights.

Corollary 1. *If* $|S| \geq |\Omega|$*, all vectors of posteriors are feasible if* $\mu_0(\omega) = 1/|\Omega|$ *for every* ω*. If* $|S| < |\Omega|$*, not all vectors of posteriors are feasible regardless of the prior.*

A stronger result holds in 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 $\mu_{0,\varepsilon} = (1/2 - \varepsilon, 1/2 + \varepsilon)$ and $\mathcal{F}_\varepsilon$ the feasibility set with respect to this prior.

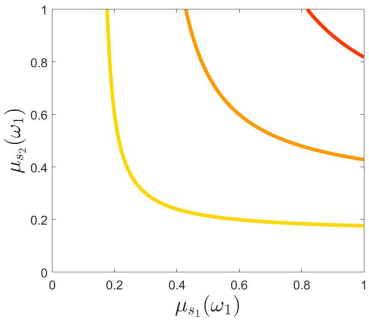


Figure 2: Frontier of the feasibility set, by prior

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

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.

There are three key restrictions faced by the sender. 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 signal realization. In what follows, I examine how these assumptions may constrain persuasion.

First, the true model cannot be manipulated. If this was allowed on top of proposing models, the sender would have the power to affect his expected value only by manipulating the probabilities of the signal realizations — the feasible vectors of posteriors would remain the same. What if he could choose: Is it more beneficial to choose the true model or propose models? The next result clarifies that proposing models always allows the sender to induce Bayes-inconsistent vectors of posteriors, that is, the support of a distribution of posteriors unattainable with Bayesian persuasion.

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

Second, the signal space is fixed. This crucial assumption restricts the sender's communication only to interpretations of observable events in S. If this were not the case and the sender could add dummy signals, he could persuade the receiver to hold any beliefs in the original space. The intuition is that if the receiver believes that 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.[20]

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. What is the impact of this constraint on the sender's expected utility? Knowing which signal the receiver observes allows the sender to communicate a tailored model inducing the desired posterior. Avoiding competition among models across signal realizations, he could induce any vector of posteriors.[21] The cost of committing ex-ante to models equals the gap between the unconstrained maximal sender's value over any vector of posteriors and the maximal sender's value over the feasible vectors of posteriors:

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

When all vectors of posteriors are feasible, there is no advantage in communicating ex-post. However, if the unconstrained maximum is unfeasible, the commitment cost is positive. This measure captures the sender's willingness to pay to learn the data available to the receiver and it can be used to comment on the value of microtargeting.[22] I discuss an example of this in Section 4.2.

4 Applications

This section discusses several applications. The first formalizes the political example outlined in the introduction, focusing on the polarizing consequences of being exposed to conflicting models. Then, I provide suggestive evidence of this mechanism. Second, a financial application illustrates the sender's optimization problem in detail. The third application discusses self-persuasion.

4.1 Firehose of Falsehood

Firehose of falsehood is a propaganda technique based on a large number of possibly contradictory and mutually inconsistent messages, defined by Paul and Matthews (2016) to describe modern

[20] This argument adds validity to only rule out models with zero-fit signals. Although Proposition 5 illustrates the consequences of manipulating the signal space, the sender has incentives to add not suppress signal realizations.

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

[22] 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`.

Russian propaganda.[23] The growing interest in understanding fake news revealed that people have a hard time distinguishing true and false stories: fake news is widely shared and believed (Allcott and Gentzkow, 2017). Spreading conflicting and fake narratives can be an effective strategy for a persuader in manipulating a target audience. However, the next example abstracts from persuasion to illustrate the consequences of exposure to conflicting models, regardless of their source.

(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 the presidency and the election outcome is soon to be reveled. 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, having 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 reliability of the election system. On the one hand, the official narrative is that mistakes in vote counting are very rare: $\pi^f(W|B) = 99\%$ and $\pi^f(W|\neg B) = 1\%$. On the other hand, 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 that initially expects Bob to be the legitimate president with high probability (Figure 3a).

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

The red point lies on the steeper isofit line and the blue point lies on the flatter isofit line: the voter would adopt f conditional on W and c conditional on $\neg W$, resulting in $\boldsymbol{\mu} = \left(\mu^f_W, \mu^c_{\neg W}\right)$ (purple point). This type of voter always recognizes Bob as the legitimate president regardless of the election outcome. In contrast, voters that expect Bob to be the legitimate president with low probability never recognize him as president (Figure 3b) because they adopt the c given W and f given $\neg W$. Interestingly, for different priors, the exposure to the same pair of models not only induces opposite actions conditional on each signal, but also polarizes beliefs.

4.1.1 Model Polarization

The previous example illustrates how exposure to conflicting models might be a strong driver of inevitable polarization. It is possible to generalize this result 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 6 (Binary Case, Polarization)**.** *For each pair of conflicting models, there exists a threshold p such that, for every signal s, it holds that (i) $\mu_s(\omega_1) < \mu_0(\omega_1)$ if $\mu_0(\omega_1) < p$, and (ii) $\mu_s(\omega_1) > \mu_0(\omega_1)$ if $\mu_0(\omega_1) > p$.*

The intuition is the following. Any pair of conflicting models induces a vector of posteriors that is not Bayes-consistent, with both posteriors higher or lower than the prior. This follows from the fact that each signal triggers the adoption of a different model. Because models are conflicting, the updating goes always in the same direction. Crucially, the prior drives in which direction the posteriors are stretched: there is a threshold such that receivers with prior higher (lower) than the threshold would hold extreme high (low) posteriors regardless of the signal realization.[24] In the 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.

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 her priors. 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,

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

leading to inevitable polarization.[25] This is in stark contrast to models with Bayesian agents with heterogeneous priors, where beliefs do not move in opposite directions regardless of the different priors if they agree on how to interpret new information (Baliga et al., 2013).

Several mechanisms have been proposed in the literature to understand the determinants of polarization. Often polarization is associated with confirmation bias, first formalized by Rabin and Schrag (1999). They assume agents misinterpret new information as supportive of current beliefs with an exogenous probability. A recent paper by Fryer et al. (2019) builds on this, assuming a similar distortion only to signals open to interpretation, and provides evidence of their predictions. They directly assume the prior to be driving the direction of polarization, while in Rabin and Schrag (1999) this role is assigned to early observed signals as agents start with a uniform prior over states. Instead, in this paper, confirmation bias is not assumed but implied by adopting the best-fitting model and can only occurs in the presence of conflicting models. Baliga et al. (2013) provide an explanation for polarization based on ambiguity aversion, in which agents hedge against uncertainty by making predictions in different directions depending on the prior after intermediate signals. Other papers illustrate how polarization arises with Bayesian updating in the presence of additional relevant features, such as high dimensionality of signal space compared to state space (Andreoni and Mylovanov, 2012) or private signals on the interpretation of evidence (Benoît and Dubra, 2019). Recent papers discuss how mistakes in source credibility could amplify polarization. Cheng and Hsiaw (2022) investigate belief distortion due to double-using the data to update beliefs on the states while also updating beliefs on source accuracy. Also, this mechanism can lead agents to disagree on how to interpret the same data. Gentzkow et al. (2021) shows how a small bias in data perception due to ideological preferences can cause divergent beliefs about both the state and the source precision, even with Bayesian updating. Unlike these papers, I contribute to this literature on polarization by highlighting why such divergence in beliefs could happen given a supply of conflicting models. Furthermore, this channel illustrate another form of polarization: agents polarize in how they interpret new data, rather than in beliefs. This form of model polarization occurs in two forms. First, agent with sufficiently different priors adopt different models to explain the same observed outcome. Second, when agents with similar priors observe different outcomes, they also adopt different models to make sense of the different information. The next

[25]There are different ways to measure polarization: ideological polarization (the extent to which the electorate has divergent beliefs on ideological issues, e.g., Dixit and Weibull, 2007), partisan sorting (the extent to which voters identify with a party, e.g., Levendusky, 2009; Mason, 2015), and affective polarization (the extent to which party members dislike members of other parties, e.g., Iyengar et al., 2019). I focus on the first: posteriors on states shift in different directions depending on the prior. This type of polarization has been documented for many decades. In the ground-breaking paper by Lord et al. (1979) and similar subsequent studies (e.g., Darley and Gross, 1983; Plous, 1991; Russo et al., 1998), subjects were asked to read the same study relative to a controversial issue (e.g., capital punishment, nuclear technology), then judge whether it provides evidence for or against the issue, and finally report how the study changed their beliefs. They all find that participants' final attitudes were either more in favor if initially favorable to the issue, or less in favor if initially opposed to the issue.

section provides suggestive evidence of this mechanism.

4.1.2 The Case of the 2020 US Presidential Election

The debate on the fairness of the 2020 US election fractured the American electorate. No evidence was found supporting the claims of widespread voter fraud, yet competing narratives on dysfunctional elections were broadly diffused. These allegations circulated for the entire election campaign before the vote. In particular, the incumbent president at the time, Donald Trump, cast doubts on the election system, especially on the mail-in ballots, well ahead of the election results. When ballots were tabulated, some voters interpret the reported election outcome using these narratives, concluding the election to be rigged.

The preemptive provision of an alternative narrative with respect to the conventional idea that the election system is fair fits well with the application discussed in the previous section. 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.[26,27]

Figure 4 shows the confidence in accurate vote count over time. Persily and Stewart (2021) report that before the election, around half of poll respondents expressed confidence that their own vote would be counted accurately, with Democrats slightly more confident than Republicans. After the release of the election outcome, the aggregate measure remained unchanged but an extreme partisan polarization occurred: the gap between Democrats and Republicans went from 10.9% to 51.7%.[28] This suggests that voters with different priors adopt different models once the election outcome is observed: after the election, Democrats adopt the narrative claiming the election system to be fair, while Republicans adopt an alternative story questioning the integrity of the process. This effect is not unique to the 2020 election, and it is also known as the "winners-losers effect":

[26] See Appendix C for more details about the distributions of priors.

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

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

after the election, supporters of the losing candidate tend to question the legitimacy of the election, while supporters of the winning candidate tend to gain confidence in the election system (Sances and Stewart, 2015; Sinclair et al., 2018). However, the 2020 gap is much wider than in previous elections (Persily and Stewart, 2021). A potential explanation is the disproportionate spread of distrustful narratives during the 2020 election campaign compared to previous elections.

Suggestive evidence about the second prediction can be found by looking at how voters' confidence in state elections changes depending on the state's reported election outcome. Figure 5 reports data on the confidence in state elections by the percentage of Trump'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 the election accuracy were right-leaning, it is reasonable to assume that Democrats discard them. People tend to ignore messages inconsistent with their view or coming from sources perceived as untrustworthy (Graber, 1984).

4.2 Financial Advice

Next, I illustrate the optimization problem for a financial advisor who wants to persuade investors to make a specific investment. It is well-known that commissions on investments could lead to a conflict of interest for the advisor. I consider the case in which the advisor knows that the investors' past financial experience influences their beliefs about the quality of new investments.[29] 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.[30] In finance, two important alternatives are that returns can exhibit predictable mean reversion or predictable continuation. When returns exhibit predictable mean reversion, high past returns predict low future returns, and a contrarian strategy — selling following high returns

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

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

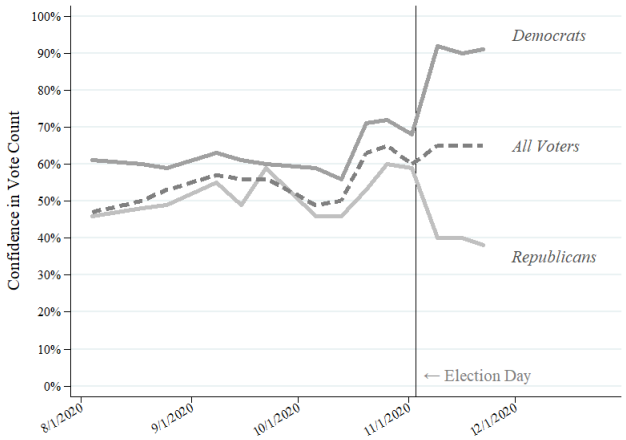
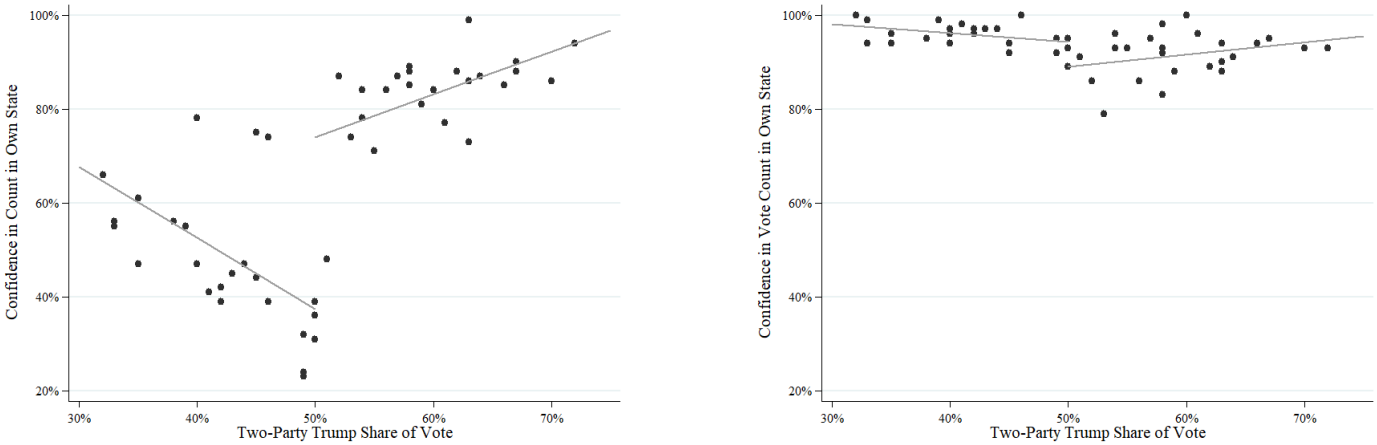


Figure 4: Accuracy of vote count (Persily and Stewart, 2021)

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



(a) Republicans (b) Democrats

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

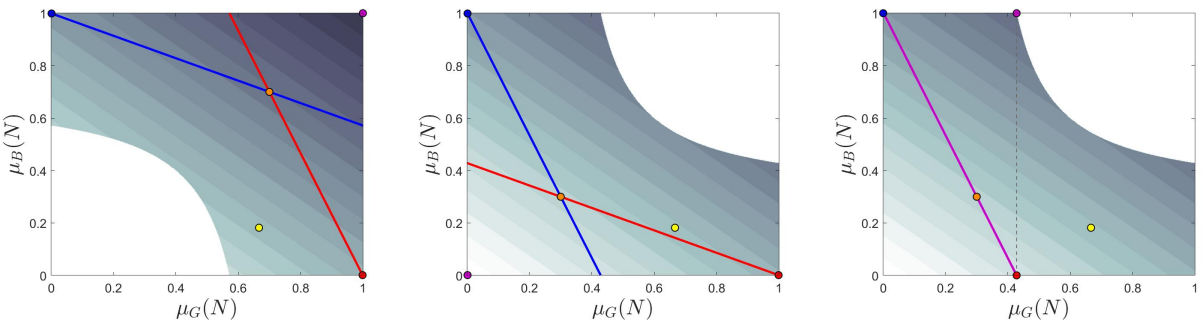
Notes: The y-axis shows the percentage answering "very confident" or "somewhat confident" in response to the question "How confident are you that votes in [state of residence] were counted as voters intended?"
Source: Survey of the Performance of American Elections (SPAE), November 2020.

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

I discuss these insights in the context of choosing a hedging strategy. Hedging aims to limit the risk of uncertain events on financial assets. Usually, it involves diversification in offsetting or opposite positions. Formally, each investor has to allocate one unit of endowment over two possible outcomes, $\Omega = \{N, \neg N\}$, where N is the event of normal conditions and $\neg N$ is the event that extreme or catastrophic circumstances occur, e.g., financial crisis or fluctuations of foreign currency in the economic domain, or flood or drought as extreme weather events. This results in the choice of $\alpha = (\alpha_N, \alpha_{\neg N})$ with $\alpha_N + \alpha_{\neg N} = 1$. All investors have the same initial beliefs and I consider two cases: *optimistic* investors expecting extreme conditions with a probability of 30%, and *pessimistic* investors expecting extreme conditions with a probability of 70%. Also, each investor had a previous experience either with normal conditions (good experience, G) or with extreme conditions (bad experience, B) and they try to understand how this previous experience impacts the future one. Assuming a logarithmic utility over the outcomes, the investor's expected utility based on her posterior μ_s is $\mathbb{E}[U^R(\alpha)] = \sum_{\omega \in \{N, \neg N\}} \mu_s(\omega) \log(\alpha_\omega)$. The investor's optimal action 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: $U^S(\alpha) = r_N\ \alpha_N + r_{\neg N}\ \alpha_{\neg N}$. Assuming $r_N > r_{\neg N} = 0$, $V(\boldsymbol{\mu}) = \sum_{s \in \{G,B\}} \Pr^t(s)\ r_N\ \mu_s(N)$. The advisor expects normal conditions with probability of 40% and knows the true model where a positive past experience positively (negatively) correlates with the success (failure) of the investment: $\pi^t(G|N) = \pi^t(B|\neg N)75\%$. Based on this, he expects 45% of investors have had a positive experience in the past, while 55% have had a negative one. Figure 6 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 6a). The financial advisor does not want to discard the investors' experiences as irrelevant. Otherwise, the investors' beliefs would remain at the prior with an investment of $\alpha_N = 70\%$. Using multiple models, the advisor can attain a higher value. The highest value for the advisor is achieved at the top-right corner, where an optimistic investor always expects normal conditions and never hedges against extreme circumstances. Intuitively, this means that the advisor can leverage any past experience of the investor and 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 conditions, i.e., $\pi^{m_1}(G|N) = \pi^{m_1}(B|\neg N) = 1$ (red point), and (ii) model m_2 suggesting a perfect negative correlation between past and future conditions, i.e., $\pi^{m_2}(B|N) = \pi^{m_2}(G|\neg N) = 1$ (blue point). These can be read as simplified versions of the momentum ("early success predicts long-run success") and mean reversion ("what goes down goes up"). Because of their optimistic initial beliefs, investors adopt the first given a good experience and the second given a bad experience. As a result, they never hedge against adverse events.



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

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

Notes: The 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 $r_N = 1$.

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 the most pessimistic model and never invests in the advisor-preferred outcome (Figure 6b).

With a pessimistic investor, the maximal value the advisor can achieve is attained at $\boldsymbol{\mu}^* = ((0.43, 0.57), (0, 1))$ at the right-top kink (Figure 6c). The optimal communication strategy is to entertain two models: (i) model m_1 such that $\pi^{m_1}(B|N) = 0$ and $\pi^{m_1}(G|\neg N) = 0.57$ (red point), and (ii) model m_2 as defined above for the optimistic investors (blue point). According to m_1, a

bad experience is a perfectly revealing signal of extreme conditions. In contrast, a past positive experience is indefinite news. This model encourages only investors with a positive experience to have $\alpha_N > 0$, and, indeed, it is tailored to those. Again, m_2 is an oversimplified version of the mean reversion, which pushes investors with a bad experience 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 N is not attainable, the advisor shifts to the second-best: convincing the largest group of investors (the ones with bad experiences) to set $\alpha_N = 100\%$ while increasing α_N for the other group (the ones with good experiences) as much as possible.

How much would the financial advisor be willing to pay to know the investors' experience? This information would allow the advisor to perfectly target each group of investors with a tailored model, 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.[31] I consider a multiple-selves setting, where the conscious mind (receiver) demands the unconscious one (sender) to supply models. This proposed mechanism to achieve self-serving beliefs can deliver the classic implications of this literature, but it also provides a bound on belief distortion.

Confirmation bias can emerge because the agent initially keeps signals open to many favorable interpretations. This could be the case of a student who thinks, and subconsciously likes, to be good at school and thus leaves the informativeness of grades open to two interpretations: grades are a good measure of own ability, or grades are based on luck. She always keeps high confidence in her abilities because she believes grades to be informative after a good grades but do not convey much information after a bad grade. In such manner, inconsistent updating across signals could result from selectively adopt 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

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

et al., 2022; Drobner and Goerg, 2021).[32]

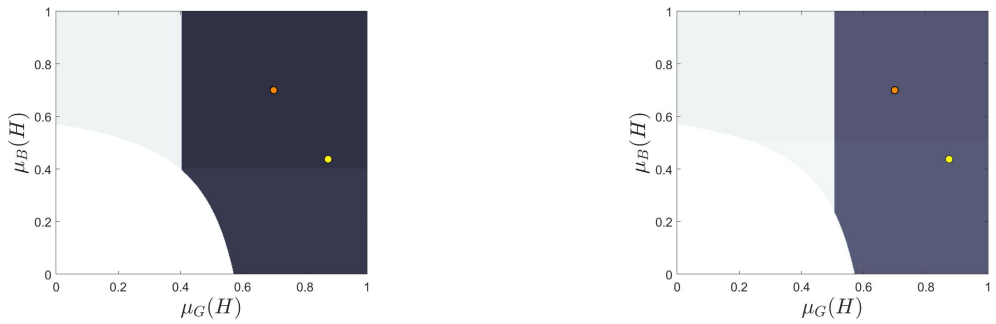
However, depending on the agent's subconscious preferences, beliefs might be distorted in any direction. The previous example assumes opposing goals for the sender and the receiver, such as higher self-confidence and accuracy. To better compare aligned and misaligned preferences, the next example focuses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

The agent can have high (H) or low (L) abilities. She receives either a good (G) or a bad (B) signal. After observing 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 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 she suffers from present bias, e.g., $\beta < 1$.[33] The agent might have the incentives to distort her own interpretations of signals to avoid a future lack of willpower.

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 7a). She does not distort her beliefs in case of aligned incentives over time. However, self-deception could be beneficial in the case of sufficiently severe present bias (Figure 7b). Before the signal, she anticipates that the imminent cost of the action will be more salient than the future reward at the moment of the decision. Thus, conditional on the bad signal, confidence in her abilities will not be high enough to act. She overcomes this

[32]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).

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



(a) No Present bias (b) Present bias

Figure 7: 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 s $c = 4$, $v = 10$, $\delta = 0.99$, and $\beta = 0.8$.

by distorting the perceived informativeness of upcoming signals — either discarding the signals as uninformative or believing only the good signal to be accurate enough. Belief manipulation allows her to stay motivated as in Bénabou and Tirole (2002), but through a different mechanism — manipulating how she interprets feedback rather than assuming memory loss or inattention.

5 Extension: Default Model

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

5.1 Feasible Vectors of Posterior Beliefs with a Default Model

Assume 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 the presence of a default model.

Theorem 2 (Default, Many Models)**.** *The set of feasible vectors of posterior beliefs 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. Indeed, the proposed models compete not only

with each other but also with the receiver's default model given each signal realization: the better the fit of the default model given a signal, the less the sender can move beliefs given that signal.[34]

Interestingly, the presence of a default model eliminates the cost of ex-ante commitment for the sender. While, in the absence of a default model, every posterior is feasible when the sender can communicate a model knowing the signal, this is not the case if the receiver is endowed with a default model. Proposition 1 of Schwartzstein and Sunderam (2021) characterizes the feasible posterior beliefs in this setting. Given 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 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) for vectors of posteriors. Because post-$\mathcal{F}^d = \mathcal{F}^d$, providing multiple models without knowing the signal (*ex-ante*) allows for the same posteriors achievable by communicating an alternative model given the realized signal (*ex-post*). Therefore, ex-ante commitment does not affect the sender's value in expectation in the presence of a default model.

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

Notice that 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 in some cases. In particular, the sender might need extra care in choosing which models to communicate if there are more than two signals.[35]

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 sets of feasible vectors of posteriors with a default model for all default models.

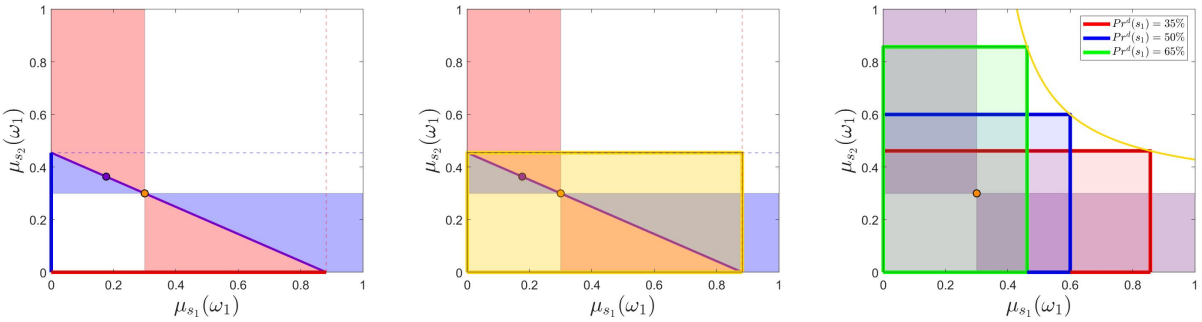
Proposition 7.

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

Figure 8 provides a graphical intuition of these results for the binary case. Consider a default

[34] A receiver endowed with a default model might be more skeptical to adopt other models and thus only switch to another model if the latter explains the data much better than the default model, e.g., $\Pr^m(s) - \Pr^d(s)$ greater than a positive threshold. This would restrict the feasibility further around the prior depending on the threshold, but the intuition would be qualitatively unchanged.

[35] 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 8: Graphical intuition of Theorem 2 and Proposition 7

model (purple point) in Figure 8a. 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 construct the feasible vectors of posteriors (yellow area) in Figure 8b. 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 8c helps build intuition for Proposition 7. The yellow line corresponds to the upper frontier of the feasibility set without a default model, while the colorful 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."

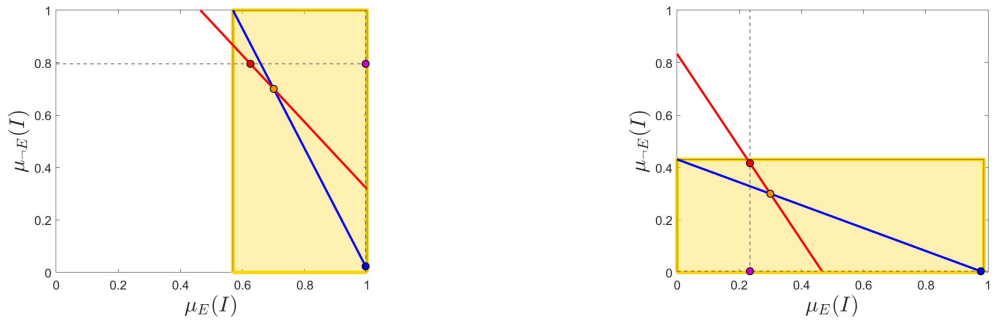
— Cigarette Executive (1969)

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

— Internal memo by The American Petroleum Institute (1998)

The strategic communication of an alternative model with respect to a commonly shared one might be used to deceive the public. This strategy was used by the tobacco industry and oil companies to challenge a well-established way of looking at the scientific evidence and to manufacture uncertainty on issues like the health effects of smoking and climate change (e.g., Michaels, 2008; Oreskes and Conway, 2011). These so-called "merchants of doubt" established a trustworthy presence in academia and media to discredit peer-reviewed articles (e.g., blaming other factors, false positive results). Their aim was to delay regulations, defeat delegations, and plant doubt in the population. Ultimately, they diluted consensus, despite the scientific community having no doubt. This strategy

guaranteed that even with new evidence emerging over time, the general public had already been exposed to competing ways of interpreting new facts and found them worthy of consideration. The following example illustrates how a shared initial model is not enough to prevent polarization in an heterogeneous population.



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

Figure 9: 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 9 shows the feasibility sets for either supportive agents with $\mu_0(I) = 70\%$ or 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 9b). By contrast, agents initially expecting the issue to be real become even more confident (Figure 9a). Strategically introducing a conflicting model promotes doubt among agents with different initial beliefs and sharing the same default model does not deter polarization.

6 Relationship to the Literature

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

Narratives Recently, there has been an effort to incorporate narratives in economics, starting with Shiller (2017, 2019). The present paper builds on the formalization of narratives as models introduced by Schwartzstein and Sunderam (2021). I also adopt the same model selection rule but investigate the strategic provisions on models without knowing the signal realization. Section 5 provides a direct comparison and discussion of the effect of this ex-ante commitment for the sender. Recent papers build on this notion of narratives and assume that best-fitted models are adopted.[36] Ichihashi and Meng (2021) sequentially combine Bayesian and model persuasion. They study a persuader who first designs and interprets information to persuade the receiver once the signal is realized. Their paper differs from mine in two ways: first, the sender chooses the signal generating process; second, they study the ex-post communication of models. Schwartzstein and Sunderam (2022) examine the social exchange of models in networks: agents start with a default shared model, but each comes up with a better interpretation after the release of data, shares it within their network, and then picks the best-fitting model among all shared ones. Social learning leads agents to have beliefs closer to the prior and feel better able to explain data than before. Moreover, recent papers take different approaches to what makes models persuasive. Ispano (2022) explores the following setting: before the signal is realized, the sender communicates a model and the receiver adopts the proposed model if it is coherent (conditional on a state, probabilities of each possible news sum to one) and compatible with her default model (the marginal distribution of news is undistorted).[37] He argues that coherence implies Bayes-consistent posteriors across signals and limits the scope of manipulation. Yang (2022) proposes a preference for "decisive models," that is, models that provide a strong recommendation regarding the best course of action.

Another influential way to formalize narratives is by describing them as causal models, expressed as directed acyclical graphs (Spiegler, 2016). 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 equi-

[36] Closely related, 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.

[37] 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 the sender proposing multiple models and how the receiver selects models across signals. To compare results, coherent and compatible models can only induce vectors of posteriors corresponding to the isofit line of the true model.

librium as a long-run distribution over narrative-policy pairs. Eliaz et al. (2021b) study to what extent a misspecified model can distort pairwise correlations between variables. In this paper, an analyst has incentives to show a strong correlation between two variables, and he can propose a (mis)specification model for estimation. The authors quantify the worst-case distortion when the proposed model is flexible, in which variables enter the model under the constraint that the estimated model cannot distort the marginal distributions of individual variables. Increasing the number of variables in the model can lead to an almost perfect correlation.[38] Directed acyclical graphs are also used by Eliaz et al. (2022) to study the proliferation of false narratives and their effect on political mobilization in a heterogeneous society of multiple social groups, and by Horz and Kocak (2022) to explore which conditions affect the effectiveness of authoritarian propaganda in reducing citizens' protests. Other formal frameworks in which narratives have been defined are Bénabou et al. (2018), where narratives are described in terms of moral value, and Izzo et al. (2021), where narratives describe the linear relation between policies and their outcome.[39]

There are recent experimental studies inspired by these formal notions of narratives (Barron and Fries, 2022; Charles and Kendall, 2022). In particular, Barron and Fries (2022) study narrative provision and adoption in a financial advice setting, building on an example discussed in Schwartzstein and Sunderam (2021). Overall, their evidence are supportive of this framework. They find that advisors with misaligned incentives communicate narratives biased from the truth and are successful in manipulating investors' beliefs in the desired direction and narratives that better fit the observed data are more persuasive.[40] Finally, there is a noteworthy line of empirical research focusing on narratives. Papers investigate their impact on beliefs and behavior (Hagmann et al., 2020; Morag and Loewenstein, 2021; Harrs et al., 2021; Graeber et al., 2022; Hillenbrand and Verrina, 2022; Bursztyn et al., forthcoming), while others document how narratives about macroeconomic phenomena are spread (Andre et al., 2021, 2022).

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

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

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

[40] In one of their treatment, the advisor does not have the opportunity to tailor the narrative to the data the investor observes. Unlike the present paper, the persuader is restricted to one narrative. Results show that advisors are less effective in moving beliefs to their target in this treatment.

nal could be manipulated. A recent paper by Gleyze and Pernoud (2022) investigates a cheap-talk game in which the receiver is not only uncertain about the state realization but also about the true model (which variables are payoff-relevant). They find that communication on models is impossible in equilibrium. Eliaz et al. (2021a) 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. As a result, full persuasion can sometimes be attained. Second, the persuader cannot influence the signal generating process. This is in stark contrast with the literature on Bayesian persuasion. Kamenica and Gentzkow (2011) and many generalizations of their framework (e.g., Alonso and Câmara, 2016; Ely, 2017; Galperti, 2019; Ball and Espín-Sánchez, 2021) are about persuasion by generating information which is then interpreted by Bayesian receivers. This restricts the sender to inducing only Bayes-plausible distributions of posteriors, unlike this paper.

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). The latter paper studies a sender who chooses an ambiguous device, that is, multiple possible signal generating processes à la Kamenica and Gentzkow (2011), one of which will be implemented with unknown probabilities unlike this paper. Also, both the sender and the receiver are assumed to be ambiguity-averse.

Biased Beliefs This paper provides alternative explanations for phenomena related to biased beliefs, such as confirmation bias and polarization — discussed in Section 4. Other papers suggest different criteria to form beliefs in uncertain settings that could lead to violation of Bayes-consistency, such as the literature on belief updating with ambiguity-aversion. For example, Epstein et al. (2021) define sensitivity to signal ambiguity as the attitude towards the uncertainty of the signal generating process and report that a fraction of subjects is averse to signal ambiguity using a lab experiment. According to their definition, being sensitive to signal ambiguity implies a violation of the Bayes-consistency property extended to preferences.

Finally, I hope to contribute also to the literature on biases in belief updating by highlighting the importance of looking at beliefs updated conditional on all possible contingencies. I show that each model maps onto a richer object, the vector of posterior beliefs. Thus, the updated posterior given a signal is compatible with many possible signal generating processes. However, in the rich literature on biased beliefs (for a survey see Benjamin, 2019), most papers look only at deviations from Bayesian updating in belief formation along one dimension — the gap between the reported

and the Bayesian posterior, conditional on the observed signal.[41]

7 Conclusion

Persuasion has typically been studied in settings where the persuader can control the information observed by the agent prior to making a decision, e.g., either by sending a persuasive message or providing new informative data. However, sometimes this is not possible. In such contexts, although the persuader cannot control or even know the information that the agent uses in making a decision, I demonstrate that the scope for persuasion using 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 the receiver is boundedly rational in choosing the interpretation of outcomes she observes, each signal realization might trigger the adoption of a different model. Thus, the sender can leverage multiple models to induce beliefs across signals that the sender cannot attain by choosing the signal generating process as in Bayesian persuasion.

Several extensions are left to be explored in future research. First, this paper focuses on a problem with only a sender and a receiver. I discuss the consequences of conflicting models in a population of receivers with different priors. Future research should develop further insights on the sender's optimization given a distribution of heterogeneous receivers, balancing the diverging effects models have. Moreover, I consider only one sender communicating multiple models. This can also be interpreted as a coordinated strategy by senders with the same incentives. The extension to the default model is the first step towards studying 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 (2022), senders communicate biased narratives to their advantage but they also display truth-telling preferences to some extent. Incorporating these frictions might lead to insightful predictions. On the other hand, receivers might consider only some types of models depending on the context. Research along this

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

line could shed light on how these restrictions harm or benefit welfare.

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

References

Allcott, Hunt and Matthew Gentzkow (2017) "Social media and fake news in the 2016 election," *Journal of economic perspectives*, 31 (2), 211–36.

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

Andre, Peter, Ingar Haaland, Chrisotpher Roth, and Johannes Wohlfart (2021) "Inflation narratives."

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022) "Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples," *The Review of Economic Studies*, 10.1093/restud/rdac008, rdac008.

Andreassen, Paul B (1990) "Judgmental extrapolation and market overreaction: On the use and disuse of news," *Journal of Behavioral Decision Making*, 3 (3), 153–174.

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

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

Baliga, Sandeep, Eran Hanany, and Peter Klibanoff (2013) "Polarization and ambiguity," *American Economic Review*, 103 (7), 3071–83.

Ball, Ian and José-Antonio Espín-Sánchez (2021) "Experimental Persuasion."

Barron, Kai (2021) "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" *Experimental Economics*, 24 (1), 31–58.

Barron, Kai and Tilman Fries (2022) "Narrative Persuasion," *Mimeo*.

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

Beauchêne, Dorian, Jian Li, and Ming Li (2019) "Ambiguous persuasion," *Journal of Economic Theory*, 179, 312–365.

Bénabou, Roland (2015) "The economics of motivated beliefs," *Revue d'économie politique*, 125 (5), 665–685.

Bénabou, Roland, Armin Falk, and Jean Tirole (2018) "Narratives, imperatives, and moral reasoning."

Bénabou, Roland and Jean Tirole (2002) "Self-confidence and personal motivation," *The quarterly journal of economics*, 117 (3), 871–915.

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

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

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

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

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

Charles, Constantin and Chad Kendall (2022) "Causal Narratives."

Chater, Nick and George Loewenstein (2016) "The under-appreciated drive for sense-making," *Journal of Economic Behavior & Organization*, 126, 137–154.

Cheng, Haw and Alice Hsiaw (2022) "Distrust in experts and the origins of disagreement," *Journal of economic theory*, 200, 105401.

Clark, Jesse and Charles Stewart, III (2021) "The Confidence Earthquake: Seismic Shifts in Trust and Reform Sentiments in the 2020 Election," *Available at SSRN*.

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

Crawford, Vincent P and Joel Sobel (1982) "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1431–1451.

Darley, John M and Paget H Gross (1983) "A hypothesis-confirming bias in labeling effects.," *Journal of Personality and Social Psychology*, 44 (1), 20.

Dempster, Arthur P (1967) "Upper and lower probability inferences based on a sample from a finite univariate population," *Biometrika*, 54 (3-4), 515–528.

DiFonzo, Nicholas and Prashant Bordia (1997) "Rumor and prediction: Making sense (but losing dollars) in the stock market," *Organizational Behavior and Human Decision Processes*, 71 (3), 329–353.

Dixit, Avinash K and Jörgen W Weibull (2007) "Political polarization," *Proceedings of the National Academy of sciences*, 104 (18), 7351–7356.

Douven, Igor and Jonah N Schupbach (2015a) "Probabilistic alternatives to Bayesianism: the case of explanationism," *Frontiers in Psychology*, 6, 459.

——— (2015b) "The role of explanatory considerations in updating," *Cognition*, 142, 299–311.

Drobner, Christoph (2022) "Motivated beliefs and anticipation of uncertainty resolution," *American Economic Review: Insights*, 4 (1), 89–105.

Drobner, Christoph and Sebastian J Goerg (2021) "Motivated belief updating and rationalization of information."

Eil, David and Justin M Rao (2011) "The good news-bad news effect: asymmetric processing of objective information about yourself," *American Economic Journal: Microeconomics*, 3 (2), 114–38.

Einhorn, Hillel J and Robin M Hogarth (1986) "Judging probable cause.," *Psychological Bulletin*, 99 (1), 3.

Eliaz, Kfir, Simone Galperti, and Ran Spiegler (2022) "False Narratives and Political Mobilization," *arXiv preprint arXiv:2206.12621*.

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

Eliaz, Kfir, Ran Spiegler, and Heidi C Thysen (2021a) "Strategic interpretations," *Journal of Economic Theory*, 192, 105192.

Eliaz, Kfir, Ran Spiegler, and Yair Weiss (2021b) "Cheating with models," *American Economic Review: Insights*, 3 (4), 417–34.

Ely, Jeffrey C (2017) "Beeps," *American Economic Review*, 107 (1), 31–53.

Epstein, Larry G, Yoram Halevy et al. (2021) "Hard-to-interpret signals."

Ertac, Seda (2011) "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback," *Journal of Economic Behavior & Organization*, 80 (3), 532–545.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel (2020) "Mental Models and Learning: The Case of Base-Rate Neglect."

Eyster, Erik (2019) "Errors in strategic reasoning," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2, 187–259.

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

Frick, Mira, Ryota Iijima, and Yuhta Ishii (2022) "Efficient learning under ambiguous information," *Mimeo*.

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

Galperti, Simone (2019) "Persuasion: The art of changing worldviews," *American Economic Review*, 109 (3), 996–1031.

Gentzkow, Matthew, Michael B Wong, and Allen T Zhang (2021) "Ideological bias and trust in information sources."

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

Gleyze, Simon and Agathe Pernoud (2022) "The Value of Model Misspecification in Communication."

Graber, Doris Appel (1984) *Processing the news: How people tame the information tide*: Longman Press.

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

Hagmann, David, Julia Minson, and Catherine Tinsley (2020) "Personal narratives build trust across ideological divides."

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

Harrs, Sören, Lara Marie Müller, and Bettina Rockenbach (2021) "How Narratives Impact Financial Behavior."

Heidhues, Paul and Botond Kőszegi (2018) "Behavioral industrial organization," *Handbook of Behavioral Economics: Applications and Foundations 1*, 1, 517–612.

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

Horz, Carlo and Korhan Kocak (2022) "How To Keep Citizens Disengaged: Propaganda and Causal Misperceptions."

Ichihashi, Shota and Delong Meng (2021) "The Design and Interpretation of Information," *Available at SSRN 3966003*.

Ispano, Alessandro (2022) "The perils of a coherent narrative."

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood (2019) "The origins and consequences of affective polarization in the United States," *Annual Review of Political Science*, 22 (1), 129–146.

Izzo, Federica, Gregory J Martin, and Steven Callander (2021) "Ideological Competition," *SocArXiv. February*, 19.

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

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

Kamenica, Emir and Matthew Gentzkow (2011) "Bayesian persuasion," *American Economic Review*, 101 (6), 2590–2615.

Kellner, Christian and Mark T Le Quement (2017) "Modes of ambiguous communication," *Games and Economic Behavior*, 104, 271–292.

——— (2018) "Endogenous ambiguity in cheap talk," *Journal of Economic Theory*, 173, 1–17.

Koehler, Derek J (1991) "Explanation, imagination, and confidence in judgment.," *Psychological bulletin*, 110 (3), 499.

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

Laibson, David (1997) "Golden eggs and hyperbolic discounting," *The Quarterly Journal of Economics*, 112 (2), 443–478.

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

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

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

Levy, Gilat and Ronny Razin (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 (2022) "Bayesian Explanations for Persuasion."

Lombrozo, Tania (2007) "Simplicity and probability in causal explanation," *Cognitive psychology*, 55 (3), 232–257.

Lombrozo, Tania and Susan Carey (2006) "Functional explanation and the function of explanation," *Cognition*, 99 (2), 167–204.

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

Malmendier, Ulrike and Stefan Nagel (2011) "Depression babies: do macroeconomic experiences affect risk taking?" *The Quarterly Journal of Economics*, 126 (1), 373–416.

Mason, Lilliana (2015) ""I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization," *American journal of political science*, 59 (1), 128–145.

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

Michaels, David (2008) *Doubt is their product: how industry's assault on science threatens your health*: Oxford University Press.

Milgrom, Paul R (1981) "Good news and bad news: Representation theorems and applications," *The Bell Journal of Economics*, 380–391.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat (2022) "Managing self-confidence: Theory and experimental evidence," *Management Science*.

Morag, Dor and George Loewenstein (2021) "Narratives and Valuations," *Available at SSRN 3919471*.

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

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

Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (forthcoming) "Competing models," *Quarterly Journal of Economics*.

Oreskes, Naomi and Erik M Conway (2011) *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*: Bloomsbury Publishing USA.

Pacheco Pires, Cesaltina (2002) "A rule for updating ambiguous beliefs," *Theory and Decision*, 53 (2), 137–152.

Paul, Christopher and Miriam Matthews (2016) "The Russian "firehose of falsehood" propaganda model," *Rand Corporation*, 2 (7), 1–10.

Pennington, Nancy and Reid Hastie (1992) "Explaining the evidence: Tests of the Story Model for juror decision making.," *Journal of personality and social psychology*, 62 (2), 189.

Persily, Nathaniel and Charles Stewart, III (2021) "The Miracle and Tragedy of the 2020 US Election," *Journal of Democracy*, 32 (2), 159–178.

Plous, Scott (1991) "Biases in the assimilation of technological breakdowns: Do accidents make us safer?" *Journal of Applied Social Psychology*, 21 (13), 1058–1082.

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

Reich, Taly and Zakary L Tormala (2013) "When contradictions foster persuasion: An attributional perspective," *Journal of Experimental Social Psychology*, 49 (3), 426–439.

Russo, J Edward, Margaret G Meloy, and Victoria Husted Medvec (1998) "Predecisional distortion of product information," *Journal of Marketing Research*, 35 (4), 438–452.

Sances, Michael W and Charles Stewart, III (2015) "Partisanship and confidence in the vote count: Evidence from US national elections since 2000," *Electoral Studies*, 40, 176–188.

Schwartzstein, Joshua and Adi Sunderam (2021) "Using models to persuade," *American Economic Review*, 111 (1), 276–323.

——— (2022) "Shared Models in Networks, Organizations, and Groups."

Seidenfeld, Teddy and Larry Wasserman (1993) "Dilation for sets of probabilities," *The Annals of Statistics*, 21 (3), 1139–1154.

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

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

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

——— (2019) *Narrative economics*: Princeton University Press Princeton.

Shishkin, Denis and Pietro Ortoleva (2021) "Ambiguous information and dilation: An experiment."

Sinclair, Betsy, Steven S Smith, and Patrick D Tucker (2018) ""It's largely a rigged system": voter confidence and the winner effect in 2016," *Political Research Quarterly*, 71 (4), 854–868.

Spiegler, Ran (2016) "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 131 (3), 1243–1290.

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

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

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

A Appendix: Proofs

Proof of Lemma 1. Consider the two statements separately.

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

Consider $\boldsymbol{\mu} \in \mathcal{B}$. 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, for each s and ω, $\pi^{\varphi}(s|\omega) = (\mu_s(\omega)\ \varphi_s)/\mu_0(\omega)$. 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 its fit given s equals to φ_s:

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

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 posterior beliefs that is Bayes-consistent $\boldsymbol{\mu}^m \in \mathcal{B}$.

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

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

Corollary 3 (Binary Signal)**.** *Let $\boldsymbol{\mu}^{\varnothing} = (\mu_0, \mu_0)$. For each $\mu \in \mathcal{B} \backslash \{\boldsymbol{\mu}^{\varnothing}\}$, 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 that $\mu_0(\omega) = \varphi\mu_{s_1}(\omega) + (1-\varphi)\mu_{s_2}(\omega)$, which results in $\varphi = (\mu_0(\omega) - \mu_{s_2}(\omega))/(\mu_{s_1}(\omega) - \mu_{s_2}(\omega)), \forall \omega$. Therefore, $(\varphi_{s_1}, \varphi_{s_2})$ is a distribution over signals if, for every ω, either (i) $\mu_{s_1}(\omega) > \mu_0(\omega) > \mu_{s_2}(\omega)$, or (ii) $\mu_{s_1}(\omega) < \mu_0(\omega) < \mu_{s_2}(\omega)$. These two conditions are equivalent to $\mu \in \mathcal{B} \backslash \{\mu^{\varnothing}\}$ for a binary signal. $\square$

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 $\mu_s(\bar{\omega})\mu_0(\omega) \geq \mu_s(\omega)\mu_0(\bar{\omega})$, for every ω, by definition of maximal movement. □

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

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

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

Proof of Theorem 1. Note that the condition of Theorem 1 can be rewritten as $\sum_s \bar{\delta}(\mu_s)^{\text{-}1} \geq 1$.

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 $\boldsymbol{\mu} \in \mathcal{F}$. 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, te corresponding model exists by Lemma 1. Last, I show that each m_k is adopted conditional on 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 $(\Pr^{m_k}(s))_{s \in S}$ because for each ω

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

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 $\bar{\delta}(\mu_{s_k})^{-1} < 1 - \sum_{i \neq k} \bar{\delta}(\mu_{s_i})^{-1}$ for every k. If it were to exist a set of models inducing $\boldsymbol{\mu}$, each tailored model m_k inducing μ_{s_k} has to be adopted given s_k, that is, $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each $j \neq k$. 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 $\mu_s(\omega) = 0$ for every $\omega \neq \underline{\omega}$. Then,

$\bar{\delta}(\mu_s)^{-1} = \min_\omega \mu_0(\omega)$ for every μ_s. Hence, $\boldsymbol{\mu} \notin \mathcal{F}$ because $\sum_s \bar{\delta}(\mu_s)^{-1} = K \min_\omega \mu_0(\omega) < 1$. □

Claim 2 (Binary Case)**.** *If $\mu_0(\omega_1) \leq 50\%$ and $\boldsymbol{\mu}$ is such that $\mu_s(\omega_1) \leq \mu_0(\omega_1)$ 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, $\bar{\delta}(\mu_s)^{-1} + \bar{\delta}(\mu_{s'}) \geq 1$ and $\boldsymbol{\mu} \in \mathcal{F}$. □

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 $\bar{\delta}(\mu_s^m)^{-1} \geq \Pr^m(s)$ for every s by Lemma 2. Hence, $\mathcal{B} \subseteq \mathcal{F}$.

It is left to show that there exists $\boldsymbol{\mu} \in \mathcal{F}$ such that $\boldsymbol{\mu} \notin \mathcal{B}$. If $\min_\omega \mu_0(\omega) \geq 1/K$, this is trivially true because 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. Add a dummy signal $s_0 \notin S$ to the signal space: $S' = S \cup \{s_0\}$. I want to show that any vector of posterior on the original signal space $\boldsymbol{\mu} \in [\Delta(\Omega)]^S$ can be induced.

Take $\boldsymbol{\mu} \notin \mathcal{F}$, 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. Note that all these are now defined in correspondence to S'.

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 for each ω

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

Each m_k is adopted given s_k since $\Pr^{m_k}(s_k) \geq \Pr^{m_j}(s_k)$ for each j. Thus, $\boldsymbol{\mu}$ is feasible on S. $\square$

Proof of Proposition 6 (Binary Case, Polarization). Take two conflicting models 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:

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

Hence, 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 $\mu_0(\omega_1) > p$. $\square$

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 $(\Pr^{m_k}(s))_{s \in S}$ 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 the posterior μ_{s_k} has to 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 it must be that $\Pr^{m_\ell}(s_\ell) \geq \Pr^d(s_\ell)$, but $\Pr^d(s_\ell) > \bar{\delta}(\mu_{s_\ell})$ by assumption. $\square$

Proof of Proposition 7. 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, for each s, $\bar{\delta}(\mu_s)^{-1} \geq 1 - \sum_{s' \neq s} \bar{\delta}(\mu_{s'})$. Hence, $\boldsymbol{\mu} \in \mathcal{F}$. □

B 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 vector of posteriors within $\mathcal{F}$ for every $M \subseteq \mathcal{M}$.

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 B

First, I formalize these rules and show that the resulting vectors of posteriors are feasible. Then, I provide graphical intuitions for these results for the binary case.

B.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 $m \in M$ such that $\sum_m \alpha^m = 1$. The resulting vector of posteriors $\boldsymbol{\mu}^{\alpha M}$ is Bayes-consistent and, hence, also 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, $\boldsymbol{\mu}^{\alpha M} \in \mathcal{B} \subseteq \mathcal{F}$ 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 and with prior over models $\rho \in \Delta(M)$. The resulting vector of posteriors $\boldsymbol{\mu}^{(\rho,M)}$ is Bayes-consistent and, hence, also feasible.*

Proof of Proposition A2. A Bayesian agent with prior over models ρ has posterior for ω and s

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

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 bias her Bayesian beliefs towards the best-fitting model's prediction.[42] If the bias was maximal, the receiver updates beliefs as in the main text; if the bias is minimal, the receiver is Bayesian as in the previous results.

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 $\beta \in [0,1]$, $\mu_s^{\beta(\rho,M)} = \beta\mu_s^{m_s^*} + (1-\beta)\mu^{(\rho,M)})_s$. 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 as in Proposition A1 where the weights are given by the prior over the model ρ.

[42] 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.

Notice that for every s and ω we have that

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

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 $\bar{\delta}\left(\mu_s^{\beta(\rho,M)}\right)^{-1} \geq \Pr^{\rho M}(s)$ for every s. This condition is verified because $\left(\max\left\{\bar{\delta}\left(\mu_s^{(\rho,M)}\right), \bar{\delta}\left(\mu_s^{m_s^*}\right)\right\}\right)^{-1} \geq \Pr^{\rho M}(s)$ for every s. To see this, consider the two cases separately. First, $\bar{\delta}\left(\mu_s^{(\rho,M)}\right)^{-1} \geq \Pr^{\rho M}(s)$ for every s by Lemma 2 and $\boldsymbol{\mu}^{\rho M} = \boldsymbol{\mu}^{(\rho,M)}$ (see the proof of Proposition A2). Second, for every s

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

where the first inequality follows from Lemma 2 and the second inequality holds by definition of best-fitting model m_s^*, i.e., $\Pr^{m_s^*}(s) \geq \Pr^m(s)$ for every m. □

Last, I consider the case in which 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. This inference rule does not change which model the receiver adopt given each signal, but it only affects inference. Once selected a model 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

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

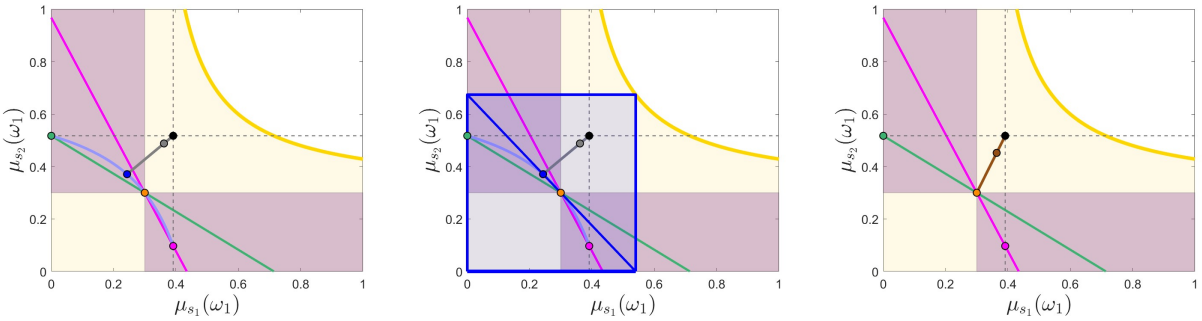
Recall that $m_s^* \in \arg\max_{m\in M} \Pr^m(s)$. Theorem 1 implies that $\boldsymbol{\mu}^M \in \mathcal{F}$ with $\sum_s \bar{\delta}\left(\mu_s^{m_s^*}\right)^{-1} \geq 1$. Since $\bar{\delta}\left(\mu_s^{m_\theta}\right)^{-1} \geq \bar{\delta}(\mu_s^m)^{-1}$, it holds that $\sum_s \bar{\delta}\left(\mu_s^{m_{s,\theta}^*}\right)^{-1}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

B.2 Graphical Intuition

Consider the receiver to be exposed to 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 by 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.

C Appendix: Additional Figures

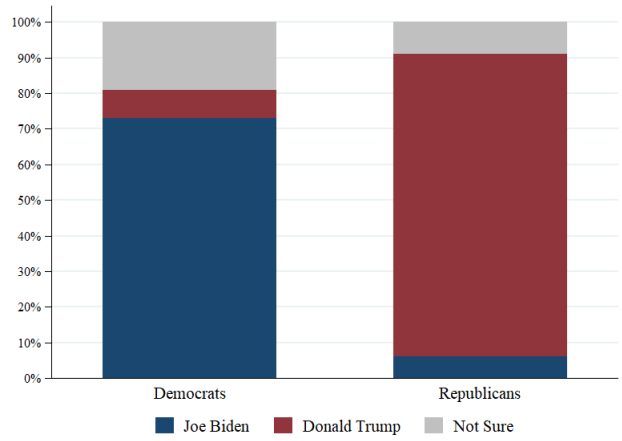


Figure A2: Priors on election winner, by party affiliation

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

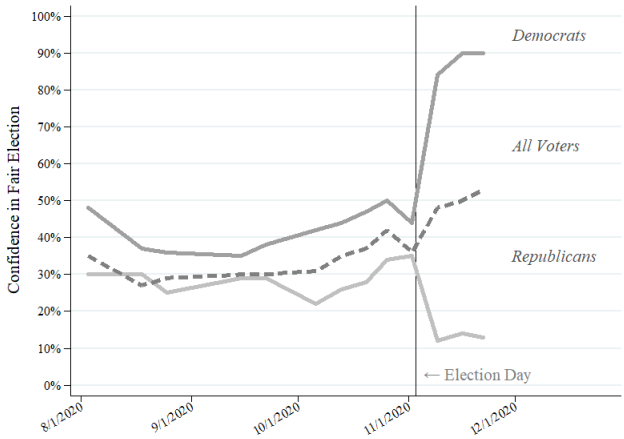


Figure A3: Confidence in fair election (Persily and Stewart, 2021)

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.