Informing agents amidst biased narratives*

Atulya Jain

November 21, 2023

Job Market Paper

Latest Version Available Here

Abstract

I study the strategic interaction between a benevolent sender (who provides data) and a biased narrator (who interprets data) who compete to persuade a boundedly rational receiver (who takes action). The receiver does not know the data-generating model. She chooses between models provided by the sender and the narrator using the maximum likelihood principle, selecting the one that best fits the data given her prior belief. The sender faces a trade-off between providing precise information and minimizing misinterpretation. Surprisingly, full disclosure can be suboptimal and even backfire. I identify a finite set of models that contain the optimal data-generating model, which maximizes the receiver's expected utility. The sender can guarantee non-negative value of information, preventing harm from misinterpretation. I apply this framework to information campaigns and employee feedback.

JEL classification: C72, D82, D83.

Keywords: Information provision, Persuasion, Narratives, Polarization.

*Department of Economics and Decision Sciences, HEC Paris. email: atulya.jain@hec.edu. I am indebted to Tristan Tomala, Nicolas Vieille, Itai Arieli and Frédéric Koessler for their advice and encouragement throughout the project.

1 Introduction

Benevolent experts guide decision-making by providing data and its interpretation. For instance, researchers supply data to assess the effectiveness of policies, and scientists provide evidence of the health risks caused by smoking. Despite this, some people support flawed policies and deny these health effects. This often occurs because people misinterpret data under the influence of biased narratives. For instance, politicians twist data to support their policies, and cigarette companies undermine the evidence against smoking. Thus, persuasion - shaping behavior through information - depends on both data provision and data interpretation. Ignoring the influence of biased narratives when providing data can unintentionally steer people toward poor decisions.[1]

I study the strategic interaction between a benevolent sender, who provides data on the state of the world, and a biased narrator, who interprets this data. Both compete to persuade a boundedly rational receiver who needs to take an action. The state and the receiver's action jointly determine the utility for all agents. The sender chooses a statistical model and generates data from it, à la Kamenica and Gentzkow (2011). The sender's choice is the true data-generating model. After observing the data and the sender's model, the narrator proposes his model (or interpretation) of how the data was generated. The receiver observes both models and the data but does not know which one is the true data-generating model. Different models can lead to varying and even conflicting interpretations of the same data. The receiver selects the model that maximizes the likelihood (or fit) of the data given her prior belief. Finally, she takes an action based on this selected model. The sender wants to maximize the receiver's expected utility, while the narrator maximizes his own. *How should a benevolent sender provide data to a boundedly rational receiver when facing a biased narrator who could misinterpret it?*

To illustrate the key findings of this paper, consider a simplified example, which I will return to throughout the paper. A voter must decide between voting for or against a strict immigration policy. The policy's effect is uncertain and complex. A researcher designs a statistical experiment and gathers data to guide the voter in making an informed decision. However, a politician always wants the voter to support the policy. He can influence the voter's choice by providing a competing interpretation of the data. Suppose that the researcher chooses a very informative experiment. If the data shows a high unemployment rate among immigrants, she recommends voting for the policy. In this case, the researcher and the politician agree, and the voter supports the policy. Conversely, if the data shows a low unemployment rate among immigrants, the researcher strongly advises against the policy. She asserts that immigrants have a positive impact on the economy. However, the politician interprets the same data in a conflicting way, arguing that immigrants are taking jobs from locals. The voter must decide between two interpretations, each advocating for opposite choices. If initially she does not strongly oppose the policy, she finds the researcher's recommendation unconvincing, making her more receptive to the politician. Therefore, the politician

[1] Baysan (2022) demonstrates that an information campaign, intended to address executive power and censorship, inadvertently led to voter polarization during a referendum in Turkey.

can persuade the voter to support the policy, regardless of whether the data supports or opposes it.

How should the researcher strategically design her experiment to minimize misinterpretations? Surprisingly, a partially informative experiment is optimal. Even though the researcher and the politician are perfectly aligned when the policy is effective, the experiment deliberately withholds full disclosure of this state. It occasionally yields low unemployment data when the policy is effective - akin to a Type I error.[2] However, it consistently produces low unemployment data when the policy is ineffective. Under this experiment, when the voter sees low unemployment data, she trusts the researcher's finding more because this data is now more likely to be generated. As a result, the politician cannot sway her with any interpretation, and she votes against the policy on seeing low unemployment data.

In general, how should the sender choose her data-generating model? She should choose a model that balances providing precise data with minimizing misinterpretation. The value of any sender's model can be decomposed into two components: data provision and data (mis)interpretation. Data provision represents the receiver's positive value from acting based on the (true) posterior belief over the prior belief. In contrast, data (mis)interpretation represents the receiver's negative value from acting based on the selected model rather than the sender's model, the true data-generating model. Notably, through interpretations, the narrator can steer the receiver's belief in any direction, even those inconsistent with the sender's model.

The sender's goal is to provide data that leads the receiver to have precise beliefs and take informed actions. As the receiver selects the model based on the fit (or likelihood) of the data, the sender should also consider the fit of her model. A sender's model that fits the data well reduces the narrator's ability to misinterpret. However, models that have a good fit have a limited ability to alter the receiver's beliefs. Thus, the sender faces a trade-off between how likely the data is under her model and its ability to alter the receiver's beliefs. Consequently, given any data, the induced action depends on both the fit and the posterior belief induced by the sender's model.

My main result (Theorem 1) identifies a finite set of models that contains the optimal data-generating model. A key element is the vector of induced actions, with each action being conditional on a data point. The result relies on two key observations: (i) the receiver's expected utility gain is linear with respect to models when the induced actions remain constant, and (ii) the set of models where the induced actions remain constant is given by a finite union of convex sets. Due to the linearity of the expected utility, I can restrict the search for the optimal model within each such set to its extreme points. These sets are determined using the preferences of both the narrator and receiver, along with the receiver's prior. The optimal model guarantees a non-negative value of information, ensuring that the receiver is at least as well off as she would be without any information. This technique is applicable across various contexts, including cases where the sender is not benevolent, the receiver correctly interprets data, and the set of allowed models is restricted.

[2]For example, Centers for Disease Control and Prevention opted to only partially disclose data on vaccination effects to mitigate the risk of misinterpretation. See https://www.nytimes.com/2022/02/20/health/covid-cdc-data.html.

In the absence of the narrator, the sender prefers to fully disclose the state. However, this is no longer true when the narrator is involved. In particular, disclosing no information is optimal when the preferences of the narrator and the receiver are perfectly misaligned (as in a zero-sum game). The narrator misinterprets any information that is provided, leading to suboptimal outcomes. Under what circumstances is full disclosure optimal? For binary states and a narrator whose utility only depends on the action, full disclosure is optimal if, given the state, the narrator induces an action that does not perform worse than the optimal action under the receiver's initial belief. This happens when the narrator does not use a conflicting model to interpret the data. Consequently, the receiver's belief in the true state is higher than her initial belief, even if it is not exact.

I apply this framework to two settings. First, in Section 3.4, I elaborate on the example of the researcher and the politician. I show that the full disclosure model can not only be suboptimal but even *backfire*: it can be worse than providing no information. Notably, a voter initially opposed to the policy can be persuaded to support the policy, irrespective of the data. The politician does this by offering two different models (or interpretations), each designed for specific data. Given supportive data, he proposes a model that aligns with the researcher's theory. Conversely, given opposing data, he proposes a model that raises doubts about the sender's model. The voter's prior determines which interpretation she perceives as more plausible. I use this example to illustrate why the researcher should provide data in a manner that aligns with the voter's prior. (i) If the voter initially strongly disapproves of the policy, the researcher should fully disclose the states. The politician cannot devise any misinterpretation that convinces the voter to always support the policy. (ii) If the voter initially slightly opposes the policy, the researcher should choose a partially informative model. This optimal model has the exact fit as the best model the politician can use to convince the voter to support the policy. Choosing a more informative model would allow the politician to misinterpret opposing data. (iii) Finally, when the voter initially supports the policy, the politician can always persuade the voter to continue supporting it, regardless of which model the researcher chooses.

Second, in Section 4.2, I consider the example of a manager providing feedback to an employee about her ability. The interpretation is clear: positive feedback boosts her confidence in her ability, and negative feedback does the opposite. However, there is uncertainty about the precision of the feedback. First, I assume the employee is an optimist who prefers to believe she has good ability. The employee acts both as the narrator and the receiver. She interprets the feedback and forms a belief about her ability. I consider a dynamic setting of my framework and illustrate how even a tiny amount of uncertainty can lead to biased learning and polarization. Despite repeated feedback, the employee incorrectly concludes that her ability is good even when it is not. She distorts her own beliefs by perceiving positive feedback to be more informative than negative feedback (Eil and Rao, 2011). Next, I demonstrate how two employees, an optimist and a pessimist, who start with the same initial beliefs and receive identical feedback, can become polarized. Regardless of their actual ability, the optimist perceives she has good ability, while the pessimist thinks the opposite. I show that the best way to provide feedback to an optimist is to provide negative feedback more frequently than positive feedback. This counters her asymmetric interpretation and ensures that she learns her actual ability.

Finally, I explore three extensions to my setting. First, I examine a natural setting where data has a straightforward interpretation, such as a bad grade resulting in reduced confidence in one's ability. While the narrator cannot alter this direction of belief change, he can alter beliefs by varying the precision. This limits the narrator's persuasive powers, but as seen in the employee feedback application, this can lead to skewed perceptions and biased learning. Second, I assume the narrator can propose multiple models (or interpretations) before observing the data. This gives his models more credibility but also introduces additional constraints. His models compete not only with the sender's model but also among themselves. Surprisingly, the timing of interpretation (before versus after) does not impact the narrator's persuasiveness. Third, I explore a scenario where the receiver treats the models of the sender and the narrator asymmetrically. If the trust in the sender's model decreases, the narrator's ability to persuade increases. If trust in the sender's model decreases, it can lead to unfavorable outcomes, possibly resulting in negative value of information for all data-generating models. In such cases, the receiver would be better off if no data was provided.

Literature review

Bayesian Persuasion: My work contributes to the literature on Bayesian persuasion (or Information design). This literature examines how a sender can influence the behavior of a rational receiver by generating data. Crucially, I assume that the receiver is unaware of the data-generating model. She does not know how to interpret the data. When provided with multiple models (or interpretations), she chooses the one that best fits the data given her prior belief. This is in stark contrast to the seminal paper of Kamenica and Gentzkow (2011) and further generalizations such as Alonso and Câmara (2016), Renault, Solan, and Vieille (2017), and Ball and Espín-Sánchez (2021). An exception is de Clippel and Zhang (2022) which considers non-Bayesian receivers. Despite this, the sender's problem can still be addressed using the standard concavification technique.

My contribution lies in developing a technique to identify the optimal model for both non-Bayesian and Bayesian receivers within a general set of allowed models. In my setting, the receiver's action, in equilibrium, depends not only on the posterior belief but also on the likelihood of the sender's model. I define a preference over the space of models as the concavification technique (Aumann, Maschler, and Stearns, 1995; Kamenica and Gentzkow, 2011) cannot be applied.[3] This technique can even be applied to settings outside my framework, where the sender is not benevolent and/or the receiver correctly interprets data. In a related paper, Ball and Espín-Sánchez (2021) also define preferences over models due to a restricted set of allowed models. However, they analyze a stylized binary model with rational receivers.

[3]Even when applicable, finding the concave envelope of a function can be difficult (see Tardella, 2004; Lipnowski and Mathevet, 2017).

The closest paper is Ichihashi and Meng (2021), which assumes that the same agent both generates data and interprets it in a stylized binary setup. On the contrary, I consider two agents: one who generates data and the other who interprets it. My focus is on their strategic interaction, given their preference misalignment. Another related paper is Eliaz, Spiegler, and Thysen (2021), where the sender, apart from providing data, also strategically provides an accurate but coarse interpretation.

Narratives: My work contributes to the literature on narratives in economics. There has been growing interest in understanding the role of narratives in shaping behavior (Shiller, 2017). This literature examines how individuals use subjective models to interpret and make sense of data. My focus is on narratives modeled as Blackwell experiments (or likelihood functions) (Schwartzstein and Sunderam, 2021; Aina, 2021; Yang, 2023; Ispano et al., 2022; Izzo, Martin, and Callander, 2023).

Schwartzstein and Sunderam (2021) formalize that the receiver prefers models that best fit the data given her prior belief. Aina (2021) builds on this framework, analyzing a setting in which the persuader commits to a menu of models before the data is observed. Yang (2023) assumes that the receiver prefers decisive models, which induce low regret. The literature assumes that the data-generating model is fixed and exogenous. My contribution is to consider a strategic and endogenous data-generating model.

There have been other approaches to formalize narratives such as directed acyclical graphs (DAGs) (Eliaz and Spiegler, 2020; Eliaz, Galperti, and Spiegler, 2022) and moral reasoning (Bénabou, Falk, and Tirole (2018)). Also, recent papers experimentally investigate the role of narratives in persuasion (Barron and Fries, 2023; Kendall and Charles, 2022). In particular, Barron and Fries (2023) provide evidence that individuals pick models with better fit.

Biased updating: My work also relates to the literature on biased updating (see Benjamin (2019) for a survey). The main contribution is to analyze the effect of information on the welfare of the receiver who uses biased updating. The framework allows explaining both prior-based and preference-biased updating. A crucial aspect is that the receiver can use different models to update given different data. This allows for reconciling behavioral biases inconsistent with updating using a single model.

Some papers analyze the receiver's welfare given a fixed data-generating model for biased updating under *all* decision problems. Braghieri (2023) provides a characterization for when the value of information is non-negative, while Frick, Iijima, and Ishii (2021) compare the welfare of the receiver for different biases. In contrast, I focus on finding the optimal data-generating model for a fixed decision problem.

Some papers in this literature investigate learning under model uncertainty. Chen (2022) assumes the receiver interprets data in a self-serving manner, while Fryer Jr, Harms, and Jackson

(2019) assume the receiver interprets data in a manner that aligns with her current beliefs. They can explain phenomena like self-serving bias, confirmation bias, and polarization. I focus on finding a data-generating model that leads to correct learning under biased updating.

Structure of the paper

Section 2 introduces the setup. Section 3 provides the main result and applies it to the setting of information campaigns. Section 4 provides three extensions: models with clear interpretations, timing of interpretation, and asymmetric trust. It also includes an application on employee feedback. Finally, I conclude in Section 5. All proofs are in the Appendix.

2 Setup

Consider a game of incomplete information between three players: the sender (S, she), the narrator (N, he), and the receiver (R, she). The sender chooses a model to generate a signal $s \in \mathcal{S}$ about an unknown state of the world $\omega \in \Omega$.[4] The narrator also chooses a model, but to interpret this signal. He posits his own model of how this signal was generated. Finally, the receiver takes an action $a \in A$ based on this signal and the two models. I assume that the set of states Ω, the set of actions A, and the set of signals $\mathcal{S}$ are finite, with $|\mathcal{S}| \geq |\Omega|$. All players share a common prior belief over the states $p \in \text{int}(\Delta\Omega)$.[5][6] For each player $i \in \{S, N, R\}$, the utility function $u_i(\omega, a)$ depends on the state of the world $\omega \in \Omega$ and the receiver's action $a \in A$.

A **model** $m : \Omega \to \Delta\mathcal{S}$ is a stochastic map that specifies the probability $m(s \mid \omega)$ of observing signal $s \in \mathcal{S}$ conditioned on state $\omega \in \Omega$.[7] Given a signal s, a model m induces posterior belief $q_s^m \in \Delta\Omega$, which is derived using Bayes' rule.[8] Let $\mathcal{M}$ denote the set of all models. Following Aina (2021), I define the **fit** of model m given signal s as the (ex-ante) likelihood:

$$\mathbb{P}_m(s) = \sum_{\omega \in \Omega} p(\omega) m(s \mid \omega). \tag{1}$$

[4]A signal can be empirical data, evidence, or even a message.

[5]$\text{int}(S)$ denotes the interior of the set S, and ΔS represents the set of all probability distributions over the set S.

[6]The common prior assumption is made for simplicity. The game can be generalized to heterogeneous priors.

[7]The term "model" is also referred to as Blackwell experiment, likelihood function, information structure, and information policy in the literature.

[8]The posterior belief $q_s^m \in \Delta\Omega$ is given by:

$$q_s^m(\omega) = \frac{p(\omega) m(s \mid \omega)}{\sum_{\omega \in \Omega} p(\omega) m(s \mid \omega)}$$

whenever Bayes' rule is applicable.

Let $\mathcal{F} \subseteq \mathcal{M}$ denote the set of feasible (or allowed) models, which is assumed to be closed and convex. Unless stated otherwise, every model is feasible ($\mathcal{F} = \mathcal{M}$). The assumptions imply that the sender can fully disclose the state if she wants.

Timing of the game:

1. Sender chooses the signal-generating model $I : \Omega \to \Delta\mathcal{S}$.

2. Nature draws the state $\omega \sim p(\cdot)$ according to the prior belief and the signal $s \sim I(\cdot \mid \omega)$ according to the sender's model. The signal s is publicly observed.

3. After observing the sender's model I and signal s, the narrator selects a model $n_s : \Omega \to \Delta\mathcal{S}$ to propose a competing interpretation of how the signal was generated.

4. Upon observing signal s, the receiver is presented with the sender's model I and the narrator's model n_s. She does not know the true signal-generating model and selects the model $m_s \in \{I, n_s\}$, which has the best fit given signal s:[9]

$$m_s := \arg\max_{m \in \{I, n_s\}} \mathbb{P}_m(s). \tag{2}$$

5. The receiver forms her posterior belief using the selected model m_s and takes action

$$a_R^*(q_s^{m_s}) := \arg\max_{a \in A} \mathbb{E}_{q_s^{m_s}}[u_R(\omega, a)] \tag{3}$$

   where, $a_R^*(q)$ denotes the receiver's optimal action given belief q. It maximizes the receiver's expected utility given her belief over the states.[10]

Given signal s, the objective (or true) posterior belief and fit are derived using the senders model, since it is the signal-generating model. In contrast, the receiver's action $a_R^*(q_s^{m_s})$ represents an equilibrium outcome. She acts as if the signal is generated according to the selected model m_s. This model, in turn, is determined by the choices made by both the sender and the narrator. The expected utility of each player i, where $i \in \{S, N, R\}$, is given by:

[9] When both models have the same fit, I assume that the receiver chooses the sender's model.

[10] In case of multiple optimal actions, I break the tie by choosing the action the narrator prefers. If there are multiple such actions, I choose an action arbitrarily.

$$\sum_{s\in\mathcal{S}} \mathbb{P}_I(s)\mathbb{E}_{q_s^I}[u_i(\omega, a_R^*(q_s^{m_s}))]. \tag{4}$$

I investigate the behavior of a biased narrator who chooses models (or interpretations) to maximize his expected utility. In contrast, the sender is benevolent ($u_S = u_R$) and chooses the signal-generating model to maximize the receiver's expected utility.

Discussion of Assumptions: First, I focus on the receiver. Crucially, I assume that she does not know the signal-generating model.[11] However, once she chooses a model, she updates her belief in a rational manner using that model and the Bayes' rule. Following Schwartzstein and Sunderam (2021), the receiver selects the model via the maximum likelihood principle. This principle is a popular way to select between parameters in statistics and economics.[12] Given a set of models, the receiver selects the model which has the best fit (or likelihood) given the observed signal and her prior. Furthermore, Barron and Fries (2023) provides experimental evidence indicating that individuals prefer models with a better fit. This assumption is also in line with the interdisciplinary work on narratives and sense-making (Fisher, 1985; Weick, 1995; Chater and Loewenstein, 2016). The receiver may deem some models infeasible but she cannot come up with her own model. Her choice is confined to the exposed models.[13] Also, unlike the sender and the narrator, the receiver is non-strategic: she does not take into account the incentives of the sender and narrator or rather deems them both equally credible. In Section 4.4, I relax this condition and allow the receiver to trust the sender and the narrator asymmetrically.

Next, I assume the narrator cannot influence the signal itself or provide an additional signal. He can only provide an interpretation of the observed signal.[14] Also, he provides his model after the signal realization, whereas the sender chooses his model before the realization. I show, in Section 4.3, that the results do not depend on whether the narrator provides his interpretations before or after observing the signal. The only caveat is that, in the ex-ante timing, the narrator provides a menu of models instead of a single one.

Finally, I assume that the sender can only interpret using the signal-generating model.[15] I do not allow her to generate the signal using one model and to interpret the signal using a different

[11] Two other popular choices for dealing with model uncertainty are the fully Bayesian approach and the maxmin approach.

[12] For example, the principle is used to select the prior under ambiguity Gilboa and Schmeidler (1993). Levy and Razin (2021) use this principle to combine forecasts. Also, recently Frick, Iijima, and Ishii (2023) show that it is a maximally efficient updating rule in learning under ambiguity aversion.

[13] This assumption is natural in many settings. Interpreting data may require expertise (finance, medicine) or leadership (politics, see: Bullock, 2011; Izzo, Martin, and Callander, 2023).

[14] For instance, a stock analyst cannot change stock prices or create new price data - he can only interpret existing price trends to guide investors.

[15] For instance, researchers have to preregister their experiment and cannot misinterpret how they collect data.

one. If she could, the receiver's expected utility would be (weakly) higher (Ichihashi and Meng, 2021). My goal is to assess and compare the influence of providing a signal and interpreting this signal to persuade a decision-maker.

3 Main Results

In this section, I state my main results. The equilibrium of the sequential game is determined by backward induction. First, I characterize the extent of persuasion by the narrator. I illustrate it using a graphical illustration for the binary case. Next, I solve the sender's problem and find the optimal signal-generating model. I do this by defining a value function over the set of feasible models. Finally, I apply my results to the setting of information campaigns.

3.1 Scope of persuasion by interpretations

In this subsection, I characterize the sets of feasible posterior beliefs and actions that the narrator can induce under a given sender's model.

There are constraints on beliefs and actions that the narrator can induce. Given sender's model I and signal s, denote B_s^I and A_s^I as the sets of feasible posterior beliefs and actions that the narrator can induce. When accounting for all signals, denote $\boldsymbol{B}^{\boldsymbol{I}} = (B_s^I)_{s\in\mathcal{S}}$, and $\boldsymbol{A}^{\boldsymbol{I}} = (A_s^I)_{s\in\mathcal{S}}$ as the sets of feasible vectors of posterior beliefs and actions. The narrator's model is selected only if it has a better fit on the signal than the sender's model. The set of feasible posterior beliefs and actions depends only on the fit of the sender's model and the prior belief.

Lemma 1. *Given the sender's model I and signal s, the sets of feasible posterior beliefs and actions that the narrator can induce are:*

$$B_s^I := \{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \forall \omega \in \Omega\} \cup \{q_s^I\}, \tag{5}$$

$$A_s^I := \{a \in A : \exists q \in B_s^I \text{ such that } a = a_R^*(q)\}. \tag{6}$$

The narrator can always induce the true belief consistent with the sender's model. For any other belief q, the key argument is that there exists a model with maximal fit, which is given by $[\max_{\omega\in\Omega} \frac{q(\omega)}{p(\omega)}]^{-1}$. The narrator cannot propose a model that induces belief q and that has a better fit than this. Thus, the narrator can induce a belief if and only if it's maximal fit surpasses that of the sender's model; otherwise, he cannot. The set of feasible beliefs is always convex.[16] An action is feasible if it corresponds to the receiver's optimal action under some feasible belief. Proposition

[16]The belief q_s^I either lies in the interior or is an extreme point of the set $\{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \forall \omega \in \Omega\}$.

1 of Schwartzstein and Sunderam (2021) can be applied to my setting to characterize the feasible posterior beliefs.

The sender's model acts as a constraint to the narrator's ability to persuade. The better the sender's model fits the signal, the less flexibility the narrator has in shifting beliefs. The narrator can only persuade the receiver to have beliefs not too far from her prior. Crucially, using interpretations, the narrator can manipulate the receivers belief in any direction, even those that are inconsistent with the signal-generating model. By proposing distinct models for different signals, he can consistently shift beliefs in the same direction for those signals. This is impossible if the receiver uses a single model, even if it is incorrect, to interpret all signals.

3.2 Binary Example: Graphical illustration

In this subsection, I graphically illustrate the narrator's extent of persuasion. I use the example of a researcher (sender), a politician (narrator), and a voter (receiver).

Consider two states: $\Omega = \{G, B\}$, where G and B are the states where the policy is good and bad, respectively. The researcher provides evidence $\mathcal{S} = \{g, b\}$, where g indicates evidence that supports the policy, and b is evidence that opposes it. The voter chooses from $A = \{a_-, a_+\}$, where a_- is to vote against the policy and a_+ is to vote for it. The utility function of the politician and the voter are given by the following matrix:

		Actions	
		a_-	a_+
States	G	$(0,1)$	$(1,2)$
	B	$(0,1)$	$(1,0)$

Table 1: Matrix of utility functions for the politician and the voter, respectively.

The voter only votes for the policy if she believes it's likely to be good, that is, $q(G) \geq \frac{1}{2}$. Otherwise, she prefers to vote against. The politician, regardless of the state, always wants her to vote for the policy.

The researcher and the politician persuade the voter's choice of action. This action depends on her belief over the states. This belief, in turn, depends on the chosen models. So, focusing on the vector of posteriors rather than the model that induces it provides useful insights.

First, I focus on the beliefs the researcher can induce without the influence of the politician. Given the binary states, let the probability of state G identify the beliefs in the example. The graph's axes represent the posterior belief on state G given evidence b and g (see Fig. 1). Each point in this graph is a vector of posterior beliefs. I represent the prior $p = \mathbb{P}(G) = 0.4$ as the vector of posterior beliefs, where each posterior equals the prior (black point). A vector of posteriors $(q_b(G), q_g(G))$ is *Bayes plausible* if and only if either: (i) $q_b(G) \geq p(G) \geq q_g(G)$ or (ii) $q_b(G) \leq p(G) \leq q_g(G)$. This condition ensures that there is a model such that the expected posterior equals the prior. The

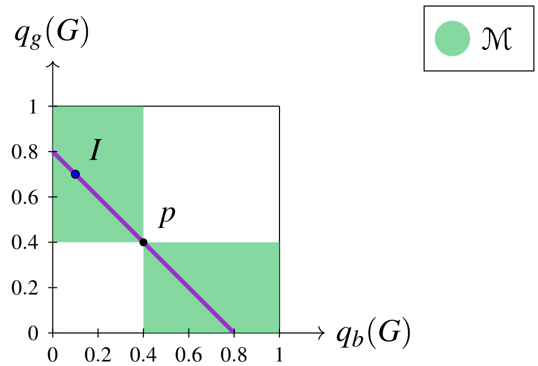


Figure 1: The set of all Bayes-plausible vector of beliefs.

set of all models $\mathcal{M}$ cover all the vectors of posterior beliefs that are Bayes plausible (green area). The researcher can induce any such vector of posteriors. This condition prevents the voter from updating beliefs in the same direction in response to both opposing and supportive evidence. In the binary case, generically, there is a one-to-one mapping between models and the Bayes plausible vector of beliefs.[17]

Suppose that the researcher chooses the model I (blue point): supportive evidence is likely generated under the good state, and vice versa for opposing evidence. Formally, $I(g \mid G) = I(b \mid B) = \frac{7}{8}$. Consider the purple line that passes through the model I (blue point) and the prior p (black point). All models on this line have the exact fit as model I on both evidences.[18] The steeper this line, the better fit the model has on evidence b ; conversely, the flatter this line, the better fit the model has on evidence g. The set of all models $\mathcal{M}$ can be divided into three subsets based on this line (see Fig. 2): (i) models that have the same fit on both evidences (purple line), (ii) models with a better fit on evidence b (blue dotted area) and (ii) models with a better fit on evidence g (red area).

What can the politician do to persuade the voter? Given evidence b, the politician can choose any model n_b in the blue dotted area and the voter will select it over the researcher's model I. Given evidence b, the politician can induce any belief on the blue line on the x-axis, which is the projection of the blue dotted area. This also includes beliefs higher than the prior, contrary to the intention of the researcher's model. Conversely, given evidence g, the politician can choose any model n_b in the red area and the voter will select it over the researcher's model I. Given evidence g, the politician can induce any belief on the red line on the y-axis, which is the projection of the

[17]The only exceptions are models that provide no information; they induce, for at least one signal, a posterior belief identical to the prior belief.

[18]For the binary case, any model m that induces the vector of beliefs $(q_b^m(G), q_g^m(G)) \in \mathcal{M}$ has the same fit as model I given evidence b (and g) if it satisfies the following condition :

$$\mathbb{P}_I(b) q_b^m(G) + (1 - \mathbb{P}_I(b)) q_g^m(G) = p.$$

This condition represents the line passing through the points I and p in Fig. 2. It is worth noting that the vector of prior beliefs always satisfies this condition.

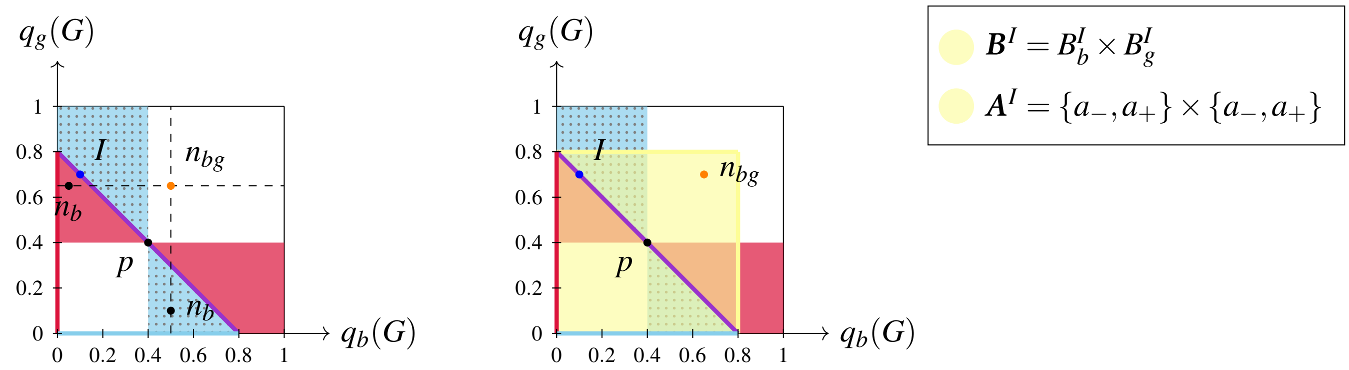


Figure 2: The posterior beliefs and actions that the politician can induce.

red area. The set of beliefs satisfying equation (5) in Lemma 1 precisely corresponds to the blue and red lines for evidence b and g, respectively

From an ex-ante perspective, the politician can induce any vector of posterior beliefs within the yellow area, which is defined as the Cartesian product of the red and blue lines. This set depends only on the fit of the researcher's model. Since all models on the purple line have the same fit, they yield identical feasible vectors of posteriors within the yellow area. Crucially, the politician can also induce vectors of posterior beliefs that are not Bayes plausible, i.e. outside the green area in Fig. 1. Indeed, under model I, the politician can persuade the voter to take any action, regardless of the evidence provided. For example, by proposing models n_b and n_b given evidence b and g, he can induce the vector of beliefs n_{bg} (depicted as the orange point in Fig. 2). The voter is more convinced that the policy is good, given both supportive and opposing evidence, compared to her prior belief. She votes for the policy with certainty. In the next section, I demonstrate how the researcher should generate evidence to prevent misinterpretation by the politician.

3.3 Optimal signal-generating model

In this section, I turn to the sender's problem. I identify a finite set of models that contain the optimal signal-generating model. I do this by defining a value function over the set of feasible models.

Let $\boldsymbol{a}^I = (a^I_s)_{s\in\mathcal{S}} \in A^{|\mathcal{S}|}$ denote the *vector of induced actions* when sender chooses the model I. If the sender chooses model I and signal s is generated, the narrator selects an action from the set of feasible actions A^I_s to maximize his expected utility.[19] I have

$$a^I_s := \arg\max_{a\in A^I_s} \mathbb{E}_{q^I_s}[u_N(\omega, a)]. \tag{7}$$

[19]If the narrator has multiple optimal actions, I break the tie by choosing the receiver's preferred action. As a result, the action a^I_s is unique.

The induced action a_s^I is the narrator's best response to the sender's model I and signal s. Define the **value function** $V : \mathcal{F} \to \mathbb{R}$ over the set of feasible models as:

$$V(I) := \sum_{s\in\mathcal{S}} \mathbb{P}_I(s)\mathbb{E}_{q_s^I}[u_R(\omega, a_s^I)] - \mathbb{E}_p[u_R(\omega, a_R^*(p))]. \tag{8}$$

The value function is given by the receiver's expected utility gain given the sender's model I. It determines the sender's preference over the feasible models taking into account the narrator's ability to misinterpret and his preferences. The goal of the sender is to find the model I^* that maximizes the value function among the set of feasible models. I call I^* the *optimal signal-generating model*. The value of any model can be decomposed into two components: information provision and information misinterpretation.

$$V(I) = \sum_{s\in\mathcal{S}} \mathbb{P}_I(s)\big(\underbrace{\mathbb{E}_{q_s^I}[u_R(\omega, a_R^*(q_s^I) - u_R(\omega, a_R^*(p))]}_{\text{information provision } \geq 0} + \underbrace{\mathbb{E}_{q_s^I}[u_R(\omega, a_s^I) - u_R(\omega, a_R^*(q_s^I))]}_{\text{information misinterpretation } \leq 0}\big). \tag{9}$$

The information provision component represents the receiver's value from using the true posterior belief over the prior belief. It is the focal point in the Bayesian persuasion literature and is always non-negative. The information misinterpretation component reflects the receiver's value when she acts based on the chosen model rather than the sender's model. This value is always non-positive. It becomes strictly negative when the receiver deviates from the action recommended by the sender's model.

When choosing a model, the sender has to take into account the trade-off between providing information and minimizing misinterpretation. To simplify the search for the optimal signal-generating model, I partition the set of all feasible models into a disjoint union of convex subsets. The vector of induced action remains fixed within each such set. Let $C_{\boldsymbol{a}} \subseteq \mathcal{F}$ denote the set of sender's models where the vector of induced actions is $\boldsymbol{a} \in A^{|\mathcal{S}|}$:

$$C_{\boldsymbol{a}} := \{I \in \mathcal{F} : \boldsymbol{a}^I = \boldsymbol{a}\}. \tag{10}$$

The collection $\mathcal{C} = \{C_{\boldsymbol{a}}\}_{\boldsymbol{a}\in A^{|\mathcal{S}|}}$, over all vectors of actions, is a finite cover of the set of feasible models $\mathcal{F}$.

Lemma 2. *The set $C_{\boldsymbol{a}}$ is a finite disjoint union of convex sets for any vector of actions $\boldsymbol{a} \in A^{|\mathcal{S}|}$.*

This follows as any set $C_{\boldsymbol{a}}$ can be written as a finite disjoint union of the intersection of finitely many half spaces. Let $\overline{C}$ and $Ext(C)$ denote the closure and the set of extreme points for any convex set C.[20] Due to the linearity of the value function when the vector vector of actions remains fixed, I can restrict the search of the optimal model within each set $\overline{C}_{\boldsymbol{a}}$ to its extreme points.[21] This technique simplifies the sender's optimization into a finite linear program.

Theorem 1. *The optimal signal-generating model*

$$I^* := \arg\max_{I \in \mathcal{F}} V(I) \tag{11}$$

corresponds to an extreme point of the set $\overline{C}_{\boldsymbol{a}}$ for some $\boldsymbol{a} \in A^{|\mathcal{S}|}$. Furthermore, $Ext(\overline{C}_{\boldsymbol{a}})$ is finite for all $\boldsymbol{a} \in A^{|\mathcal{S}|}$.

By virtue of the theorem, one can pinpoint finite candidate models in the search for the optimal one. This vastly simplifies the sender's optimization problem because the space of all models is very large (see green area in Fig. 1 for the binary case). The set of candidate models is obtained by taking the union of the set of extreme points $Ext(\overline{C}_{\boldsymbol{a}})$ over all possible vector of actions $\boldsymbol{a} \in A^{|S|}$. Each set $C_{\boldsymbol{a}}$ is determined by the preferences of the narrator and the receiver, in addition to the common prior. This technique for identifying the optimal signal generating model is applicable in various settings. For example, it can be used even when the sender is not benevolent, when there are restrictions on the set of allowed models, or when the receiver always correctly interprets the signal.

Consider any candidate model that does not fully disclose the states. This model either (i) results in a posterior belief where the narrator or the receiver is indifferent between multiple actions and/or (ii) matches the fit of another model, where the narrator can induce a different vector of actions. If such a candidate model is optimal, opting for a more informative model results in more misinterpretation. It changes the induced vector of actions, making it worse for the receiver. If not for this, the sender would want to give more information, and this model would not be the best choice.

A model that is always a candidate model is the no disclosure model I_{ND_s}, defined for any $s \in \mathcal{S}$. This model unambiguously sends the signal s, that is, $I_{ND_s}(s \mid \omega) = 1$ for all $\omega \in \Omega$. When the preferences of the narrator and receiver are perfectly misaligned, akin to a zero-sum game, the no disclosure model is optimal. Furthermore, in this setting, any optimal model uniquely induces the receiver's optimal action under her prior.

Proposition 1. *If $u_N = -u_R$, then for any $s \in \mathcal{S}$, the no disclosure model I_{ND_s} is optimal. Additionally, any optimal model induces the unique action $a^*_R(p)$.*

As the no disclosure model I_{ND_s} sends the signal s with probability 1, it has the maximal fit among the set of all models for the signal s. It results in the posterior being identical to the prior

[20] The set $C_{\boldsymbol{a}}$ can be an open set as I break the tie between models with equal fit in favor of the receiver.

[21] Lipnowski and Mathevet (2017) use a similar property to identify candidate beliefs rather than candidate models.

when signal s is observed, that is, $q_s^{I_{ND_s}} = p$. Thus, this model provides no information and subsequently there is no scope for misinterpretation, that is, $V(I_{ND_s}) = 0$. In case of perfectly misaligned preferences, any information provided by the sender will be misinterpreted by the narrator. Crucially, this model ensures that the value of information under the optimal model is non-negative, that is, $V(I^*) \geq 0$.

The full disclosure model I_{FD} is always a candidate model. If there was no narrator, it would be the benevolent sender's optimal choice. The more information the receiver has, the better action she can take. As seen in the example before, this is no longer true in the presence of a biased narrator. However, is it ever optimal for the sender to fully disclose the states? If so, then when? A narrator has state-independent utility $u_N(a)$ if his utility depends only on the receiver's action and not the state.[22] For example, politicians want to get elected, investors want to sell high-fee products, lobbyists want favorable policies. As $|\mathcal{S}| \geq |\Omega|$, I can assume that the set of signals contains a copy of the set of states, that is, $\Omega \subseteq \mathcal{S}$. Formally, the model I_{FD} fully discloses (or reveals) the states, that is, $I_{FD}(\omega \mid \omega) = 1$ for all $\omega \in \Omega$.

Proposition 2. *For binary states and a narrator with state-independent utility, the full disclosure model I_{FD} is optimal if $u_R(\omega, a_\omega^{I_{FD}}) \geq u_R(\omega, a_R^*(p))$ for all $\omega \in \Omega$.*

The proposition states that full disclosure is optimal if the narrator either cannot or does not want to induce an action worse than the optimal action at the prior belief. This means that the narrator does not use a conflicting model to interpret the signal. This ensures that the receiver's posterior on the disclosed state is higher than the prior. However, the subsequent corollary demonstrates that when the narrator has state-independent utility, the full disclosure model cannot be universally optimal for all prior beliefs. For the next result, to avoid generic situations, I assume that there are at least two actions, which are uniquely optimal for the receiver under some beliefs and that the narrator is not indifferent between them.

Corollary 2. *Given any narrator with state-independent utility, there exists a prior belief $p \in int(\Delta\Omega)$ such that the full disclosure model I_{FD} is not optimal.*

To see why, suppose that the receiver has a prior belief, where the narrator's most preferred action is not optimal but it is very close to being optimal. If the sender fully discloses the state, the narrator can induce his most preferred action, with probability 1, due to its proximity to the prior belief. However, since this action is not the receiver's optimal choice given her prior belief, it is better for the sender to provide no information.

3.4 Application: Information campaigns

Experts use information campaigns to influence health behavior, policy attitudes, and voter turnout (Haaland, Roth, and Wohlfart, 2023). However, there are cases where a campaign not only fails

[22]This is an often studied case in the literature, see Lipnowski and Ravid (2020).

to provide information but may also increase misperceptions.[23] In particular, I demonstrate that information can backfire; specifically, the receiver might shift her belief in the opposite direction to that intended. The *backfire effect* has been observed empirically in information campaigns (Nyhan and Reifler, 2010; Hart and Nisbet, 2012; Baekgaard et al., 2019; Baysan, 2022). In particular, Baekgaard et al. (2019) provide experimental evidence that (i) the farther the prior belief is from the target belief, the greater the chance of misinterpretation and that (ii) more information can paradoxically lead to a higher chance of misinterpretation.

For the example of the researcher and the politician, I characterize the optimal signal-generating model. First, the set of all models can be partitioned based on the vector of induced actions (Fig. 3). Using Theorem 1, I limit my search to the finite set of extreme points for each set $C_{\boldsymbol{a}}$ (black nodes). My focus is exclusively on the models in the top-left quadrant, as the models in the bottom-right quadrant are obtained simply by swapping the labels g and b. In total, there are only six candidate models, including the full disclosure and no disclosure models.

If there was no politician, the researcher would choose the full disclosure model, that is, $I_{FD}(g \mid G) = I_{FD}(b \mid B) = 1$. However, when evidence can be misinterpreted, it is suboptimal to do this. Given opposing evidence b, the politician can choose the model n_b (blue point) where $n_b(b \mid G) = 1$ and $n_b(b \mid B) = \frac{2}{3}$. This model has better fit than the full disclosure model for opposing evidence: $\mathbb{P}_{I_{FD}}(b) = \frac{3}{5} < \frac{4}{5} = \mathbb{P}_{n_b}(b)$. This can also be seen graphically as the purple line passing through the points n_b and p is steeper than the dashed line passing through the points I_{FD} and p. Given evidence b, the politician is able to persuade the voter to support the policy, as the model n_b induces a posterior belief equal to 0.5. Infact, the full disclosure model performs worse than the no disclosure model. The voter initially prefers to vote against the policy. However, after the full disclosure model, the politician can convince her to vote for the policy with certainty. Therefore, it is better for the researcher to provide no information.

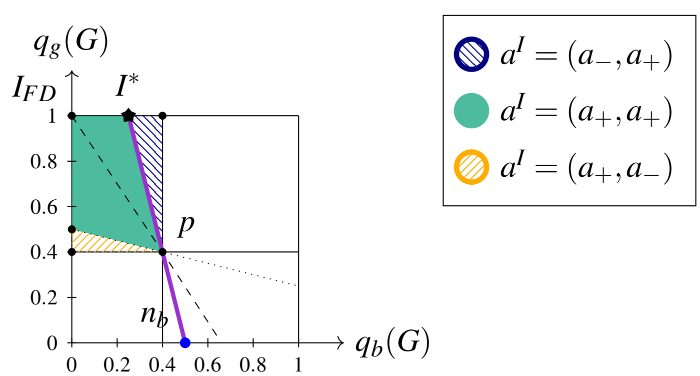


Figure 3: Partition of the set of models based on the vector of induced actions.

I have shown that fully disclosing the states is not optimal. How should the researcher provide evidence? The researcher chooses a partially informative model that better fits the opposing

[23] It is important to distinguish between misinformation (false information) and misperceptions (false beliefs). My focus is on the cases where correct information can lead to false beliefs due to misinterpretation.

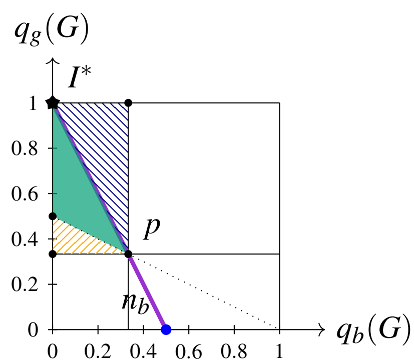
evidence. The (unique) optimal signal-generating model I^* (black star) is given by:

$$I^*(b \mid B) = 1, \qquad I^*(b \mid G) = \frac{1}{2},$$
$$I^*(g \mid B) = 0, \qquad I^*(g \mid G) = \frac{1}{2}.$$

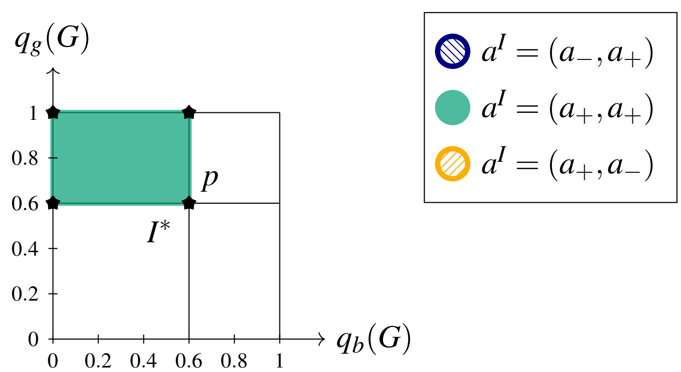
The model n_b has the maximal fit among all models that induce belief $q_b(G) = 0.5$. The fit of the optimal model I^* and the model n_b exactly match: $\mathbb{P}_{n_b}(b) = \frac{4}{5} = \mathbb{P}_{I^*}(b)$. This can also be inferred graphically, as both models lie on the same purple line passing through the prior belief. The politician cannot propose any model that has a better fit on the opposing evidence than the optimal model and that persuades the voter to support the policy. If the researcher chose a more informative model, the politician can misinterpret the opposing evidence. All vectors of beliefs on the line segment between I^* and I_{FD} represent models more informative than I^*. On the other hand, if she chose a less informative model, at worst, it lowers the likelihood of correctly matching the state and the action. Thus, it is the unique optimal model.

Surprisingly, even though the politician and the researcher are perfectly aligned when the policy is good, the optimal signal-generating model does not disclose state G. Instead, it pools state G with state B where their preferences are misaligned, to make the opposing evidence more plausible. This pooling is to prevent the voter from misinterpreting the opposing evidence.

The prior belief of the voter plays an important role in determining the partition and optimal model (see Fig. 4). The optimal model is of three types based on the prior : (a) full disclosure for low prior, (b) partially informative for mid prior, and (c) no disclosure for high prior.



(a) Low prior ($p = \frac{1}{3}$)



(b) High prior ($p = 0.6$)

Figure 4: The partition and the optimal model for different prior beliefs.

For low prior , ($p \leq \frac{1}{3}$), the full disclosure model I_{FD} is optimal. Given opposing evidence b, any model of the politician that better fits the evidence cannot convince her to vote for the policy. The threshold $p = \frac{1}{3}$ is precisely the prior for which the full disclosure model I_{FD} lies on the line passing through the prior p and the model n_b (purple line in Fig. 4a).

For a high prior ($p \geq 0.5$), any signal-generating model including the no disclosure model is optimal. Given evidence $e \in \{g, b\}$, the politician can always choose the no disclosure model I_{ND_e} that sends evidence e with probability 1. This model has the maximal fit on evidence e among all models, and it keeps the voter's posterior fixed at the prior. Essentially, the politician can always confirm the voter's prior belief. As $p \geq 0.5$, the voter's optimal action under the prior is to vote for the policy. Thus, the politician can always persuade the voter to support the policy, irrespective of the researcher's model (Fig. 4b).

4 Extensions

In this section, I introduce three extensions to the base model: a restricted set of feasible models, ex-ante interpretation and asymmetric trust. I also provide an application on employee feedback.

4.1 Models with clear interpretation

In this section, I consider a setting in which the signals have a clear interpretation. For example, a bad grade can only make the agent think that their ability is worse, not better. This restricts the set of feasible models. I assume that the set of signals is a copy of the set of states $\mathcal{S} = \Omega$ and that each signal ω is likely generated under the state ω, that is, $I(\omega \mid \omega) \geq I(\omega \mid \tilde{\omega})$ for all states $\tilde{\omega} \neq \omega$. This prevents the narrator from proposing conflicting models that move the receiver's belief in a direction opposite to what was intended. The set of models with a clear interpretation $\mathcal{M}_C \subset \mathcal{M}$ is given by:

$$\mathcal{M}_C := \{I : \Omega \to \Delta\mathcal{S} : I(\omega \mid \omega) \geq I(\omega \mid \tilde{\omega}) \forall \tilde{\omega} \neq \omega\}$$

For this section, I assume the set of feasible models $\mathcal{F} = \mathcal{M}_C$. This imposes a constraint on the set of receiver's belief that can be induced. The space of beliefs can be partitioned into a collection of convex subsets $\{P_\omega\}_{\omega \in \Omega}$ such that given a signal ω, the posterior belief must be in P_ω.

$$P_\omega := \{q \in \Delta\Omega : \frac{q(\omega)}{p(\omega)} \geq \frac{q(\tilde{\omega})}{p(\tilde{\omega})} \quad \text{for all } \tilde{\omega} \in \Omega\}. \tag{12}$$

The signals have a clear interpretation. On seeing the signal ω, the change in the likelihood of the state ω is greater than any other state $\tilde{\omega} \neq \omega$. Ichihashi and Meng (2021) also impose this restriction on the set of feasible models. Now, I characterize the set of feasible beliefs and actions when the narrator is restricted to choosing models from $\mathcal{M}_C$.

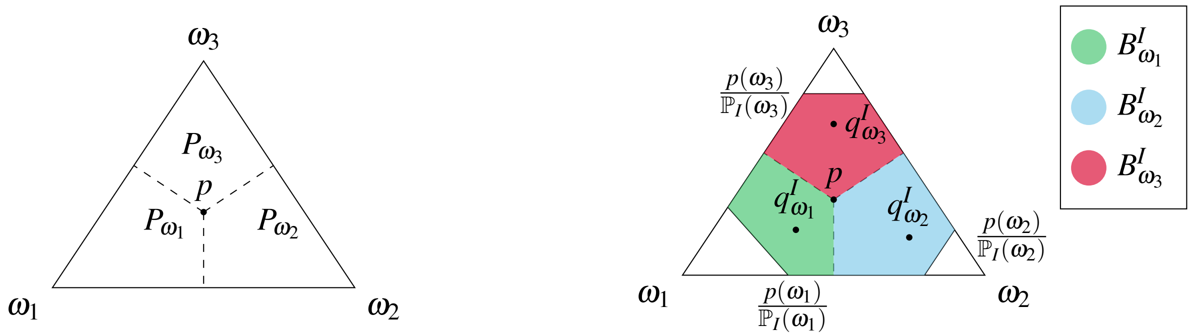


Figure 5: (a) The partition of the belief space and (b) the set of feasible posterior beliefs given sender's model I.

Proposition 3. *If $\mathcal{F} = \mathcal{M}_C$, the sets of feasible posterior beliefs and actions that the narrator can induce given sender's model $I \in \mathcal{M}_C$ and the signal ω are given by*

$$B^I_\omega := \{q \in P_\omega : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(\omega)\} \cup \{q^I_\omega\}, \tag{13}$$

$$A^I_\omega := \{a \in A : \exists q \in B^I_\omega \text{ such that } a = a^*_R(q)\}. \tag{14}$$

The narrator can induce either the true belief or any belief that aligns in the correct direction but is less precise. This ensures that the belief has a higher probability on the true state than before. The narrator can garble the signal but cannot change the direction of interpretation. The key difference from Lemma 1 is that the condition in equation (13) for the signal ω applies only to the most likely state ω, not to all states. One can still use Theorem 1 to find the optimal signal-generating model. The only caveat is that the set C_a will differ, contingent on the set of feasible models. Nonetheless, the fundamental property of convexity remains applicable, as the set of feasible models is a closed and convex set. This permits the partitioning of the set of feasible models into finite convex sets and allows for a search within these candidate models to determine the optimal one.

If the voter in the example of the researcher and the politician knew the direction of each evidence, the full disclosure model would be optimal for all priors. While one might expect that in this setting of clear interpretations, the narrator's role would be insignificant, it can still have adverse consequences. I illustrate this in the subsequent application.

4.2 Application: Employee Feedback

In this section, I consider the setting where a manager provides feedback to her optimistic employee. The employee wants to believe that her ability is good. She interprets the direction of feedback correctly but not the precision. The focus will be on whether the employee learns her true ability. First, I show that a small amount of uncertainty in precision can lead to biased learning. Furthermore, I show that two employees, an optimist and a pessimist, who start with the same initial beliefs and observe the same (infinite) sequence of feedback can be polarized.[24] Finally, I show that giving bad feedback more frequently than good can ensure that the optimist employee always learns her true ability.

Consider the states $\Omega = \{G, B\}$, where G and B refer to the state when the employee's ability is good and bad, respectively. The manager provides feedback about the ability using signals $\mathcal{S} = \{g, b\}$, where g refers to good news and b refers to bad news. The signals have a clear interpretation: good news is more likely generated when her ability is good, and vice versa. Suppose that the manager provides feedback according to model I (black point in Fig. 6a):

$$I(g \mid G) = I(b \mid B) = \kappa \tag{15}$$

where, $\kappa > 0.5$ is the precision of the news. She symmetrically provides good and bad news.

The employee correctly interprets the direction of the news, but he is uncertain about the precision. The set of feasible models $\mathcal{F} = \mathcal{M}_\varepsilon \subset \mathcal{M}_C$ have precision in the range of $[\kappa - \varepsilon, \kappa + \varepsilon]$ (green area in Fig. 6a):

$$\mathcal{M}_\varepsilon := \{m : \Omega \to \Delta\mathcal{S} : m(g \mid G) \in [\kappa - \varepsilon, \kappa + \varepsilon], m(b \mid B) \in [\kappa - \varepsilon, \kappa + \varepsilon]\}. \tag{16}$$

where, $\kappa - \varepsilon \geq 0.5$ and $\kappa + \varepsilon \leq 1$.

The level of uncertainty is given by ε. The lower the value ε, the more certain the employee is about the precision of the news.

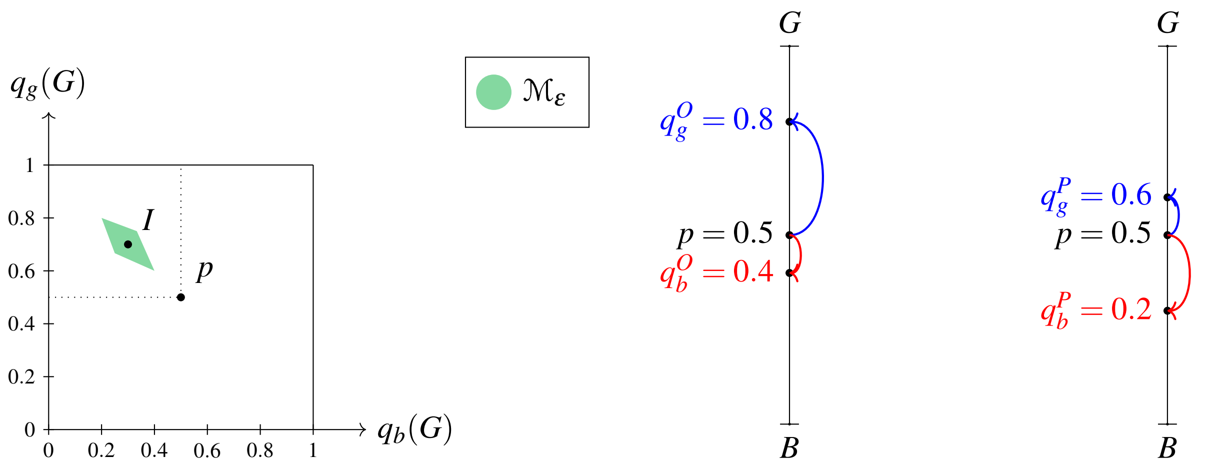
I assume that the employee is an **optimist (O)** who overestimates the likelihood of experiencing positive outcomes and underestimates the likelihood of experiencing negative events (Hey, 1984). Most people tend to be optimistic (Sharot, 2011).

Directed (or motivated) reasoning posits that people interpret news (often unconsciously) in the direction they find attractive. But even an optimist cannot interpret the news in any direction she wants. She faces a trade-off between accuracy and directional motives. In my setting, the accuracy goal corresponds to the model fit: how well the news fits the model. The employee can adopt a biased interpretation only if it has a better fit than the manager's model.

The employee here acts as both the narrator and the receiver. I consider a dual-self framework, where the unconscious mind (directional motives) is the narrator and the conscious mind (accuracy motives) is the receiver.[25] The narrator wants to interpret the news in the direction of his biased

[24] There's a growing theoretical literature that offers explanations, both Bayesian and non-Bayesian, on why and when polarization occurs. (Benoit and Dubra, 2016; Dixit and Weibull, 2007; Baliga, Hanany, and Klibanoff, 2013; Acemoglu, Chernozhukov, and Yildiz, 2016; Fryer Jr, Harms, and Jackson, 2019; Chen, 2022).

[25] Formally, one can assume the narrator (unconscious mind) has a belief based utility which is an increasing function



(a) The set of feasible models $\mathcal{M}_\varepsilon$.

(b) Biased updating by the optimist and the pessimist ($p = 0.5, \kappa = 0.7, \varepsilon = 0.1$).

Figure 6: Employee Feedback: Optimist (O) and Pessimist (P)

state G.

On receiving good news, the employee interprets using the model n_g:

$$n_g(g \mid G) = \kappa + \varepsilon \qquad\qquad n_g(b \mid B) = \kappa + \varepsilon. \tag{17}$$

This model has a (weakly) better fit than the manager's model for good news: $\mathbb{P}_{n_g}(g) \geq \mathbb{P}_I(g)$.[26] The employee interprets the news to be very informative and overreacts to it (see Fig. 6b).

On receiving the bad news, the employee interprets using the model n_b:

$$n_b(g \mid G) = \kappa - \varepsilon \qquad\qquad n_b(b \mid B) = \kappa - \varepsilon. \tag{18}$$

The model has a (weakly) better fit than the manager's model for bad news: $\mathbb{P}_{n_b}(b) \geq \mathbb{P}_I(b)$. The employee interprets the news to not be very informative and underreacts to it (see Fig. 6b).

Thus, the employee reacts asymmetrically to good and bad news (Eil and Rao, 2011; Möbius et al., 2022). My model provides a possible explanation for the *good news-bad news effect*, where the employee does not stray away from Bayesian updating, but instead uses different models to interpret different news.

What happens when the employee receives repeated feedback? Does she learn her true ability? I assume that in each round, she interprets each piece of news individually rather than considering

in his belief $q(G)$.

[26] A slight perturbation of n_g ensures a strictly better fit than I.

the entire sequence. The past news sequence only influences her prior belief for that specific round. I show if the amount of uncertainty ε is sufficiently large, the optimistic employee learns her biased state G, almost surely.

Lemma 3. *For any prior belief $p \in (0,1)$, the optimist (asymptotically) learns the biased state G almost surely if*

$$\left[\frac{\kappa-\varepsilon}{1-\kappa+\varepsilon}\right]^{\kappa} < \left[\frac{\kappa+\varepsilon}{1-\kappa-\varepsilon}\right]^{(1-\kappa)}. \tag{19}$$

This condition holds for example, when $\kappa = 0.7$ and $\varepsilon = 0.1$. Thus, even if the employee's ability is bad, despite repeated feedback, she ends up confident that her ability is good. Camerer and Lovallo (1999) shows that misinterpretation can lead to overconfidence when starting new businesses, often resulting in failure. Weeks et al. (1998) show that due to optimism bias, patients choose wrong treatments despite an accurate prognosis. Massey, Simmons, and Armor (2011) show that people exhibit optimism bias and misinterpret accurate signals despite repeated feedback.

Next, I consider two employees, an optimist (O) and a pessimist (P). Contrary to an optimist, a **pessimist** underestimates the likelihood of an favorable outcome and overestimates the likelihood of unfavorable outcomes (Hey, 1984).[27] One can also imagine two news outlets that twist the same news story to support their preferred political position. I show that the two employees despite having the same prior and observing the same (infinite) sequence of news can be polarized. In the long run, both employees become confident in their biased states and disagree with each other.

On receiving good news, the optimist interprets the signal to be very informative and overreacts to it, while the pessimist interprets it to be very uninformative and underreacts to it. And vice versa when receiving bad news (see Fig. 6b).

When presented with a balanced set of good news and bad news, the employees' beliefs are polarized, that is, $q^O_{gb}(G) > p > q^P_{gb}(G)$ where q^O_{gb} and q^P_{gb} refer to posterior belief of the optimist and pessimist after observing g and b, respectively.[28] The employees shift their beliefs in opposite directions after observing the same balanced set of good and bad news (Taber and Lodge, 2006; Bolsen, Druckman, and Cook, 2014). In the long run, under sufficient uncertainty, each employee always learns her biased state.

Corollary 3. *For any common prior belief $p \in (0,1)$, an optimist and pessimist learn their biased state G and B respectively almost surely if equation (19) holds.*

Given employees distort the news, how should a manager provide feedback to her optimist (or pessimist) employee? The manager should also provide news in an asymmetrical manner to

[27] Strunk, Lopez, and DeRubeis (2006) show that individuals suffering from depression tend to exhibit pessimism bias.
[28] The order of news does not affect the posterior belief: $q^i_{gb} = q^i_{bg}$ for $i = O, P$.

counter the asymmetric interpretation. This ensures that the employee's belief is close to being accurate.

Proposition 4. *For any prior belief $p \in (0,1)$, the optimist (asymptotically) learns the correct state almost surely under the optimal model $I^* \in \mathcal{M}_\varepsilon$, where the model I^* is given by:*

$$I^*(g \mid G) = \kappa - \varepsilon \qquad\qquad I^*(b \mid B) = \kappa + \varepsilon. \tag{20}$$

The optimal model I^* provides good and bad news in an asymmetric way to counter the asymmetric interpretation of signals. When providing feedback to an optimist employee, the manager should provide bad news more often than good news.[29] Thus, the optimal way to provide feedback to the employee is crucially dependent on the direction of bias they exhibit and the degree of uncertainty. This has implications on how to provide feedback or design test results. For example, medical tests can vary in their ability to rule in or rule out disease, and human resource departments can tailor their feedback style accordingly.

4.3 Ex-ante interpretation of signals

In this section, I consider a setting where the narrator has to provide his models ex-ante, before the signal has been observed, instead of ex-post, after the signal has been observed. However, he chooses his menu of models after observing the sender's choice. Schwartzstein and Sunderam (2021) focuses on ex-post interpretation, while Aina (2021) analyzes ex-ante interpretation. I assume that the narrator can provide multiple models instead of a single one in this setting. Communicating models before observing the signal enhances credibility. However, providing a menu of models imposes additional constraints for the narrator, as each model not only competes with the sender's model but also with the other models in the menu. Surprisingly, the narrator can attain the same vector of posterior beliefs and actions with ex-ante or ex-post provision of models. Thus, the results do not depend on whether the narrator communicates the models before or after the signals have been observed.

Let the menu of models the narrator provides be denoted by $\mathcal{N} = \bigcup_{s \in \mathcal{S}} n_s$.[30] The narrator does not need to provide more than $|\mathcal{S}|$ models. This is because the receiver selects one model for each signal. Consequently, the narrator tailors model n_s to correspond with each signal s.

The receiver observes the signal s and the set of models provided by the narrator and the sender. He adopts the model that has the best fit given the signal s.

[29] Similarly, when providing feedback to a pessimist employee, a manager should provide good news more often than bad news.

[30] The models, which can be conflicting, need not be provided by the same agent but by a collusion of agents to maintain credibility. For example, different members of a political party or different news shows on the same network (Bursztyn et al., 2020)

$$m_s := \underset{m \in \mathcal{N} \bigcup \{I\}}{\arg\max} \sum_{\omega \in \Omega} p(\omega) m(s \mid \omega). \tag{21}$$

Lemma 4 (ex-ante interpretation)**.** *Given the sender's model I and signal s, the set of feasible posterior beliefs and actions that the narrator can induce are:*

$$B_s^I := \{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \quad \forall \omega \in \Omega\} \cup \{q_s^I\}, \tag{22}$$

$$A_s^I := \{a \in A : \exists q \in B_s^I \text{ such that } a = a_R^*(q)\}. \tag{23}$$

Surprisingly, whether the model is provided ex-ante or ex-post does not impact the narrator's persuasive ability. Each model, denoted as n_s, is tailored for a specific signal, inducing the desired posterior belief q_s. And this model n_s has a worse fit than model n_t for any signal $t \neq s$. Theorem 2 of Aina (2021) can be applied to my setting to characterize the set of feasible posterior beliefs.

4.4 Asymmetric Trust

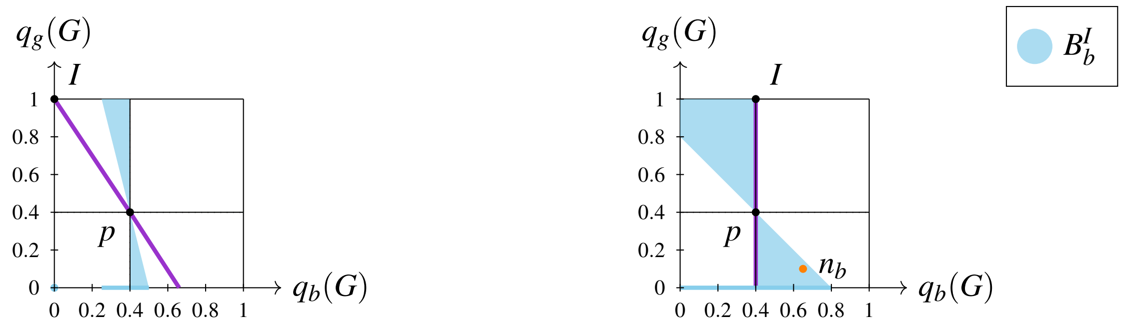
In this section, I consider a setting where the receiver can evaluate the models of the sender and narrator asymmetrically. This can happen due to differences in trust or credibility. For example, in the case of the researcher and the politician, a voter who firmly believes in science might need stronger evidence to accept the politician's model, while a skeptic may require less evidence.

I define the **trust coefficient** $\eta \in [0, \infty)$ as the trust ratio between the sender and the narrator. When the likelihood of the narrator's model exceeds this threshold, the receiver adopts the narrator's model:

$$\frac{\mathbb{P}_n(s)}{\mathbb{P}_I(s)} \geq \eta. \tag{24}$$

In the base model, η equals 1, meaning the receiver treats both models alike. When $\eta \geq \overline{\eta}$, where $\overline{\eta} = \min_{\omega \in \Omega} \frac{1}{p(\omega)}$, the receiver never adopts the narrator's model when the sender fully discloses the states. Conversely, when $\eta = 0$, the receiver always adopts the narrator's model. Let's analyze the example of the researcher and the politician for varying η.

If $\eta > 1$, then the voter trusts the researcher much more than the politician. She only chooses the politician's model when it significantly better explains the signal than the researcher's model. Specifically, when $\eta = 2$, the full disclosure model I_{FD} is optimal (see Fig. 7a). On seeing opposing evidence, the politician's influence to convince the voter is diminished. He cannot convince her to vote for the policy. The blue line on the x-axis represents the set of feasible beliefs he can induce given opposing evidence.



(a) For $\eta = 2$ and $I = I_{FD}$ **(b)** For $\eta = 0.5$ and $I = I_{ND_b}$

Figure 7: The set of feasible beliefs that the narrator can induce for prior belief $p = 0.4$.

If $\eta < 1$, then the voter trusts the researcher less than the politician. She chooses the politician's model, even if it has a lower likelihood to generate the signal than the researcher's model. Suppose the researcher chooses the no disclosure model I_{ND_b} that reports opposing evidence with certainty. This model has the maximal fit among all models. However, when $\eta = 0.5$, the politician can convince the voter to support the policy using the model n_b (orange point in Fig. 7b). The blue line on the x-axis represents the set of feasible beliefs that he can induce given opposing evidence. In this scenario, unlike the baseline setting, the value of information for any model can be strictly negative.

Given the sender model I and η, let $B_s^I(\eta)$ and $A_s^I(\eta)$ denote the set of feasible posterior beliefs and actions conditional on signal s and trust coefficient η.

Proposition 5. *If $\eta_1 > \eta_2$, then $B_s^I(\eta_1) \subseteq B_s^I(\eta_2)$ and $A_s^I(\eta_1) \subseteq A_s^I(\eta_2)$ for all $s \in \mathcal{S}$ and $I \in \mathcal{F}$.*

The proposition states that the narrator has a higher ability to persuade when the parameter η is lower. The higher the trust the receiver places on the narrator, the greater the persuasive ability he has. Two receivers with the same prior belief but different trust coefficients can interpret the same signal using different models. Also, notice that the set of feasible beliefs the narrator can induce does not have to be a convex set.

5 Conclusion

This paper analyzes the competing role of information provision and information interpretation in persuading an agent. If the agent does not know how data is generated, she can be persuaded to adopt biased interpretations over the correct one. In particular, full disclosure can backfire and lead to outcomes that are worse than providing no information. Hence, it is imperative to consider the narratives an agent might be exposed to when providing information. A novel technique is developed to find the optimal data-generating model. This model balances between providing precise information and minimizing misinterpretations. The optimal approach is to provide information in a manner that aligns with the agent's initial beliefs, effectively mitigating the risk of misinterpretation. This approach ensures that the agent always derives positive value from the information.

On a theoretical level, this paper presents a novel method to find the optimal data-generating model for both Bayesian and non-Bayesian agents, even in situations where the set of allowed models is restricted. The findings hold relevance in crafting information campaigns that influence public welfare, encompassing health, policy, and voting decisions, while also facilitating tailored feedback for agents who process information in a biased manner.

In terms of future research directions, there are several avenues to explore. One direction could involve examining competition between agents who can both provide and interpret information. Another interesting area to investigate is the impact of persuasion on a population of agents with heterogeneous prior beliefs. Additionally, alternative model selection rules, such as a convex combination of proposed models, could be considered, with coefficients reflecting the fitness ratio among the models.

References

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz. 2016. "Fragility of asymptotic agreement under Bayesian learning." *Theoretical Economics* 11 (1):187–225.

Aina, Chiara. 2021. "Tailored Stories." Tech. rep., Mimeo.

Alonso, Ricardo and Odilon Câmara. 2016. "Bayesian persuasion with heterogeneous priors." *Journal of Economic Theory* 165:672–706.

Aumann, Robert J, Michael Maschler, and Richard E Stearns. 1995. *Repeated games with incomplete information*. MIT press.

Baekgaard, Martin, Julian Christensen, Casper Mondrup Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen. 2019. "The role of evidence in politics: Motivated reasoning and persuasion among politicians." *British Journal of Political Science* 49 (3):1117–1140.

Baliga, Sandeep, Eran Hanany, and Peter Klibanoff. 2013. "Polarization and ambiguity." *American Economic Review* 103 (7):3071–3083.

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

Barron, Kai and Tilman Fries. 2023. "Narrative persuasion." .

Baysan, Ceren. 2022. "Persistent polarizing effects of persuasion: Experimental evidence from turkey." *American Economic Review* 112 (11):3528–3546.

Bénabou, Roland, Armin Falk, and Jean Tirole. 2018. "Narratives, imperatives, and moral reasoning." Tech. rep., National Bureau of Economic Research.

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

Benoit, Jean-Pierre and Juan Dubra. 2016. “A theory of rational attitude polarization.” *Available at SSRN 2754316* .

Bolsen, Toby, James N Druckman, and Fay Lomax Cook. 2014. “The influence of partisan motivated reasoning on public opinion.” *Political Behavior* 36:235–262.

Braghieri, Luca. 2023. “Biased Decoding and the Foundations of Communication.” *Available at SSRN 4366492* .

Bullock, John G. 2011. “Elite influence on public opinion in an informed electorate.” *American Political Science Review* 105 (3):496–515.

Bursztyn, Leonardo, Aakaash Rao, Christopher P Roth, and David H Yanagizawa-Drott. 2020. “Misinformation during a pandemic.” Tech. rep., National Bureau of Economic Research.

Camerer, Colin and Dan Lovallo. 1999. “Overconfidence and excess entry: An experimental approach.” *American economic review* 89 (1):306–318.

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

Chen, Jaden Yang. 2022. “Biased learning under ambiguous information.” *Journal of Economic Theory* 203:105492.

de Clippel, Geoffroy and Xu Zhang. 2022. “Non-bayesian persuasion.” *Journal of Political Economy* 130 (10):2594–2642.

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

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):114–138.

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. 2021. “Strategic interpretations.” *Journal of Economic Theory* 192:105192.

Fisher, Walter R. 1985. “The narrative paradigm: An elaboration.” *Communications Monographs* 52 (4):347–367.

Frick, Mira, Ryota Iijima, and Yuhta Ishii. 2021. “Welfare comparisons for biased learning.” .

———. 2023. “Efficient Learning Under Ambiguous Information.” Tech. rep., Mimeo.

Fryer Jr, Roland G, 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.

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

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. “Designing information provision experiments.” *Journal of economic literature* 61 (1):3–40.

Hart, P Sol and Erik C Nisbet. 2012. “Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies.” *Communication research* 39 (6):701–723.

Hey, John D. 1984. “The economics of optimism and pessimism: a definition and some applications.” *Kyklos* 37 (2):181–205.

Ichihashi, Shota and Delong Meng. 2021. “The Design and Interpretation of Information.” *Available at SSRN 3966003* .

Ispano, Alessandro et al. 2022. “The perils of a coherent narrative.” Tech. rep., THEMA (THéorie Economique, Modélisation et Applications), Université de .

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

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

Kendall, Chad W and Constantin Charles. 2022. “Causal narratives.” Tech. rep., National Bureau of Economic Research.

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

Lipnowski, Elliot and Laurent Mathevet. 2017. “Simplifying bayesian persuasion.” *Unpublished Paper, Columbia University.[642]* .

Lipnowski, Elliot and Doron Ravid. 2020. “Cheap talk with transparent motives.” *Econometrica* 88 (4):1631–1660.

Massey, Cade, Joseph P Simmons, and David A Armor. 2011. “Hope over experience: Desirability and the persistence of optimism.” *Psychological Science* 22 (2):274–281.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2022. “Managing self-confidence: Theory and experimental evidence.” *Management Science* 68 (11):7793–7817.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32 (2):303–330.

Ok, Efe A. 2007. *Real analysis with economic applications*, vol. 10. Princeton University Press.

Renault, Jérôme, Eilon Solan, and Nicolas Vieille. 2017. "Optimal dynamic information provision." *Games and Economic Behavior* 104:329–349.

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

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.

Strunk, Daniel R, Howard Lopez, and Robert J DeRubeis. 2006. "Depressive symptoms are associated with unrealistic negative predictions of future life events." *Behaviour research and therapy* 44 (6):861–882.

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American journal of political science* 50 (3):755–769.

Tardella, Fabio. 2004. *On the existence of polyhedral convex envelopes*. Springer.

Weeks, Jane C, E Francis Cook, Steven J O'Day, Lynn M Peterson, Neil Wenger, Douglas Reding, Frank E Harrell, Peter Kussin, Neil V Dawson, Alfred F Connors Jr et al. 1998. "Relationship between cancer patients' predictions of prognosis and their treatment preferences." *Jama* 279 (21):1709–1714.

Weick, Karl E. 1995. *Sensemaking in organizations*, vol. 3. Sage.

Yang, Jeffrey. 2023. "A Criterion of Model Decisiveness." *Available at SSRN 4425088* .

A Appendix: Proofs

Lemma 1. *Given the sender's model I and signal s, the sets of feasible posterior beliefs and actions that the narrator can induce are:*

$$B_s^I := \{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \forall \omega \in \Omega\} \cup \{q_s^I\}, \tag{5}$$

$$A_s^I := \{a \in A : \exists q \in B_s^I \textit{ such that } a = a_R^*(q)\}. \tag{6}$$

Proof. First, assume $q \in B_s^I$. I construct a model for the narrator n that results in posterior belief q and has a better fit than the sender's model I under signal s, that is, $q_s^n = q$ and $\mathbb{P}_n(s) > \mathbb{P}_I(s)$. The model n only sends two signals s and $\neg s$ with positive probability, where $\neg s \neq s$. An equivalent way of writing the condition in equation (5) is that $[\max_{\tilde{\omega}\in\Omega} \frac{q(\tilde{\omega})}{p(\tilde{\omega})}]^{-1} > \mathbb{P}_I(s)$. Let $\lambda = [\max_{\tilde{\omega}\in\Omega} \frac{q(\tilde{\omega})}{p(\tilde{\omega})}]^{-1}$ and the model n be given by:

$$n(s \mid \omega) = \frac{\lambda q(\omega)}{p(\omega)}, \qquad n(\neg s \mid \omega) = 1 - \frac{\lambda q(\omega)}{p(\omega)}. \tag{25}$$

First, as $\lambda \leq \frac{p(\omega)}{q(\omega)}$ for all $\omega \in \Omega$, I have $n(s \mid \omega) \leq 1$. Next, I show that the model n induces posterior belief q under the signal s.

$$\mathbb{P}_n(s) = \sum_{\omega\in\Omega} \lambda q(\omega) = \lambda, \qquad q_s^n(\omega) = \frac{p(\omega)n(s \mid \omega)}{\mathbb{P}_n(s)} = q(\omega). \tag{26}$$

By assumption, from equation (5), I have $\lambda > \mathbb{P}_I(s)$, so the receiver chooses model n over I under signal s. Thus, the narrator can induce posterior belief q under the signal s.

For the converse, I prove this by contradiction. Let q be a feasible posterior belief conditional on the signal s that does not satisfy equation (5). So, for some $\omega^* \in \Omega$, I have

$$\frac{p(\omega^*)}{q(\omega^*)} \leq \mathbb{P}_I(s). \tag{27}$$

As q is feasible, there exists a model n such that $q_s^n = q$ and $\mathbb{P}_n(s) > \mathbb{P}_I(s)$.

$$\mathbb{P}_I(s) < \mathbb{P}_n(s) = \frac{n(s \mid \omega)p(\omega)}{q(\omega)} \quad \forall \omega \in \Omega. \tag{28}$$

But from equation (27), I have

$$\frac{p(\omega^*)}{q(\omega^*)} < \frac{n(s \mid \omega)p(\omega)}{q(\omega)} \quad \forall \omega \in \Omega. \tag{29}$$

But this implies $n(s \mid \omega^*) > 1$, which is a contradiction.

The condition for feasible actions in equation (6) follows directly from the condition on the feasible posterior beliefs. The narrator can induce an action if he can induce a posterior belief under which the action is optimal for the receiver.

□

Lemma 2. *The set $C_{\boldsymbol{a}}$ is a finite disjoint union of convex sets for any vector of actions $\boldsymbol{a} \in A^{|\mathcal{S}|}$.*

Proof. Fix a vector of subsets of actions $\boldsymbol{R} = (R_s)_{s \in \mathcal{S}} \in \mathcal{P}(A)^{|\mathcal{S}|}$.[31] Let $C_{\boldsymbol{a},\boldsymbol{R}}$ denote the subset of models where the vector of induced and feasible actions are given by $\boldsymbol{a}$ and $\boldsymbol{R}$, respectively.

$$C_{\boldsymbol{a},\boldsymbol{R}} = \{I \in \mathcal{F} : a_s^I = a_s, A_s^I = R_s \forall s \in \mathcal{S}\} \subset C_{\boldsymbol{a}}. \tag{30}$$

First, I show that the set $C_{\boldsymbol{a},\boldsymbol{R}}$ is a convex set for any pair $(\boldsymbol{a},\boldsymbol{R})$. Assume $C_{\boldsymbol{a},\boldsymbol{R}}$ is non-empty. Let I_1 and I_2 belong to $C_{\boldsymbol{a},\boldsymbol{R}}$. Let $I_\alpha = \alpha I_1 + (1-\alpha)I_2$ denote the convex combination of the models I_1 and I_2, where $\alpha \in (0,1)$.[32] I show that $I_\alpha \in C_{\boldsymbol{a},\boldsymbol{R}}$ for all $\alpha \in (0,1)$. First notice, that the fit of the model I_α lies in between the model I_1 and I_2.

$$\mathbb{P}_{I_\alpha}(s) = \alpha \mathbb{P}_{I_1}(s) + (1-\alpha)\mathbb{P}_{I_2}(s) \quad \forall s \in \mathcal{S}. \tag{31}$$

To see that $C_{\boldsymbol{a},\boldsymbol{R}}$ is convex, let $a \in R_s$, I show $a \in A_s^{I_\alpha}$. From Eq. (27), I know there exist q such that $a_R^*(q) = a$ and such that

$$[\max_{\omega \in \Omega} \frac{q(\omega)}{p(\omega)}]^{-1} > \mathbb{P}_{I_i}(s) \text{ for } i = 1,2, \tag{32}$$

$$\Rightarrow [\max_{\omega \in \Omega} \frac{q(\omega)}{p(\omega)}]^{-1} > \max_{i=1,2} \mathbb{P}_{I_i}(s) > \mathbb{P}_{I_\alpha}(s). \tag{33}$$

Thus, this implies $a \in A_s^{I_\alpha}$. Now, I show it is also optimal for the narrator to induce the action a when the sender's model is I_α and the signal is s, that is, $a \in a^{I_\alpha}$. Recall as I_1 and I_2 belong to $C_{\boldsymbol{a},\boldsymbol{R}}$, the vector of induced action is given by $\boldsymbol{a}$. Thus, I have

$$a_s \in \arg\max_{a \in R_s} \mathbb{E}_{q_s^{I_i}}[u_N(\omega, a)] \text{ for } i = 1,2. \tag{34}$$

[31] Here $\mathcal{P}(A)$ refers to the power set of the set A.

[32] Formally, $I_\alpha(s \mid \omega) = \alpha I_1(s \mid \omega) + (1-\alpha)I_2(s \mid \omega)$ for all $\omega \in \Omega$ and $s \in \mathcal{S}$.

However, since I_α is a convex combination of I_1 and I_2, this implies $q_s^{I_\alpha} \in (q_s^{I_1}, q_s^{I_2})$. Thus, this implies $a_s \in a_s^{I_\alpha}$ for all $s \in \mathcal{S}$. So, I have shown that the set $C_{\boldsymbol{a},\boldsymbol{R}}$ is convex.

To complete the proof, take the union of all vector of subsets of actions where the induced vector of action is $\boldsymbol{a}$.

$$C_{\boldsymbol{a}} = \bigcup_{\boldsymbol{R} \in \mathcal{P}(A)^{|\mathcal{S}|}} C_{\boldsymbol{a},\boldsymbol{R}}$$

As the power set of the set A is finite, this a union over finitely many subsets. Thus, I have shown that the set $C_{\boldsymbol{a}}$ is a disjoint finite union of convex sets.

□

Theorem 1. *The optimal signal-generating model*

$$I^* := \arg\max_{I \in \mathcal{F}} V(I) \tag{11}$$

corresponds to an extreme point of the set $\overline{C}_{\boldsymbol{a}}$ for some $\boldsymbol{a} \in A^{|\mathcal{S}|}$. Furthermore, $Ext(\overline{C}_{\boldsymbol{a}})$ is finite for all $\boldsymbol{a} \in A^{|\mathcal{S}|}$.

Proof. Recall that the set $C_{\boldsymbol{a},\boldsymbol{R}}$ denotes the subset of models where the vector of induced and feasible actions are given by $\boldsymbol{a}$ and $\boldsymbol{R}$ respectively.

$$C_{\boldsymbol{a},\boldsymbol{R}} := \{I \in \mathcal{F} : a_s^I = a_s, A_s^I = R_s \forall s \in \mathcal{S}\} \subseteq C_{\boldsymbol{a}}. \tag{35}$$

First, I find the optimal policy within each set $\overline{C}_{\boldsymbol{a},\boldsymbol{R}}$ for a given $\boldsymbol{a} \in A^{|\mathcal{S}|}$ and $\boldsymbol{R} \in \mathcal{P}(A)^{|\mathcal{S}|}$. Note that the value function is linear within each such set, as it given by the receiver's expected utility given the vector of induced actions. As the vector induced actions remains the same, it is linear, or in general, convex. By the Bauer maximum principle (Ok, 2007, p. 658), the optimal model can be found at some extreme point of the closed convex set $\overline{C}_{\boldsymbol{a},\boldsymbol{R}}$.

Similarly, as the set $C_{\boldsymbol{a}}$ is given by a finite disjoint union of convex sets, I can restrict the search for each $\boldsymbol{a}$ to the extreme points of all possible convex sets $\overline{C}_{\boldsymbol{a},\boldsymbol{R}}$.

$$\begin{aligned} Ext(\overline{C}_{\boldsymbol{a}}) &= \{I \in \overline{C}_a : I \in Ext(\overline{C}_{\boldsymbol{a},\boldsymbol{R}}) \text{ whenever } I \in \overline{C}_{\boldsymbol{a},\boldsymbol{R}}\}, && (36) \\ &= \bigcup_{\boldsymbol{R} \in \mathcal{P}(A)^{|\mathcal{S}|}} Ext(\overline{C}_{\boldsymbol{a},\boldsymbol{R}}). && (37) \end{aligned}$$

To find the overall optimal model, one needs to take the union over all possible induced vectors of actions. All that is left to show is that the set of such extreme points is finite. To do so, I show that any set $\overline{C}_{\boldsymbol{a},\boldsymbol{R}}$ is the intersection of the finite collection of closed half spaces and thus must have finite extreme points.

$$\overline{C}_{\boldsymbol{a},\boldsymbol{R}} := \bigcap_{s \in \mathcal{S}} \bigcap_{b \in R_s} \{I \in \mathcal{F} : \mathbb{E}_{q_s^I}[u_N(\omega, a_s)] \geq \mathbb{E}_{q_s^I}[u_N(\omega, b)]\}. \tag{38}$$

The set of signals and the sets of feasible vectors of actions are finite. Therefore, each set $\overline{C}_{\boldsymbol{a}}$ has finitely many extreme points.

$\square$

Proposition 1. *If $u_N = -u_R$, then for any $s \in \mathcal{S}$, the no disclosure model I_{ND_s} is optimal. Additionally, any optimal model induces the unique action $a_R^*(p)$.*

Proof. I show that both agents can guarantee the utility (or outcome) corresponding to the no disclosure model I_{ND_s} for some $s \in \mathcal{S}$.

First, I show the sender can secure non-negative value of information by using the model I_{ND_s}, that is, $V(I_{ND_s}) = 0 \quad \forall s \in \mathcal{S}$. The no disclosure model I_{ND_s} has the maximal fit among the set of all models for the signal s. This signal is observed with certainty and the narrator cannot come up with any interpretation with a better fit.

On the other hand, assume a model I is optimal which leads to a different outcome than the no disclosure model. This implies that $V(I) \geq 0$. So, there exists some signal (which is observed with positive probability) under which the action a is induced which does strictly better than $a_R^*(p)$, that is $\mathbb{E}_{q_s^I}[u_N(\omega, a_s^I)] > \mathbb{E}_{q_s^I}[u_N(\omega, a_R^*(p))]$.

However as the narrator's utility is perfectly misaligned with the receiver's, this implies the narrator's expected utility is negative, that is, $\mathbb{E}_{q_s^I}[u_N(\omega, a_s^I)] < \mathbb{E}_{q_s^I}[u_N(\omega, a_R^*(p))]$. But, the narrator can choose the no disclosure model I_{ND_s} on observing signal s. This model is chosen over the sender's model I (as it is not no disclosure) and is a profitable deviation for the narrator. This again results in the induced action $a_R^*(p)$. Thus, I have shown that the no disclosure model I_{ND_s} is optimal when the preferences of the narrator and the receiver are perfectly misaligned. Also, the induced outcome is unique under any optimal model.

$\square$

Proposition 2. *For binary states and a narrator with state-independent utility, the full disclosure model I_{FD} is optimal if $u_R(\omega, a_\omega^{I_{FD}}) \geq u_R(\omega, a_R^*(p))$ for all $\omega \in \Omega$.*

Proof. From assumption, the full disclosure model leads to an expected utility higher than that of providing no information. Therefore, the optimal signal-generating model I^* is at least partially informative.

Assume I^* is not full disclosure and let $\Omega \subseteq \mathcal{S}$. From Lemma 2, the optimal model can be found at an extreme point of the set $C_{\boldsymbol{a}}$. For binary states and state-independent preferences of the narrator, this implies that atleast one state will be fully disclosed, i.e., $q_\omega^{I^*} = \delta_\omega$ for some $\omega \in \Omega = \{\omega_0, \omega_1\}$. Without loss of generality, assume that this state is ω_0.

As I^* is obtained by pooling ω_0 and ω_1 from I_{FD}, I have $\mathbb{P}_{I^*}(\omega_0) \leq \mathbb{P}_{I_{FD}}(\omega_0)$. However, the set of feasible beliefs still includes all beliefs $q \in [\delta_{\omega_0}, p]$. This is because the narrator can always induce any belief in between by taking the convex combination of I^* and the no disclosure model $I_{ND_{\omega_0}}$ which sends signal ω_0 with probability 1. This ensures the combined model has a better fit than I^* under ω_0 and induces the belief that lies in between. So, I have $u_R(\omega_0, a_\omega^{I^*}) \leq u_R(\omega_0, a_\omega^{I_{FD}})$. The action for signal ω_0 performs at worst no better than the full disclosure model.

Now, for I^* to be optimal we need that $u_R(\omega_1, a_{\omega_1}^{I^*}) \geq u_R(\omega_1, a_R^*(p))$. If this does not hold then full disclosure model would be a profitable deviation. So, the chosen action is optimal at a belief q_1^* closer to the state ω_1 than p. As $\mathbb{P}_{I^*}(\omega_1) \geq \mathbb{P}_{I_{FD}}(\omega_1)$, I have $A_{\omega_1}^{I^*} \subseteq A_{\omega_1}^{I_{FD}}$. So, if $a_{\omega_1}^{I_{FD}} \in A_{\omega_1}^{I^*}$, then the narrator would choose it. This implies that $a^{I_{FD}}_{\omega_1} \notin A^{I^*}_{\omega_1}$. But this implies the action $a^{I_{FD}}_{\omega_1}$ is closer to the state ω_1 than the action $a^{I^*}_{\omega_1}$. But then the sender can deviate profitably by providing more information and induce action $a^{I_{FD}}_{\omega_1}$. But then the induced actions are exactly the same as the full disclosure model. But as the induced actions perform better than the action at the prior belief and the linearity of the value function, the sender prefers the full disclosure model. Thus, I have shown that any model I^* that does not fully disclose must be suboptimal.

□

Corollary 2. *Given any narrator with state-independent utility, there exists a prior belief $p \in int(\Delta\Omega)$ such that the full disclosure model I_{FD} is not optimal.*

Proof. Assume the narrator's most preferred action among the set of actions that are optimal for the receiver at some belief is $\overline{a}$. From assumption, this action is not optimal for all beliefs. Let $\overline{p}$ be the (interior) belief, such that $\overline{a} \notin a^*_R(\overline{p}+\varepsilon)$ for any $\varepsilon > 0$. The assumption $\varepsilon > 0$ is without loss of generality, one can also derive conditions for $\varepsilon < 0$.

I will derive conditions for ε such that the full disclosure model is not optimal for the prior belief $\overline{p}+\varepsilon$. From Lemma 1, I can verify that the narrator can induce his preferred action $\overline{a}$ with probability 1 if

$$\frac{1-\overline{p}}{1-\overline{p}-\varepsilon} > \overline{p} \quad \text{and} \quad \frac{\overline{p}}{\overline{p}+\varepsilon} > 1-\overline{p}. \tag{39}$$

Both the conditions is satisfied if $\varepsilon < \frac{\overline{p}^2}{1-\overline{p}}$. Thus, the narrator is able to induce the action $\overline{a}$ with probability 1. But recall this is not the receiver's optimal action given her prior, that is, $\overline{a} \notin a^*_R(\overline{p}+\varepsilon)$. Thus, providing no information such that the receiver's belief stays fixed at $\overline{p}+\varepsilon$ is a profitable deviation. Thus, I_{FD} is not the optimal model.

□

Proposition 3. *If $\mathcal{F} = \mathcal{M}_C$, the sets of feasible posterior beliefs and actions that the narrator can induce given sender's model $I \in \mathcal{M}_C$ and the signal ω are given by*

$$B^I_\omega := \{q \in P_\omega : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(\omega)\} \cup \{q^I_\omega\}, \tag{13}$$

$$A^I_\omega := \{a \in A : \exists q \in B^I_\omega \text{ such that } a = a^*_R(q)\}. \tag{14}$$

Proof. Assume $q \in B^I_\omega$. I construct a model for the narrator $n \in \mathcal{M}_C$ that results in posterior beliefs q and has a better fit than the correct under signal ω, that is, $q^n_\omega = q$ and $\mathbb{P}_n(\omega) > \mathbb{P}_I(\omega)$. The model n only sends two signals ω and $\tilde{\omega}$ with positive probability, where $\tilde{\omega} \neq \omega$.

$$n(\omega \mid \tilde{\omega}) = \frac{q(\tilde{\omega})p(\omega)}{p(\tilde{\omega})q(\omega)} \qquad n(\neg\omega \mid \tilde{\omega}) = 1 - \frac{q(\tilde{\omega})p(\omega)}{p(\tilde{\omega})q(\omega)} \text{for all } \tilde{\omega} \in \Omega. \tag{40}$$

First, as $n \in \mathcal{M}_C$, I have $\frac{q(\tilde{\omega})}{p(\tilde{\omega})} \leq \frac{q(\omega)}{p(\omega)}$ for all $\tilde{\omega} \in \Omega$. So, I have $n(\omega \mid \tilde{\omega}) \leq 1$ for all $\omega \in S$. Next, I show that the model n induces posterior belief q under signal ω.

$$\mathbb{P}_n(\omega) = \frac{p(\omega)}{q(\omega)} \qquad\qquad q^n_\omega(\tilde{\omega}) = \frac{p(\tilde{\omega})n(\omega \mid \tilde{\omega})}{\mathbb{P}_n(\omega)} = q(\tilde{\omega}). \tag{41}$$

From assumption, I have $\mathbb{P}_n(\omega) > \mathbb{P}_I(s)$, so the receiver chooses model n over I under signal ω. Thus, the narrator can induce posterior belief q under the signal ω.

For the converse, I prove this by contradiction. (a) Suppose q is feasible but $q(\omega) \notin P_\omega$. So, there exists a model $n \in \mathcal{M}_C$ such that $q^n_\omega = q$ and $\mathbb{P}_n(\omega) > \mathbb{P}_I(\omega)$. But this implies that

$$\frac{\frac{q(\omega)}{q(\tilde{\omega})}}{\frac{p(\omega)}{p(\tilde{\omega})}} = \frac{n(\omega \mid \omega)}{n(\omega \mid \tilde{\omega})}, \tag{42}$$

$$\Rightarrow 1 > \frac{n(\omega \mid \omega)}{n(\omega \mid \tilde{\omega})} \text{ for some } \tilde{\omega} \in \Omega. \tag{43}$$

But this is a contradiction as the model $n \in \mathcal{M}_C$. So, the signal ω is most likely to be generated in the state ω.

(b) Suppose q is feasible but $q(\omega) > \frac{p(\omega)}{\mathbb{P}_I(\omega)}$. As q is feasible, there exists a model n such that $q^n_\omega = q$ and $\mathbb{P}_n(\omega) > \mathbb{P}_I(\omega)$.

$$\mathbb{P}_I(\omega) < \mathbb{P}_n(\omega) = \frac{n(\omega \mid \omega)p(\omega)}{q(\omega)}. \tag{44}$$

But by assumption, I have

$$\frac{p(\omega)}{q(\omega)} < \frac{n(s \mid \omega)p(\omega)}{q(\omega)}. \tag{45}$$

But this implies $n(\omega \mid \omega) > 1$, which is a contradiction.

□

Lemma 3. *For any prior belief $p \in (0,1)$, the optimist (asymptotically) learns the biased state G almost surely if*

$$\left[\frac{\kappa - \varepsilon}{1 - \kappa + \varepsilon}\right]^{\kappa} < \left[\frac{\kappa + \varepsilon}{1 - \kappa - \varepsilon}\right]^{(1-\kappa)}. \tag{19}$$

Proof. I show that even when the state is B , the optimist employee's belief converges to the (incorrect) state G almost surely.

In any round $n \in \mathbb{N}$, the news s_n is generated according to the model $I(\cdot \mid B)$, where $I(g \mid G) = I(b \mid B) = \kappa$. This implies that in the long run, he observes bad news b in κ fraction of the rounds and good news g in $(1-\kappa)$ fraction of the rounds. Let $s^n = (s_1, ...s_n)$ denote the sequence of news observed in the first n rounds. Formally, I want to show

$$\lim_{n \to \infty} q^O_{s^n} = \delta_G \quad \mathbb{P}_{I(\cdot|B)} - a.s. \tag{46}$$

where, $q^O_{s^n}$ is the posterior belief of the optimist employee given the sequence of news s^n

First, I show a useful property that the order of sequence of news does not impact the posterior belief. Consider the sequence of news g, b and b, g respectively. I have

$$\mathbb{P}(\omega \mid g, b) = \frac{p(\omega) n_g(g \mid \omega) n_b(b \mid \omega)}{\mathbb{P}_{n_g}(g) \cdot \mathbb{P}_{n_b}(b)}, \tag{47}$$

$$= \frac{\mathbb{P}_{n_b}(\omega \mid b) n_g(g \mid \omega)}{\mathbb{P}_{n_g}(g)} = \mathbb{P}(\omega \mid b, g). \tag{48}$$

The key aspect is that, irrespective of the order, the receiver uses fixed models n_g and n_b to process good and bad news, respectively. Thus, one can choose any sequence of order as long as the proportion of good and bad news remains the same.

Assume the employee observes n signals, of which κn are bad news and $(1-\kappa)n$ good news, where κn and $(1-\kappa)n$ are natural numbers. If her posterior belief on state G after observing the n sequence of news is greater than the prior belief, then in the long run her beliefs will converge to good state δ_G. Assume that the employee first observes the κn sequence of bad news and then the $(1-\kappa)n$ sequence of good news.

Let $x = \frac{1-q}{q}$ denote the likelihood ratio of the belief after observing the κn sequence of bad news. I derive the condition that after observing $(1-\kappa)n$ sequence of good news, her posterior belief is higher than the prior belief p.

$$\frac{(\kappa+\varepsilon)^{(1-\kappa)n}}{(\kappa+\varepsilon)^{(1-\kappa)n} + x(1-\kappa-\varepsilon)^{(1-\kappa)n}} > p,$$

$$(\frac{k+\varepsilon}{1-\kappa-\varepsilon})^{(1-\kappa)n} \cdot (\frac{1-p}{p}) > x.$$

Now, in place of x, I substitute the likelihood ratio that I get after observing the κn sequence of bad news, so I have

$$x = (\frac{1-p}{p}) \cdot (\frac{\kappa-\varepsilon}{1-\kappa+\varepsilon})^{\kappa n}.$$

Thus, I have the following condition:

$$(\frac{\kappa-\varepsilon}{1-\kappa+\varepsilon})^{\kappa} < (\frac{k+\varepsilon}{1-\kappa-\varepsilon})^{(1-\kappa)}.$$

□

Proposition 4. *For any prior belief $p \in (0,1)$, the optimist (asymptotically) learns the correct state almost surely under the optimal model $I^* \in \mathcal{M}_\varepsilon$, where the model I^* is given by:*

$$I^*(g \mid G) = \kappa - \varepsilon \qquad\qquad I^*(b \mid B) = \kappa + \varepsilon. \tag{20}$$

Proof. The difficult part of the proof is to show that when the state is B, the optimistic employee's belief converges to the correct state B almost surely.

In any round $n \in \mathbb{N}$, the news s_n is generated according to the model $I^*(\cdot \mid B)$. This implies that in the long run, he observes bad news b in the $\kappa+\varepsilon$ fraction of the rounds and good news g in $(1-\kappa-\varepsilon)$ fraction of the rounds. Let $s^n = (s_1, ...s_n)$ denote the sequence of news observed in the first n rounds. Formally, I want to show

$$\lim_{n\to\infty} q^O_{s^n} = \delta_B \quad \mathbb{P}_{I^*(\cdot|B)} - a.s. \tag{49}$$

where, $q^O_{s^n}$ is the posterior belief of the optimist given the sequence of news s^n

Observing bad news b, the employee updates her beliefs using the true signal-generating model I^*. This follows, as no model $n \in \mathcal{M}_\varepsilon$ has a better fit than I^* on bad news b. While observing good news g, the employee interprets using the model n_g which has precision $\kappa+\varepsilon$.

Assume that the employee observes n signals, of which $(\kappa+\varepsilon)n$ signals are bad and $(1-\kappa-\varepsilon)n$ signals are bad. If her posterior belief on state G after observing the n sequence of news is greater than the prior, then in the long run her beliefs will converge to good state δ_G. Assume that the employee first observes the $(\kappa+\varepsilon)n$ sequence of bad news and then the $(1-\kappa-\varepsilon)n$ sequence of good news.

Let $x = \frac{1-q}{q}$ denote the likelihood ratio of the belief after observing the $(\kappa+\varepsilon)n$ sequence of bad news. I derive the condition that after observing $(1-\kappa-\varepsilon)n$ sequence of good news, her posterior belief is higher than the prior belief p.

$$\frac{(\kappa+\varepsilon)^{(1-\kappa-\varepsilon)n}}{(\kappa+\varepsilon)^{(1-\kappa-\varepsilon)n} + x(1-\kappa-\varepsilon)^{(1-\kappa-\varepsilon)n}} > p,$$
$$(\frac{\kappa+\varepsilon}{1-\kappa-\varepsilon})^{(1-\kappa)n} \cdot (\frac{1-p}{p}) > x.$$

Now, in place of x, I substitute the likelihood ratio that I get after observing the $(\kappa+\varepsilon)n$ sequence of bad news, so I have

$$x = \Big(\frac{1-p}{p}\Big) \cdot \Big(\frac{\kappa+\varepsilon}{1-\kappa-\varepsilon}\Big)^{(\kappa+\varepsilon)n}.$$

However, this happens when the following condition holds:

$$\Big(\frac{\kappa+\varepsilon}{1-\kappa-\varepsilon}\Big)^{(\kappa+\varepsilon)} < \Big(\frac{\kappa+\varepsilon}{1-\kappa-\varepsilon}\Big)^{(1-\kappa-\varepsilon)}.$$

But this inequality does not hold for any values of κ and ε. This ensures that the employee learns the correct state almost surely. Thus, to counter the asymmetric reaction by the receiver, the sender sends bad news with a higher frequency compared to bad news.

□

Lemma 4 (ex-ante interpretation)**.** *Given the sender's model I and signal s, the set of feasible posterior beliefs and actions that the narrator can induce are:*

$$B_s^I := \{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \quad \forall \omega \in \Omega\} \cup \{q_s^I\}, \tag{22}$$
$$A_s^I := \{a \in A : \exists q \in B_s^I \text{ such that } a = a_R^*(q)\}. \tag{23}$$

Proof. First, assume $q = (q_s)_{s\in\mathcal{S}} \in B^I$. I construct a menu of models $\mathcal{N} = \bigcup_{s\in\mathcal{S}} n_s$ for the narrator such that given the signal s, the model n_s results in the posterior belief q_s and it has a better fit than other models i.e., $q_s^{n_s} = q$ and $\mathbb{P}_{n_s}(s) > \mathbb{P}_m(s)$ for $m \in \{I\} \bigcup_{t\neq s}\{n_t\}$. Let $\lambda_s = [\max_{\omega\in\Omega} \frac{q_s(\omega)}{p(\omega)}]^{-1}$.

$$n_s(s \mid \omega) = \frac{\lambda_s q_s(\omega)}{p(\omega)}, \qquad n_s(t \mid \omega) = \Big(\frac{\lambda_t}{\sum_{r\neq s}\lambda_r}\Big)\Big(1 - \frac{\lambda_s q_s(\omega)}{p(\omega)}\Big) \text{for all } t \neq s. \tag{50}$$

Note, this ensures that on receiving signal s and updating using model n_s, the receiver's posterior belief is equal to q_s. Also, from assumption it has a better fit than sender's model I for signal s.

$$\mathbb{P}_{n_s}(s) = \sum_{\omega\in\Omega} p(\omega)\lambda_s q_s(\omega) = \lambda_s, \qquad q_s^{n_s}(\omega) = \frac{p(\omega) n_s(s \mid \omega)}{\mathbb{P}_{n_s}(s)} = q_s(\omega). \tag{51}$$

Now, I show that it also has a better fit than other models of the narrator in the menu. For any other model n_t, I have

$$\mathbb{P}_{n_t}(s) = \sum_{\omega} p(\omega)\Big(\frac{\lambda_s}{\sum_{r\neq t}\lambda_r}\Big)\big(1 - \frac{\lambda_t q_t(\omega)}{p(\omega)}\big), \tag{52}$$

$$= \lambda_s\Big(\frac{1-\lambda_t}{\sum_{r\neq t}\lambda_r}\Big), \tag{53}$$

$$\leq \lambda_s = \mathbb{P}_{n_s}(s). \tag{54}$$

So, I have shown that the model n_s has a better fit than the model n_t for signal s.

Now, assume $\boldsymbol{q} \notin B^I$. Assume the inequality is not satisfied for signal s. It follows, from Lemma 1, that the narrator cannot come up with a model n_s that induces belief q_s and has a fit greater than $[\max_{\omega\in\Omega}\frac{q_s(\omega)}{p(\omega)}]^{-1}$. But from assumption, I have

$$\mathbb{P}_I(s) \geq [\max_{\omega\in\Omega}\frac{q_s(\omega)}{p(\omega)}]^{-1}. \tag{55}$$

Thus, the narrator cannot come up with a model n_s such that $q_s^{n_s} = q_s$ and has better fit than the sender's model I.

□

Proposition 5. *If $\eta_1 > \eta_2$, then $B_s^I(\eta_1) \subseteq B_s^I(\eta_2)$ and $A_s^I(\eta_1) \subseteq A_s^I(\eta_2)$ for all $s \in \mathcal{S}$ and $I \in \mathcal{F}$.*

Proof. Assume $\eta_1 > \eta_2$ and $q \in B_s^I(\eta_1)$. I will show that $q \in B_s^I(\eta_2)$. As $q \in B_s^I(\eta_1)$, $\exists n$ such that $q_s^n = q$ and $\mathbb{P}_n(s) \geq \eta_1\mathbb{P}_I(s)$.

But, as $\eta_1 > \eta_2$, this implies that $\mathbb{P}_n(s) \geq \eta_2\mathbb{P}_I(s)$. The narrator can use the same model n to induce belief q. Thus, $q \in B_s^I(\eta_2)$. This also implies that if any action a belongs to $A_s^I(\eta_1)$ then it also belongs to the set $A_s^I(\eta_2)$.

□