Informing agents amidst biased narratives

Atulya Jain

HEC Paris[*]

November 7, 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 and must choose between models provided by the sender and the narrator. The receiver chooses the model 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 information and minimizing misinterpretation. To find the optimal data-generating model, which maximizes the receiver's expected utility, I restrict the search to a finite set of models. This set depends solely on the preferences of the narrator and receiver, along with the receiver's prior belief. I show that fully informative model can be sub-optimal and even backfire. Finally, I apply this framework to information campaigns and providing feedback.

JEL classification: C72, D82, D83.

Keywords: Information provision, Persuasion, Narratives, Polarization, Bounded Rationality.

[*] email: atulya.jain@hec.edu

1 Introduction

Benevolent experts guide decision-making by providing data and it's interpretation. For instance, researchers supply data to assess the effectiveness of policies, and scientists provide evidence of climate change. Despite this, some people support flawed policies and deny climate change. This often occurs because people misinterpret data under the influence of biased narratives. For instance, politicians twist data to support their policies, and oil companies dispute climate change evidence. 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. For example, an information campaign to address executive power and combat censorship inadvertently led to voter polarization during a referendum in Turkey (Baysan, 2022).

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. After observing the data and the sender's model, the narrator proposes his model on how to interpret the data. The receiver observes both models and the data, but she does not know which one is the actual data-generating model. Different models can lead to varying and even conflicting interpretations of the same data. The receiver chooses the model that maximizes the likelihood (or fit) of the data given her prior belief. Finally, she takes an action based on this chosen model. *How should a benevolent sender provide data when a biased narrator 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 indicate 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 she does not strongly oppose the policy initially, she finds the researcher's recommendation implausible, which makes her more receptive to the politician. Therefore, the politician 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 guide the voter while minimizing misinterpretations? Surprisingly, a partially informative experiment is optimal. This experiment occasionally yields low unemployment data even when the policy is effective - akin to a Type I error.[1] While, 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. As a result, the politician cannot sway her with any interpretation, and she votes against the policy.

In general, how should the sender choose her data-generating model? To answer this, I first examine the narrator's ability to misinterpret the data. I characterize the set of feasible posterior beliefs and actions the narrator can induce given a fixed data-generating model. The sender's model constrains the narrator's ability to persuade. A sender's model that fits the data well limits the beliefs the narrator can induce by misinterpretation. Also, the farther a belief is from the prior, the harder it is for the narrator to induce it because any model that results in it has a bad fit. Notably, through interpretations, the narrator can steer the receiver's belief in any direction, even in those inconsistent with the true data-generating model.

Taking into account the narrator's influence, the sender must find a balance between providing information and minimizing misinterpretation. There is a trade-off between how well a model fits the data and how much it can alter the receiver's belief when given that data. Given any data, the induced action depends on both the posterior belief and fit of the sender's model. So, I focus on the space of models rather than the space of beliefs.

My main result shows that I can identify a finite set of models that contains the optimal data-generating model. This relies on two key observations: (i) the receiver's expected utility is linear if the induced action remains unchanged and (ii) the set of models where the induced action remains fixed is given by a finite union of convex sets. These sets are defined based on the preferences of both the narrator and receiver, along with the receiver's prior belief. This technique is applicable in various settings, including cases where the sender is not benevolent or the receiver interprets data correctly.

Without the narrator, the sender would choose the full disclosure model, which reveals the state completely. However, in the presence of a biased narrator, this model can be suboptimal. In the case of binary states and a narrator whose utility does not depend on the state, it is still optimal when the narrator does not choose a conflicting model to interpret the data. This results in the receiver's belief in the true state being higher than before, even if it is not exact.

I apply this framework to two settings. First, 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 might even be worse than providing no information. Notably, even a voter initially opposed to the policy can be persuaded to support the policy, irrespective of the data. The politician achieves this by providing two models: one based on scientific theory (supportive data) and one based on doubt and skepticism (opposing data). The prior belief of the voter influences which interpretation she finds to be 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 choose the full disclosure model. The politician cannot devise any interpretation to convince the voter to support the policy. (ii) If the voter initially slightly opposes the policy, the researcher should choose a partially informative model.

[1] 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

The optimal model has the same fit as the best model the politician could use to convince the voter to support the policy. Choosing a more informative model would allow the politician to misinterpret opposing data. (iii) Finally, if the voter initially favors the policy, any model by the researcher cannot prevent the politician from convincing the voter to support the policy.

Next, I consider the example of a manager providing feedback to an optimistic employee about her ability. The interpretation is clear: positive feedback boosts her confidence about her ability, and negative feedback does the opposite. However, the employee is uncertain about the precision of the feedback. I illustrate how even a small amount of uncertainty can lead to biased learning and polarization. The employee distorts her own beliefs by making the positive feedback seem more informative than the negative ones (Eil and Rao, 2011). Despite observing an infinite sequence of feedback, she incorrectly concludes that her ability is good even when it is not. 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 true 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 the asymmetric interpretation by the optimist and ensures that she learns her true ability.

I explore three extensions to my setting. First, I impose a natural restriction on the set of models, where the receiver correctly interprets the most likely state that generates the data. This reduces the narrator's ability to persuade. While the narrator cannot change the direction of the belief, he can alter the precision. I demonstrate that he can induce any belief with a lower probability of the most likely state than the sender's model. Second, I assume that the narrator can choose multiple models (or interpretations) before the data is generated. This makes his models more credible, but also adds additional constraints. His models compete not only with the sender's model, but also among themselves. Nevertheless, I show that the timing of interpretation (ex-post versus ex-ante) does not impact his ability to persuade. Third, I consider the setting where the receiver treats the models of the sender and the narrator asymmetrically. I show that as trust in the sender's model increases, the narrator's persuasiveness decreases.

1.1 Literature review

Bayesian Persuasion: My work contributes to the literature on Bayesian persuasion.[2] The literature of Bayesian persuasion 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 by Kamenica and Gentzkow (2011) and further generalizations such as Alonso and Câmara (2016), de Clippel and Zhang (2022), Ball and Espín-Sánchez (2021).

On the theoretical front, I come up with a technique to search for the optimal model for non-Bayesian receivers and restricted set of models. In my setting, the receiver's action depends not

[2] My work more broadly contributes to the literature on information provision. This includes cheap talk (Crawford and Sobel, 1982), verifiable disclosure (Milgrom, 1981) and Bayesian persuasion (Kamenica and Gentzkow, 2011).

only on the posterior belief but also on the likelihood of the model. I define a preference over the space of experiments because the concavification technique (Aumann, Maschler, and Stearns (1995), Kamenica and Gentzkow (2011)) cannot be applied. The key takeaway is that, using the preferences of the receiver and narrator alone, one can restrict the search of the optimal model to a finite set. This technique can even be applied to general settings, where the sender is not benevolent and/or the receiver knows the true data-generating model. In a related paper, Ball and Espín-Sánchez (2021) also define preferences over experiments due to a restricted set of models. However, they assume a stylized binary model with rational receivers.

In a closely related paper, Ichihashi and Meng (2021) examines a stylized binary setup in which the same agent generates and interprets data with a restricted set of models. On the contrary, I consider two agents, one who generates data and another who interprets it, and impose no restriction on the set of models. Another related paper is by Eliaz, Spiegler, and Thysen (2021), where the sender strategically provides an accurate but coarse interpretation of the signal.

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. There have been different approaches to formalize narratives such as Blackwell experiments (or likelihood functions) (Schwartzstein and Sunderam (2021), Aina (2021), Yang (2023), Ispano et al. (2022), Izzo, Martin, and Callander (2023)), directed acyclical graphs (DAGs) (Eliaz and Spiegler (2020)) and moral reasoning (Bénabou, Falk, and Tirole (2018)).

This paper builds on the first approach. The receiver is exposed to more than one interpretation. Schwartzstein and Sunderam (2021) formalize that the receiver prefers models that best fit the observed data given her previous belief. Aina (2021) builds on this framework, analyzing a setting in which the persuader commits to a set of models before the data is observed. Yang (2023) adopts a different model selection rule. He assumes that the receiver prefers decisive models, which provide a strong recommendation and induce low regret. The literature assumes that the data-generating model is exogenous and fixed. My contribution is to consider a strategic and endogenous data-generating model.

Recent papers experimentally test the role of narratives in persuasion (Barron and Fries (2023), Kendall and Charles (2022)). In particular, Barron and Fries (2023) provide experimental evidence that individuals pick models with better fit in the context of financial advising.

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 to model both prior-based and preference-biased updating. A crucial aspect of the framework is that the receiver can use different models to update across different signals. This allows reconciling behavioral biases that are not consistent with a single model.

The closest papers in the literature are Braghieri (2023) and Frick, Iijima, and Ishii (2021). They analyze the welfare given a fixed data-generating model for biased updating in *all* decision problems. Braghieri (2023) provide 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 a fixed decision problem and finding the optimal data-generating model model that maximizes the receiver's welfare.

1.2 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. 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; the narrator also chooses a model, but to interpret this signal. A signal can be empirical data, evidence, or even a message. Finally, the receiver takes an action based on this signal and the two models. Each player i, where $i \in \{S,R,N\}$, has a utility function $u_i : \Omega \times A \to \mathbb{R}$ that depends on the state of the world $\omega \in \Omega$ and the receiver's action $a \in A$. The state is unknown to the players who share a common prior belief $p \in \text{int}(\Delta\Omega)$.[3][4]

A **model** $m : \Omega \to \Delta\mathcal{S}$ is a stochastic map that specifies the probability $m(s \mid \omega)$ of the observing signal $s \in \mathcal{S}$ conditioned on state $\omega \in \Omega$.[5] Let $\mathcal{M}$ denote the set of all models. Given a signal s, a model m induces posterior belief $q_s^m \in \Delta\Omega$, which is derived via Bayes' rule.[6] Following Aina (2021), I define the **fit** of model m given signal s as the (ex-ante) likelihood:

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

I assume that the set of states Ω, the set of actions A and the set of signals $\mathcal{S}$ are finite, where $|\mathcal{S}| \geq |\Omega|$. Let $\mathcal{F} \subseteq \mathcal{M}$ denote the set of feasible (or allowed) models, which is assumed to be closed and convex. I assume every model is feasible ($\mathcal{F} = \mathcal{M}$), unless stated otherwise. These assumptions together imply that the sender can fully disclose the state if she wants.

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

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

[5] The term "model" is also referred to as information structure, likelihood function, Blackwell experiment or information policy in the literature. Also, keeping with the literature, I stick to the term signal instead of data.

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

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

whenever Bayes'rule is applicable.

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 the 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 the 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 the observed signal s:[7]

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

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

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

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.[8]

For a given signal s, the 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-generating model is 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:

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

I investigate the behavior of a biased narrator who strategically select models (or interpretations) to maximize his expected utility. In contrast, I assume that the sender is benevolent, that is, $u_S = u_R$. The sender's objective is to choose a signal-generating model that maximizes 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.[9] However, once she chooses a model, she updates her belief

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

[8] 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.

[9] The other two popular choices to capture uncertainty over models are the fully Bayesian approach and the maxmin approach.

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.[10] Given a set of models, the receiver selects the model that best fits the observed signal given her prior. Barron and Fries (2023) provide experimental evidence for this assumption. I assume the receiver cannot come up with her own model. She may deem some models infeasible but her choice is confined to the exposed models.[11] I assume 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 4.3, I relax this condition and allow the receiver to trust the sender and the narrator in an asymmetric manner.

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.[12] Also, he provides his model after the signal realization, whereas the sender chooses his model before the realization. I show, in Section 4.2, that my results do not depend on whether the narrator provides his interpretations *before* or *after* observing the signal.

Finally, I assume that the sender can only interpret using the signal-generating model.[13] I do not allow her to generate the signal using one model and to interpret the signal using a different one. If she could, the receiver's expected utility would be (weakly) higher (see Ichihashi and Meng (2021)). My aim is to analyze the contrasting impacts of signal provision and signal interpretation in persuading a decision-maker.

3 Main Results

In this section, I state and prove my main results. The equilibrium of the game is determined through a backward induction approach. 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 misinterpretation

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

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

[11] 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)).

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

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

Denote F_s^I and A_s^I as the sets of feasible posterior beliefs and actions that the narrator can induce, given the sender's model I and the signal s. Incorporating all the signals, I denote the set of feasible vectors of posterior beliefs and actions as $F^I = (F_s^I)_{s\in\mathcal{S}}$, and $A^I = (A_s^I)_{s\in\mathcal{S}}$ respectively. The narrator's model is selected only if it better fits the observed signal than the sender's model does. 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, the set of feasible posterior beliefs and actions conditional on the signal s are:*

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

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

The key argument is that for any belief q, a model with maximal fit exists. The maximal fit is given by the left term in equation (5), that is, $\min\limits_{\omega\in\Omega} \frac{p(\omega)}{q(\omega)}$. The narrator cannot propose a model that induces the posterior belief q and simultaneously has a better fit than this particular model. The narrator can induce this belief if its maximal fit surpasses that of the sender's model; otherwise, he cannot. Proposition 1 of Schwartzstein and Sunderam (2021) can be applied to my setting to characterize the feasible posterior beliefs obtained in Lemma 1.

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. Also, he 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 different models for different signals, he can even induce vector of beliefs in the same direction. This is impossible if the receiver used a single model, even if it is incorrect, to interpret all the signals.

3.2 Binary Example: Graphical illustration

In this section, I present a graphical illustration of the extent of persuasion by the narrator, exemplified by the interactions among a researcher (sender), a politician (narrator), and a voter (receiver).

Consider two states: $\Omega = \{G, \neg G\}$, where G and $\neg G$ are the states where the policy is good and bad, respectively. The researcher provides evidence $\mathcal{S} = \{E, \neg E\}$ that supports the policy (E) or opposes it $(\neg E)$. The voter has two choices, $A = \{X, V\}$, where she can vote against (X) or for the policy (V). The utility function of the politician and the voter is given by the following matrix:

		Actions	
		X	V
States	G	$(0,1)$	$(1,2)$
	$\neg G$	$(0,1)$	$(1,0)$

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

The voter only wants to vote for the policy when the state is likely good, that is, when her belief is $q(G) \geq \frac{1}{2}$. Otherwise, she prefers to vote against. The narrator, regardless of the state, always wants her to vote for the policy.

The researcher and the politician persuade the voter to take a specific action. This action depends on her posterior belief over the states. So, focusing on the induced vector of posterior beliefs rather than the model provides useful insights. 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 conditional on the evidence $\neg E$ and E. Each point in this graph is a vector of posterior beliefs (see Fig. 1). I represent the prior belief $p = \mathbb{P}(G) = 0.4$ as the vector of posterior beliefs where the posterior belief is equal to the prior belief (black point). A vector of posterior beliefs $(q_{\neg E}(G), q_E(G))$ are *Bayes plausible* if and only if either: (i) $q_{\neg E}(G) \geq p(G) \geq q_E(G)$ or (ii) $q_{\neg E}(G) \leq p(G) \leq q_E(G)$. This condition ensures that there is a model such that the expected posterior belief equals the prior belief. This prevents updating belief in the same direction. The set of all models $\mathcal{M}$ represents the vectors of posterior beliefs that satisfy the Bayes plausibility condition (green area).

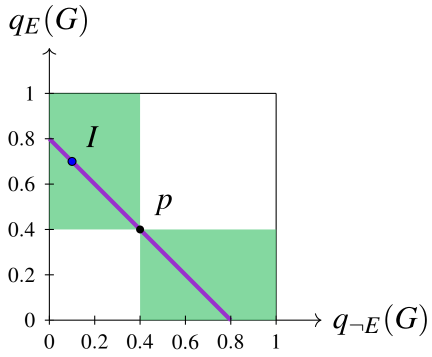


Figure 1: The set of all Bayes-plausible vector of beliefs $\mathcal{M}$.

Suppose that the researcher chooses the model I (blue point): supportive evidence is likely generated under the good state and vice versa for the opposing evidence. Formally, $I(E \mid G) = I(\neg E \mid \neg G) = \frac{7}{8}$. Consider the purple line that passes through the vector of posteriors of model I (blue point) and the prior p (black point). All models in this line have the same fit as the researcher's model I on both evidences. The steeper (flatter) the slope of this line, the better the fit given evidence $\neg E$ (E). The set of all models $\mathcal{M}$ can be partitioned 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 $\neg E$ (blue dotted area) and (ii) models with a better fit on evidence E (red area).

What can the politician do to persuade the voter? If the evidence is $\neg E$, the politician can choose any model $n_{\neg E}$ in the blue dotted area and the receiver will choose it over the researcher's model I. Given evidence $\neg E$, the politician can induce any posterior belief on the blue line on the x-axis. This blue line is the projection of the blue dotted area on the x-axis. This also includes beliefs higher than the prior belief, which is opposite to the intention of the researcher's model. Similarly, if the evidence is E, the narrator can choose any model n_E in the red area and the receiver will choose it over the researcher's model I. Given evidence E, the politician can induce any posterior belief on the red line on the y-axis. This red line is the projection of the red area on the y-axis. The set of posterior beliefs satisfying equation (5) in Lemma 1 are precisely represented by the blue and red lines for the evidence $\neg E$ and E, respectively

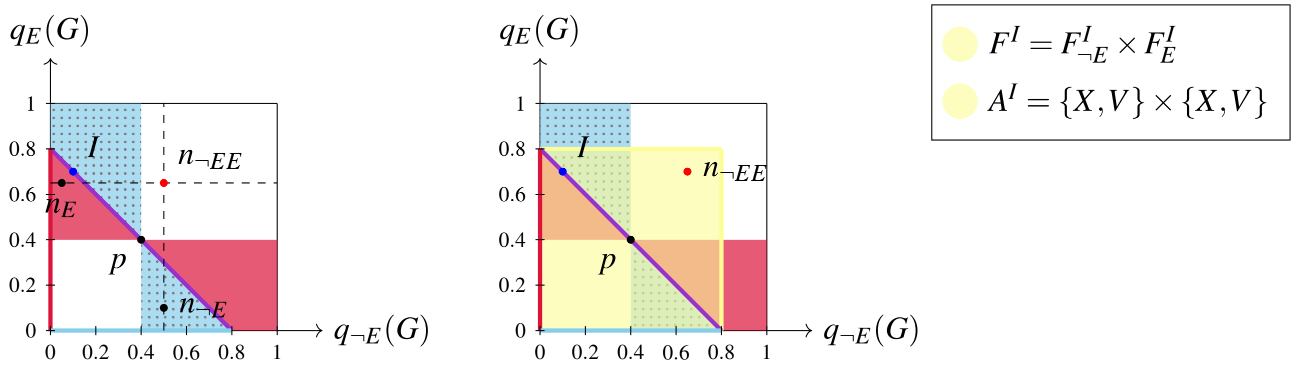


Figure 2: The set of feasible posterior beliefs and actions that the politician can induce.

From an ex-ante perspective, the politician can induce any vector of posterior beliefs in the yellow area, given by the Cartesian product of the green and blue lines. This set depends only on the fit of the researcher's model. All models that lie on the purple line have the same fit, and as a result, they generate an identical vector of beliefs within the yellow area. Crucially, the narrator can also induce vectors of posterior beliefs that are not Bayes plausible, i.e. outside the red area in Fig. 1. In fact, under this model, the politician can convince the voter to take any action under any evidence. For example, by providing the models $n_{\neg E}$ and n_E given evidence $\neg E$ and E respectively, he can induce the vector of posterior beliefs $n_{\neg EE}$ (red point in Fig. 2). The voter is more convinced that the policy is good, under both supportive and opposing evidence, compared to her prior belief.

The politician successfully influences the voter to support the policy using his interpretations of the evidence. The voter ends up voting for the policy, with probability 1, regardless of whether the policy is good or bad. 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 show the optimal signal-generating model can be found by searching within a finite set of models. I identify this set by defining a value function on the set of models.

Let $\boldsymbol{a}^I = (a_s^I)_{s\in\mathcal{S}} \in A^{|\mathcal{S}|}$ denote the *vector of induced actions* when sender chooses the model I. If the sender chooses model I and signal s is observed, the narrator selects an action from the set of feasible actions A_s^I to maximize his expected utility. This expected utility is computed using the posterior belief resulting from signal-generating model I.[14] I have

$$a_s^I := \underset{a\in A_s^I}{\arg\max}\, \mathbb{E}_{q_s^I}[u_N(\omega, a)]. \tag{7}$$

The induced action a_s^I results from the narrator's optimal model in response to the sender's model I. Now, I 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 parts: 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 component of information provision is the value to the receiver of using the true posterior belief over the prior belief. It is the focal point in the Bayesian persuasion literature and always has a non-negative value. It assumes the receiver knows the signal-generating model. The component of information misinterpretation is the value of the receiver when she acts based on her belief under the selected model instead of the sender's model. This component always has a non-positive value. If the narrator's model is selected and it lead to a different action, the value is negative. 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

[14] If the narrator has multiple optimal actions, I break the tie by choosing the receiver's preferred action.

within each set. I denote by $\boldsymbol{C_a} \subseteq \mathcal{F}$ the set of sender's models where the vector of induced action 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.[15] Due to the linearity of the value function for fixed vector of actions, I can restrict the search of the optimal model within each set $\overline{C}_{\boldsymbol{a}}$ to its extreme points.[16] 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 setup). 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 prior belief. This technique to find the optimal signal-generating model is applicable in many settings, for instance, even when the sender is not benevolent and 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 an interior candidate model is optimal, opting for a more informative model results in increased misinterpretation. It changes the induced vector of actions in a way that is worse for the receiver. Had this not been the case, the interior model would not be optimal.

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 $\omega \in \Omega$. All no

[15] 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.

[16] Lipnowski and Mathevet (2017) use the same property to identify candidate beliefs rather than candidate models.

disclosure models map to the same vector of posterior beliefs, where the posterior belief given any signal equals the prior belief. The model assures that the sender's value of information under the optimal model is non-negative, that is, $V(I^*) \geq 0$. When the preferences of the narrator and receiver are perfectly misaligned, akin to a zero-sum game, I show that the no disclosure model is optimal. Furthermore, in this setting, any optimal model induces the unique action $a_R^*(p)$, the receiver's best response under her prior belief.

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 $a_R^*(p)$.*

This model sends the signal s with probability 1 and results in the posterior belief $q_t^{I_{NDs}} = p$ for all $t \in \mathcal{S}$. The model I_{ND_s} has the maximal fit among the set of all models for signal s. This model provides no information and subsequently there is no scope for misinterpretation, that is, $V(I_{ND_s}) = 0$ for all $s \in \mathcal{S}$. In case of perfectly misaligned preferences, any information provided by the sender will be misinterpreted by the narrator and so the no disclosure model is optimal.

If there were no narrator, the benevolent sender prefers to fully disclose the states. The more information the receiver has, the better action she can take. 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.[17] 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, $\Omega \subseteq \mathcal{S}$. The full disclosure model always belongs to the set of candidates models. Formally, the model $I_{FD} \in \mathcal{M}$ fully discloses (or reveals0 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 belief on the true state is higher than the prior belief. 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 optimal for the receiver under some (interior) belief and that the narrator is not indifferent between them.

Corollary 2. *For 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, let us suppose 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 it's proximity to the

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

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. This implies that fully discloseing the state cannot be optimal

3.4 Application: Information campaigns

Experts use information campaigns to influence health behavior, policy attitudes, and voter turnout (Haaland, Roth, and Wohlfart, 2023). However, in some cases the campaign not only fails to provide information but can even increase misperception.[18] In particular, I show that information can backfire, that is, the receiver may shift his beliefs in the opposite direction as intended. The *backfire effect* has been observed empirically in information campaigns (see Nyhan and Reifler (2010), Hart and Nisbet (2012), Baysan (2022)). In particular, Baekgaard et al. (2019) provide experimental evidence that (i) the further the prior belief is from the intended posterior belief, the greater the chance of misinterpretation and that (ii) increasing the amount of information can lead to a higher chance of misinterpretation.

I characterize the The set of all models can be partitioned based on the vector of induced actions (Fig. 3). Using Theorem 1, I only need to search among the finite set of extreme points for each set C_a (black nodes). I focus solely on the models in the top left quadrant, as the models in the bottom right quadrant can be derived by simply swapping the labels E and $\neg E$. In total, I only need to search among the six candidate models. Note, the full disclosure model I_{FD} and the no disclosure model I_{ND} belong to this set.

If there was no politician, the researcher would choose the full disclosure model, that is, $I_{FD}(E \mid G) = I_{FD}(\neg E \mid \neg G) = 1$. However, when evidence can be misinterpreted, it is suboptimal to do this. Given opposing evidence $\neg E$, the narrator can choose the model $n_{\neg E}$ (blue point) where $n_{\neg E}(\neg E \mid G) = 1$ and $n_{\neg E}(\neg E \mid \neg G) = \frac{2}{3}$. This model has better fit than the full disclosure model for opposing evidence: $\mathbb{P}_{I_{FD}}(\neg E) = \frac{3}{5} < \frac{4}{5} = \mathbb{P}_{n_{\neg E}}(\neg E)$. This can also be seen graphically (see Fig. 3) as the purple line passing through the points $n_{\neg E}$ and p is steeper than the dashed line passing through the points I_{FD} and p. The model $n_{\neg E}$ induces a posterior belief of 0.5 for the opposing evidence, that is, $q_{\neg E}^{n_{\neg E}}(G) = 0.5$. Thus, the narrator persuades the voter to choose the action V. Infact, the full disclosure model performs worse than providing no information, that is, the vector of belief p. 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.

I have shown that fully discloseing the states is not optimal. How should the sender provide evidence? The sender chooses a partially informative model that better fits the opposing evidence. Under this evidence, the preferences of the researcher and the politician are misaligned. The (unique) optimal signal-generating model I^* (black star) is given by:

[18] 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.

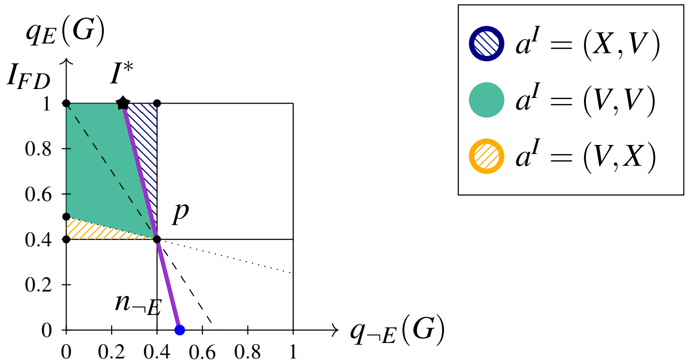


Figure 3: Partition of the set of models based on the vector of induced actions.

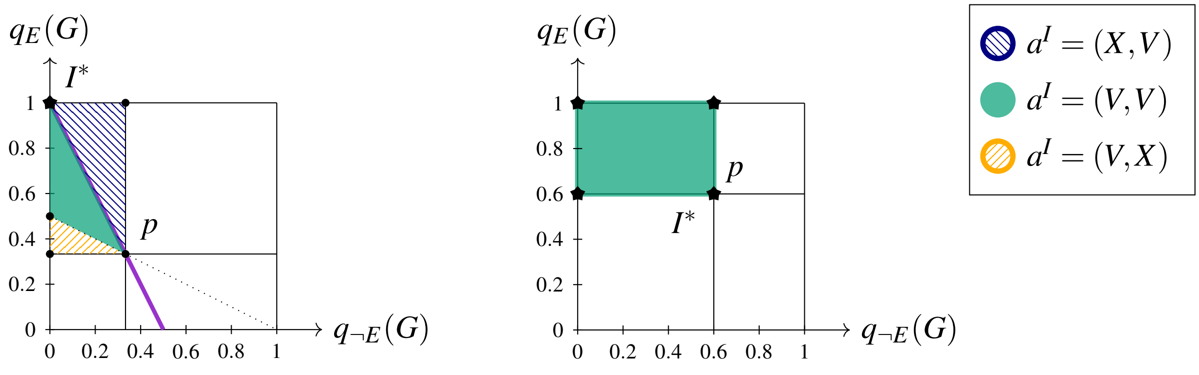
$$\begin{aligned} I^*(\neg E \mid \neg G) &= 1, & I^*(\neg E \mid G) &= \frac{1}{2}, \\ I^*(E \mid \neg G) &= 0, & I^*(E \mid G) &= \frac{1}{2}. \end{aligned}$$

The model $n_{\neg E}$ has the maximal fit among all models that induce belief $q_{\neg E}(G) = 0.5$. The fit of the optimal model I^* and the model $n_{\neg E}$ exactly match: $\mathbb{P}_{n_{\neg E}}(\neg E) = \frac{4}{5} = \mathbb{P}_{I^*}(\neg E)$. This can also be inferred from the figure, 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 given opposing evidence than the optimal model and that persuades the voter to support the policy. All the vectors of beliefs that lie on the line segment joining I^* and I_{FD} represent models that are more informative than the model I^*. If the researcher chose a more informative model, the politician could misinterpret the opposing evidence. 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, I^* 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 pools that state G with the state $\neg G$ where their preferences are misaligned. This pooling is to prevent the voter from misinterpreting the opposing evidence. She pools the states to make the opposing evidence more plausible under her model.

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 belief: (a) full disclosure for low prior belief, (b) partially informative for mid prior belief, and (c) no disclosure for high prior belief.

For low prior belief, $(p \leq \frac{1}{3})$, the full disclosure model I_{FD} is optimal. Given opposing evidence $\neg E$, 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 belief for which the full disclosure model I_{FD} lies on the line passing through the prior belief and the model $n_{\neg E}$ (purple line in Fig. 4).



(a) Low prior belief $(p = \frac{1}{3})$ **(b)** High prior belief $(p = 0.6)$

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

For a high prior belief $(p \geq 0.5)$, any signal-generating model including the no disclosure model is optimal. Given any evidence e, the politician can always choose the no disclosure model that sends the evidence with probability 1, that is, $U_e(e \mid G) = U_e(e \mid \neg G) = 1$. This model has the maximal fit in the set of all models and it keeps the voter's belief fixed at the prior belief. Essentially the politician can always confirm the voter's prior. As the prior belief is greater than 0.5, the voter's optimal action is to vote for the policy. Thus, the politician can always persuade the voter to support the policy, irrespective of the researcher's model.

4 Extensions

Now, I introduce three extensions to the base model: a restricted set of feasible models, ex-ante interpretation and asymmetric trust.

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 and moving 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

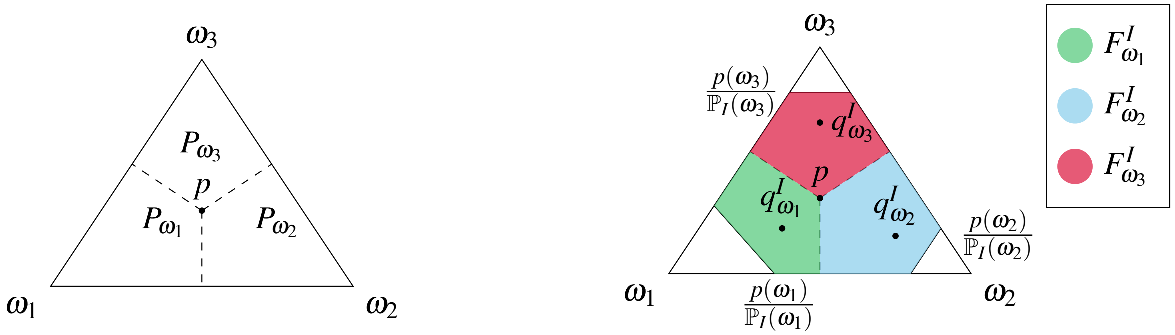


Figure 5: (a) The partition of the belief space and (b) the feasible set of posterior beliefs given sender's model I.

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 their set of feasible models. Now, I characterize the set of feasible beliefs and actions when the narrator is restricted to choosing models in $\mathcal{M}_C$.

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

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

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

The narrator can induce any belief that is less precise than the true belief. However, this belief still has a higher belief on the true state than before. In a way, 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 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 provided that the set of feasible models are 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 signal, the full disclosure model would be optimal for all prior beliefs. The politician would not be able to convince the voter otherwise. One might hope that in this setting of clear interpretations, the full disclosure experiment would always be optimal. However, even though the extent of misinterpretation is severely limited, it can lead to bad outcomes. I exemplify this in the following application.

4.1.1 Application: Providing 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.[19] Finally, I show that giving good feedback more frequently than bad 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. 6):

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

where, κ 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 $[\kappa - \varepsilon, \kappa + \varepsilon]$ (see Fig. 6):

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

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.

[19] 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)).

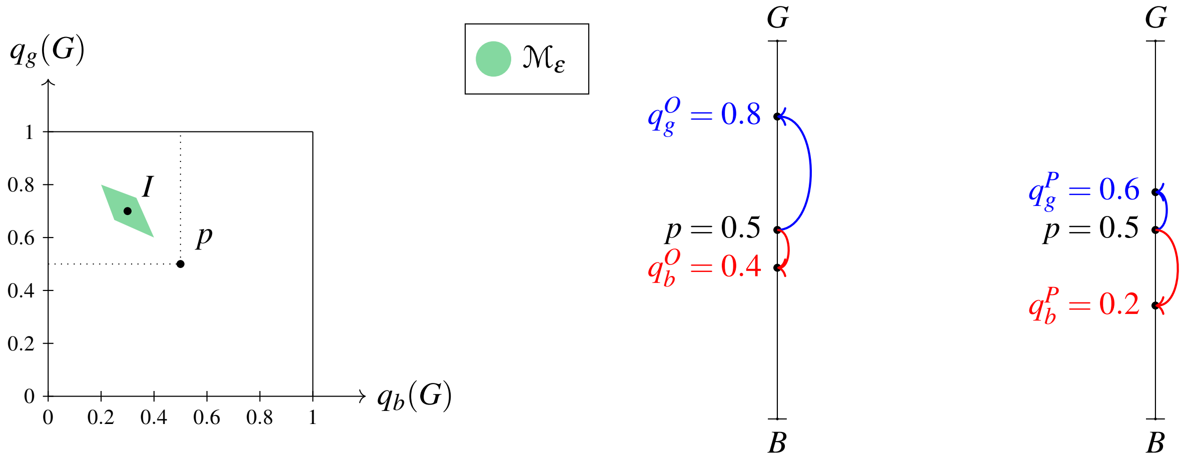


Figure 6: (a) The set of feasible models $\mathcal{M}_\varepsilon$ (b) Biased updating by the optimist and the pessimist ($p = 0.5, \kappa = 0.7, \varepsilon = 0.1$).

I assume that the employee is an **optimist** 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). (see Fig. 6)

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.[20,21] The narrator wants to interpret the news in the direction of his biased 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: $\mathbb{P}_{n_g}(g) \geq \mathbb{P}_I(g)$.[22] The employee interprets the news to be very informative and overreacts to it (see Fig. 6).

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

[20] Formally, one can assume the narrator (unconscious mind) has a belief based utility $\hat{u}_N(q)$ which is a increasing function in his belief $q(G)$.

[21] Note, the results of Theorem 1 cannot be applied here directly as the set of actions A is not finite. But since my focus is on asymptotic learning, the receiver's belief converges to either δ_G or δ_B.

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

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

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. 6).

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 at each round she chooses how to interpret each piece of news. The sequence of previous news only affects her prior belief for that round. I show if the amount of uncertainty ε is sufficiently large, the 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*

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

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, he 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).[23] One can also imagine two news outlets that can twist the same news story to support their preferred political position. I show that the two employees despite having the same prior belief 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. 6).

When presented with a balanced set of good news and bad news, the employees' beliefs are polarized, that is, $q^O_{gb}(G) > \frac{1}{2} > q^P_{gb}(G)$ where q^O_{gb} and q^P_{gb} refer to posterior belief of the

[23] (Strunk, Lopez, and DeRubeis, 2006) show that individuals suffering from depression tend to exhibit pessimism bias.

optimist and pessimist after observing g and b, respectively.[24] 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 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 model $I^* \in \mathcal{M}_\varepsilon$:*

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

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.[25] 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.

4.2 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 communication while Aina (2021) analyzes ex-ante communication. For example, the receiver might be skeptical if a model is provided after the signal has been observed. I assume that the narrator can provide multiple models instead of a single one in this setting. Ex-ante interpretation poses an additional constraint for the narrator, as each model competes not only with the sender's model but also with the other models that he provided. I show that the narrator can attain the same vector of posterior beliefs and actions with ex-ante or ex-post provision of models. Thus, my results do not depend on whether the narrator communicates the models before or after the signals have been observed.

[24] The order of news does not affect the posterior belief: $q^i_{gb} = q^i_{bg}$ for $i = O, P$.

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

Let the menu of models the narrator provides be denoted by $\mathcal{N} = \bigcup_{s\in\mathcal{S}} n_s$.[26] The narrator does not need to provide more than $|\mathcal{S}|$ models. This is because the receiver selects one model for each signal. Thus, the narrator customizes a model n_s for 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.

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

Lemma 4 (ex-ante interpretation)**.** *Given sender's model I, the set of feasible posterior beliefs for each signal $s \in \mathcal{S}$ is given by*

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

$$A_s^I := \{a \in A : \exists q \in F_s^I \text{ such that } a \in \underset{a\in A}{\arg\max}\, \mathbb{E}_q[u_R(\omega, a)]\}. \tag{23}$$

Theorem 2 of Aina (2021) can be applied to my setting to characterize the set of feasible posterior beliefs. Surprisingly, the timing of model provision, ex-ante versus ex-post, does not impact the narrator's ability to persuade the receiver. The reason is that the model n_s is customized for signal s. This model induces the desired posterior belief q_s, and it has a worse fit than the model n_t for signal $t \neq s$. The narrator could use the same model n_s as in the main theorem but must adjust the fit for signals $t \neq s$, ensuring each n_s model has the best fit for its respective signal s. I show this is always possible and thus does not affect his ability to induce any target belief.

4.3 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:

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

[26] 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)

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. Conversely, when $\eta = 0$, the receiver always adopts the narrator's model.

If $\eta = 2$, then the voter trusts the researcher much more than the politician (Fig. 7). She only chooses the politician's model when it significantly better explains the signal than the researcher's model. In this case, the full disclosure experiment is optimal. When the voter sees supporting evidence, the narrator cannot mislead her in any way. However, if she sees opposing evidence, the narrator can attempt to persuade her, though his influence is limited (blue line). He cannot convince her to vote for the policy.

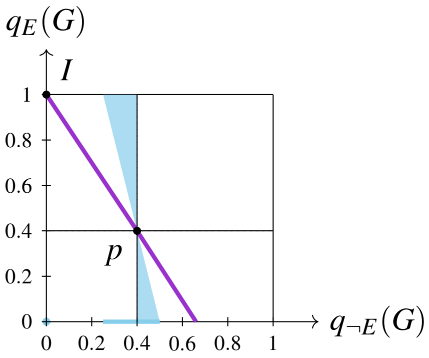


Figure 7: The set of feasible beliefs the narrator can induce given sender's model I and $\eta = 2$.

Given the sender model I and η, let $F^I_s(\eta)$ and $A^I_s(\eta)$ 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 $F^I_s(\eta_1) \subseteq F^I_s(\eta_2)$ and $A^I_s(\eta_1) \subseteq A^I_s(\eta_2)$ 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 is no longer a convex set, as it always includes the belief resulting from the signal-generating model.

5 Conclusion

A large focus of the economic literature on persuasion has been on the role of information provision where the decision maker understands how the data is generated. In such a setting, the value of information is always positive. But in many situations, the decision maker may not know how to interpret signals and can be persuaded by providing biased interpretations. I demonstrate in such a setting, fully discloseing information may backfire. When providing information, it is important to take into account the interpretations or narratives to which agents might be exposed. I show that agents are more likely to interpret information correctly if it aligns with their priors.

Several extensions could be explored in future research. First, I assume the sender cannot provide incorrect interpretations and that the narrator cannot generate an additional signal. One future direction could be to look at competition between agents who can both provide and interpret information. Another interesting direction could be to look at the impact of persuasion on a population of receivers with heterogeneous priors. For example, the narratives used by politicians are public and impact all voters. Also, an interesting direction could be to consider a different model selection rule. For example, one could even consider a convex combination of models proposed by the sender and the narrator, where the coefficient corresponds to the ratio of fit between 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.

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

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

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 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.

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.

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

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

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

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.

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.

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

A Proofs

Lemma 1. *Given the sender's model I, the set of feasible posterior beliefs and actions conditional on the signal s are:*

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

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

Proof. First, assume $q \in F_s^I$. 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 $\neg s \neq s$. 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 corresponding 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 $\boldsymbol{R} = (R_s)_{s\in\mathcal{S}} \in \mathcal{P}(A)^{|\mathcal{S}|}$.[27] 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 $\alpha \in (0,1)$.[28] 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

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

[28] 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 $s \in \mathcal{S}$.

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

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 some $\boldsymbol{R} \in \mathcal{P}(A)^{|\mathcal{S}|}$. 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. By the Bauer maximum principle, 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.

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

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.

□

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 fully informative 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 $\mathbb{E}_{q^I_s}[u_N(\omega,a^I_s)] > \mathbb{E}_{q^I_s}[u_N(\omega,a^*_R(p))]$.

However as the narrator's utility is perfectly misaligned with the receiver's, this implies the narrator's expected utility is negative, that is, $\mathbb{E}_{q^I_s}[u_N(\omega,a^I_s)] < \mathbb{E}_{q^I_s}[u_N(\omega,a^*_R(p))]$. 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.

□

Proposition 2. *For binary states and a narrator with state-independent utility, the full disclosure model I_{FD} is optimal if $u_R(\omega, a^{I_{FD}}_\omega) \geq u_R(\omega, a^*_R(p))$ 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., $q^{I^*}_\omega = \delta_\omega$ 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 between δ_{ω_0} and 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 $a_{\omega_1}^{I_{FD}} \notin A_{\omega_1}^{I^*}$. But this implies the action $a_{\omega_1}^{I^{FD}}$ lies is closer to the state ω_1 than the action $a_{\omega_1}^{I^*}$. But then the sender would like to provide more information atleast till the point that action $a_{\omega_1}^{I_{FD}}$ is induced. 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, the sender prefers the most informative model. Thus, I have shown that any model I^* that does not fully disclose cannot be the optimal model.

□

Corollary 2. *For 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 set of feasible posterior beliefs and actions that a narrator can induce given sender's model $I \in \mathcal{M}_C$ and the signal ω is given by*

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

$$A_\omega^I := \{a \in A : \exists q \in F_\omega^I \textit{ such that } a_R^*(q) = a\}. \tag{14}$$

Proof. Assume $q \in F_\omega^I$. 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, $q_\omega^n = q$ and $\mathbb{P}_n(\omega) > \mathbb{P}_I(\omega)$. The model n only sends two signals ω and $\tilde{\omega}$ with positive probability, where $\tilde{\omega} \neq \omega$.

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

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.

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

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 $q_\omega^n = q$ 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 $q_\omega^n = q$ 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*

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

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 model $I^* \in \mathcal{M}_\varepsilon$:*

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

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

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

However, this happens when the following condition holds:

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

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 sender's model I, the set of feasible posterior beliefs for each signal $s \in \mathcal{S}$ is given by*

$$F_s^I := \{q \in (\Delta\Omega) : \frac{p(\omega)}{q(\omega)} > \mathbb{P}_I(s) \quad \forall \omega \in \Omega\}, \tag{22}$$
$$A_s^I := \{a \in A : \exists q \in F_s^I \text{ such that } a \in \arg\max_{a \in A} \mathbb{E}_q[u_R(\omega, a)]\}. \tag{23}$$

Proof. First, assume $q = (q_s)_{s\in\mathcal{S}} \in F^I$. 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 $\lambda_s = [\max_{\omega\in\Omega} \frac{q_s(\omega)}{p(\omega)}]^{-1}$.

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

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

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

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

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

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

Now, assume $\boldsymbol{q} \notin F^I$. 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 $F_s^I(\eta_1) \subseteq F_s^I(\eta_2)$ 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 $q \in F_s^I(\eta_1)$. I will show that $q \in F_s^I(\eta_2)$. As $q \in F_s^I(\eta_1)$, $\exists n$ 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, $q \in F_s^I(\eta_2)$. This also implies that if any action a belongs to $A_s^I(\eta_1)$ then it also belongs to the set $A_s^I(\eta_2)$.

$\square$