Misinterpreted Persuasion

Mengxi Sun*

August 2025

link to current version

Abstract

This paper studies a behavioral model of persuasion where the receiver may misinterpret information designed by the sender. Instead of assuming perfect observation, we allow the receiver to confuse one message for another with some probability. This misinterpretation limits the sender's ability to influence beliefs and disrupts the standard concavification method of Kamenica and Gentzkow (2011). We also study a second departure in non-Bayesian updating: naïveté, where the receiver fails to account for misinterpretation. Misinterpretation weakly harms the sender without affecting the receiver, but naïveté can benefit the sender at the receiver's expense. In a binary example, we illustrate how misinterpretation reduces the sender's optimal payoffs, and when naïveté lowers the receiver's demand for information. We then extend the analysis to confirmation bias, where misinterpretation depends endogenously on beliefs and strategy. Our analysis complements prior work on non-Bayesian inference (de Clippel and Zhang, 2022) and symmetric noise (Tsakas and Tsakas, 2021), by focusing on structured misunderstanding in persuasion.

*Ph.D. candidate, Department of Economics, University of Pittsburgh. Email: mengxisun@pitt.edu. Personal website: mengxis.github.io

1 Introduction

Communication hinges not only on what is said, but also on how it is understood. A policymaker, investor, voter, or juror may be presented with carefully designed information, yet reach an unintended conclusion. This paper studies a sender-receiver communication game in which such misunderstandings arise: the sender chooses how to present information, but the receiver may perceive a message different from what was intended.

We build on the canonical model of Kamenica and Gentzkow (2011) (henceforth, KG), in which a sender (she) persuades a receiver (he) by designing an information policy—a mapping from states to distributions over messages. In the standard model, both observe the realized message as intended. Here, we relax this assumption: the sender observes the true realization, but the receiver may misinterpret it. Each informatively distinct message has some chance of being confused for another, introducing variation in how the receiver interprets what has been realized.

Such disagreement over outcomes often arises because irrelevant factors intrude on our information processing. For example, when complex ideas resist simple labeling, consensus can be elusive. Jurors may misunderstand testimony; voters may misread policy effects; investors may diverge in what an analyst's report indicates. In other cases, interpretive heterogeneity stems from biases tied to social identity, preferences, or beliefs. Stereotypes skew individual assessments based on group impressions; motivated reasoning lets preferences mold evaluation of new information; confirmation bias colors perception through the lens of what one already believes. Regardless of source, understanding how probabilistic misinterpretation alters strategic communication has broad implications.

Misinterpretation unravels a core tool used to characterize optimal persuasion: the concavification technique introduced by KG and extended by de Clippel and Zhang (2022) to accommodate a variety of non-Bayesian updating rules in the literature. In their framework, the sender's payoff from inducing each posterior belief is independent of the other beliefs that she might induce. By perturbing posterior beliefs, this clean separation no longer holds: the

value of a belief depends on how other messages may be misinterpreted as it. Hence, even slight misinterpretation can sharply reduce the sender's persuasive power.

We formalize misinterpretation as a probabilistic mapping over signal realizations, represented by a row-stochastic matrix. The receiver's effective information is a noisy version of what the sender designed. We consider two receiver types: sophisticated, who accounts for misinterpretation when updating beliefs, and naïve, who ignores it. The naïve receiver exhibits a second deviation—naïve misspecification—on top of misinterpretation. Unlike misinterpretaton, naïveté can sometimes help the sender, by reducing how much evidence the receiver needs to change actions. Our framework accommodates other forms of misspecification (or non-Bayesian updating[1]). We focus on naïveté as it is especially relevant to our primary departure. Decomposing into misinterpretation and misspecification allows us to isolate welfare effects due to each behavioral assumption.

Our main analysis proceeds in two components: (1) misinterpretation structurally transforms information environments, and (2) naïve misspecification systematically distorts posterior beliefs. First, we examine how misinterpretation affects which posterior distributions the sender can induce. While misinterpretation breaks the usual concavification approach, we show that the belief-based approach still applies. Misinterpretation reduces the sender's optimal persuasion value by constraining implementability and alters how the receiver should ideally update beliefs according to Bayes' rule.

Second, we introduce naïveté. In contrast to misinterpretation limiting implementable posterior distributions, naïveté restores the set of implementable beliefs. It systematically distorts[2] the receiver's posterior beliefs such that he demands less information to shift decisions. In some cases, this benefits the sender at the receiver's expense, through violation of the martingale property so that the sender can induce her preferred action more frequently, without revealing more information than in KG.

We illustrate these forces using the stylized Judge-Prosecutor example from KG, featur-

[1] as they are formally equivalent in distorting posterior beliefs; see Bohren and Hauser, 2022

[2] as per de Clippel and Zhang (2022), but is still outside their scope because this systematic distortion is with respect to the receiver's Bayesian posterior beliefs, not the sender's.

ing two states and two signal realizations[3]. We characterize the sender's optimal strategy under both sophistication and naïveté, and show how the interaction between misinterpretation and naïve misspecification shapes persuasion outcomes. We then extend the binary example to confirmation bias, where the way messages are misinterpreted depends on what the receiver already believes and what the sender induces the receiver to believe. This extension captures a richer form of misinterpretation observed in practice and demonstrates that our core insights are robust to endogeneity.

Suppose a lawyer wants to persuade a jury to acquit her client. Both the lawyer and the jury begin with a common prior on the suspect's innocence. While the lawyer always prefers acquittal, the jury aims to acquit the innocent and convict the guilty. Information is conveyed through the realization of an expert's testimony. In KG, the lawyer can significantly improve the likelihood of acquittal by committing to an information policy. For instance, the jury optimally acquits 60% of suspects despite knowing that only 30% are innocent—doubling the acquittal rate relative to prior beliefs. We use the KG outcome as benchmark.

First, we consider a sophisticated receiver. Misinterpreting the testimonies always weakens the lawyer's ability to steer beliefs and thereby reduces the value of persuasion, but how it reduces value differs in the direction:

- Favorable misinterpretation (e.g., the jury misinterprets a guilty testimony as exculpatory) limits the lawyer's ability to push beliefs toward innocence. It shuts down the persuasion channel entirely at low priors, where shifting the jury to acquittal requires substantial evidence. Anticipating this interpretive noise, the sophisticated jury becomes skeptical of seemingly innocent testimony and demands more evidence to acquit. At low priors, this demand explodes: the amount of information demanded for acquittal scales up skepticism for innocent testimony, and persuasion fails when full revelation does not suffice. For example, with 20% chance of misinterpreting the guilty testimony, the lawyer's value from persuasion drops from 30% in KG to 0 at a prior

[3]The results shown in the main section assume *infrequent* misinterpetation without loss of generality. The appendix includes results for *frequent* misinterpretation, comparative statics, and some discussion on policy implication in debiasing effort.

belief of 15% in innocence. At priors where persuasion remains feasible, favorable misinterpretation doesn't bite directly.

- Unfavorable misinterpretation (e.g., the jury misinterprets an innocent testimony as incriminating) limits her ability to push beliefs toward the guilty state. This weakens the guilty testimony, and thereby erodes the lawyer's ex-ante expected payoffs. The sophisticated jury becomes skeptical of testimony indicating guilt, but the lawyer cannot compensate by providing more information because the guilty testimony already reveals the guilty state in KG. For instance, with 20% chance of misinterpreting the innocent testimony, the lawyer's value from persuasion falls from 60% in KG to 48%, at a prior belief of 30% in innocence.

Despite these effects on the sender's welfare and equilibrium, the sophisticated receiver continues to make optimal decisions and remains indifferent in equilibrium.

Next, we analyze a naïve receiver who not only misinterprets but also misspecifies. Naïveté helps the lawyer only with favorable misinterpretation, since the jury overestimates the informativeness of exculpatory testimony and thus requires less evidence to acquit. It leads to a violation of the martingale property, and hence the lawyer can induce acquittal more often by revealing the same amount of information as in KG's optimal policy. However, with unfavorable misinterpretation, naïveté doesn't benefit. Although it still restores implementable beliefs, the martingale violation now works against the sender and cancels out this advantage.

Lastly, we extend the binary example to confirmation bias, where the direction of misinterpretation depends on the receiver's prior belief and the sender's strategy—endogenizing the interpretation process. When persuasion requires substantial evidence (i.e., at low priors), confirmation bias undermines the sender, and naïveté offers no advantage. By contrast, at high priors, where biased updating steers beliefs toward acquittal in equilibrium, naïveté helps the sender get even higher payoffs than in KG. Moreover, it broadens the range of priors over which favorable misinterpretation can be exploited.

Related Literature

Before laying out model, we highlight our contribution in relation to the existing literature.

We contribute to the persuasion literature by considering behavioral deviations that introduce interdependence across posterior beliefs in the support. In de Clippel and Zhang (2022) and Alonso and Câmara (2016), sender and receiver may disagree in posterior beliefs, yet the receiver's posterior can still be expressed as a function of the sender's posterior belief induced by the same realization. KG's concavification technique remains robust to such deviations. One key property is that the sender's payoff from each posterior belief is independent of what else in the support. Hence, the sender's (possibly distorted) indirect utility function can still be evaluated pointwise in posterior beliefs at a given prior. Then, concavifying this function under Bayes-plausibility yields the sender's optimal value.

However, with misinterpretation, the concavification characterization fails even at arbitrarily small perturbations. A posterior belief's value to the sender can depend on any realizations that could be misinterpreted as the one inducing it. For example, suppose a guilty testimony fully reveals the guilty state for the lawyer, but the jury has a positive chance of mistaking it for an innocent testimony. Then, knowing the client being guilty for sure after a guilty realization, the lawyer has either 0 probability of acquitting if the jury doesn't misinterpret, or some positive probability of acquitting if the jury misinterprets. This interdependence within posterior support renders concavification unhelpful for determining the optimal value, despite the posterior beliefs satisfying Bayes-plausibility. Nevertheless, we can still use the belief approach by establishing connection between the sender's and the receiver's posterior distributions by row-stochastic matrices. Our welfare analysis of misinterpretation and naïve misspecification complements the welfare analysis of the systematic distortion as per de Clippel and Zhang (2022) by Bordoli (2024).

The most closely related paper is Tsakas and Tsakas (2021). Both papers study noisy perturbation in persuasion, but with different motivations and emphases. Tsakas and Tsakas (2021) model noise as implementation errors. If we think of the data-generating process that the sender commit to as a machine, they focus on a broken machine that adds symmetric

noise when spitting out information. Their sender benefits from complicating the signal to dilute the noise's impact across realizations inducing the same posterior. By contrast, we model misinterpretation. Here, semantic proximity matters: we treat synonyms as one, and focus on misinterpretation that only occurs across meaningfully distinct posterior beliefs. Our sender does not benefit from complicating the signal. But both models share the insight that noise hurts the sender.

Relatedly, Eliaz et al. (2021) study multidimensional persuasion motivated by real-world communication complexity. In their model, the sender designs a "decipher" that alters how messages are processed. Our sender, by contrast, is constrained by interpretive flexibility on the receiver's side. This flexibility undermines the sender's ability to induce receiver's posterior beliefs.

We also add to the growing literature on behavioral communication games that examine specific cognitive patterns, such as base-rate neglect (Benjamin et al., 2019), correlation neglect (Levy et al., 2022), and wishful thinking (Augias and Barreto, 2023). Our analysis of confirmation bias, as defined by Rabin and Schrag (1999), provides a tractable foundation for studying this widely studied phenomenon in a stylized setting. This framework serves as a stepping stone toward richer dynamics explored in the psychological literature (Lord et al., 1979; Plous, 1991; Darley and Gross, 1983) and in work by economists and political scientists (Klayman, 1995; Nickerson, 1998; Taber and Lodge, 2006; Del Vicario et al., 2017; Kim, 2015; Knobloch-Westerwick et al., 2020; Falck et al., 2014). Finally, our analysis offers a potential explanation for why generic debiasing interventions sometimes fall short in the field (Alesina et al., 2024). Misinterpretation personalizes the information environment: without individualized feedback, naïve misinterpretation makes it hard to steer behavior away from suboptimal actions.

The remainder of the paper is organized as follows. Section 2 introduces the model and main analysis. Section 3 presents a binary example and an extension to illustrate optimal persuasion and welfare implications of misinterpretation and naïve misspecification. Section 4 discusses relaxation of commitment and concludes.

2 Model

A sender (she) and a receiver (he) communicate about the state of the world drawn from a finite set $\omega \in \Omega$. They share a common prior belief $\mu_0 \in int(\Delta(\Omega))$. The receiver chooses an action $a \in \mathcal{A}$ that affects both the sender's and receiver's payoffs, given by continuous utility functions $v(a, \omega)$ and $u(a, \omega)$ respectively. The sender can influence the receiver's beliefs by designing and committing[4] to an information policy π before observing the state. The information policy encodes information in finite realizations $s \in \mathcal{S}$ generated according to conditional distributions $\{\pi(s \mid \omega)\}_{\omega \in \Omega, s \in \mathcal{S}}$.

Behavioral Departures

Unlike KG, the receiver may interpret a realized signal $s \in supp(\pi) \subseteq \mathcal{S}$ as a different one $\tilde{s} \in \mathcal{S}$. We denote the probability of misinterpreting an information policy π as $\gamma_\pi : supp(\pi) \to \Delta(\mathcal{S})$. This paper focuses on *errors of meaning*. That is, for any information policy π such that no two realizations in the support of π are sent with the same conditional probability in all state, there is a structural transformation of π, denoted as γ, representing how receiver misinterprets sender-designed information [5]. γ can be represented by a row-stochastic matrix Γ of size $|supp(\pi)| \times |\mathcal{S}|$, where each row is a probability distribution of interpreting a realized $s \in supp(\pi)$ as a receiver-perceived $\tilde{s} \in \mathcal{S}$.

The effective information environment for the sender is (π, s). Misinterpretation perturbs

[4] In Section 4, we discuss relaxing the commitment assumption. Commitment is not essential to our main results on misinterpretation, which operate in the belief space. The difference between with and without commitment is whether we allow the induced posterior beliefs to generate different values to the sender in equilibrium. This distinction is geometrically captured by concavification and quasi-concavification (Lipnowski and Ravid, 2020).

[5] We abuse the notation here and omit π in the subscript. We do need to consider how γ varies with information policy π in order to characterize the equilibrium solution with misinterpretation. The reason that we do not even assume γ constant with cardinality $|\mathcal{S}|$ is that we don't want the misinterpretation probabilities to be attached to realization labels so that the sender can affect the receiver's behavior by manipulating the labels. Additionally, any uninformative information policy doesn't noise the receiver's posterior away from the prior belief. For the case where the sender can manipulate labels, Tsakas and Tsakas (2021) offers some insight.
For now, we only need γ to be errors of meaning. The current result in the section doesn't depend on the specific values. In Section 3, we solve equilibrium and analyze welfare effects by further restricting the realization space $\mathcal{S}$ to be binary and the misinterpretation probabilities γ to be constant.

this environment, leading to a less informative signal structure for the receiver, given by $\phi(\tilde{s} \mid \omega) = \sum_s \pi(s \mid \omega)\gamma(\tilde{s} \mid s)$. If both players correctly specify their own effective information and apply Bayes's rule accordingly, given a common prior μ_0 and a pair of effective information environments (π, s) and $(\phi, \tilde{s})$, then

$$\begin{aligned}
\text{the sender's Bayesian posterior belief arrives at} \quad & \mu_s(\omega; \pi) = \frac{\pi(s \mid \omega)\mu_0(\omega)}{\sum_{\omega'} \pi(s \mid \omega')\mu_0(\omega')} \\
\text{with probability} \quad & \tau_1(\mu) = \sum_{\omega'} \pi(s \mid \omega')\mu_0(\omega'); \\
\text{the receiver's Bayesian posterior belief arrives at} \quad & \tilde{\mu}_{\tilde{s}}(\omega; \phi) = \frac{\phi(\tilde{s} \mid \omega)\mu_0(\omega)}{\sum_{\omega'} \phi(\tilde{s} \mid \omega')\mu_0(\omega')} \\
\text{with probability} \quad & \tau_2(\tilde{\mu}) = \sum_{\omega'} \phi(\tilde{s} \mid \omega')\mu_0(\omega'),
\end{aligned}$$

where $\tau_1(\mu)$ and $\tau_2(\tilde{\mu})$ denote the sender's and the receiver's marginals of the joint posterior distribution $\tau(\mu, \tilde{\mu})$ over the belief pairs. In KG, the marginals are perfectly correlated since γ is the identity. With misinterpretation, $\tau_1(\mu)$ and $\tau_2(\tilde{\mu})$ are partially correlated by the γ, which captures the full dependence structure. It allows us to write the joint distribution $\tau(\mu, \tilde{\mu})$ as a γ-transformation of sender's posterior distribution $\tau_1(\mu)$. Consequently, despite interdependence among posterior beliefs in the support, the sender's problem can be reformulated entirely in terms of her own posterior distribution, without referring to signal realizations or the information policy.

Furthermore, we allow the receiver to systematically deviate [6] from his Bayesian posterior beliefs and analyze a specific form particularly salient in our context—naïve misspecification. Let $\hat{\mu}_{\tilde{s}}$ denote the receiver's subjective belief. We say that the receiver is sophisticated if he correctly specifies his effective information, $\hat{\mu}_{\tilde{s}} = \tilde{\mu}_{\tilde{s}}$. He is naïve if he mistakes the sender's

[6] as per systematic distortion in de Clippel and Zhang (2022). However, solving for optimal persuasion with naïve misspecification in this paper is still outside their scope. The systematic distortion is with respect to the receiver's Bayesian posterior beliefs. But since the sender's and receiver's Bayesian posterior beliefs are partially correlated by γ, we cannot rewrite receiver's subjective posterior beliefs as a function of sender's posterior beliefs without reference to irrelevant realizations due to misinterpretation.
We can incorporate a variety of non-Bayesian belief updating rules in their definition. We decide to focus on naïveté because this non-Bayesian belief updating is interactive with our main interest in misinterpretation. In Section 3, we take a closer look at how they interact.

Bayesian beliefs for his own, $\hat{\mu}_{\tilde{s}} = \mu_s$.

Note that although the sophisticated and naïve receivers hold different subjective posterior beliefs, they arrive at them with the same probability, $\tau_2(\cdot) = \sum_{\omega'} \phi(\tilde{s} \mid \omega')\mu_0(\omega')$. This departure only affects the subjectively optimal action taken but not the probability of taking the respective action for each receiver types.

Sender's Problem

Given posterior $\hat{\mu}_{\tilde{s}}$, the receiver chooses an action that maximizes expected utility, breaking ties in favor of the sender

$$a^*(\hat{\mu}_{\tilde{s}}) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\hat{\mu}_{\tilde{s}}} u(a, \omega).$$

Then for each pair of posterior beliefs $(\mu_s, \hat{\mu}_{\tilde{s}})$, the sender evaluates her expected utility at her posterior belief μ_s

$$\hat{v}(\mu_s, \hat{\mu}_{\tilde{s}}) = \mathbb{E}_{\mu} v(a^*(\hat{\mu}_{\tilde{s}}), \omega).$$

The sender's problem is to design an information policy that induces a joint posterior distribution that maximizes expected utility

$$V(\mu_0, \gamma) = \sup_{\tau} \mathbb{E}_{\tau(\mu_s, \hat{\mu}_{\tilde{s}})} \hat{v}(\mu_s, \hat{\mu}_{\tilde{s}}),$$

where each element in the support of the joint distribution arises with probability $\tau(\mu_s, \hat{\mu}_{\tilde{s}})$ and the marginal distributions are correlated by γ.

Misinterpretation: structural transformation of information environment

First, let us examine the effect of misinterpretation by comparing a sophisticated receiver to a rational receiver in KG.

Note that misinterpretation violates the independence of irrelevant alternatives, a key condition for applying the concavification technique. Although the sender's problem can still be expressed in terms of Bayes-plausible posterior beliefs μ, we cannot use concavification to find optimal value because posterior beliefs are no longer payoff-separable. The

sender's utility from a given posterior μ_s now depends on the set of posterior beliefs that may be confounded with it via γ within the support, i.e., $\{\mu_{s'} \mid \gamma(s \mid s') > 0\} \subset supp(\tau_1)$. Although we cannot solve the sender's problem without further assumptions on how γ varies with π, we can still conclude that the sender's optimal persuasion value must be weakly smaller with misinterpretation than without, due to bounded implementability. The intuition is straightforward: any receiver's posterior distribution implementable in misinterpreted persuasion can be implemented in Bayesian persuasion.

Remark 1. *(Bayes-plausibility with misinterpretation)*

Given the receiver's misinterpretation behavior captured γ, the pair $(\mu, \tilde{\mu})$ is Bayes-plausible if each marginal expected posterior probability equals the prior: $\sum_{supp(\tau_1)} \mu\tau_1(\mu) = \mu_0$ and $\sum_{supp(\tau_2)} \tilde{\mu}\tau_2(\tilde{\mu}) = \mu_0$, where $\tau_2(\tilde{\mu}) = \tau_1(\mu)\gamma$.

Remark 2. *The belief pair $(\mu, \tilde{\mu})$ is implementable by a sender-designed information policy π if and only if it is Bayes-plausible.*

We say γ is non-identity if there exists π and $s \in supp(\pi)$ such that $\gamma(s \mid s) < 1$.

Lemma 1. *(Bounded Implementability) For any non-identity γ, the set of implementable receiver's posterior distribution is a subset of that in KG.*

Proposition 1. *Given non-identity γ, for state-independent $v(a)$, the sender's optimal value from misinterpreted persuasion is weakly smaller than KG.*

Naïveté: systmatic distortion of posterior beleifs

Now, consider the effect of naïveté, by comparing a naive receiver to a sophisticated receiver. A naïve receiver over-infers from his information environment because his effective information policy—ϕ is less informative than what he thinks. It makes persuasion easier, but may not always translate to positive welfare gain for the sender.

Note that naïveté violates the martingale property. This violation could work either against or for the sender depending on the problem. Either it cancels out the sender's benefit from the receiver's over-inference; or it reallocates communication surplus from the

receiver to the sender. In Section 3, we will showcase when naïveté strictly benefits the sender and hurts the receiver, and when it has no effect. We will also describe the composite effect of misinterpretation and naïveté. Persuading a naïve reciever may leave the sender better or worse off relative to the KG benchmark.

3 Binary Example

This section focuses on the canonical Prosecutor-Judge example in KG. We restrict our attention to binary realization space and solve for the sender's optimal value from misinterpreted persuasion and naïvely misinterpreted persuasion.

This example demonstrates the proposed behavioral decomposition into two components: (1) misinterpretation, structurally distorting information policy, and (2) non-Bayesian updating rules or misspecification, systematically distorting posterior beliefs. We analyze the interaction between misinterpretation and a specific form of misspecification—naïveté.

We highlight an extension of the binary example to confirmation bias. It shows that the insights about the decomposition are robust to endogeneity in confirmation bias.

3.1 Setup

Suppose a lawyer (she, the sender) defending a suspect who is either guilty (L) or innocent (H), $\omega \in \Omega = \{L, H\}$, tries to persuade a jury (he, the receiver) for acquittal. The jury chooses whether to convict (a_l) or to acquit (a_h) the suspect, $a \in \mathcal{A} = \{a_l, a_h\}$. The lawyer and the jury share a common prior belief in innocence at $\mu_0 \coloneqq Prob.(\omega = H) \in (0, 1)$.

The lawyer can influence the jury's belief through information design. She invites a forensic expert to testify, generating either a guilty (l) or innocent (h) testimony, the signal realizations $s \in \mathcal{S} = \{l, h\}$. Regardless of state, the lawyer gets $v(a_h) = 1$ if the jury acquits her client and $v(a_l) = 0$ if the jury convicts her client. However, the jury aims to match action to state. He prefers to convict if the suspect is guilty, $u(a_l, L) > u(a_h, L)$, and to acquit if innocent, $u(a_h, H) \geq u(a_l, H)$. He is indifferent between conviction and acquittal if

he believes that the probability of innocence is $\bar{\mu} \coloneqq \frac{\big(u(a_l,L)-u(a_h,L)\big)}{\big(u(a_h,H)-u(a_l,H)\big)+\big(u(a_l,L)-u(a_h,L)\big)} \in (0,1]$.

3.1.1 Benchmark

In the KG, the lawyer's optimal strategy is characterized by the concavification of the lawyer's indirect utility function. Given a prior $\mu_0 < \bar{\mu}$, her best ex-ante expected value from persuasion is $\frac{\mu_0}{\bar{\mu}}$, through inducing posterior beliefs to 0 w.p. $1-\frac{\mu_0}{\bar{\mu}}$ and to $\bar{\mu}$ w.p. $\frac{\mu_0}{\bar{\mu}}$.

3.1.2 Misinterpretation

Let π_ω represent the probability of sending realization h in each state $\omega \in \{L,H\}$. The lawyer-designed information policy in matrix form is $\Pi = \begin{bmatrix} 1-\pi_L & \pi_L \\ 1-\pi_H & \pi_H \end{bmatrix}$. The lawyer observes the realization $s \in \{l,h\}$ generated according to Π. But the jury (mis)interprets s as $\tilde{s} \in \{l,h\}$ with probabilities $\{\gamma(\tilde{s} \mid s)\}_{s,\tilde{s}\in\{l,h\}}$, represented by $\Gamma = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. γ_l is the probability of misinterpreting the realization (l) intended to lower belief in innocence as the realization (h) intended to raise belief; γ_h is the probability of misinterpreting h as l. We refer $\gamma_h > 0$ as the unfavorable misinterpretation and $\gamma_l > 0$ as the favorable misinterpretation. Without loss of generality, we assume *infrequent* misinterpretation, $1-\gamma_l-\gamma_h > 0$[7].

The jury's effective information policy Φ is less informative than the lawyer-designed policy Π. Γ captures the correlation between the sender's and the receiver's effective policies, $\Phi \coloneqq \Pi\Gamma = \begin{bmatrix} 1-\phi_L & \phi_L \\ 1-\phi_H & \phi_H \end{bmatrix}$, where $\phi_\omega \coloneqq \pi_\omega(1-\gamma_h-\gamma_l)+\gamma_l$ is the probability that the jury perceives realization $\tilde{h}$ when the state is ω. Correspondingly, the lawyer's Bayesian posterior beliefs $\mu_s \coloneqq \{\mu_l,\mu_h\}$[8] mean-preserving spreads around the prior with respect to jury's Bayesian posterior beliefs $\tilde{\mu}_{\tilde{s}} \coloneqq \{\tilde{\mu}_{\tilde{l}},\tilde{\mu}_{\tilde{h}}\}$[9].

[7] For result with *frequent* misinterpretation, see Appendix B.

[8] $\mu_h = \mu^B(H \mid h;\Pi) \coloneqq \frac{\pi_H\mu_0}{\pi_H\mu_0+\pi_L(1-\mu_0)}$; $\mu_l = \mu^B(H \mid l;\Pi) \coloneqq \frac{(1-\pi_H)\mu_0}{(1-\pi_H)\mu_0+(1-\pi_L)(1-\mu_0)}$.

[9] $\tilde{\mu}_h = \mu^B(H \mid \tilde{h};\Phi) \coloneqq \frac{\phi_H\mu_0}{\phi_H\mu_0+\phi_L(1-\mu_0)}$; $\tilde{\mu}_l = \mu^B(H \mid \tilde{l};\Phi) \coloneqq \frac{(1-\phi_H)\mu_0}{(1-\phi_H)\mu_0+(1-\phi_L)(1-\mu_0)}$.

Given the vector of prior beliefs $P = \begin{bmatrix} 1-\mu_0 & \mu_0 \end{bmatrix}$, the lawyer's Bayesian posterior beliefs are distributed as

$$\begin{bmatrix} \tau_1^l & \tau_1^h \end{bmatrix} \coloneqq P\Pi = \begin{bmatrix} 1-\big(\mu_0\pi_H + (1-\mu_0)\pi_L\big) & \mu_0\pi_H + (1-\mu_0)\pi_L \end{bmatrix}$$

and the jury's Bayesian posterior beliefs are distributed as

$$\begin{bmatrix} \tau_2^{\tilde{l}} & \tau_2^{\tilde{h}} \end{bmatrix} \coloneqq P\Phi = P\Pi\Gamma = \begin{bmatrix} 1-\big(\mu_0\phi_H + (1-\mu_0)\phi_L\big) & \mu_0\phi_H + (1-\mu_0)\phi_L \end{bmatrix}.$$

Both marginal posteriors are Bayes-plausible, $\tau_1^l\mu_l + \tau_1^h\mu_h = \mu_0$ and $\tau_2^{\tilde{l}}\tilde{\mu}_{\tilde{l}} + \tau_2^{\tilde{h}}\tilde{\mu}_{\tilde{h}} = \mu_0$.

3.1.3 Naïve misspecification

Beyond misinterpretation, the jury may also misspecify the effective information environment he's in. To characterize his level of knowledge about his information, we denote the jury's subjective posterior beliefs as $\hat{\mu} \coloneqq \{\hat{\mu}_l, \hat{\mu}_h\}$.

The jury is *sophisticated* if he knows his effective information structure Φ and applies Bayes rule accordingly. Therefore, a sophisticated jury's subjective posterior beliefs coincide with the jury's Bayesian posterior beliefs, $\hat{\mu} = \tilde{\mu}$ and satisfy the martingale property.

The jury is *naïve* if he mistakenly believes that the lawyer's announced signal structure Π is the one he perceives, and updates using Bayes rule based on this misspecified model. Therefore, a naïve jury's subjective posterior beliefs coincide with the lawyer's Bayesian posterior beliefs, $\hat{\mu} = \mu$. But these beliefs $\{\mu_l, \mu_h\}$ still arise with probability $\begin{bmatrix} \tau_2^{\tilde{l}} & \tau_2^{\tilde{h}} \end{bmatrix}$. Thus, the naïve jury's posterior beliefs violate the martingale property: $\tau_2^{\tilde{l}}\mu_l + \tau_2^{\tilde{h}}\mu_h \neq \mu_0$.

3.2 Persuading a Sophisticated Receiver

A sophisticated receiver correctly specifies his effect information environment $(\Phi, \tilde{s})$ and updates to his Bayesian posterior beliefs $\tilde{\mu}$. For $\mu_0 < \bar{\mu}$, the sender solves

$$\max_{\substack{\mu_l \in [0,\mu_0),\\ \mu_h \in (\mu_0,1]}} \tau_2^h(\mu_l,\mu_h)$$
$$\text{s.t. } \tilde{\mu}_h(\mu_l,\mu_h) \geq \bar{\mu} \qquad (O^S)$$

where
$$\tau_2^h(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\gamma_h-\gamma_l)+\gamma_l$$
$$\tilde{\mu}_h(\mu_l,\mu_h) = \frac{(1-\gamma_h)(\mu_0-\mu_l)\mu_h+\gamma_l(\mu_h-\mu_0)\mu_l}{(1-\gamma_h)(\mu_0-\mu_l)+\gamma_l(\mu_h-\mu_0)}$$

3.2.1 Solution and Welfare Analysis

In the KG benchmark, the sender benefits from persuasion for any prior $\mu_0 \in (0,\bar{\mu})$. Given the game, favorable misinterpretation limits how far the sender can raise the receiver's posterior beliefs, shrinking the range of priors where she can benefit from persuasion. The mathematical proof is in Appendix A.

Proposition 2. *Given $\mu_0 < \bar{\mu}$, γ_l, and γ_h, the sender benefits from misinterpreted persuasion if and only if the common prior is large enough so that it is possible to persuade the receiver to switch actions, $\mu_0 \geq \frac{\gamma_l\bar{\mu}}{(1-\gamma_h)(1-\bar{\mu})+\gamma_l\bar{\mu}} =: \underline{\mu_0} \geq 0$ (strict inequality if $\gamma_l > 0$).*

Intuitively, with low priors, the effect of favorable misinterpretation gets amplified by the amount of information required to switch action so that even full information revelation is not informative enough for the receiver to be persuaded. Full information revelation is always a feasible strategy for the sender. She benefits from misinterpreted persuasion if she can switch action with full revelation. Conversely, if the sender cannot persuade the receiver even with full revelation, then no strategy can.

In a sender-preferred PBE, the sender extracts full communication surplus. The receiver is always (subjectively) indifferent when switching actions in equilibrium. Misinterpretation reduces the sender's optimal value and the total communication surplus from persuasion.

Proposition 3. *When the sender benefits from misinterpreted persuasion with a sophisticated receiver, an optimal information policy induces the receiver's Bayesian posterior to the indifference threshold, $\tilde{\mu}_h(\mu_l^*, \mu_h^*) = \bar{\mu}$. In equilibrium, the sender reveals strictly more information than in KG*[10] *if and only if there is favorable misinterpretation:*

$$\{\mu_l^*, \mu_h^*\} = \left\{0, \frac{\bar{\mu}\mu_0(1-\gamma_h-\gamma_l)}{\mu_0(1-\gamma_h-\gamma_l)-\gamma_l(\bar{\mu}-\mu_0)}\right\} \text{ mean-preserving spreads } \{0, \bar{\mu}\} \Leftrightarrow \gamma_l > 0.$$

Proposition 3 implies that a sophisticated receiver still makes the optimal decisions because he switches to the sender-preferred action at the correct indifferent threshold $\bar{\mu}$. For the sender, misinterpretation reduces payoff relative to KG due to bounded implementability. Figure 1 shows the sender's best ex-ante expected value from persuasion with and without misinterpretation.

For low priors ($\mu_0 < \underline{\mu_0}$), misinterpretation hurts the sender because favorable misinterpretation ($\gamma_l > 0$) reduces the sender's ability to pull posterior belief too far away from the prior μ_0 towards the action-switching belief threshold $\bar{\mu}$. For high priors ($\mu_0 > \underline{\mu_0}$), misinterpretation hurts the sender because unfavorable misinterpretation ($\gamma_h > 0$) reduces the sender's ability to pull posterior belief too far away from the prior μ_0 towards 0 that maximizes the ex-ante probability of sending h realization. Formally,

Corollary 1. *(Welfare effects of misinterpretation)*

Misinterpretation has no welfare effect on the receiver and strictly reduces the sender's welfare by impairing her ability to implement the receiver's posterior distributions.

- *The range of prior that the sender benefits from persuasion is strictly smaller than KG if and only if there is a favorable misinterpretation: $\underline{\mu_0} > 0 \Leftrightarrow \gamma_l > 0$.*
- *The sender's gain from misinterpreted persuasion is strictly less than that in KG if and only if there is an unfavorable misinterpretation: $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \frac{\mu_0}{\bar{\mu}} \Leftrightarrow \gamma_h > 0$.*

[10] Remind that in an equilibrium of KG, the sender induces posterior beliefs to $\{0, \bar{\mu}\}$ for any $\mu_0 \in (0, \bar{\mu})$

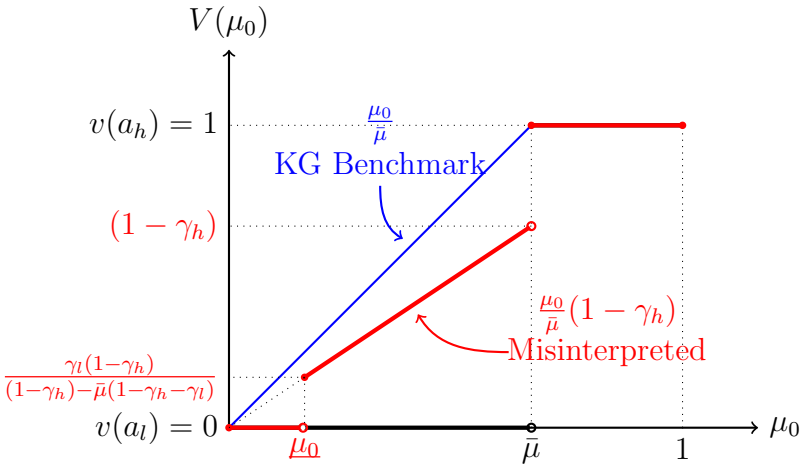


Figure 1: Value function comparison

— with *infrequent* misinterpretation

— without misinterpretation [11]

3.3 Persuading a Naïve Receiver

The receiver we've studied in the previous subsection is so sophisticated that he knows the exact probability that he misinterprets the information policy. What if the receiver lacks this level of sophistication? This subsection investigates a naïve receiver who misspecifies his information environment. He thinks the perceived realization comes from the announced policy $(\Pi, \tilde{s})$, even though the true environment is $(\Phi, \tilde{s})$. Hence, instead of the his Bayesian posteriors $\tilde{\mu}$, the naïve receiver arrives at misspecified posterior beliefs equal to the sender's Bayesian posterior belief $\mu = (\mu_l, \mu_h)$, but still with probability τ_2. In addition to misinterpretation breaking the independence among posterior beliefs, we further lose Bayes-plausibility to this naïve misspecification.

The sender still maximizes the probability of the receiver taking her preferred action a_h but is subject to a different constraint. With infrequent misinterpretations ($\frac{\gamma_l}{1-\gamma_h} < 1$), the

[11] Results for *frequent* misinterpretation is in Appendix B.1, which are quantitatively different but qualitative the same. Comparative statics is in Appendix C.1.1 and some policy implication based on comparative statics in Appendix C.2.

sender's problem is a positive linear transformation of that in KG:

$$\max_{\substack{\mu_l \in [0,\mu_0),\\ \mu_h \in (\mu_0,1]}} \tau_2^h(\mu_l,\mu_h) = \tau_1^h(\mu_l,\mu_h)(1-\gamma_h-\gamma_l)+\gamma_l$$

$$\text{s.t. } \mu_h \geq \bar{\mu} \qquad (O^N)$$

3.3.1 Solution and Welfare Analysis

Maïveté reallocates communication surplus through suboptimal decision-making and violation of martingale property. It only affects persuasion if in equilibrium the naïve receiver takes a different action than the sophisticated receiver would have.

Proposition 4. *Naïveté restores implementability back to the same as in KG. When the sender benefits from naively misinterpreted persuasion ($\mu_0 \in (0,\bar{\mu})$), an optimal information policy induces the naïve receiver's misspecified posterior (μ_h) to the indifference threshold ($\bar{\mu}$). For $\gamma_l > 0$, the naïve receiver demands sub-optimally less information to be persuaded:*

$$\tilde{\mu}_h(\mu_l^*,\mu_h^*) = \frac{\mu_0(1-\gamma_h)}{\frac{\mu_0}{\bar{\mu}}(1-\gamma_h-\gamma_l)+\gamma_l} \leq \bar{\mu} \text{ equality with } \gamma_l = 0.$$

Unlike the sophisticated receiver, the naïve receiver switches to sender-preferred action sub-optimally. He should take a_h when his Bayesian posterior is weakly greater than the indifference threshold $\bar{\mu}$. But in equilibrium, the sender only needs to bring the naïve Receiver's misspecifed subjective posterior μ_h to $\bar{\mu}$, which is weakly easier since $\tilde{\mu}_h \leq \mu_h$ (with equality if no favorable misinterpretation $\gamma_l = 0$) under the same distribution. Thus, naïveté benefits the sender not by increasing surplus, but by shifting it from the receiver.

Corollary 2. *(Welfare effects of naïve misspecification)*

1. *Naïve misspecification weakly hurts the receiver.*

   The receiver is strictly worse off if and only if he favors the sender ($\gamma_l > 0$) AND is unaware of his favorable misinterpretation.

2. *Naïve misspecification weakly benefits the sender.*

   - *Naïveté recovers the sender's implementabilty back to KG.*
   - *The Sender gets all the surplus from the receiver's sub-optimal decision due to naively favorable misinterpretation.*

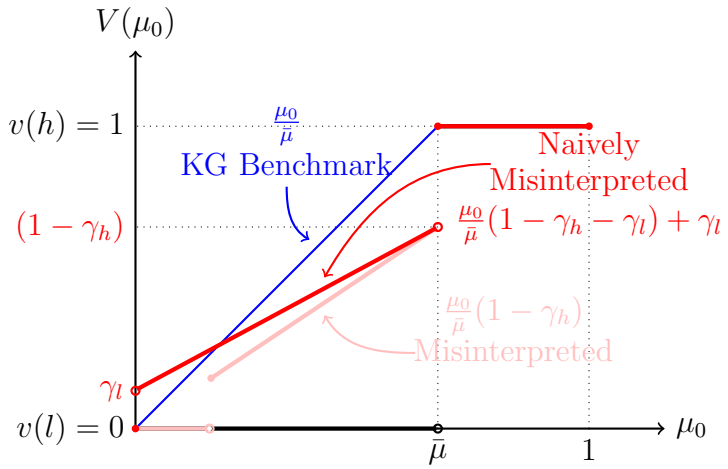


Figure 2: Value function comparison
— with *infrequent* misinterpretation and naïveté
— with *infrequent* misinterpretation and sophistication
— without misinterpretation [12]

Combining misinterpretation and naïve misspecification, the sender may strictly benefit relative to KG with the naïve receiver who needs a lot of information to switch actions (low prior). Misinterpretation imposes friction. It hurts the sender through both unfavorable misinterpretation ($\gamma_h > 0$) restricting the sender's ability to lower beliefs and favorable misinterpretation ($\gamma_l > 0$) restricting the sender's ability to raise beliefs. Naïveté, in contrast, exposes the asymmetry in the effects of misinterpretation. It benefits the sender only through favorable misinterpretation ($\gamma_l > 0$). As a result, the lower the prior belief is, the more persuasion needed, the more sub-optimal the naïve receiver's equilibrium action is, and hence the larger benefits from naïveté. Eventually, the gain from naïveté eventually outweighs the cost of being misinterpreted.

[12] Results for *frequent* misinterpretation is in Appendix B.2. Comparative statics is in Appendix C.1.2 and some policy implication in Appendix C.2.

Corollary 3. *(Composite welfare effects of misinterpretation and naïveté misspecification)*

1. *For low priors ($\mu_0 < \frac{\gamma_l \bar{\mu}}{\gamma_l + \gamma_h}$), the sender is better off persuading a naively misinterpreted receiver than persuading a rational receiver in KG.*

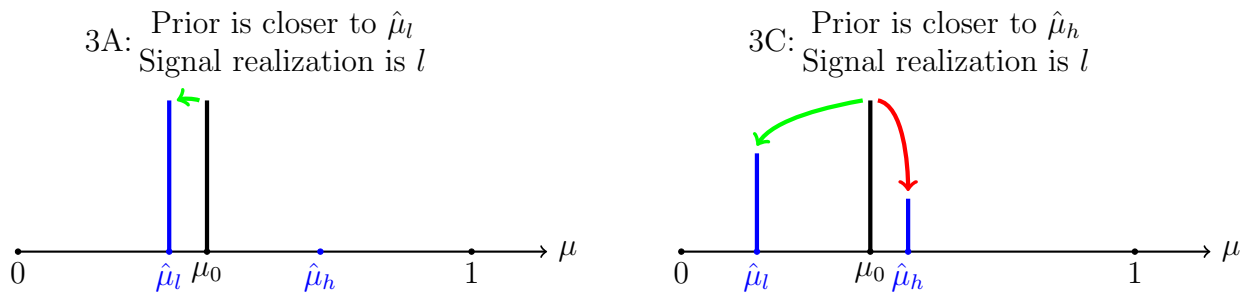
2. *For high priors ($\mu_0 > \frac{\gamma_l \bar{\mu}}{\gamma_l + \gamma_h}$), the sender is worse off persuading a naively misinterpreted receiver than persuading a rational receiver in KG.*

3.4 Binary Extension — Confirmation Bias

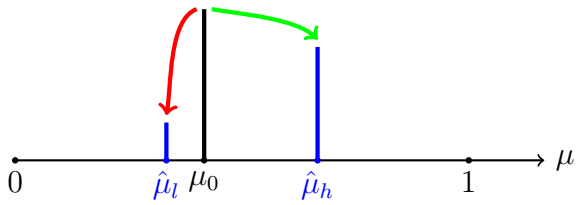
This subsection extends the binary example to consider confirmation bias, which endogenizes the direction of misinterpretation. The setup remains unchanged, except that misinterpretation now depends on how the information aligns with the jury's prior belief. The result echoes two special cases of the binary example.

Specifically, a confirmation-biased jury misinterprets in one of two distinct patterns. (1) When the lawyer designs an information policy inducing high posterior closer to the prior, the jury may misinterpret guilty testimony (l) as innocent (h), but never the reverse. When the lawyer designs an informative policy inducing low posterior closer to the prior, the jury may misinterpret innocent testimony (h) as guilty (l), again never the reverse. Figure 3 illustrates the confirmation bias visually. The associated misinterpretation matrices are $\Gamma_h = \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ for the case on the left (Figure 3A and 3B) and $\Gamma_l = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix}$ for the case on the right (Figure 3C and 3D).

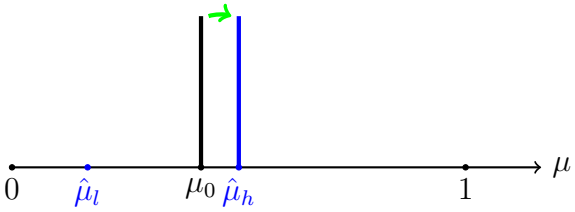
Figure 3: Direction of Misinterpretation



3B: Prior is closer to $\hat{\mu}_l$
Signal realization is h



3D: Prior is closer to $\hat{\mu}_h$
Signal realization is h



To formalize confirmation bias, we make a few modeling choices that do not substantively affect the results. The effective direction of bias depends on the relative distance between the prior μ_0 and the receiver's subjective posterior, which (1) equates to receiver's Bayesian posterior $\tilde{\mu}$ if he is sophisticated or (2) equates to sender's Bayesian posterior μ if the receiver is naïve. We also assume the cutoff rule associated to Γ_h.

Definition 1. *(Confirmation Bias)*

For a given prior μ_0, suppose the sender implements π to induce Sender's Bayesian posterior $(\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$.

1. *The sophisticated receiver with confirmation bias exhibits errors represented by Γ^{SCB}.*

- *If $\gamma_h < \frac{1}{2}$,* $\Gamma^{SCB} = \begin{cases} \Gamma_h \coloneqq \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}; \\ \Gamma_l \coloneqq \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 > \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}. \end{cases}$
- *If $\gamma_h \geq \frac{1}{2}$,* $\Gamma^{SCB} = \Gamma_h \coloneqq \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ *for any (μ_l, μ_h).*

2. *The naïve receiver with confirmation bias exhibits errors represented by Γ^{NCB}.*

$$\Gamma^{NCB} = \begin{cases} \Gamma_h \coloneqq \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + \mu_l}{2}\right\}; \\ \Gamma_l \coloneqq \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 > \frac{\mu_h + \mu_l}{2}\right\}. \end{cases}$$

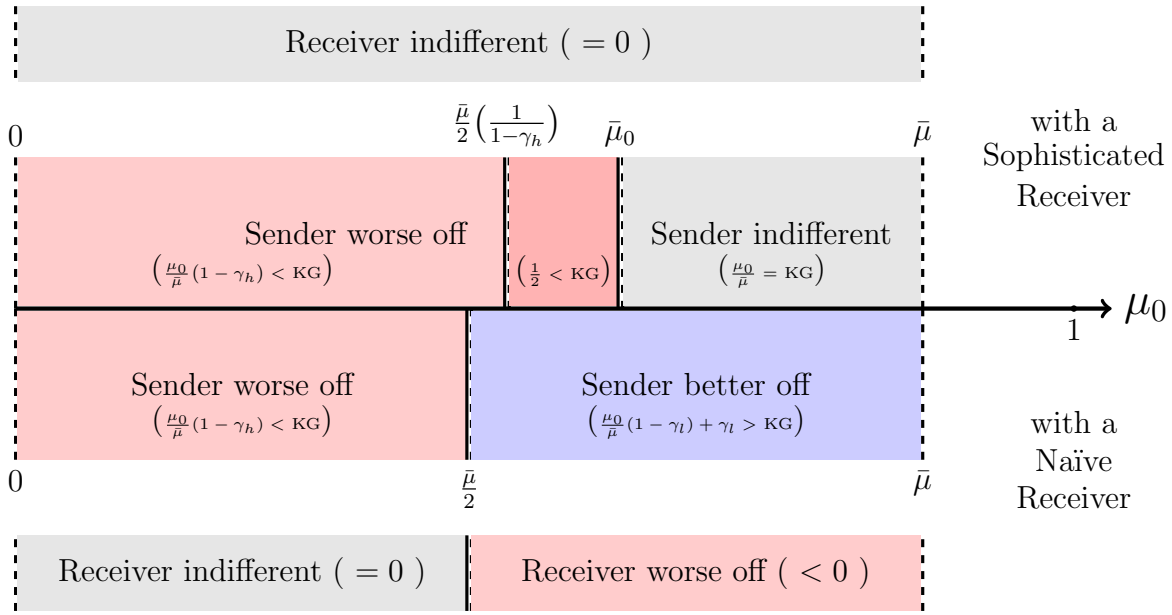
Given a problem with indifference threshold $\bar{\mu}$, prior μ_0, and misinterpretation patterns as defined, the sender solves the optimal strategy in two steps. First, she searches for a solution under each misinterpretation matrix Γ_h or Γ_l in the corresponding posterior beliefs set; then, she selects the best of the two if both corresponding posterior sets are non-empty.

Insights from the binary example remain robust.Figure 4 summarizes welfare implications. For low priors, the sender is strictly worse off than in the KG benchmark. For high priors, the outcome depends on the receiver's type: if the receiver is sophisticated, the sender achieves KG's value; if the receiver is naïve, the sender strictly outperforms KG.

Because the sender extracts all surplus in equilibrium, the receiver is made subjectively indifferent. But when the receiver is naïve, he over-infers and unknowingly incurs an objectively negative payoff.

In the lawyer-jury example, confirmation bias harms the jury only when he is naïve and the prior is high (i.e., little persuasion is needed). In equilibrium, only high priors activate favorable misinterpretation; low priors do not. As a result, the lawyer profits from confirmation bias compared to KG when the prior is high, in contrast to Corollary 3. This reversal stems from that the direction of misinterpretation is strategy-dependent.

Figure 4: Welfare Effects of Confirmation Bias in Comparison to KG



In the next subsections, we formally characterize the sender's optimal persuasion strategy. The solution procedure is identical whether the receiver is sophisticated or naïve; the only difference lies in the posterior belief constraints.

3.4.1 Persuading a Sophisticated Receiver with Confirmation Bias

Proposition 5. *(Persuasion with Sophisticated Confirmation Bias)*

Suppose a confirmatory biased receiver is sophisticated and misinterprets according to Γ^{SCB}*. Fixing an indifference threshold* $\bar{\mu}$*, there exists a prior belief threshold*

$$\bar{\mu}_0 = \max\left\{\frac{\bar{\mu}}{2(1-\gamma_h)}\Big(1+\gamma_l(1-2\gamma_h)\Big), \frac{\gamma_l\bar{\mu}}{\gamma_l\bar{\mu}+1-\bar{\mu}}\right\}$$

such that in equilibrium

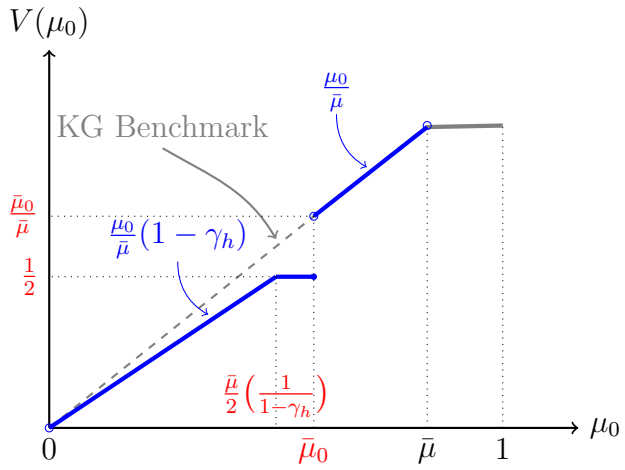
- *For low priors* ($\mu_0 \leq \bar{\mu}_0$)*, the receiver misinterprets against the sender (that is,* Γ_h *is effective). Compared to KG, the sender reveals the same amount of information but less is transmitted to the receiver. The receiver still switches action at* $\bar{\mu}$ *and gets the same* 0 *expected payoffs as in KG. However, the sender gets* $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) \leq \frac{1}{2}$*, which is strictly less than* $\frac{\mu_0}{\bar{\mu}}$ *in KG.*

- *For high prior* μ_0 *(above* $\bar{\mu}_0$*), the receiver misinterprets in favor of the sender (that is, the effective error matrix is* Γ_l*). Compared to KG, the sender reveals more information to compensate for the informational loss due to misinterpretation. Both the sender and the receiver get the same expected payoffs as in KG, respectively* $\frac{\mu_0}{\bar{\mu}}$ *and* 0*.*

The outcome follows directly from Corollary 1 under Γ_h for low prior and Γ_l for high prior respectively. Figure 5 [13] shows the sender's value function. The flat region in the middle arises because the cutoff for Γ_l lies above the cutoff for Γ_h.

[13]The sender's problem is just a combination of two special cases of the baseline model with a sophisticated receiver, with an additional constraint on the posterior beliefs. The posterior condition restricts the solution to a half-space of posterior beliefs, which doesn't change the nature of the optimization problem. We show the mathematical solution in Appendix A.2.1

Figure 5: Value Function with Sophisticated Confirmation Bias



3.4.2 Persuading a Naïve Receiver with Confirmation Bias

This subsection states the naïve equivalent of Proposition 5 in the previous subsection.

Proposition 6. *(Persuasion with Naïve Confirmation Bias)*

Suppose a confirmatory biased receiver is fully naïve and misinterprets according to Γ^{NCB}. *Fixing an indifference threshold* $\bar{\mu}$, *there exists a prior belief threshold* $\frac{\bar{\mu}}{2}$ *such that in equilibrium*

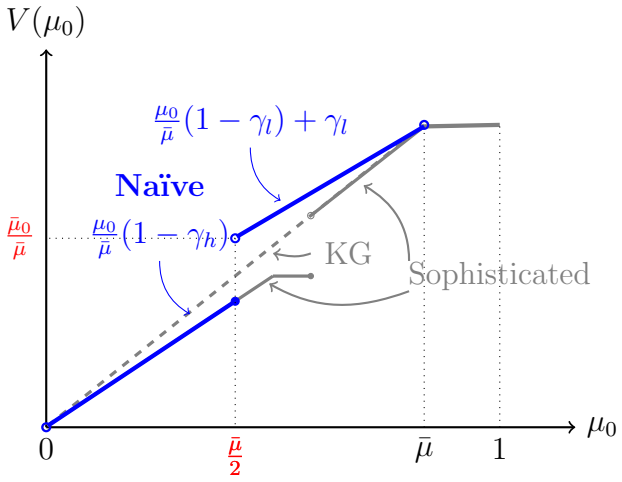
- *For low priors* ($\mu_0 \leq \frac{\bar{\mu}}{2}$), *the receiver misinterprets against the sender (that is, the effective error matrix is* Γ_h*). Compared to KG and the sophisticated confirmation bias, the sender reveals the same amount of information but less is transmitted to the receiver. Both the sender and the receiver get the same payoffs as in the sophisticated case; that is, the sender is worse off than in KG and the receiver remains indifferent as in KG.*

- *For high prior* ($\mu_0 > \frac{\bar{\mu}}{2}$), *the receiver misinterprets in favor of the sender (that is, the effective error matrix is* Γ_l*). The sender reveals the same amount of information compared to KG and less information compared to the sophisticated case. The receiver switches action before reaching* $\bar{\mu}$ *and thus gets strictly less payoff than in KG and the*

sophisticated benchmarks. However, the sender gets a strictly higher payoff than in KG. Compared to the sophisticated case, the sender gains from naïveté; she profits the most from naïveté for intermediate priors $\mu_0 \in \left(\frac{\bar{\mu}}{2}, \bar{\mu}_0\right]$.

As with the sophisticated case, the outcome follows directly from Proposition 4 under Γ_h for low prior and Γ_l for high prior respectively. The seemingly contradictory result compared to Corollary 3 arises because the direction of misinterpretation is endogenously evoked: unfavorable at low priors, favorable at high. The resulting sender value function is illustrated in Figure 6[14].

Figure 6: Value Function with Naïve Confirmation Bias



4 Discussion

4.1 Relaxing Commitment

Our analysis centers on persuasion with sender commitment, but the core implications of misinterpretation extend naturally to settings without commitment. This robustness arises

[14]Again, the sender's problem combines two special cases of the binary example with a naïve receiver, constrained to posterior belief half-spaces. These restrictions preserve the convex structure. See Appendix A.2.2 for the formal solution.

because misinterpretation operate within the belief space, and its effect on implementability does not depend on the shape of sender's value as function of beliefs.

In the setting with finite state and state-independent sender utility, the value of commitment links to the gap between concavification and quasi-concavification of the sender's indirect utility (Lipnowski and Ravid, 2020). Without commitment, equilibrium requires posterior beliefs in the support to yield equal sender utility, achievable by randomization at threshold beliefs. Analogously, misinterpretation reduces the sender's payoffs through bounded implementability by excluding beliefs and hence actions, even absent commitment.

Yet, Tsakas and Tsakas (2021) show that noise can improve informativeness when sender-receiver preferences are partially aligned but not too close or opposed. Our findings, where misinterpretation always hurts the sender, pose a puzzle: Is this difference due to the equilibrium structures in discrete versus continuous state spaces? Why partial alignment, when paired with noise, create new channels for pareto improvement? whether and how does the sender's payoff dependence on the state weigh in? These questions open fruitful avenues for further research.

4.2 Conclusion

This paper studies how structural misinterpretation shapes strategic communication by decomposing behavioral deviations into misinterpretation and misspecification (or non-Bayesian updating). This framework broadens the behavioral landscape for information design.

Misinterpretation, whether arising from complexity, bias, or other sources, limits the sender's ability to implement desired posterior distributions, shrinking communication surplus. Conversely, naïveté about misinterpretation partly benefits the sender at the receiver's expense, through sub-optimal decision-making and violation of martingale property. We extend these insights to richer behaviors such as confirmation bias, where the direction of misinterpretation depends endogenously on beliefs. While the sender can evoke different bias directions, she cannot manipulate equilibrium strategies beyond choosing optimal posteriors under effective misinterpretation.

Our results push literature forward by capturing a class of real-world communication frictions that standard models abstract away. More broadly, this paper spotlights the critical role of epistemic friction: not merely what agents learn, but how they interpret what they receive. As digital and social communication becomes increasingly mediated, these frictions are likely to grow, raising the stakes for how we understand, design, and regulate information environments. Our theoretical analysis lays groundwork for navigating and, where possible, mitigating misinterpretation's impact in practical communication settings.

References

Alberto Alesina, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review*, 114(7):1916–48, July 2024. doi: 10.1257/aer.20191184. URL https://www.aeaweb.org/articles?id=10.1257/aer.20191184.

Ricardo Alonso and Odilon Câmara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706, 2016. doi: https://doi.org/10.1016/j.jet.2016.07.006.

Victor Augias and Daniel M. A. Barreto. Persuading a wishful thinker. 2023.

Dan Benjamin, Aaron Bodoh-Creed, and Matthew Rabin. Base-rate neglect: Foundations and implications. 2019.

J. Aislinn Bohren and Daniel N. Hauser. The behavioral foundations of model misspecification: a decomposition. 2022.

Davide Bordoli. Non-bayesian updating and value of information. 2024.

J. M. Darley and P. H. Gross. A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20–33, 1983. doi: https://doi.org/10.1037/0022-3514.44.1.20.

Geoffroy de Clippel and Xu Zhang. Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022. doi: 10.1086/720464.

Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports*, 7, 01 2017. doi: 10.1038/srep40391.

Kfir Eliaz, Rani Spiegler, and Heidi Christina Thysen. Strategic interpretations. *Journal of Economic Theory*, 192:105192, 2021.

Oliver Falck, Robert Gold, and Stephan Heblich. E-lections: Voting behavior and the internet. *American Economic Review*, 104(7):2238–65, 2014. doi: 10.1257/aer.104.7.2238.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Yonghwan Kim. Does disagreement mitigate polarization? how selective exposure and disagreement affect political polarization. *Journalism & Mass Communication Quarterly*, 92(4):915–937, 2015. doi: 10.1177/1077699015596328. URL https://doi.org/10.1177/1077699015596328.

Joshua Klayman. Varieties of confirmation bias. volume 32 of *Psychology of Learning and Motivation*, pages 385–418. Academic Press, 1995. doi: https://doi.org/10.1016/S0079-7421(08)60315-1. URL https://www.sciencedirect.com/science/article/pii/S0079742108603151.

Silvia Knobloch-Westerwick, Cornelia Mothes, and Nick Polavin. Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, 47(1):104–124, 2020. doi: 10.1177/0093650217719596. URL https://doi.org/10.1177/0093650217719596.

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

2022. doi: 10.1257/aeri.20210007. URL https://www.aeaweb.org/articles?id=10.1257/aeri.20210007.

Elliot Lipnowski and Doron Ravid. Cheap talk with transparent motives. *Econometrica*, 88(4):1631–1660, 2020. doi: https://doi.org/10.3982/ECTA15674. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15674.

Charles Lord, Lee Ross, and Mark Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109, 11 1979. doi: 10.1037/0022-3514.37.11.2098.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175. URL https://doi.org/10.1037/1089-2680.2.2.175.

S. Plous. Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, 21(13):1058–1082, 1991. doi: https://doi.org/10.1111/j.1559-1816.1991.tb00459.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1991.tb00459.x.

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

Charles S. Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769, 2006. URL http://www.jstor.org/stable/3694247.

Elias Tsakas and Nikolas Tsakas. Noisy persuasion. *Games and Economic Behavior*, 130:44–61, 2021. doi: https://doi.org/10.1016/j.geb.2021.08.001.

Appendices

A Proofs

A.1 General Model

Proof. of Lemma 1

Given an arbitrary prior $\mu_0 \in int(\Delta(\Omega))$, let T^M and T^B denote the sets of implementable receiver's posterior distributions with and without misinterpretation respectively. WTS $T^M \subseteq T^B$. Any $\tau^M \in T^M$ is Bayes plausible. Any Bayes plausible posterior distribution is implementable in KG. So $\tau^M \in T^B$. ■

Proof. of Proposition 1 [in progress]

Given an arbitrary prior $\mu_0 \in int(\Delta(\Omega))$, let T^M and T^B denote the sets of implementable receiver's posterior distributions with and without misinterpretation respectively. Let $\tau^{B*} \in T^B$ denote a posterior distribution that gives the sender ex-ante expected value on the concave envelope of her value function, $\mathbb{E}_{\tau^{B*}(\mu)} v\big(a^*(\mu)\big) = V^*(\mu_0)$. Since $T^M \subseteq T^B$, for any implementable $\tau \in T^M$, $\mathbb{E}_{\tau(\mu)} v\big(a^*(\mu)\big) \leq V^*(\mu_0)$.

[strict part in progress] ■

A.2 Binary Baseline

A.2.1 Sophisticated receiver

Proof. of Proposition 2

Sketch: To show sufficiency, suppose $\mu_0 \geq \underline{\mu_0}$, equivalently $\tilde{\mu}_h(0,1) \geq \bar{\mu}$. If the sender does nothing, the receiver always takes action a_l and the sender gets 0. If the sender reveals full information, then the receiver takes the sender-preferred action a_h at his high posterior belief. The sender is strictly better off by revealing full information and gets $\mu_0(1-\gamma_h-\gamma_l)+\gamma_l > 0$. For necessity, the receiver's high posterior $\tilde{\mu}_h$ is decreasing in $\mu_l \in [0, \mu_0)$ and increasing $\mu_h \in (\mu_0, 1]$, and thus bounded from above by full information revelation, $\tilde{\mu}_h(\mu_l, \mu_h) \leq \tilde{\mu}_h(0,1) \forall (\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$. Thus, the sender cannot get a strictly better payoff

through misinterpreted persuasion for priors so low that it is impossible to persuade the receiver to take a_h.

"$\Rightarrow$" Revealing full information to the sender, $\mu = (\mu_l, \mu_h) = (0, 1)$, is always implementable as long as the posterior distribution τ_1 over μ average back to the prior. When the receiver's high posterior belief is greater than the belief threshold of indifference $\tilde{\mu}_h(0,1) \geq \bar{\mu}$, the receiver taking action a_h when perceiving $\tilde{h}$.

Thus, when $\tilde{\mu}_h(0,1) \geq \bar{\mu}$, Sender gets $\tau_2(0,1) = \mu_0(1-\gamma_h-\gamma_l)+\gamma_l > 0$. So Sender benefits from persuasion when it is possible to induce the receiver to take the sender-preferred action $\tilde{\mu}_h(0,1) \geq \bar{\mu}$.

$$\begin{aligned} \tilde{\mu}_h(0,1) = \frac{(1-\gamma_h)\mu_0}{(1-\gamma_h)\mu_0 + \gamma_l(1-\mu_0)} &\geq \bar{\mu} \\ \Leftrightarrow \qquad \mu_0(1-\bar{\mu})(1-\gamma_h) &\geq \bar{\mu}(1-\mu_0)\gamma_l \\ \Leftrightarrow \qquad \mu_0 &\geq \frac{\gamma_l\bar{\mu}}{(1-\gamma_h)(1-\bar{\mu})+\gamma_l\bar{\mu}} \end{aligned}$$

"$\Leftarrow$" NTS Sender cannot benefit from persuasion when $\mu_0 > \bar{\mu}$ or $\tilde{\mu}_h(0,1) < \bar{\mu}$.

For $\mu_0 > \bar{\mu}$. The Receiver takes action a_h at prior μ_0. The Sender gets the maximum payoff $v(a_h) = 1$ without persuasion.

For $\hat{\mu}_h(0,1) < \bar{\mu}$, NTS $\tilde{\mu}_h(\mu_l, \mu_h) < \bar{\mu}$ $\forall(\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$.

Applying the quotient rule to find the partial derivatives of the receiver's high posterior

belief with respect to each posterior belief of the sender,

$$\begin{aligned}
\frac{\partial \tilde{\mu}_h}{\partial \mu_h} &= \frac{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)\Big((\mu_0-\mu_l)(1-\gamma_h)+\mu_l\gamma_l\Big) - \gamma_l\Big((\mu_0-\mu_l)\mu_h(1-\gamma_h)+(\mu_h-\mu_0)\mu_l\gamma_l\Big)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2} \\
&= \frac{(\mu_0-\mu_l)(1-\gamma_h)\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l-(\mu_h-\mu_l)\gamma_l\Big)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2} \\
&= \frac{-(\mu_0-\mu_l)^2(1-\gamma_h)(1-\gamma_h-\gamma_l)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \tilde{\mu}_h}{\partial \mu_l} &= \frac{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)\Big(-\mu_h(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big) - \Big(-(1-\gamma_h)\Big)\Big((\mu_0-\mu_l)\mu_h(1-\gamma_h)+(\mu_h-\mu_0)\mu_l\gamma_l\Big)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2} \\
&= \frac{(\mu_h-\mu_0)\gamma_l\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l-(\mu_h-\mu_l)(1-\gamma_h)\Big)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2} \\
&= \frac{-(\mu_h-\mu_0)^2\gamma_l(1-\gamma_h-\gamma_l)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2}
\end{aligned}$$

With *infrequent* misinterpretation $\frac{\gamma_l}{1-\gamma_h} < 1$, $\frac{\partial \tilde{\mu}_h}{\partial \mu_h} > 0$ and $\frac{\partial \tilde{\mu}_h}{\partial \mu_l} < 0$. Thus, the receiver's high posterior is bounded from above by $\tilde{\mu}_h(0,1)$. If full informative revelation cannot convince the receiver who misinterprets to move posterior belief above $\bar{\mu}$ to switch to the high action a_h, then no information strategy can.

■

Proof. of Proposition 3

Both of $\tau_2(\mu_l,\mu_h)$ and $\tilde{\mu}_h(\mu_l,\mu_h)$ are quasiconcave in $(\mu_l,\mu_h) \in [0,\mu_0)\times(\mu_0,1]$. Applying Karush-Kuhn-Tucker Theorem, the Lagrangian is $\mathcal{L}(\mu_l,\mu_h,\lambda) = \tau_2(\mu_l,\mu_h)+\lambda\big(\tilde{\mu}_h(\mu_l,\mu_h)-\bar{\mu}\big)$

and the FOCs are

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu_l} &= \frac{\partial \tau_2}{\partial \mu_l} + \lambda \frac{\partial \tilde{\mu}_h}{\partial \mu_l} \leq 0 \text{ with equality if } \mu_l > 0 \\
\frac{\partial \mathcal{L}}{\partial \mu_h} &= \frac{\partial \tau_2}{\partial \mu_h} + \lambda \frac{\partial \tilde{\mu}_h}{\partial \mu_h} \leq 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \tilde{\mu}_h - \bar{\mu} \geq 0 \\
\lambda &\geq 0 \\
\lambda(\tilde{\mu}_h - \bar{\mu}) &= 0
\end{aligned}$$

WTS the constraint always binds at optimality, $\tilde{\mu}_h(\mu_l^*, \mu_h^*) = \bar{\mu}$.

Proof by contradiction. Suppose that the constraint doesn't bind. Then the complementary slackness implies $\lambda = 0$. $\frac{\partial \mathcal{L}}{\partial \mu_h} = \frac{\partial \tau_2}{\partial \mu_h} = -\frac{(\mu_0 - \mu_l)}{(\mu_h - \mu_l)^2}(1 - \gamma_h - \gamma_l) < 0$. Then, $\mu_h^* = \min\{\mu_h \in (\mu_0, 1] | \tilde{\mu}_h \geq \bar{\mu}\}$, which contradict with assumption since $\frac{\partial \tilde{\mu}_h}{\partial \mu_h} = \frac{(\mu_0 - \mu_l)^2(1-\gamma_h)(1-\gamma_h-\gamma_l)}{\Big((\mu_0-\mu_l)(1-\gamma_h)+(\mu_h-\mu_0)\gamma_l\Big)^2} > 0$ for *infrequent* misinterpretation. ■

Proof. of Corollary 1 (Welfare effects of misinterpretation)

1. Given a prior μ_0, a Receiver who misinterprets still switches to the high action a_h at the exact belief threshold that makes the receiver indifferent, like in the KG without interpretative errors. So, the receiver gets zero ex-ante payoffs with or without misinterpretation.

2. Given Proposition 2 that the constraint always binds in equilibrium, we have

$$\tilde{\mu}(\mu_l, \mu_h^*) = \bar{\mu} \Rightarrow \mu_h^* = \frac{\bar{\mu}(\mu_0 - \mu_l)(1 - \gamma_h - \gamma_l) - \mu_l \gamma_l (\bar{\mu} - \mu_0)}{(\mu_0 - \mu_l)(1 - \gamma_h - \gamma_l) - \gamma_l(\bar{\mu} - \mu_0)}.$$

   Substituting μ_h^* into the sender's problem, it reduces to

$$\max_{\mu_l} \tau_2(\mu_l) = \frac{\mu_0 - \mu_l}{\bar{\mu} - \mu_l}(1 - \gamma_h)$$

   Then, $\tau_2^{'} < 0$ for any $\mu_l \in [0, \mu_0)$ implies $\mu_l^* = 0$. Then, $\mu_h^* = \frac{\bar{\mu}}{1 - \frac{\gamma_l(\bar{\mu} - \mu_0)}{\mu_0(1-\gamma_h-\gamma_l)}} \leq 1$. The

optimal Sender's posterior $\mu^* = (\mu_l^*, \mu_h^*)$ are valid beliefs for $\mu_0 \geq \frac{\gamma_l \bar{\mu}}{(1-\gamma_h)(1-\bar{\mu})+\gamma_l \bar{\mu}}$.

The Sender's value from (*infrequently*) Misinterpreted Persuasion is

$$\begin{cases} 0 & \text{for } \mu_0 \in [0, \underline{\mu_0}) \\ \frac{\mu_0}{\bar{\mu}}(1-\gamma_h) & \text{for } \mu_0 \in [\underline{\mu_0}, \bar{\mu}), \text{ where } \underline{\mu_0} = \frac{\gamma_l \bar{\mu}}{(1-\gamma_h)(1-\bar{\mu})+\gamma_l \bar{\mu}} > 0 \text{ for } \gamma_l > 0. \\ 1 & \text{for} \mu_0 \in [\bar{\mu}, 1] \end{cases}$$

Compared to Sender's value from Bayesian persuasion $\begin{cases} 0 & \text{for } \mu_0 = 0 \\ \frac{\mu_0}{\bar{\mu}} & \text{for } \mu_0 \in (0, \bar{\mu}), \\ 1 & \text{for } \mu_0 \in [\bar{\mu}, 1] \end{cases}$ the favorable noise $\gamma_l > 0$ hurts the sender by enlarging the region of prior that renders persuasion useless; the unfavorable noise $\gamma_h > 0$ hurts the sender by shrinking the profit from persuasion.

■

A.2.2 Naïve receiver

Proof. of Proposition **??**ïveOptimalityNaïveOptimalityh naïveté misspecification, the sender's problem with *infrequent* misinterpretation is a *positive* linear transformation of the KG problem[15]. As a result, the equilibrium strategy remains the same as in KG, and so is the range of prior where the sender can benefit.

For $\mu_0 \in (0, \bar{\mu})$, the optimal Sender's posterior beliefs arrive at $(0, \bar{\mu})$ with probability $\tau_1^* = \begin{pmatrix} \tau_1^{l*} & \tau_1^{h*} \end{pmatrix} = \begin{pmatrix} 1 - \frac{\mu_0}{\bar{\mu}} & \frac{\mu_0}{\bar{\mu}} \end{pmatrix}$. But the Naïve Receiver's misspecified posterior beliefs arrive at $(0, \bar{\mu})$ with probability $\tau_2^* = \begin{pmatrix} \tau_2^{l*} & \tau_2^{h*} \end{pmatrix} = \tau_1^* \Gamma = \begin{pmatrix} \frac{\mu_0}{\bar{\mu}}(\gamma_h + \gamma_l - 1) + (1 - \gamma_l) & \frac{\mu_0}{\bar{\mu}}(1 - \gamma_h - \gamma_l) + \gamma_l \end{pmatrix}$. The Naïve Receiver's Bayesian posterior beliefs in equilibrium are

$$\tilde{\mu}^* = (\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(\frac{\mu_0 \gamma_h}{\frac{\mu_0}{\bar{\mu}}(\gamma_h + \gamma_l - 1) + (1 - \gamma_l)}, \frac{\mu_0 (1 - \gamma_h)}{\frac{\mu_0}{\bar{\mu}}(1 - \gamma_h - \gamma_l) + \gamma_l} \right),$$

[15] With *frequent* misinterpretation, this is instead a *negative* linear transformation of the KG problem.

which are Bayes-plausible with respect to τ_2^* and the receiver should have arrived at if he is correctly specified (Sophisticated/Bayesian). So, the Naïve Receiver switches to higher action a_h before his Bayesian posterior reaches the indifference belief $\bar{\mu}$. This happens if and only if there is favoritism:

$$\tilde{\mu}_h^* = \frac{\mu_0(1-\gamma_h)}{\mu_0(1-\gamma_h)+\gamma_l(\bar{\mu}-\mu_0)}\bar{\mu} < \bar{\mu} \Leftrightarrow \gamma_l > 0$$

■

Proof. of Corollary 2 (Welfare effects of naïveté misspecification)

From Proposition **??**ïveOptimalityNaïveOptimalitynow that for a prior $\mu_0 \in (0, \bar{\mu})$, the sender's optimal strategy is to induce her Bayesian posterior and the receiver's misspecified posterior to $\mu^* = (\mu_l^*, \mu_h^*) = (0, \bar{\mu})$. Therefore, if the receiver is Bayesian about the misinterpretation mistakes, he should have arrived at his Bayesian posteriors

$$\tilde{\mu}^* = (\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(\frac{\mu_0\gamma_h}{\frac{\mu_0}{\bar{\mu}}(\gamma_h+\gamma_l-1)+(1-\gamma_l)}, \frac{\mu_0(1-\gamma_h)}{\frac{\mu_0}{\bar{\mu}}(1-\gamma_h-\gamma_l)+\gamma_l} \right),$$

1. Receiver's welfare in equilibrium:

   Denote $\hat{a}(\cdot) : \Delta(\Omega) \to \mathcal{A}$ as the receiver's best response function to a belief. The Naïve Receiver's welfare from being persuaded is calculated as the objective expected payoffs

from the misspecified posterior beliefs:

$$\begin{aligned}
\mathbb{E}_{\tilde{\mu}} u\big(\hat{a}(\mu),\omega\big) =&\tau_2^h\Big(\tilde{\mu}_h u(a_h,H)+(1-\tilde{\mu}_h)u(a_h,L)\Big)+\tau_2^l\Big(\tilde{\mu}_l u(a_l,H)+(1-\tilde{\mu}_l)u(a_l,L)\Big)\\
=&\mu_0(1-\gamma_h)\Big(u(a_h,H)-u(a_h,L)\Big)+\tau_2^h u(a_h,L)\\
&+\mu_0\gamma_h\Big(u(a_l,H)-u(a_l,L)\Big)+\tau_2^l u(a_l,L)\\
=&\mu_0\Big(u(a_h,H)-u(a_h,L)\Big)-\mu_0\gamma_h\Big(u(a_h,H)-u(a_h,L)-u(a_l,H)+u(a_l,L)\Big)\\
&+u(a_l,L)-\tau_2^h\Big(u(a_l,L)-u(a_h,L)\Big)\\
=&\mu_0\Big(u(a_h,H)-u(a_h,L)\Big)-\mu_0\gamma_h\frac{1}{\bar{\mu}}\Big(u(a_l,L)-u(a_h,L)\Big)\\
&+u(a_l,L)-\tau_2^h\Big(u(a_l,L)-u(a_h,L)\Big)
\end{aligned}$$

The first equality spells out the ex-ante expected payoffs for the receiver, who best responds to misspecified posterior beliefs μ but he should've best responded to his Bayesian posterior $\tilde{\mu}$. The second equality is due to Bayes-plausibility. The third equality rearranges the terms. The fourth equality replaces some of the terms using the following indifference condition at $\bar{\mu}$:

$$\Big(u(a_h,H)-u(a_l,H)\Big)+\Big(u(a_l,L)-u(a_h,L)\Big)=\frac{1}{\bar{\mu}}\Big(u(a_l,L)-u(a_h,L)\Big).$$

In equilibrium, we evaluate the above equation at $\mu^*=(0,\bar{\mu})$,

$$\begin{aligned}
\mathbb{E}_{\tilde{\mu}^*} u(\hat{a}(\mu^*),\omega) =&\mu_0\Big(u(a_h,H)-u(a_h,L)\Big)-\mu_0\gamma_h\frac{1}{\bar{\mu}}\Big(u(a_l,L)-u(a_h,L)\Big)\\
&+u(a_l,L)-\Big(\frac{\mu_0}{\bar{\mu}}(1-\gamma_h-\gamma_l)+\gamma_l\Big)\Big(u(a_l,L)-u(a_h,L)\Big)\\
=&\mu_0\Big(u(a_h,H)-u(a_h,L)-u(a_l,H)+u(a_l,L)\Big)+\mu_0 u(a_l,H)+(1-\mu_0)u(a_l,L)\\
&-\Big(\frac{\mu_0}{\bar{\mu}}(1-\gamma_l)+\gamma_l\Big)\Big(u(a_l,L)-u(a_h,L)\Big)\\
=&\underbrace{\mu_0 u(a_l,H)+(1-\mu_0)u(a_l,L)}_{\text{welfare at prior}}+\underbrace{\Big(\frac{\mu_0}{\bar{\mu}}-1\Big)\gamma_l\Big(u(a_l,L)-u(a_h,L)\Big)}_{<0 \text{ iif } \gamma_l>0}
\end{aligned}$$

The first equality substitutes τ_2^h in equilibrium. The second equality adds zero-sum terms ($\pm\mu_0 u(a_l, H)$) and rearranges terms. The last equality again uses the indifference condition at $\bar{\mu}$.

From Corollary 1, we know that neither favorable ($\gamma_l > 0$) nor unfavorable noise ($\gamma_h > 0$) affects the Sophisticated Receiver, who is always made indifferent in equilibrium between the prior and ex-ante at posteriors, like in the KG. Compared to KG and Misinterpreted only, naïveté misspecification has no welfare effect on the receiver if there is no favoritism ($\gamma_l = 0$). Moreover, the receiver is strictly worse off if and only if there is favoritism ($\gamma_l > 0$) AND the receiver is naïve about it.

2. Sender's welfare in equilibrium:

   The Sender's optimal profit from naively misinterpreted persuasion is

$$\begin{cases} 0 & \text{for } \mu_0 = 0 \\ \frac{\mu_0}{\bar{\mu}}(1 - \gamma_h - \gamma_l) + \gamma_l & \text{for } \mu_0 \in (0, \bar{\mu}) \\ 1 & \text{for } \mu_0 \in [\bar{\mu}, 1] \end{cases}.$$

   Compared to Misinterpreted only, the sender is strictly better off for the range of prior that the sender benefits from naively misinterpreted persuasion, $\mu_0 \in (0, \bar{\mu})$.

■

Proof. of Corollary 3 (Composite welfare effects of misinterpretation and naïveté misspecification)

If the receiver misinterprets and is also naively misspecified, the sender can do better

than KG when the prior is small,

$$\begin{aligned}\frac{\mu_0}{\bar{\mu}}(1-\gamma_h-\gamma_l)+\gamma_l &> \frac{\mu_0}{\bar{\mu}}\\ \gamma_l &> \frac{\mu_0}{\bar{\mu}}(\gamma_h+\gamma_l)\\ \frac{\gamma_l\bar{\mu}}{\gamma_l+\gamma_h} &> \mu_0\end{aligned}$$

Conversely, the sender is strictly worse off than in KG when the prior is large, $\mu_0 \in \left(\frac{\gamma_l\bar{\mu}}{\gamma_l+\gamma_h}, \bar{\mu}\right)$. ■

A.3 Confirmation Bias

A.3.1 Sophisticated Confirmation Bias

Proof. of Proposition 5

1. **Step 1 Case 1:**

   First, we search for a solution in $\left\{(\mu_l,\mu_h) \mid \mu_0 \leq \frac{\mu_h+(1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}$ under $\Gamma_h \coloneqq \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint of the posterior beliefs, which imposes the posterior beliefs to a half-space in (μ_l,μ_h).

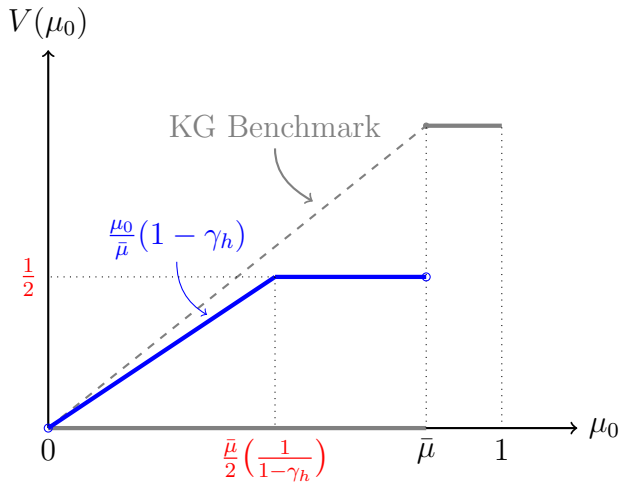
   With Sophistication, the receiver updates to his Bayesian posterior $\tilde{\mu}$. Under Γ_h, the receiver's high posterior $\tilde{\mu}_h$ equals to Sender's high posterior μ_h. The Sender solves the following problem:

$$\begin{aligned}\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\tau_h)&\\ \text{s.t. } \mu_h \geq \bar{\mu} &\qquad (O_1^S)\\ \mu_0 \leq \frac{\mu_h+(1-2\gamma_h)\mu_l}{2(1-\gamma_h)} &\qquad (CB_1^S)\end{aligned}$$

   Without the confirmation bias constraint on the posterior beliefs (CB_1^S), an optimal information policy induces Sender's posterior to $(0,\bar{\mu})$ by Corollary 1 and Sender gets

$\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$. For $\mu_0 \in \left(0, \frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right)\right]$, the CB_1^S constraint doesn't bind at the optimal Sender posterior $(0, \bar{\mu})$. For $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}\right)$, to satisfy the optimality (O_1^S) and the posterior (CB_1^S) constraints simultaneously, Sender can still induce $\hat{\mu}_h = \bar{\mu}$ by increasing μ_l so that CB_1^S is exactly satisfied. Then, Sender gets $\frac{1}{2}$. Figure 7A depicts the sender's value function with the Sophisticated Receiver in Case 1.

Figure 7A: Case 1 Value Function with Sophisticated Receiver



2. **Step 1 Case 2:**

   Next, we search for a solution in $\left\{(\mu_l, \mu_h) \mid \mu_0 > \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}$ under $\Gamma_l = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix}$. The additional posterior constraint (CB_2^S) restricts the solution to the other half-space in (μ_l, μ_h), as opposed to CB_1^S in Case 1.

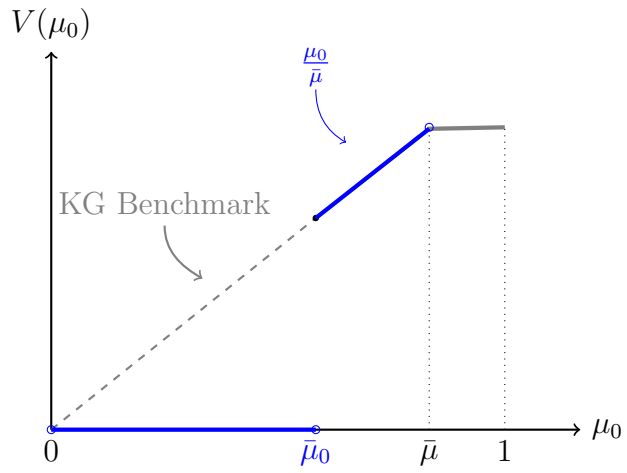
   With Sophistication, the receiver updates to his Bayesian posterior $\tilde{\mu}$. Under Γ_h, the receiver's high posterior $\tilde{\mu}_h$ is strictly less than the sender's high posterior μ_h. The

Sender solves the following problem:

$$\begin{aligned}
&\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\gamma_l)+\gamma_l \\
&\text{s.t. } \tilde{\mu}_h(\mu_l,\mu_h) = \frac{(\mu_0-\mu_l)\mu_h+\gamma_l(\mu_h-\mu_0)\mu_l}{(\mu_0-\mu_l)+\gamma_l(\mu_h-\mu_0)} \geq \bar{\mu} && (O_2^S) \\
&\mu_0 > \frac{\mu_h+(1-2\gamma_h)\mu_l}{2(1-\gamma_h)} && (CB_2^S)
\end{aligned}$$

When both the confirmation bias (CB_2^S) constraint and the optimality (O) constraint are satisfied, the sender can achieve the concavification value as in the KG benchmark. When either constraint is violated, the sender cannot benefit from persuasion since no information policy can induce the receiver to take the sender-preferred action a_h. Given a problem with indifference threshold $\bar{\mu}$, prior μ_0, and bias parameters γ_l and γ_h, each of the CB_2^S and O constraints produces a belief cutoff at optimal: $\bar{\mu}_0^{CB_2^S} := \frac{\bar{\mu}}{2(1-\gamma_h)}\Big(1+\gamma_l(1-2\gamma_h)\Big)$ and $\bar{\mu}_0^{O_2^S} := \frac{\gamma_l\bar{\mu}}{\gamma_l\bar{\mu}+1-\bar{\mu}}$[16] respectively. If either is violated, no strategy can induce the receiver to take the a_h action and the sender always gets 0. Therefore the cutoff belief $\bar{\mu}_0$ of the value function is just the larger of $\bar{\mu}_0^{CB_2^S}$ and $\bar{\mu}_0^{O_2^S}$.

Figure 7B: Case 2 Value Function with Sophisticated Receiver

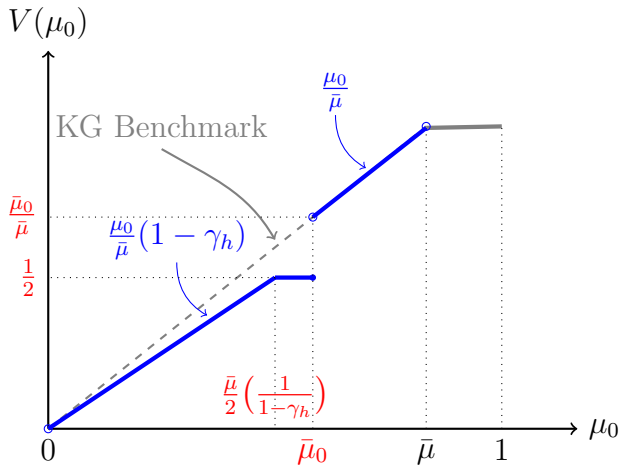


[16]Note that $\bar{\mu}_0^{O_2^S}$ is just a special case of $\underline{\mu_0}$ in the binary model.

3. **Step 2 best of the two cases:**

Now, we have solved the two cases separately. Given a prior μ_0, the sender can affect the effective direction of the bias by choosing different posterior pair (μ_l, μ_h). So, she chooses the better between the two cases at each prior. For low priors below $\bar{\mu}_0$, Γ_h takes effect and the receiver misinterprets against the sender in equilibrium; for high priors above $\bar{\mu}_0$, Γ_l takes effect and the receiver misinterprets in favor of the sender in equilibrium. The following figure summarizes the sender's value at optimal with a Sophisticated confirmatory biased Receiver in Proposition 5.

Figure 7: Value Function with Sophisticated Confirmation Bias



■

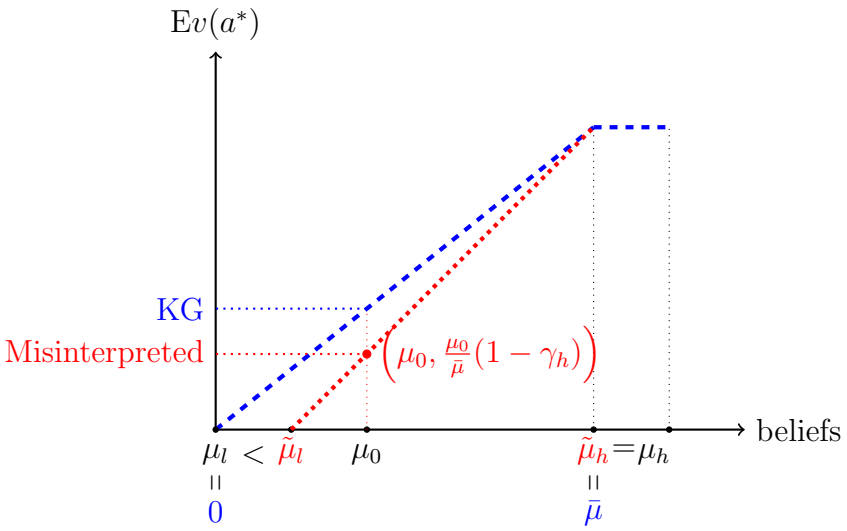
Example Solutions:

In the remainder of this subsection, we showcase representative solutions at prior μ_0 in each of the three intervals. From these examples, we can see that the receiver always makes the optimal decisions by switching to higher action at the correct indifference belief threshold, $\tilde{\mu}_h = \bar{\mu}$. If you are eager to learn the impact of naïve misspecification on top of confirmatory biased misinterpretation, skip to the next subsection.

(1) For $\mu_0 \in (0, \frac{\bar{\mu}}{2}(\frac{1}{1-\gamma_h})]$, the receiver misinterprets under Γ_h in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- The sender updates to the sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = (0, \bar{\mu})$;
- The receiver updates to the receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(\frac{\gamma_h \mu_0 \bar{\mu}}{\gamma_h \mu_0 + \bar{\mu} - \mu_0}, \bar{\mu}\right)$;
- The sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$.

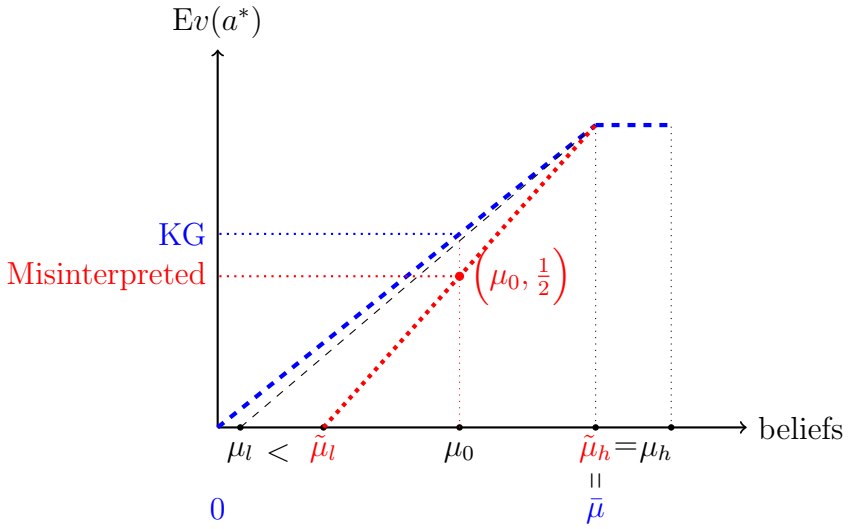
Figure 8A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}(\frac{1}{1-\gamma_h})]$



(2) For $\mu_0 \in \left(\frac{\bar{\mu}}{2}(\frac{1}{1-\gamma_h}), \bar{\mu}_0\right]$, the receiver misinterprets under Γ_h in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- The sender updates to the sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = (\frac{2\mu_0 - \bar{\mu} + \gamma_h \bar{\mu}}{1+\gamma_h}, \bar{\mu})$;
- The receiver updates to the receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(2\mu_0 - \bar{\mu}, \bar{\mu}\right)$;
- The sender gets $\frac{1}{2}$.

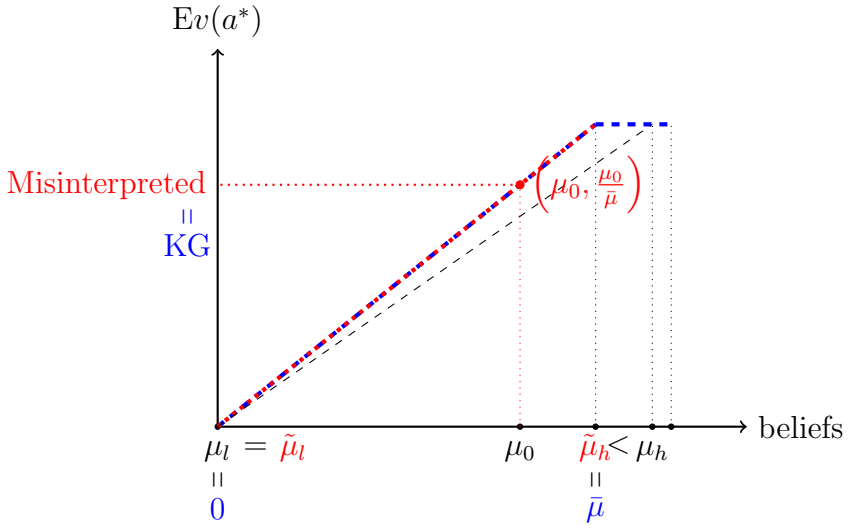
Figure 8B: solution at $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}_0\right]$



(3) For $\mu_0 \in (\bar{\mu}_0, \bar{\mu})$, the receiver misinterprets under Γ_l in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- The sender updates to the sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = \left(0, \frac{\bar{\mu}}{1-\frac{\gamma_l(\bar{\mu}-\mu_0)}{\mu_0(1-\gamma_l)}}\right)$;
- The receiver updates to the receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = (0, \bar{\mu})$;
- The sender gets $\frac{\mu_0}{\bar{\mu}}$.

Figure 8C: solution at $\mu_0 \in (\bar{\mu}_0, \bar{\mu})$



A.3.2 Naïve Confirmation Bias

Proof. of Proposition 6

1. **Step 1 Case 1:**

   First, we search for a solution in $\left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + \mu_l}{2}\right\}$ under $\Gamma_h = \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint on the posterior beliefs, which imposes solutions to a half-space in (μ_l, μ_h).

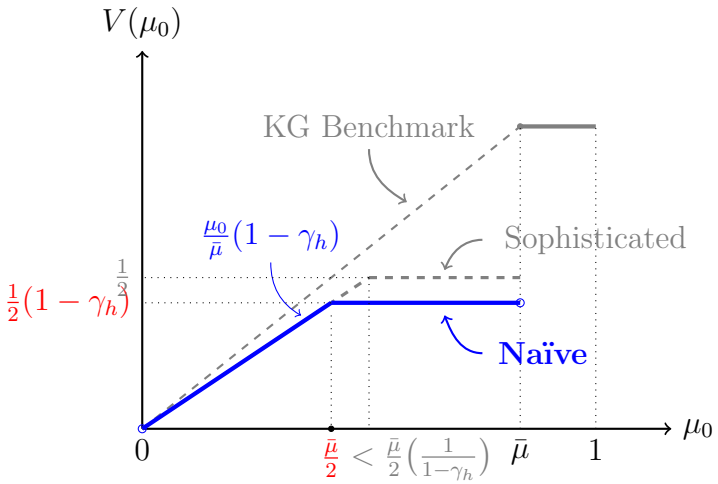
   With Naïveté misspecification, the receiver updates to a misspecified posterior coinciding with the sender's Bayesian posterior μ. Under Γ_h, the receiver's Bayesian high posterior $\tilde{\mu}_h$ equals the sender's high posterior μ_h. Thus, the receiver makes optimal decisions in equilibrium even with misspecification.

The Sender solves the following problem:

$$\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\tau_h)$$
$$\text{s.t. } \mu_h \geq \bar{\mu} \qquad (O^N)$$
$$\mu_0 \leq \frac{\mu_h+\mu_l}{2} \qquad (CB_1^N)$$

Without the confirmation bias constraint on the posterior beliefs (CB_1^N), an optimal information policy induces Sender's posterior to $(0,\bar{\mu})$ by Corollary 2 and Sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$. For low priors $\mu_0 \in \left(0, \frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right)\right]$, the CB_1^N constraint doesn't bind at the optimal Sender's posterior $(0,\bar{\mu})$. For high priors $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}\right)$, to satisfy the persuasion (O^N) and the posterior (CB_1^N) constraints simultaneously, Sender can still induce Receiver's misspecified posterior μ_h to $\bar{\mu}$ by increasing μ_l so that CB_1^N is exactly satisfied. So, Sender gets $\frac{1}{2}(1-\gamma_h)$ in equilibrium at high priors. Figure 9A depicts the sender's value function with a naïve confirmatory biased Receiver in Case 1.

Figure 9A: Case 1 Value Function with Naïve Receiver



2. **Step 1 Case 2:**

Next, we search for a solution in $\left\{(\mu_l, \mu_h) \mid \mu_0 > \frac{\mu_h+\mu_l}{2}\right\}$ under $\Gamma_l = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix}$. The additional posterior constraint (CB_2^N) restricts solutions to the other half-space in (μ_l, μ_h), as opposed to CB_1^N in Case 1.

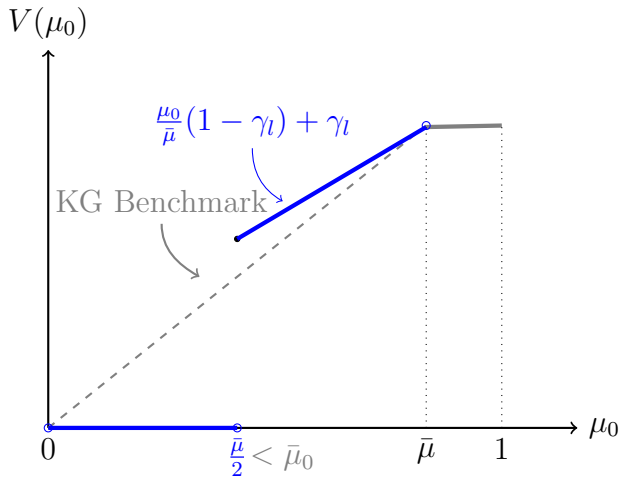
With Naïveté misspecification, the receiver updates to misspecified posterior coinciding with the sender's posterior μ like in Case 1. But the receiver's Bayesian high posterior $\tilde{\mu}_h$ is strictly less than his misspecified high posterior μ_h under Γ_l. Thus, the receiver makes a sub-optimal decision at his misspecified high posterior in equilibrium.

The Sender solves the following problem:

$$\begin{aligned} \max_{\mu_l, \mu_h} \tau_2(\mu_l, \mu_h) &= \frac{\mu_0 - \mu_l}{\mu_h - \mu_l}(1-\gamma_l) + \gamma_l \\ \text{s.t. } \mu_h &\geq \bar{\mu} && (O^N) \\ \mu_0 &> \frac{\mu_h + \mu_l}{2} && (CB_2^N) \end{aligned}$$

When both the confirmation bias (CB_2^N) constraint and the persuasion (O^N) constraint are satisfied, the sender can achieve better than the concavification value as in the KG benchmark. When either constraint is violated, the sender cannot benefit from persuasion since no information policy can induce the receiver to take the sender-preferred action a_h. Since the receiver is Naïve, only CB_2^N produces a prior cutoff in equilibrium: $\frac{\bar{\mu}}{2}$. For prior below the cutoff, no strategy can induce the receiver to take the a_h action and the sender always gets 0.

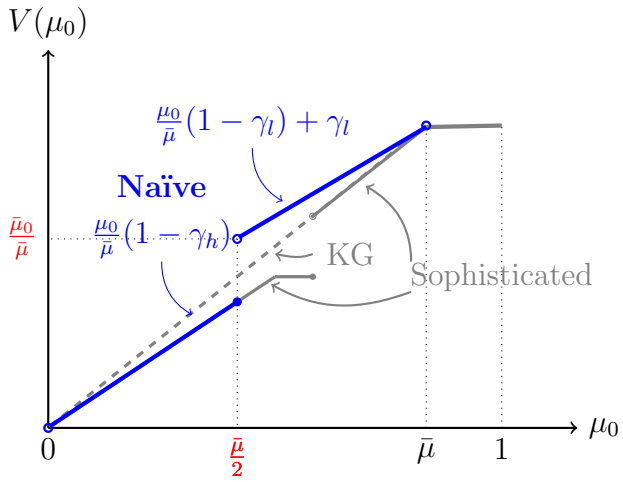
Figure 9B: Case 2 Value Function with Naïve Receiver



3. **Step 2 best of the two cases:**

   Now, we have solved the two cases separately. Given a prior μ_0, the sender can decide the effective direction of the bias by choosing between the posterior pairs (μ_l, μ_h). So, she induces the posterior that produces a better expected payoff for her at each prior. The Naïve confirmatory biased Receiver still misinterprets against the sender for low priors and misinterprets in favor of the sender for high priors in equilibrium. But the Naïve Receiver's prior range that favors the sender is larger than the Sophisticated Receiver's. The following figure summarizes the sender's value at optimal with a Naïve confirmatory biased Receiver in Proposition 6.

Figure 9: Value Function with Naïve Confirmation Bias



■

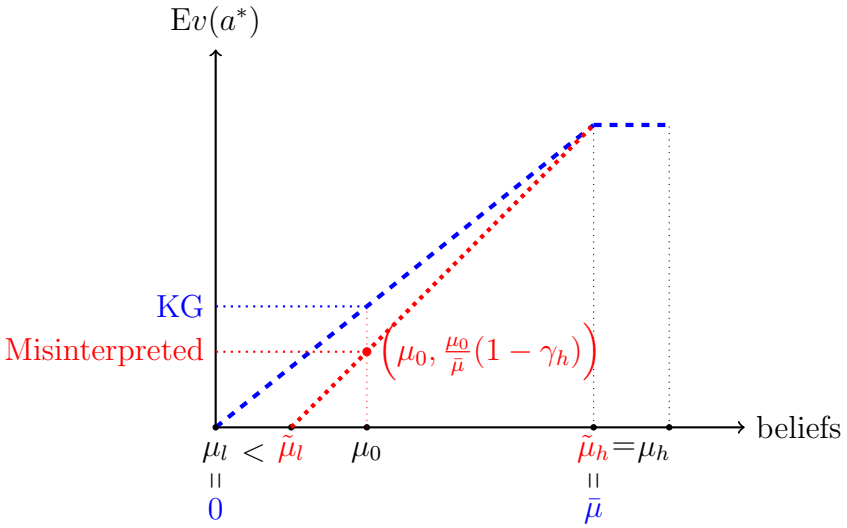
Example Solutions:

Like in the sophisticated case, the remainder of this subsection showcases example solutions with a naïve receiver for μ_0 in each interval. These examples demonstrate that the naïve receiver is worse off if and only if there are favorable misinterpretations in equilibrium.

(1) For $\mu_0 \in (0, \frac{\bar{\mu}}{2}]$, the receiver misinterprets under Γ_h in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- Both the sender and the (misspecified) receiver update to the sender's Bayesian posteriors at $(0, \bar{\mu})$.
- The receiver Bayesian posteriors should arrive at $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(\frac{\gamma_h \mu_0 \bar{\mu}}{\gamma_h \mu_0 + \bar{\mu} - \mu_0}, \bar{\mu}\right)$;
- The sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$.

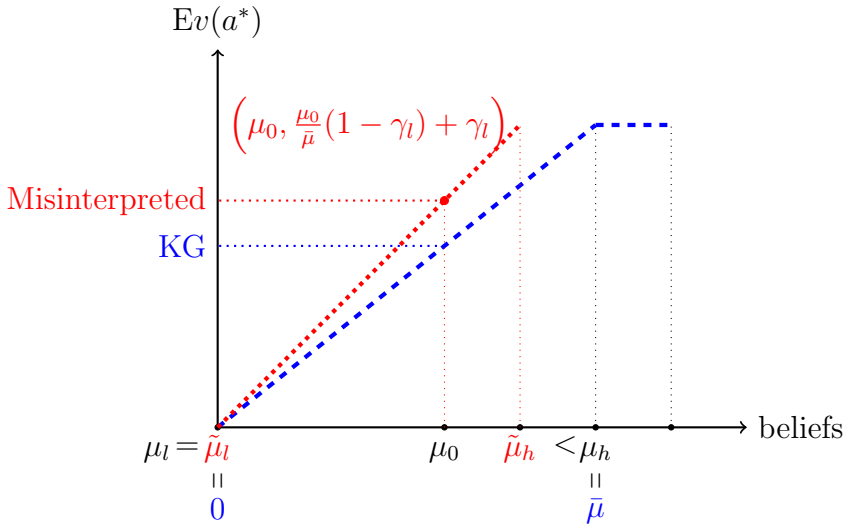
Figure 10A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}]$



(2) For $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$, the receiver misinterprets under Γ_l in equilibrium and makes sub-optimal decision ($\tilde{\mu}_h^* < \bar{\mu}$).

- Both the sender and the (misspecified) receiver update to the sender's Bayesian posteriors at $(0, \bar{\mu})$.
- The receiver Bayesian posteriors should arrive at $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(0, \frac{\bar{\mu}}{1+\gamma_l(\frac{\bar{\mu}}{\mu_0}-1)}\right)$;
- The sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_l) + \gamma_l$.

Figure 10B: solution at $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$



B Results for Frequent Misinterpretation

B.1 Sophisticated Receiver who Frequently Misinterprets

With *frequent* misinterpretations $\left(\frac{\gamma_l}{1-\gamma_h} > 1\right)$, the meaning of the realizations flips between the sender and the receiver. Suppose the sender updates to $(\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$. The realizations are flipped for the receiver's Bayesian posteriors, $(\tilde{\mu}_h, \tilde{\mu}_l) \in [0, \mu_0) \times (\mu_0, 1]$.

For $\mu_0 \in (0, \bar{\mu})$, the sender solves

$$\max_{\substack{\mu_l \in [0,\mu_0),\\ \mu_h \in (\mu_0,1]}} \tau_2^l(\mu_l, \mu_h)$$

$$\text{s.t. } \tilde{\mu}_l(\mu_l, \mu_h) \geq \bar{\mu} \qquad (O_f^S)$$

where

$$\tau_2^l(\mu_l, \mu_h) = \frac{\mu_0 - \mu_l}{\mu_h - \mu_l}(\gamma_h + \gamma_l - 1) + (1 - \gamma_l)$$
$$\tilde{\mu}_l(\mu_l, \mu_h) = \frac{\gamma_h(\mu_0 - \mu_l)\mu_h + (1 - \gamma_l)(\mu_h - \mu_0)\mu_l}{\gamma_h(\mu_0 - \mu_l) + (1 - \gamma_l)(\mu_h - \mu_0)}$$

We solve the above problem using the same method as in the *infrequent* misinterpretation case. In equilibrium, the sender still wants to induce the receiver's Bayesian posterior to equal the indifference threshold $\bar{\mu}$. Given a prior $\mu_0 \in [\underline{\mu_0^f}, \bar{\mu})$, the optimal Sender's posterior beliefs are at $(\mu_l^*, \mu_h^*) = \left(0, \frac{\frac{\gamma_h}{1-\gamma_l} - 1}{\frac{\gamma_h}{1-\gamma_l} - \frac{\bar{\mu}}{\mu_0}}\bar{\mu}\right)$. Similarly, $\underline{\mu_0^f}$ is calculated from the condition that Sender's posterior belief has to be valid probability:

$$\mu_h^* = \frac{\frac{\gamma_h}{1-\gamma_l} - 1}{\frac{\gamma_h}{1-\gamma_l} - \frac{\bar{\mu}}{\mu_0}}\bar{\mu} \leq 1$$
$$\Updownarrow$$
$$\underline{\mu_0^f} \coloneqq \frac{(1 - \gamma_l)\bar{\mu}}{\gamma_h(1 - \bar{\mu}) + (1 - \gamma_l)\bar{\mu}} \leq \mu_0.$$

The Receiver knows that the realizations mean the opposite of what the sender designed to be. He arrives at his Bayesian posterior beliefs $(\tilde{\mu}_h^*, \tilde{\mu}_l^*) = \left(\frac{\mu_0(1-\gamma_h)}{1-\frac{\mu_0}{\bar{\mu}}\gamma_h}, \bar{\mu}\right)$ with probabilities $\tau_2^* = (1 - \frac{\mu_0}{\bar{\mu}}\gamma_h, \frac{\mu_0}{\bar{\mu}}\gamma_h)$. So the sender's value from *frequently* Misinterpreted Persuasion is

$$\begin{cases} 0 & \text{for } \mu_0 \in [0, \underline{\mu_0^f}) \\ \frac{\mu_0}{\bar{\mu}}\gamma_h & \text{for } \mu_0 \in [\underline{\mu_0^f}, \bar{\mu}) \\ 1 & \text{for} \mu_0 \in [\bar{\mu}, 1] \end{cases},$$

where $\underline{\mu_0^f} = \frac{(1-\gamma_l)\bar{\mu}}{\gamma_h(1-\bar{\mu})+(1-\gamma_l)\bar{\mu}} > 0$ for $\gamma_l < 1$.

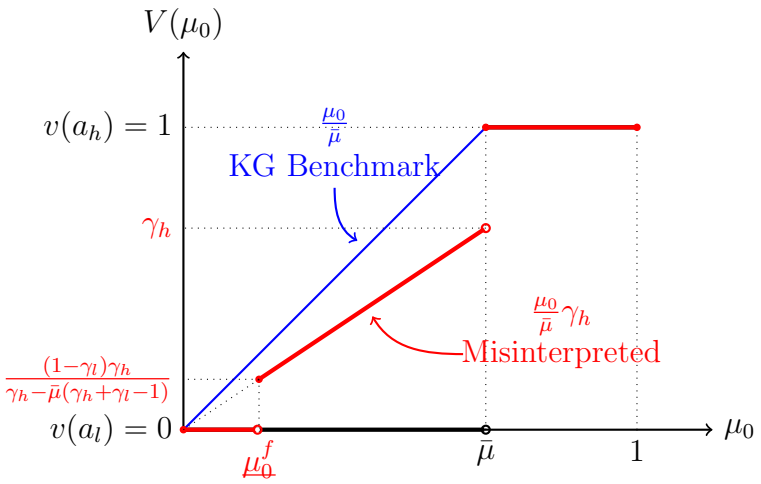


Figure 1^f: Value function comparison
— with *frequent* misinterpretation
— without misinterpretation

B.2 naïve Receiver who Frequently Misinterprets

If the receiver is naïve, he doesn't know that the Bayesian meaning of the realizations is flipped. The Sender solves the same problem as in the *infrequent* naïve misinterpretation case under a different condition of the parameters.

Suppose the sender updates to $(\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$. Then, the receiver updates to misspecified posterior beliefs $(\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$, but he should have flipped the meaning of the realizations and updated to the receiver's Bayesian posteriors, $(\tilde{\mu}_h, \tilde{\mu}_l) \in [0, \mu_0) \times (\mu_0, 1]$.

With *frequent* misinterpretations ($\frac{\gamma_l}{1-\gamma_h} > 1$), the sender's problem is a *negative* linear transformation of the KG problem. For $\mu_0 \in (0, \bar{\mu})$, the sender solves

$$\max_{\substack{\mu_l \in [0, \mu_0), \\ \mu_h \in (\mu_0, 1]}} \tau_2^h(\mu_l, \mu_h) = \tau_1^h(\mu_l, \mu_h)(1 - \gamma_h - \gamma_l) + \gamma_l$$

$$\text{s.t. } \mu_h \geq \bar{\mu} \qquad (O^N)$$

The optimal strategy induces the posterior distribution to minimize $\tau_1^h(\mu_l, \mu_h) = \frac{\mu_0 - \mu_l}{\mu_h - \mu_l}$. So, the solution with *frequent* misinterpretation flips μ_l^* and μ_h^* of the solution with *infrequent* naïve misinterpretation[17]. Thus, for $\mu_0 \in (0, \bar{\mu})$, the sender's optimal profit from *frequent* naively misinterpreted persuasion induces the receiver's Bayesian posterior distribution to $\tau_2^* = \begin{pmatrix} \tau_2^{l*} & \tau_2^{h*} \end{pmatrix} = \begin{pmatrix} \frac{\mu_0}{\bar{\mu}}(1 - \gamma_h - \gamma_l) + \gamma_h & \frac{\mu_0}{\bar{\mu}}(\gamma_h + \gamma_l - 1) + (1 - \gamma_h) \end{pmatrix}$ over the posterior beliefs $\mu^* = (\mu_l^*, \mu_h^*) = (\bar{\mu}, 0)$. In summary, the sender's value function is

$$\begin{cases} 0 & \text{for } \mu_0 = 0 \\ \frac{\mu_0}{\bar{\mu}}(\gamma_h + \gamma_l - 1) + (1 - \gamma_h) & \text{for } \mu_0 \in (0, \bar{\mu}) \\ 1 & \text{for } \mu_0 \in [\bar{\mu}, 1] \end{cases}.$$

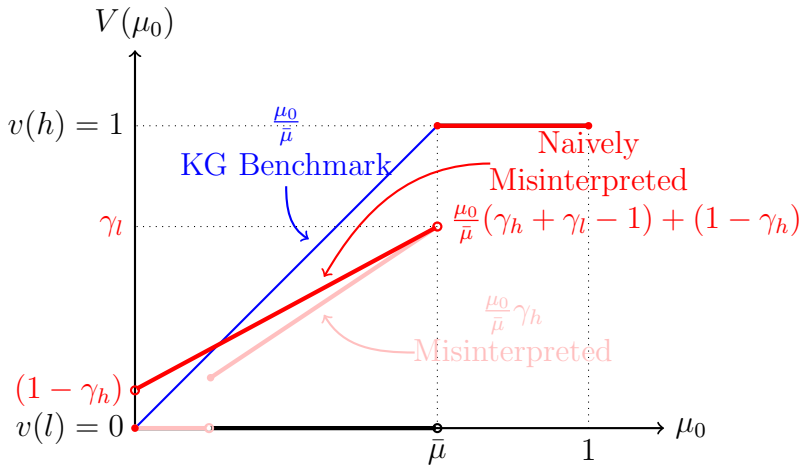


Figure 4^f: Value function comparison
- with *frequent* misinterpretation and naïveté
- with *frequent* misinterpretation and sophistication
- without misinterpretation

[17] Remember that the solution with *infrequent* naïve misinterpretation induces the receiver's Bayesian posterior distribution to $\tau_2^* = \begin{pmatrix} \tau_2^{l*} & \tau_2^{h*} \end{pmatrix} = \begin{pmatrix} \frac{\mu_0}{\bar{\mu}}(\gamma_h + \gamma_l - 1) + (1 - \gamma_l) & \frac{\mu_0}{\bar{\mu}}(1 - \gamma_h - \gamma_l) + \gamma_l \end{pmatrix}$ over the posterior beliefs $\mu^* = (\mu_l^*, \mu_h^*) = (0, \bar{\mu})$.

C Comparative Statics and Policy Implications

C.1 Comparative Statics

C.1.1 Sophisticated Receiver

In the main section, we have solved and analyzed the effects of misinterpretation in persuasion. We attribute the effect of each direction to different channels. A natural question would be how these effects change with the variation of parameters. Combining results from *frequent* misinterpretation in Appendix B, this subsection shows comparative statics of misinterpretation with a sophisticated receiver.

From the previous analysis, γ_l negatively affects the sender by limiting her ability to raise the sophisticated receiver's posterior belief. With *infrequent* misinterpretations ($\gamma_l + \gamma_h < 1$), the sender can benefit from misinterpreted persuasion for $\mu_0 \geq \underline{\mu_0}$. With *frequent* misinterpretations ($\gamma_l + \gamma_h > 1$), the sender can benefit from misinterpreted persuasion for $\mu_0 \geq \underline{\mu_0^f}$. As γ_l increases but the total noise is infrequent such that the realizations don't indicate opposite meaning to the sender and the receiver, more misinterpretation hinders information transmission: $\underline{\mu_0}$ increases in γ_l. However, as γ_l continues to increase passing the point where the meaning of realizations flips between the sender and the receiver, more misinterpretation starts to ameliorate the negative impact because the opposite meaning gets more informative: $\underline{\mu_0^f}$ decreases in γ_l. The following graph plots the range of priors where the sender can benefit from misinterpreted persuasion against γ_l.

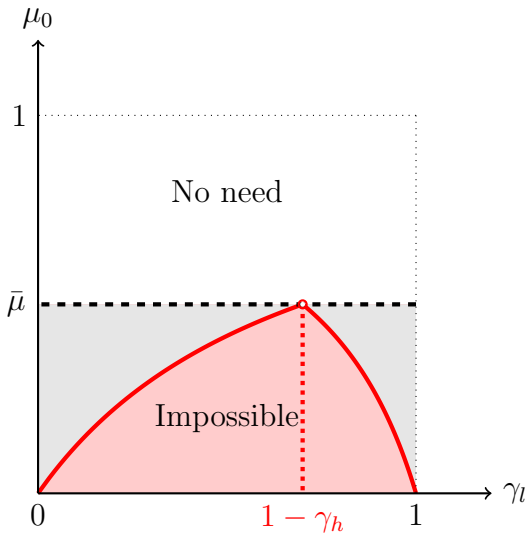


Figure 11: γ_l's Impact on Range of Prior that Sender Benefits
The graph shown fixing $\gamma_h = 0.3$ and $\bar{\mu} = 0.5$
persuasion channel shuts down due to γ_l
the range of priors where the sender benefits
the upper bound of persuasion: $\bar{\mu}$ exclusive
the lower bound of persuasion: $\underline{\mu_0}(\gamma_l)$ and $\underline{\mu_0^f}(\gamma_l)$ inclusive

For a large enough prior that the sender benefits from misinterpreted persuasion[18], γ_l has no impact on the persuasion profit and γ_h negatively affects the sender by limiting her ability to lower the sophisticated receiver's posterior belief. With infrequent misinterpretations, γ_h increasingly compresses the sender's profit; with frequent misinterpretations, γ_h gradually restores the sender's profit. The following graph depicts the effect of γ_h on the sender's persuasion profit at two examples of prior μ_0.

Fixing prior μ_0 and γ_l, for *infrequent* misinterpretation ($\gamma_h < 1-\gamma_l$), γ_h has to be small enough for the persuasion channel to be possible: $\mu_0 \geq \underline{\mu_0} \Leftrightarrow \gamma_h \leq 1 - \gamma_l \frac{\bar{\mu}(1-\mu_0)}{\mu_0(1-\bar{\mu})} =: \bar{\gamma}_h$. For *frequent* misinterpretation ($\gamma_h > 1-\gamma_l$), γ_h has to be large enough for the persuasion channel to be possible: $\mu_0 \geq \underline{\mu_0^f} \Leftrightarrow \gamma_h \geq (1-\gamma_l)\frac{\bar{\mu}(1-\mu_0)}{\mu_0(1-\bar{\mu})} =: \bar{\gamma}_h^f$.

[18] $\mu_0 \geq \underline{\mu_0}$ with *infrequent* misinterpretation and $\mu_0 \geq \underline{\mu_0^f}$ with *frequent* misinterpretation

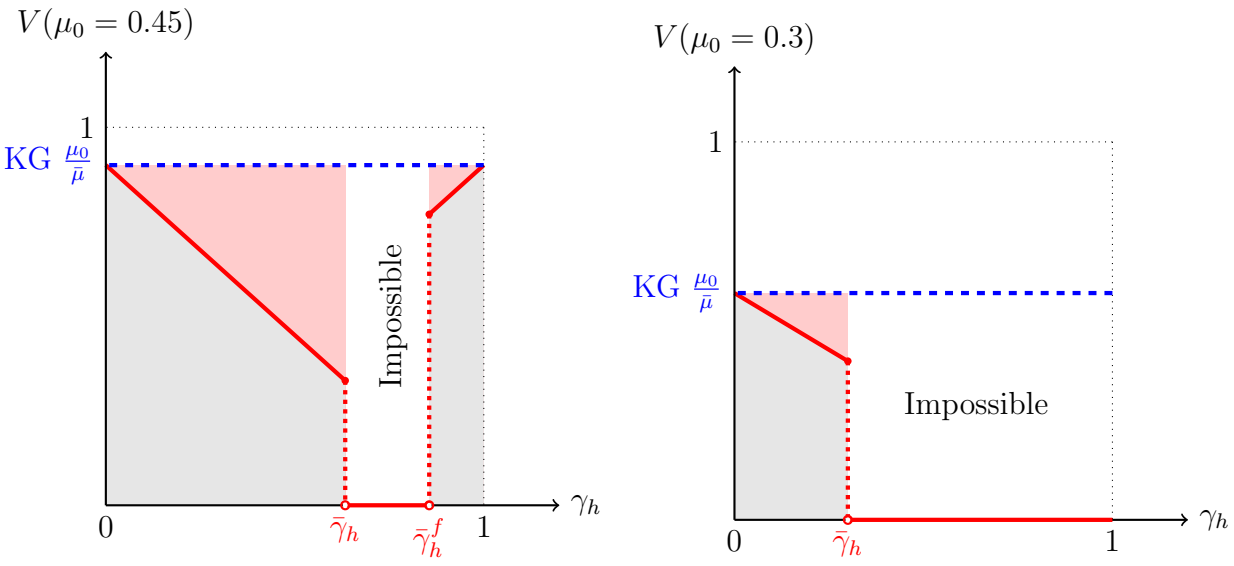


Figure 12: γ_h's Impact on Sender Profit from Misinterpreted Persuasion
The graph shown fixing $\gamma_l = 0.3$ and $\bar{\mu} = 0.5$
informational loss due to γ_h
Sender's profit from persuasion with misinterpretation
the optimal value from Bayesian Persuasion
the optimal value from Misinterpreted Persuasion

C.1.2 naïve Receiver

Similar to the case of the sophisticated receiver, we also want to know how the effects change with misinterpretation parameters in naively misinterpreted persuasion.

With naïve misspecification, the receiver doesn't respond to the parameters of misinterpretations. Thus, γ_l doesn't restrict the range of priors where the sender can benefit from persuasion. However, it does affect how beneficial the naïve misspecification is to the sender because the larger γ_l is, the more sub-optimal the receiver's decision is.

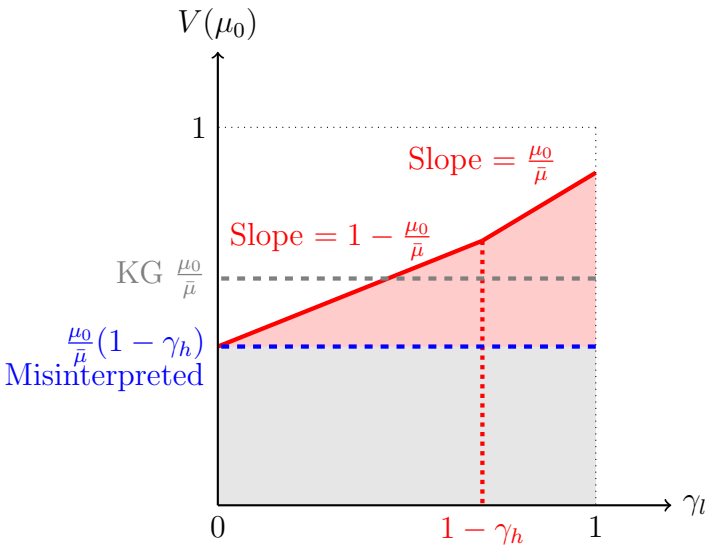


Figure 13: γ_l's Impact on Sender Profit from Naively Misinterpreted Persuasion

The graph shown fixing $\gamma_h = 0.3$, $\bar{\mu} = 0.5$, and $\mu_0 = 0.3$

- Sender's gain from naïveté due to γ_l
- the optimal value from Naively Misinterpreted Persuasion
- Sender's profit from persuasion with misinterpretation only
- the optimal value from Misinterpreted Persuasion
- the optimal value from Bayesian Persuasion

For the other direction of misinterpretation, as γ_h increases, the sender's value from naively misinterpreted persuasion decreases. When misinterpretation is *frequent*, the sender may lose from naïveté when the prior is large enough for more information to get through with sophisticated misinterpretation. This is because if the receiver is sophisticated, he infers the opposite meaning of the realizations and takes the high action a_h more often with more perturbation in γ_h.

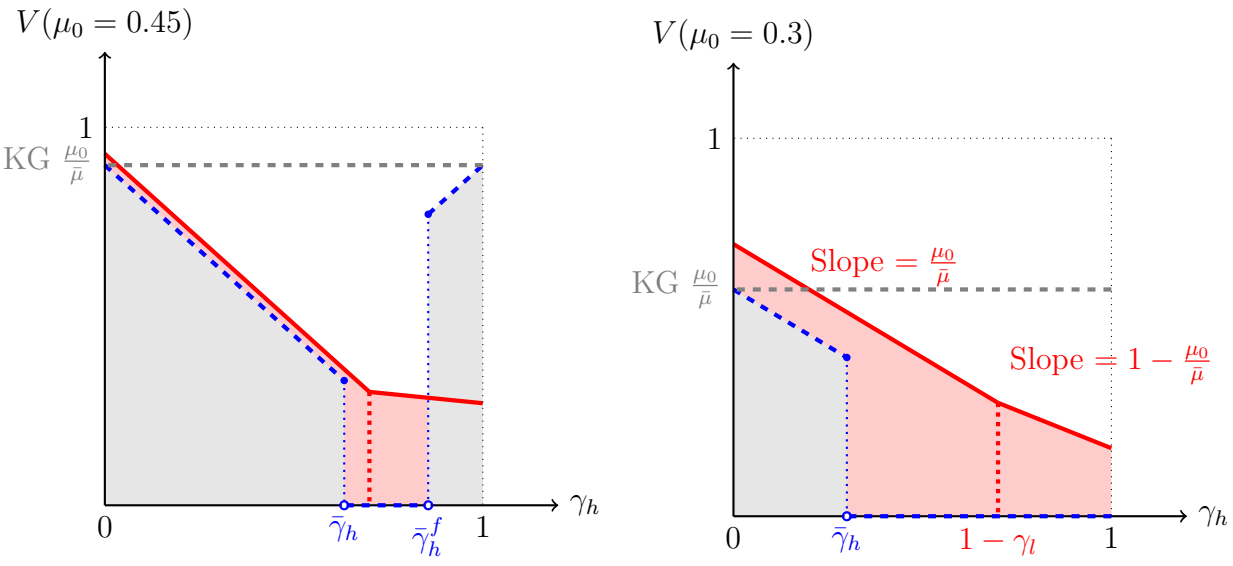


Figure 14: γ_h's Impact on Sender Profit from Naively Misinterpreted Persuasion

The graph shown fixing $\gamma_l = 0.3$, $\bar{\mu} = 0.5$, and $\mu_0 = 0.3$

- Sender's gain from naïveté
- the optimal value from Naively Misinterpreted Persuasion
- Sender's profit from persuasion with misinterpretation only
- the optimal value from Misinterpreted Persuasion
- the optimal value from Bayesian Persuasion

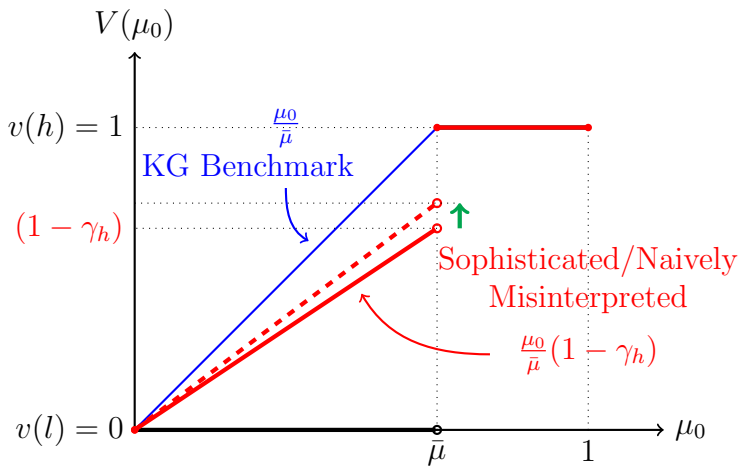
C.2 Policy Implications

Let us return to the lawyer-jury example and think about what comparative statics means from the perspective of the acquittal rate. Suppose a jury misinterprets the testimony unfavorably against a minority lawyer or a minority suspect ($\gamma_h > 0$ and $\gamma_l = 0$). Then, the probability of a minority lawyer winning a case or the probability of a minority suspect getting exonerated is lower than the KG rational benchmark across the board (for $\mu_0 \in (0, \bar{\mu})$) with either a sophisticated or naïve jury. Suppose we want to improve the average probability of a minority acquittal. How can we achieve this?

C.2.1 Unfavorable noise γ_h

If we were able to improve interpretation precision by reducing the unfavorable noise γ_h, then we not only increase the average minority acquittal rate but also get closer to the KG benchmark across the board. If we take the KG benchmark as statistically fair, then reducing γ_h achieves equality and fairness at the same time.

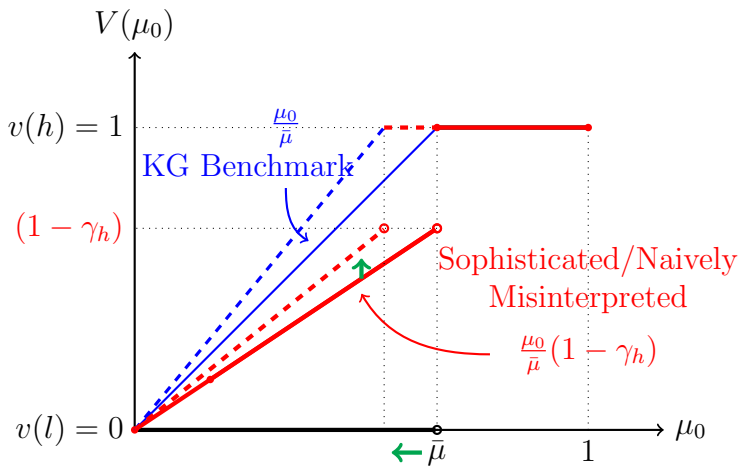
Figure 15A: reducing unfavorable noise $\gamma_h \downarrow$



C.2.2 Belief threshold

Sometimes, it is difficult to directly improve precision ($\downarrow \gamma_h$). How about we drop the bar by reducing standards ($\downarrow \bar{\mu}$)? The minority acquittal rate increases on average. However, it doesn't help in closing the gap between the rational benchmark and the misinterpreted outcome. Relaxing the standards disproportionally benefits the more fortunate individuals of the group.

Figure 15B: relaxing standard $\bar{\mu} \downarrow$

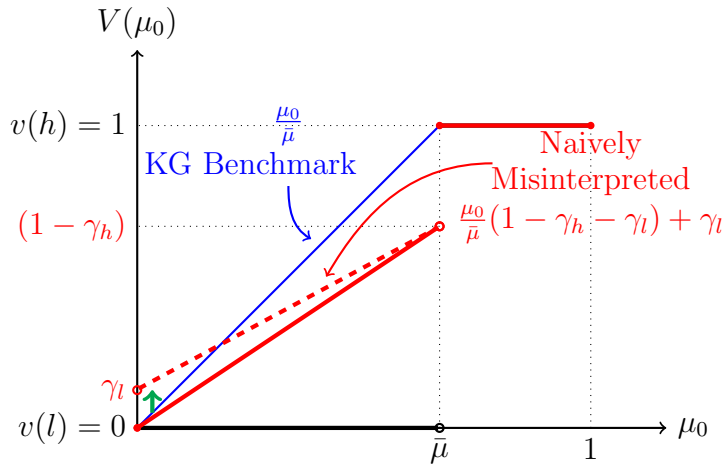


C.2.3 Favorable noise γ_l

Lastly, we may (accidentally) aggravate imprecision by introducing favorable noise towards the minority ($\uparrow \gamma_l$). The probability of misinterpreting l guilty testimony as h innocent testimony sounds favorable to the lawyer and the suspect, but it is the most dangerous approach of the three.

If the jury is a naïve receiver, then having favorable misinterpretation increases the minority acquittal rate on average. It helps the most disadvantaged (low priors) in the group more and helps the more fortunate (high priors) in the group less. However, the misinterpreted outcome gets distorted away from the rational benchmark. The detrimental consequence is that the average quality of minority verdicts decreases, which fuels the statistical discrimination against minority lawyers or suspects.

Figure 15C: introducing favorable noise $\gamma_l \uparrow$ with Naïve Receiver



If the jury is a sophisticated receiver, then favorable misinterpretation destroys the persuasion channel for the most disadvantaged individuals in the group. For low priors $\mu_0 \in (0, \underline{\mu_0})$ ($\neq \emptyset$ with $\gamma_l > 0$), it becomes impossible for these suspects to get exonerated at all.

Figure 15D: introducing favorable noise $\gamma_l \uparrow$ with Sophisticated Receiver