Misinterpreted Persuasion

Mengxi Sun*

July 2025

link to current version

Abstract

This paper examines a behavioral model of persuasion in which the receiver misinterprets the sender-designed information policy with probabilities. The perturbation unravels the concavification technique, but we preserve belief approach by attaching misinterpretation probabilities to induced beliefs. We find that misinterpretation reduces the sender's optimal value from persuasion, and total communication surplus, through bounded implementability. The receiver is unaffected by misinterpretation as long as being Bayesian with respect to the effective information environment. In the binary case, we solved for the sender's optimal value from misinterpreted persuasion. We also looked at when the receiver is unaware of misinterpretation and analyzed interplay between misinterpretation and this naïve misspecification. We find that if there is misinterpretation favoring the sender, receiver's naïveté impairs optimal decision-making, benefiting the sender by reducing the receiver's equilibrium demand for information. Finally, we extend the binary case to incorporate confirmation bias, where the direction of misinterpretation is endogenously determined.

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

1 Introduction

Real-world communication is shaped not only by strategic information design, but also by how information is interpreted. This paper studies a behavioral model of persuasion, in which the receiver misinterprets the sender's information policy via a structured transformation.

Misinterpretation can arise from various sources. Informationally irrelevant factors often intrude on our interpretive processes. When we exchange complex ideas that resist reduction to simple labels, consensus can be elusive. Jurors may misunderstand testimonies; voters may misread policy effects; investors may disagree on what an analyst's report implies. In other cases, interpretive heterogeneity stems from biases tied to social identity, preferences, or beliefs. Stereotypes skew individual assessments based on impression of a group; motivated reasoning lets preferences slant the 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.

Building on Kamenica and Gentzkow (2011) (henceforth, KG), we consider a setting where a sender (she) seeks to persuade a receiver (he) by strategically choosing an information policy, which consists of conditional likelihood of sending different realizations in each state. In the classic model, both the sender and receiver observe the same signal realization generated from this policy. Here, by contrast, our receiver may confuse one realization for another with positive probability. This deviation unravels the concavification technique that underpins much of the modern information design literature.

In the finite setting, we formalize misinterpretation as a probabilistic perturbation, represented by a row-stochastic matrix. It structurally transforms the sender-designed information policy into the receiver's effective one. To preserve the belief-based approach, misinterpretation asscociates with information policies which each realization in the support induces a distinct posterior belief. This ensures that each receiver's posterior belief is a convex combination of sender's posterior beliefs. We show that misinterpretation reduces both the sender's optimal value and the total communication surplus through bounded implementability.

In the binary setting, where the realization space contains only two elements, we analyze the model in closed form to highlight its core mechanics. We also propose a decomposition of the receiver's behavior into two components: (1) misinterpretation that structurally perturbs information environments, and (2) misspecification or non-Bayesian updating that systematically distorts posterior beliefs[1]. We study both a sophisticated and a naive receiver. A sophisticated receiver misinterprets but correctly specifies his effective information environment, updating using Bayes' rule correctly. A naïve receiver not only misinterprets but also misspecifies his information as precisely designed by the sender. Unlike misinterpretation, naïveté reallocates communication surplus from the receiver to the sender through suboptimal decision-making and violation of the martingale property. Lastly, we extend the binary example to accommodate confirmation bias allowing misinterpretation process to depend endogenously on belief updating. The core insights remain robust.

We demonstrate our detailed analysis[2] using the Judge-Prosecutor example from KG. Consider a lawyer persuading 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 revealed through the realization of a forensic expert's testimony. In KG, the lawyer can substantially improve the likelihood of acquittal by ex-ante designing and committing to an information policy. For example, a rational jury optimally acquits 60% of suspects despite knowing that only 30% are innocent.

In our model, the jury may misinterpret the testimony. There are two possible testimonies: one intended to increase the belief in innocence and another to decrease it. We define favorable misinterpretation as the probability that the jury mistakes a "guilt" signal for innocence, and unfavorable misinterpretation as the probability of misinterpreting a "innocence" signal as guilt. Both reduce the lawyer's expected probability of acquittal, but manifest in different forms.

[1] as per de Clippel and Zhang (2022).

[2] The results shown in the main section assume infrequent misinterpetation. The appendix includes results for frequent misinterpretation, comparative statics, and some discussion on policy implication in debiasing effort.

Favorable misinterpretation limits the lawyer's ability to push beliefs toward innocence. It shuts down the persuasion channel entirely at low priors, where a large amount of information is required for the jury to switch to acquittal. Since a sophisticated jury anticipates favorable misinterpretation, he becomes skeptical of testimony suggesting innocence, and demands more information to acquit. At low priors, this demand inflates: the information required for acquittal amplifies skepticism, and persuasion fails when even full revelation does not suffice. For example, with 20% chance of misinterpreting the guilty testimony, the lawyer's ex-ante expected gain from persuasion drops from 30% to 0, at a prior belief of 15% in innocence.

Unfavorable misinterpretation limits her ability to push beliefs toward the guilty state. This weakens the informativeness of the guilty testimony, and thereby erodes the lawyer's ex-ante expected probability of successful persuasion. The sophisticated jury becomes skeptical of testimony indicating guilt, but unlike the previous case, the lawyer cannot respond by providing more information. For instance, with 20% chance of misinterpreting the innocent testimony, the lawyer's ex-ante expected gain from persuasion falls from 60% to 48%, at a prior belief of 30% in innocence.

From the perspective of verdict quality, a sophisticated jury matches decision with state as often as a rational one. What about naive jury? We then analyzed naively misinterpreted persuasion and the interaction between misinterpretation and naiveté. We find that the sender gains from the receiver's naïve misspecification only with favorable misinterpretation, but not with unfavorable misinterpretation. This is because naïveté affects where to switch actions, which is also where unfavorable misinterpretation takes effect. In equilibrium, a naïve jury who misinterprets guilty testimony as indicating innocence demands less information for acquittal, and the lawyer is able to induce acquittal more often due to a violation of martingale property in this non-Bayesian behavior.

So far, in the binary example, misinterpretation is entirely exogenous. What happens to persuasion when misinterpretation becomes endogenous to belief updating? We study a specific form: confirmation bias. This behavior first drew attention from psychologists (Lord et al., 1979; Plous, 1991; Darley and Gross, 1983) and later from 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). It could play an important role in individual learning, political accountability, selective exposure, opinion polarization, etc. Our analysis in this stylized communication game offers a tractable foundation for understanding confirmation bias in more complex dynamics.

The core ideas extend. A jury with confirmation bias may misinterpret a disconfirming testimony as confirming with probability. The direction of misinterpretation depends on the lawyer's strategy. In equilibrium, confirmation bias of a sophisticated jury only harms the lawyer when information demands are high (i.e., low prior), and has no effect when little persuasion is needed (i.e., high prior). It does not affect the likelihood that the jury reaches a verdict matching state. However, naïve confirmation bias benefits the lawyer and harms the jury at high priors, where a naive jury biases in favor of the lawyer in equilibrium. Moreover, naïveté further expands the range of priors over which favorable misinterpretation is active in equilibrium play.

Related Literature

Before laying out the 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 evaluabted 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 system 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 contribute to the growing literature on behavioral models focused on 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). Finally, our analysis offers a potential explanation for why generic debiasing interventions often 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 paper proceeds as follows. Section 2 introduces the model with misinterpretation. Section 3 presents a binary example illustrating persuasion and welfare implications of misinterpretation and naïve misspecification. Section 4 discusses relaxation of commitment and concludes.

2 Model (in progress)

A sender (she) and a receiver (he) communicate about a state of the world in a finite set $\omega \in \Omega$. They share a common prior belief $\mu_0 \in int(\Delta(\Omega))$. The receiver makes a decision $a \in \mathcal{A}$ that affects both the sender's and receiver's payoffs, represented by continuous utility functions $v(a, \omega)$ and $u(a, \omega)$ respectively. The sender can influence the receiver's beliefs by designing and committing[3] to an information policy π before observing the state. The information policy encodes and transmits information with finite realizations $s \in \mathcal{S}$ generated according to $\{\pi(s \mid \omega)\}_{\omega \in \Omega, s \in \mathcal{S}}$.

In contrast to KG, the receiver may misinterpret the realization $s \in supp(\pi) \subseteq \mathcal{S}$ sent according to π as another realization $\tilde{s} \in \mathcal{S}$. We denote the probability of misinterpreting an information policy π as $\gamma_\pi : supp(\pi) \to \Delta(\mathcal{S})$. The paper focuses on the *errors of meaning.* That is, for any information policy π such that no two realizations in the support of π are

[3] In the section 4, we discuss relaxing the commitment assumption. In short, commitment is not essential to our result for misinterpretation because the effect stems 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, geometrically characterized by concavification and quasi-concavification (Lipnowski and Ravid, 2020).

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[4]. γ can be written as a row-stochastic matrix Γ in size $|supp(\pi)| \times |\mathcal{S}|$, where each row is a probability distribution of interpreting $s \in supp(\pi)$ as $\tilde{s} \in \mathcal{S}$.

The effective information environment for the sender is (π, s). Misinterpretation perturbs the information policy, leading to less informative effective information policy for the receiver, denoted as $\phi(\tilde{s} \mid \omega) = \sum_s \pi(s \mid \omega)\gamma(\tilde{s} \mid s)$. For now, we assume both the sender and the receiver are Bayesian in the sense that they correctly incorporate their own effective information, (π, s) and $(\phi, \tilde{s})$ respectively, when applying Bayes's rule[5]. Given prior μ_0 and a pair of informative environments (π, s) and $(\phi, \tilde{s})$,

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

where $\tau_1(\mu)$ and $\tau_2(\tilde{\mu})$ are marginals with respect to the first and second components of the joint posterior distribution $\tau(\mu, \tilde{\mu})$. In KG, the marginals are perfectly correlated since γ is an identity function. Here $\tau_1(\mu)$ and $\tau_2(\tilde{\mu})$ are partially correlated by the γ, which

[4] 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. If you are interested in the case where the sender can manipulate the 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.

[5] We can relax this assumption to incorporate a variety non-Bayesian belief updating rules. In section 3, we take a closer look at naive misspecification, where the receiver is unaware of misinterpretation and mistaken his effective information environment as $(\pi, \tilde{s})$. This subjective deviation only affects optimal action taken but not the probability of taking the optimal action for each receiver's Bayesian posterior belief.

captures the full dependence structure. It means that we can write the joint distribution $\tau(\mu, \tilde{\mu})$ with sender's posterior distribution $\tau_1(\mu)$ and misinterpretation γ. As a result, even though there is interdependence among the possible posterior beliefs, we can still write the sender's problem in terms of sender's posterior belief distribution without reference to signal realizations or any part of the information policy.

Given posterior $\tilde{\mu}$, the receiver optimizes the expected utility with tie-breaking in favor of the sender

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

Then for each pair of possible posterior beliefs $(\mu, \tilde{\mu})$, the sender evaluates expected utility at her posterior belief μ

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

To find the optimal value of misinterpreted persuasion, the sender solves

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

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

Note that misinterpretation violates the independence of irrelevant alternative required for the concavification technique. Even though we can still rewrite the sender's problem in the sender's Bayes-plausible posterior beliefs μ, we cannot use the concavification technique to find optimal value since the posterior beliefs are no longer payoff-separable. The value of each posterior belief μ_s for the sender depends on all other posterior beliefs, $\mu_{s'} \in \{\mu_{s'} \mid \gamma(s \mid s') > 0\} \subset supp(\tau_1)$, which could be misinterpreted to within the support. Although we cannot solve the sender's problem without disciplining 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 implemented 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 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; it is strictly smaller whenever bounded implementability is binding.*

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 in this case.

We also propose a decomposition of behaviors into two components: (1) misinterpretation, structurally distorting information policy, and (2) non-Bayesian updating rules or misspecification, systematically distorting posterior beliefs[6]. We analyze the interaction between misinterpretation and a specific form of misspecification–naiveté.

Lastly, we highlight an extension of binary example to confirmation bias. We demonstrate that the insights about the decomposition are robust to endogeneity in confirmation bias.

[6] Systematic distortion as per de Clippel and Zhang (2022). However, solving for optimal persuasion with naive misspecification in this paper is still outside their scope. The systematic distortion is with respect to the reciver'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.

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 could decide to either convict (a_l) or 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 $\mu_0 := Prob.(\omega = H) \in (0, 1)$.

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–signal realizations, $s \in \mathcal{S} = \{l, h\}$. Regardless of the client being guilty or innocent, the lawyer gets $v(a_h) = 1$ if the jury acquits her client and $v(a_l) = 0$ if the jury decides to convict her client. However, the jury wants to make the correct verdict, matching state with action: convict the guilty suspect, $u(a_l, L) > u(a_h, L)$, and acquit the innocent suspect, $u(a_h, H) \geq u(a_l, H)$. The jury is indifferent between conviction and acquittal if he believes that the probability of innocence is $\bar{\mu} := \frac{\big(u(a_l,L)-u(a_h,L)\big)}{\big(u(a_h,H)-u(a_l,H)\big)+\big(u(a_l,L)-u(a_h,L)\big)} \in (0, 1]$.

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 $\mu_0 < \bar{\mu}$, the lawyer's best ex-ante expected value from persuasion is $\frac{\mu_0}{\bar{\mu}}$, by 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 can be represented in the matrix form as $\Pi = \begin{bmatrix} 1-\pi_L & \pi_L \\ 1-\pi_H & \pi_H \end{bmatrix}$. The lawyer will observe a realization $s \in \{l, h\}$ generated according to Π. But the jury (mis)interprets s as $\tilde{s} \in \{l, h\}$ with probabilities $\{\gamma(\tilde{s} \mid s)\}_{s,\tilde{s}\in\{l,h\}}$, in row-stochastic matrix form as $\Gamma = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. γ_l is the probability of misinterpreting the realization (l) intended to induce low posterior belief as the realization (h) intended to induce high posterior belief; γ_h is the probability of misinterpreting the realization (h) intended to induce high posterior belief as the realization (l) intended to induce low posterior belief. We denote $\gamma_h > 0$ the unfavorable misinterpretation and $\gamma_l > 0$ the favorable misinterpretation. Without loss of generality, we assume *infrequent* misinterpretation, $1 - \gamma_l - \gamma_h > 0$[7].

The jury's effective information policy, denoted as Φ, is less informative than the lawyer-designed information Π. Γ captures the correlation between the sender's effective policy Π and the receiver's effective policy $\Phi := \Pi\Gamma = \begin{bmatrix} 1-\phi_L & \phi_L \\ 1-\phi_H & \phi_H \end{bmatrix}$, where $\phi_\omega := \pi_\omega(1-\gamma_h-\gamma_l)+\gamma_l$ is the probability that the jury gets realization $\tilde{h}$ when the state is ω.

Corresponding to the effective information policies, we denote the lawyer's Bayesian posterior beliefs as $\mu := \{\mu_l, \mu_h\}$[8] and the jury's Bayesian posterior beliefs as $\tilde{\mu} := \{\tilde{\mu}_l, \tilde{\mu}_h\}$[9].

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

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

and the jury's Bayesian posterior distribution is calculated as

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

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

3.1.3 Naïve misspecification

In addition to misinterpretation, the jury may also not have the full knowledge about the effective information environment he's in. In order to characterize the jury's level of knowledge about his information, we denote the jury's subjective posterior beliefs as $\hat{\mu} := \{\hat{\mu}_l, \hat{\mu}_h\}$.

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

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

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

We say the jury is *sophisticated* if he knows his effective information Φ and updates beliefs according to Bayes rule with respect to his effective information Φ. Therefore, a sophisticated jury's subjective posterior beliefs coincide with the jury's Bayesian posterior beliefs, $\hat{\mu} = \tilde{\mu}$. Given the vector of prior beliefs $P = \begin{bmatrix} 1-\mu_0 & \mu_0 \end{bmatrix}$, the sophisticated jury arrives at $\{\tilde{\mu}_l, \tilde{\mu}_h\}$ with probability $\begin{bmatrix} \tau_2^l & \tau_2^h \end{bmatrix}$. The sophisticated jury's posterior beliefs satisfy the martingale property: $\tau_2^l \tilde{\mu}_l + \tau_2^h \tilde{\mu}_h = \mu_0$.

We say the jury is *naive* if he mistakenly takes the announced lawyer's information policy Π as his effective information and updates beliefs according to Bayes rule with respect to this misspecified model of transmitted information. Therefore, a naive jury's subjective posterior beliefs coincide with the lawyer's Bayesian posterior beliefs, $\hat{\mu} = \mu$. But the naive jury arrives at $\{\mu_l, \mu_h\}$ still with probability $\begin{bmatrix} \tau_2^l & \tau_2^h \end{bmatrix}$. Therefore, the naive jury's posterior beliefs violate the martingale property: $\tau_2^l \mu_l + \tau_2^h \mu_h \neq \mu_0$

3.2 Persuading a Sophisticated Receiver

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

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

$$\text{s.t. } \tilde{\mu}_h(\mu_l, \mu_h) \geq \bar{\mu} \tag{O^S}$$

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 $\mu_0 \in (0, \bar{\mu})$. With misinterpretation, the favorable misinterpretation reduces the sender's ability to raise the receiver's posterior beliefs and hence narrows the range of prior 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, $\mu_0 \geq \frac{\gamma_l \bar{\mu}}{(1-\gamma_h)(1-\bar{\mu})+\gamma_l \bar{\mu}} =: \underline{\mu_0}$.*

The intuition is that, for 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 persuade the receiver to switch action. Conversely, if the sender cannot persuade the receiver even with full information revelation, then no strategy can.

In a sender-preferred perfect Bayesian equilibrium, the sender extracts all communication surplus from the receiver. Thus, the receiver is always (subjectively) indifferent when switching actions in equilibrium.

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:*

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

A direct welfare implication of Proposition 3 is that the receiver with misinterpretation still makes the optimal decisions as long as he is Bayesian w.r.t. his effective information policy Φ. He switches to the sender-preferred action at the optimal indifferent belief $\bar{\mu}$. For the sender, compared to KG, misinterpretation reduces her payoff due to bounded implementability. The figure below shows the sender's best ex-ante expected value from persuasion with and without misinterpretation.

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

For low priors ($\mu_0 < \underline{\mu_0}$), misinterpretation hurts the sender because favorable misinterpretation ($\gamma_l > 0$) by the sophisticated receiver reduces the sender's ability to move the high posterior too far away from the prior μ_0 to the action-switching belief threshold $\bar{\mu}$. For high priors ($\mu_0 > \underline{\mu_0}$), misinterpretation hurts the sender because unfavorable misinterpretation ($\gamma_h > 0$) by the sophisticated receiver reduces the sender's ability to move the low posterior too far away from to prior μ_0 to 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:* $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \frac{\mu_0}{\bar{\mu}} \Leftrightarrow \gamma_h > 0$.

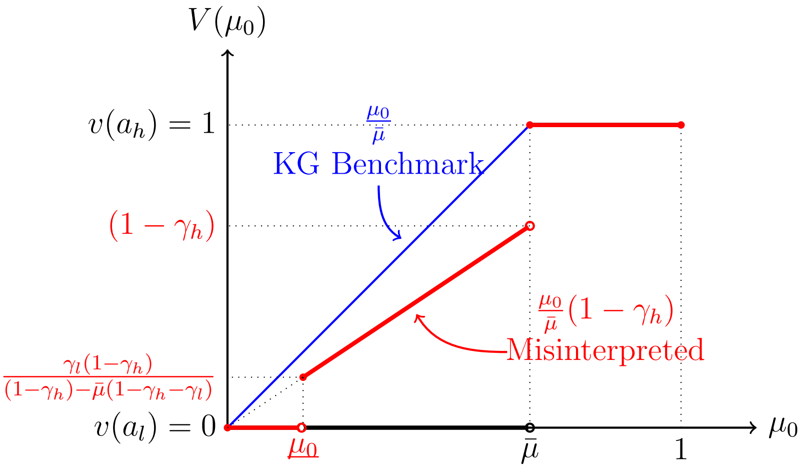


Figure 1: Value function comparison

🟥 with *infrequent* misinterpretation

🟦 without misinterpretation [11]

[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.

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 happens if the receiver doesn't have this level of sophistication? This subsection investigates a naïve receiver who misspecifies his information environment to be what the sender has announced $(\Pi, \tilde{s})$, despite his effective information environment being $(\Phi, \tilde{s})$. Hence, instead of the receiver's 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.

Now, 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 the sender-preferred action a_h but is subject to a different constraint. With infrequent misinterpretations ($\frac{\gamma_l}{1-\gamma_h} < 1$), the sender's problem is a positive linear transformation of that in KG:

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

$$\text{s.t. } \mu_h \geq \bar{\mu} \tag{O^N}$$

3.3.1 Solution and Welfare Analysis

The naiveté reallocate communication surplus through suboptimal decision-making or violation of martingale property. It only affects persuasion if in equilibrium the naive 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 less information to be persuaded than he optimally* *does:*

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

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 $\gamma_l = 0$). Distinct from misinterpretation reducing total surplus, naïveté shifts surplus from the receiver to the sender.

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.*

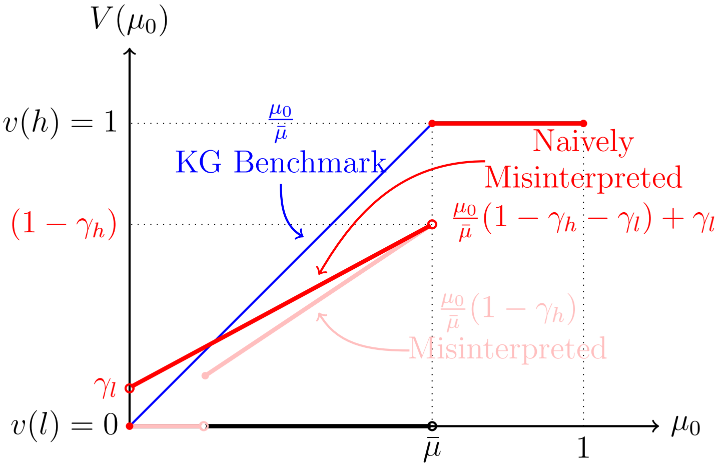


Figure 2: Value function comparison

— with *infrequent* misinterpretation and naïveté

— with *infrequent* misinterpretation and sophistication

— without misinterpretation [12]

Combining the effects of both misinterpretation and naïve misspecification, the sender can do better than in KG with the naïve receiver who needs a lot of information to switch actions (low prior). On the one hand, misinterpretation 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. On the other hand, naïveté benefits the sender only through favorable misinterpretation ($\gamma_l > 0$) leading to the easiness of being persuaded. 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. With low priors, the sender's gain from naïveté eventually outweighs the cost of being misinterpreted.

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.*

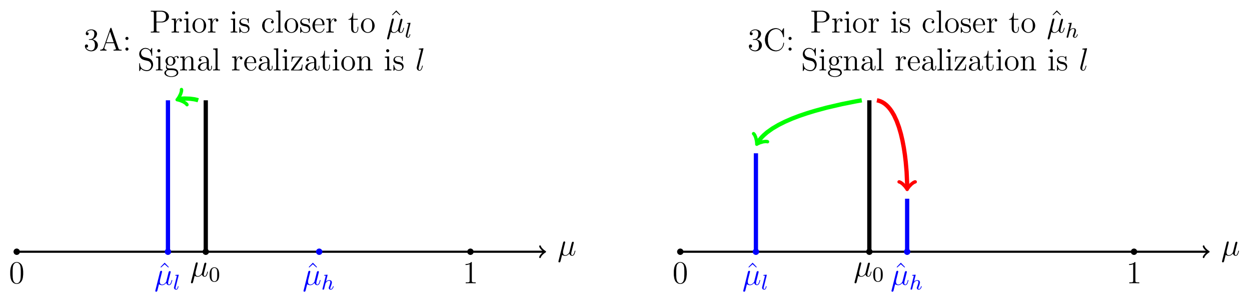
[12] Results for *frequent* misinterpretation is in Appendix B.2. Comparative statics is in Appendix C.1.2 and some policy implication in Appendix C.2.

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 section extends the binary example to include confirmation bias. The setup is the same as the binary example, but confirmation bias endogenizes the direction of misinterpretation. Instead of constant misinterpretation, a jury with confirmation bias misinterprets depending on the information policy. The jury may misinterpret realizations incongruent to his prior beliefs. The result is almost a combination of two special cases of the binary example.

Specifically, a jury with confirmation bias misinterprets in two separate cases. On the one hand, when the lawyer designs an informative policy inducing high posterior closer to the prior, the jury may misinterpret guilty testimony (l) as innocent testimony (h) but never misinterpret innocent (h) as guilty (l). On the other hand, when the lawyer designs an informative policy inducing low posterior closer to the prior, the jury may misinterpret innocent testimony (h) as guilty testimony (l) but never misinterpret guilty (l) as innocent (h). Figure 3 illustrates the confirmation bias visually. We can write the misinterpretation matrices as $\Gamma_h = \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ 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).


3A:
Prior is closer to $\hat{\mu}_l$
Signal realization is l
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ
3C:
Prior is closer to $\hat{\mu}_h$
Signal realization is l
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ

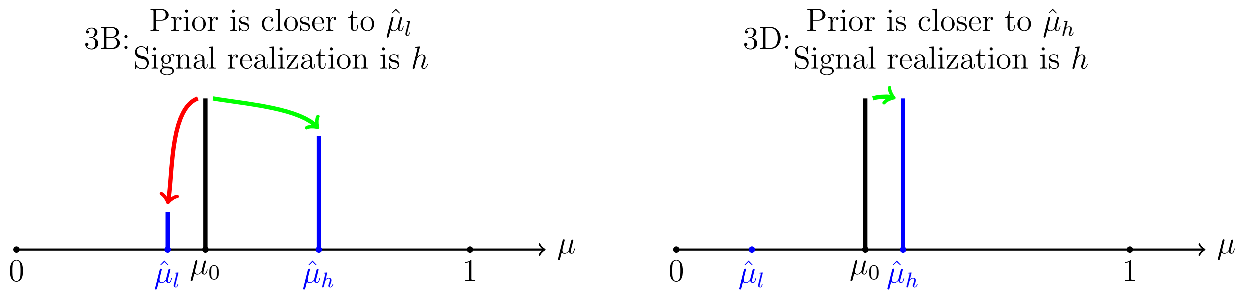

3B:
Prior is closer to $\hat{\mu}_l$
Signal realization is h
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ
3D:
Prior is closer to $\hat{\mu}_h$
Signal realization is h
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ

Figure 3: Direction of Misinterpretation

⟶ Interpret as designed w.p. $1-\gamma_s$

⟶ Misinterpret w.p. γ_s

To formalize confirmation bias, we made a few choices that insubstantially affect the results. The effective direction of bias is determined by 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) coincides to sender's Bayesian posterior μ if the receiver is naïve. We also take the cutoff rule to be the one under 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 $\gamma_h < \frac{1}{2}$,* $$\Gamma^{SCB} = \begin{cases} \Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}; \\ \Gamma_l := \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & \text{, for } \left\{(\mu_l, \mu_h) \mid \mu_0 > \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}. \end{cases}$$

   - *If $\gamma_h \geq \frac{1}{2}$,* $\Gamma^{SCB} = \Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ *for any (μ_l, μ_h).*

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

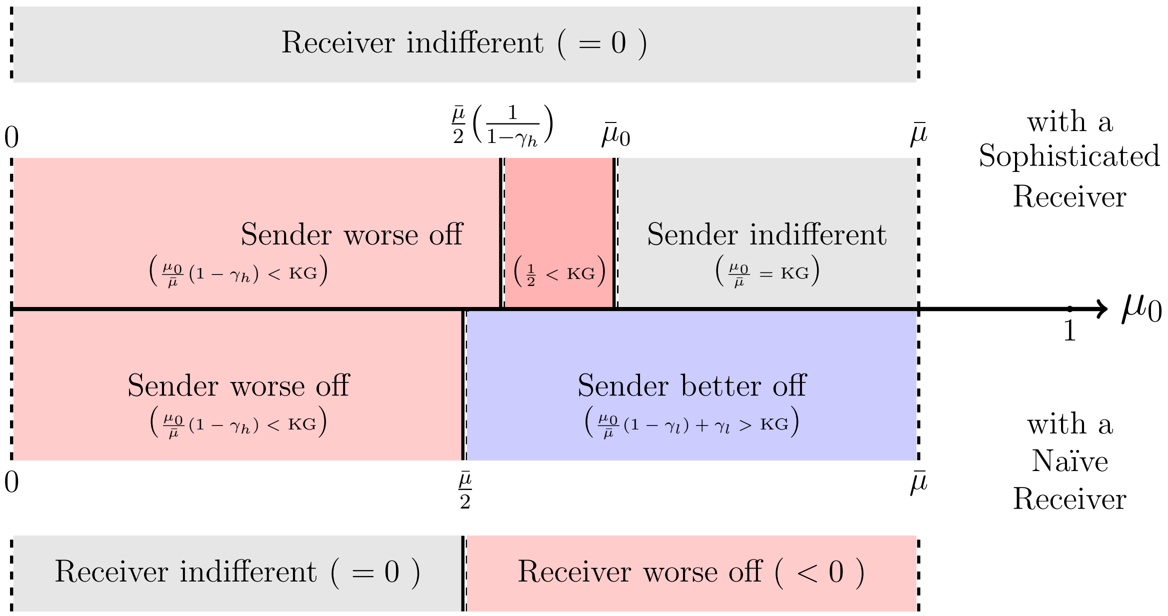
$$\Gamma^{NCB} = \begin{cases} \Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & , \textit{for } \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h+\mu_l}{2}\right\}; \\ \Gamma_l := \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & , \textit{for } \left\{(\mu_l, \mu_h)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

Given a problem with indifference threshold $\bar{\mu}$, prior μ_0, and misinterpretation parameters γ_l and γ_h, a sender persuading a receiver with confirmation bias 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 about misinterpretation and naive misspecification are robust. Figure 4 overviews the welfare effects of confirmation bias, with equilibrium payoff and comparison to the KG benchmark in parenthesis. The sender is always worse off than the KG benchmark for low priors. For high priors, the sender achieves the KG benchmark value if the receiver is sophisticated. Moreover, the sender profits from the receiver's naïveté and does even better than in the KG benchmark if the receiver is naïve. Since the sender in persuasion models extracts all communication surplus, the receiver is made subjectively indifferent in equilibrium. When the receiver is naive, he makes a sub-optimal decision at high posterior belief by being over-precise/naïve.

Let us briefly return to the lawyer-jury example. Confirmation bias only hurts the jury when he is naïve and little persuasion is needed (high prior). In equilibrium, only high prior activates favorable misinterpretation but low prior doesn't. As a consequence, the lawyer profits from confirmation bias compared to rationality in KG when the prior is high, which is opposite to Corollary 3. This is again due to the specific behavior that the direction of misinterpretation depends on the equilibrium strategy.

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



Receiver indifferent (= 0)
0
$\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right)$
$\bar{\mu}_0$
$\bar{\mu}$
with a Sophisticated Receiver
Sender worse off
$\left(\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \text{KG}\right)$
$\left(\frac{1}{2} < \text{KG}\right)$
Sender indifferent
$\left(\frac{\mu_0}{\bar{\mu}} = \text{KG}\right)$
μ_0
1
Sender worse off
$\left(\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \text{KG}\right)$
Sender better off
$\left(\frac{\mu_0}{\bar{\mu}}(1-\gamma_l) + \gamma_l > \text{KG}\right)$
with a Naïve Receiver
0
$\frac{\bar{\mu}}{2}$
$\bar{\mu}$
Receiver indifferent (= 0)
Receiver worse off (< 0)

In the following subsections, we state the sender's optimal persuasion formally. The steps to find a solution with a sophisticated or a naïve receiver with confirmation bias is the same. The difference is just in the additional 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

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

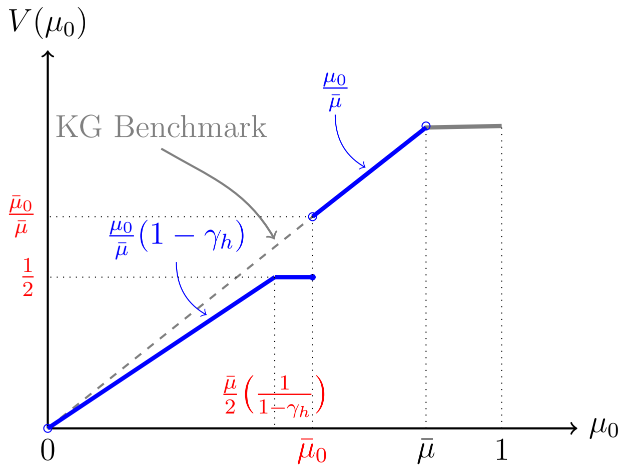
such that in equilibrium

- *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 amount gets 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 under sophisticated confirmation bias is almost direct applications of Corollary 1 under Γ_h for low prior and Γ_l for high prior respectively. The sender's value from persuading a sophisticated confirmatory biased receiver is illustrated in Figure 5[13]. Confirmation bias with sophistication confines the solutions to half-spaces in (μ_l, μ_h), which generates the flat region in the middle because the cutoff for Γ_l lies above the cutoff for Γ_h.

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.

[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 that doesn't change the nature of the convex optimization. We show the mathematical solution in Appendix A.2.1

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*

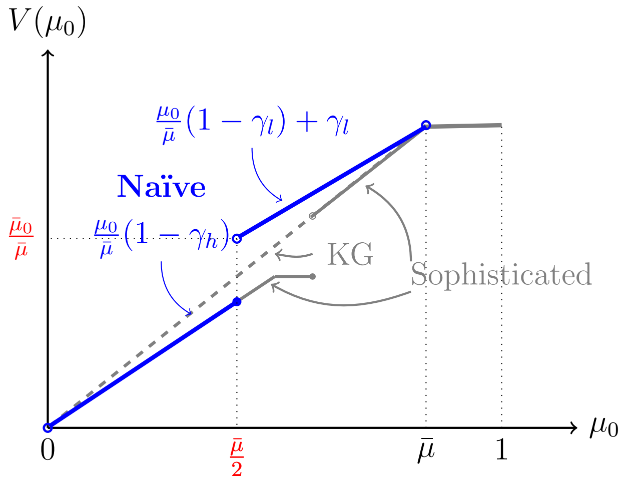
- *For low priors* ($\mu_0 \leq \frac{\bar{\mu}}{2}$)*, the receiver misinterprets against the sender (that is, the effective error matrix is* Γ_h*). Compared to KG and the sophisticated confirmation bias, the sender reveals the same amount of information but less amount gets transmitted to the receiver. 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* $\mu_0 \in \left(\frac{\bar{\mu}}{2}, \bar{\mu}_0\right]$*.*

Similarly, the outcome under naïve confirmation bias is also an almost direct application of Proposition 4 under Γ_h for low prior and Γ_l for high prior respectively. The seemingly contradictory result as opposed to Corollary 3 stems from the equilibrium strategy evoking different directions of misinterpretation (unfavorable misinterpretation with low prior and favorable misinterpretation with high prior). With naïve misspecification, confirmation bias also confines the solutions to half-spaces in (μ_l, μ_h). As a result, the sender's value from persuading a naïve confirmatory biased receiver is illustrated in Figure 6[14].

[14] Similarly, the sender's problem is just a combination of two special cases of the baseline model with a naïve receiver, with an additional constraint on the posterior beliefs. The posterior condition still restricts the solution to a half-space that doesn't change the nature of the convex optimization. We show the detailed solution in Appendix A.2.2

Figure 6: Value Function with Naïve Confirmation Bias



4 Discussion and Conclusion

4.1 Relaxing Commitment

Our analysis is grounded in a persuasion framework with sender commitment, but the implications of misinterpretation logically extend to the case without commitment. This robustness stems from the fact that the behavioral deviation operates in the belief space, and its impact on implementability is independent of the shape of sender's value.

In the finite-state case with state-independent sender utility, the value of commitment is tied to the distinction between concavification and quasi-concavification of the sender's indirect utility function (Lipnowski and Ravid, 2020). Without commitment, equilibrium requires that all posterior beliefs in the support generate the same utility to the sender. This can be achieved by some randomization device at threshold beliefs. Similarly, misinterpretation should reduce the sender's value from communication without commitment through bounded implementability.

However, Tsakas and Tsakas (2021) finds that noise could improve equilibrium informativeness for preferences neither too aligned nor too opposed. In our analysis, misinterpretation cannot benefit. This contrast raises natural questions: Does this difference arise from the equilibrium structures in discrete versus continuous state spaces? Why partial alignment, when paired with noise, create new channels for pareto improvement? These are promising directions for future inquiry.

4.2 Conclusion

This paper investigates how structural misinterpretation affects strategic communication. By proposing a novel behavioral decomposition into misinterpretation and misspecification/non-Bayesian, we expand behaviors in consideration for information design.

Misinterpretation, whether driven by complexity, bias, or other frictions, limits the sender's ability to implement desired posterior distributions. Conversely, naïveté about misinterpretation partly benefits the sender at the receiver's expense, through sub-optimal decision-making and violation of martingale property. We show that this framework extends to richer behaviors such as confirmation bias, where interpretation depends endogenously on belief. Even though the sender can evoke different directions of bias, she cannot manipulate the equilibrium strategy at a given prior beyond choosing optimal posterior distribution under the effective misinterpretation.

These results push literature by capturing a class of real-world communication frictions that standard models abstract away. The model accommodates a variety of phenomena. At a broader level, our framework highlights the importance of modeling 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. By outlining the robust theoretical insights, our analysis offers a foundation for understanding how to navigate and, where possible, mitigate the effects of misintepretation in relevant communication contexts.

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.

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.

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.

"⇒" 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,

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

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

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 4

With 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

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

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 4, we know 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 $\mu^* = (\mu_l^*, \mu_h^*) =$

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

$(0, \bar{\mu})$. 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:**

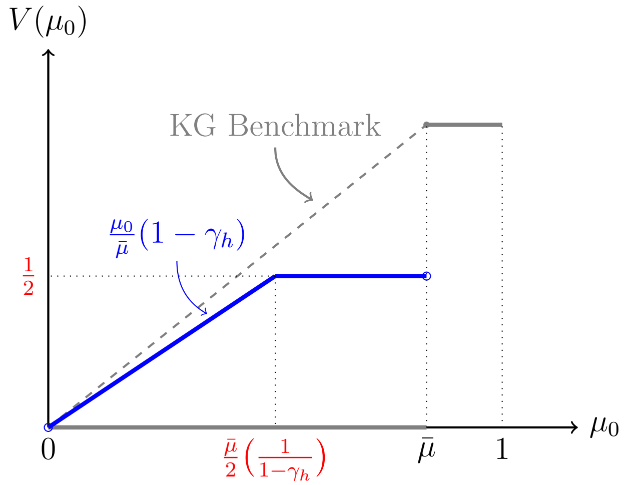
   First, we search for a solution in $\left\{(\mu_l,\mu_h) \mid \mu_0 \leq \frac{\mu_h+(1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}$ under $\Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint of the posterior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

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

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.

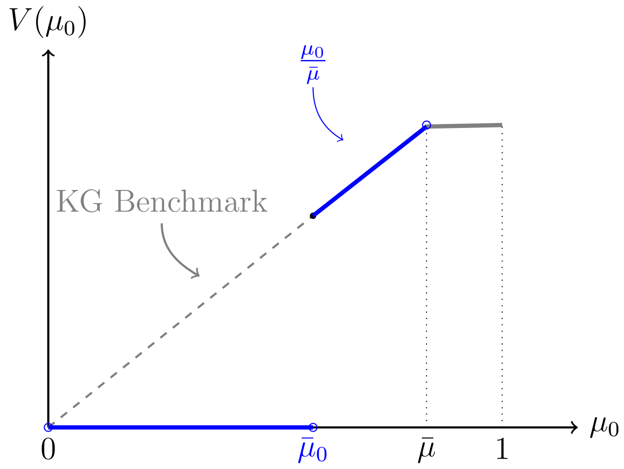
With Sophistication, the receiver updates to his Bayesian posterior $\tilde{\mu}$. Under Γ_h, the receiver's high posterior $\tilde{\mu}_h$ is strictly less than the sender's high posterior μ_h. The Sender solves the following problem:

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

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 $\bar{\mu}_0^{O_2^S}$.

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

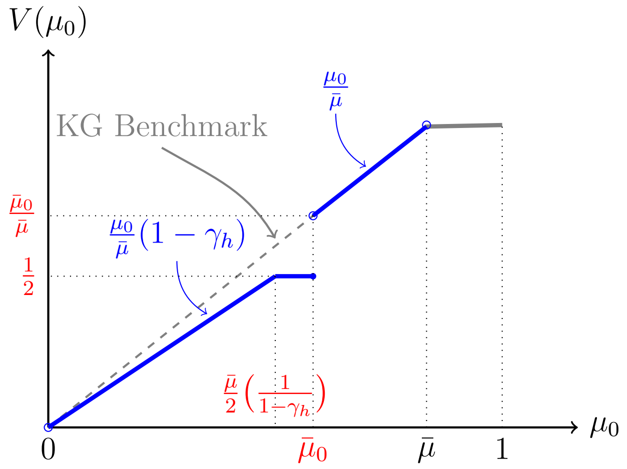
Figure 7B: Case 2 Value Function with Sophisticated Receiver



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

   Now, we have solved the two cases separately. Given a prior μ_0, the sender can 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



∎

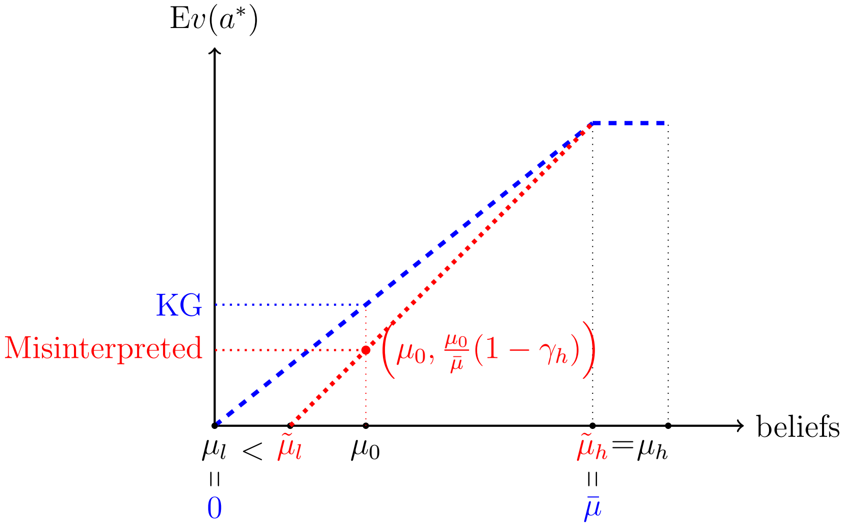
Example Solutions:

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 $\mu_0 \in (0, \frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right)]$, 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 $(\mu_l^*, \mu_h^*) = (0, \bar{\mu})$;
- 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 $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$.

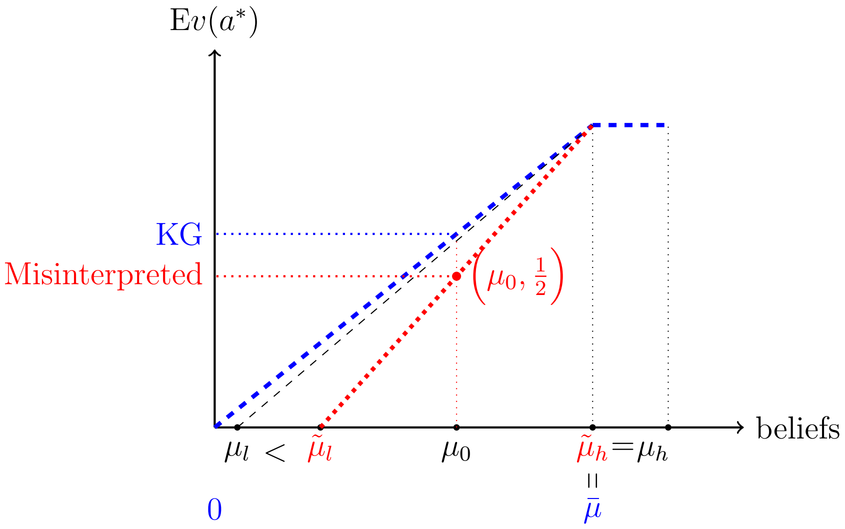
Figure 8A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big)]$



(2) For $\mu_0 \in \Big(\frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big), \bar{\mu}_0\Big]$, 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 $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \Big(2\mu_0 - \bar{\mu}, \bar{\mu}\Big)$;
- The sender gets $\frac{1}{2}$.

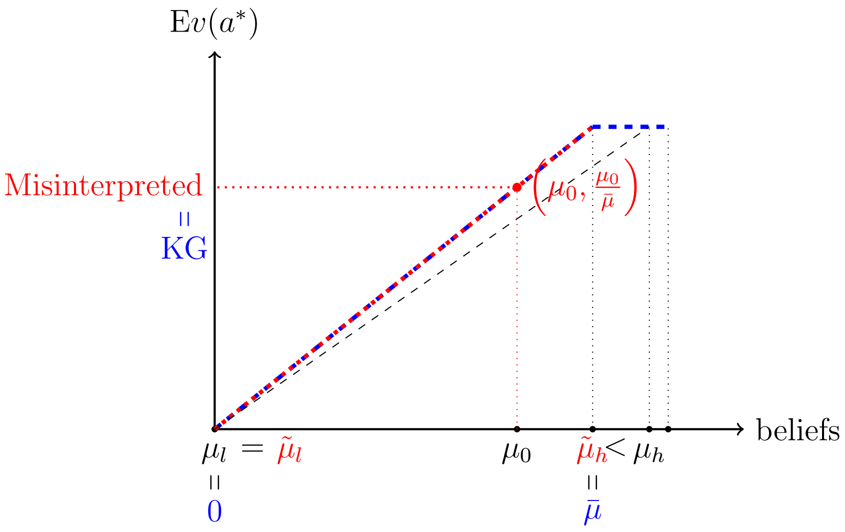
Figure 8B: solution at $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}_0\right]$



(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:

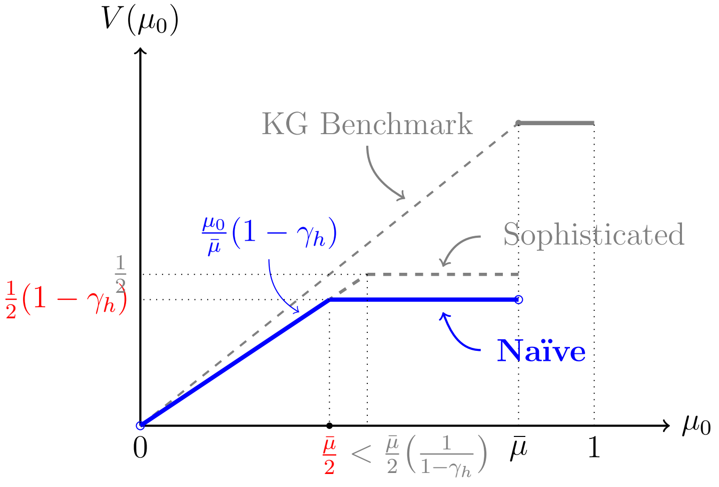
$$\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\tau_h)$$

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

$$\mu_0 \leq \frac{\mu_h+\mu_l}{2} \qquad (CB_1^N)$$

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 Naive 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.

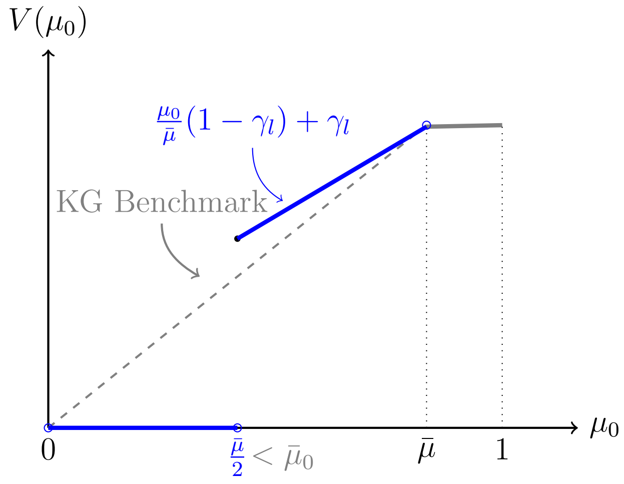
With Naïveté misspecification, the receiver updates to misspecified posterior coinciding with the sender's posterior μ like in Case 1. But the receiver's Bayesian high posterior $\tilde{\mu}_h$ is strictly less than his misspecified high posterior μ_h under Γ_l. Thus, the receiver makes a sub-optimal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

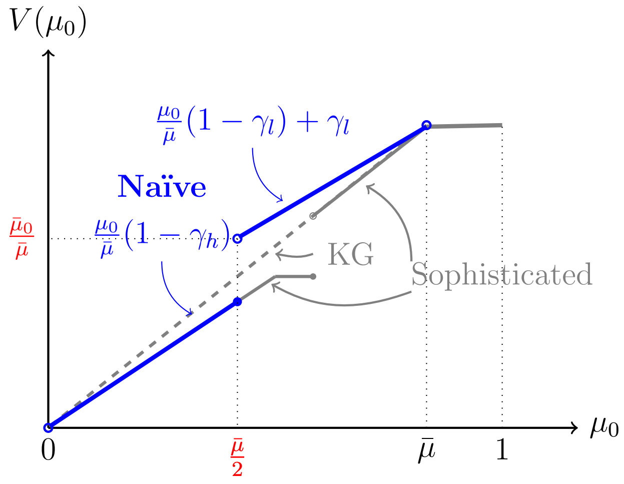
Figure 9B: Case 2 Value Function with Naïve Receiver



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



■

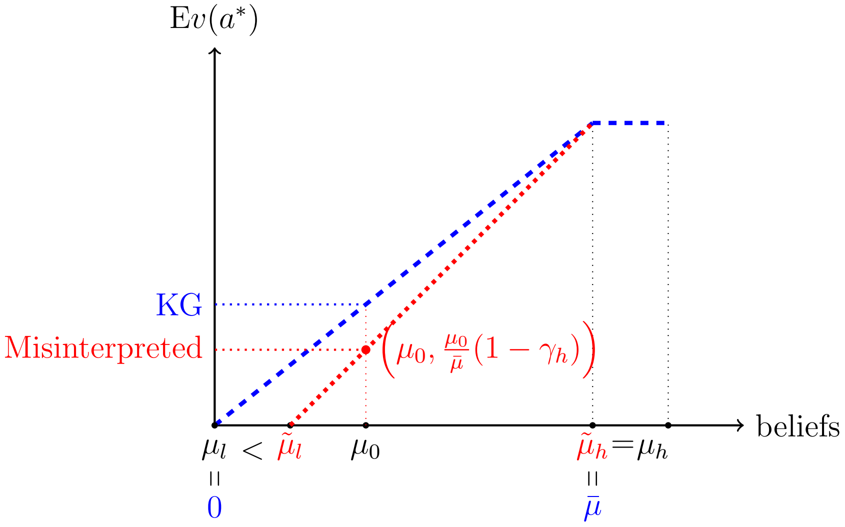
Example Solutions:

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 $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$.

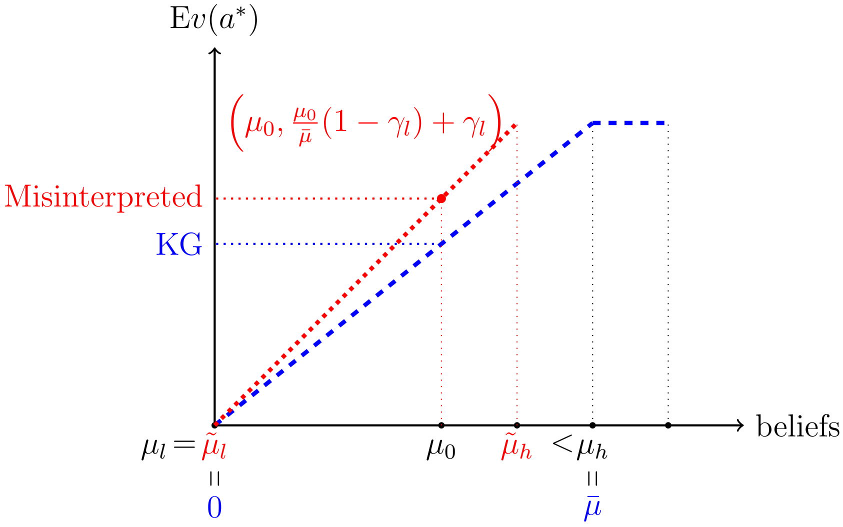
Figure 10A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}]$



(2) For $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$, the receiver misinterprets under Γ_l in equilibrium and makes suboptimal 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

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

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:

$$\mu_h^* = \frac{\frac{\gamma_h}{1-\gamma_l}-1}{\frac{\gamma_h}{1-\gamma_l}-\frac{\bar{\mu}}{\mu_0}}\bar{\mu} \leq 1$$

$$\Updownarrow$$

$$\underline{\mu}_0^f := \frac{(1-\gamma_l)\bar{\mu}}{\gamma_h(1-\bar{\mu})+(1-\gamma_l)\bar{\mu}} \leq \mu_0.$$

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 $\gamma_l < 1$.

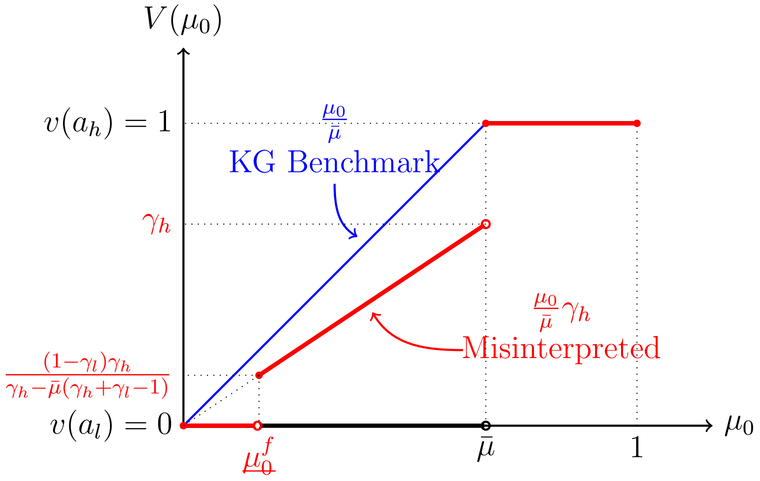


Figure 1^f: Value function comparison

— with *frequent* misinterpretation

— without misinterpretation

B.2 Naive 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

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

$$\text{s.t. } \mu_h \geq \bar{\mu} \tag{O^N}$$

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

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

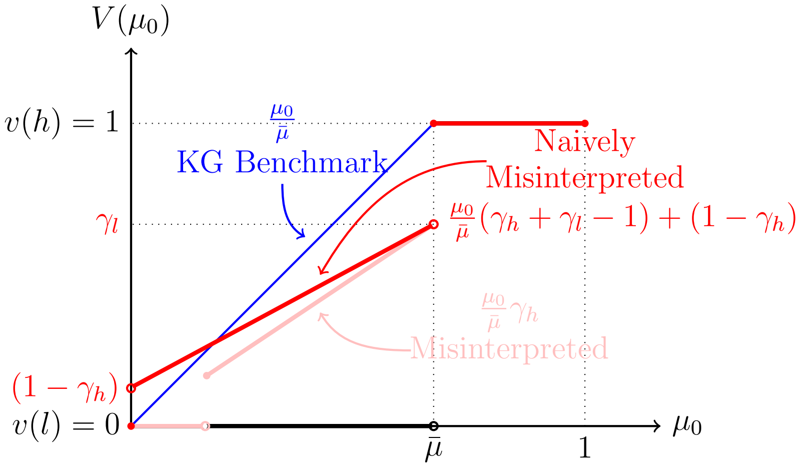


Figure 4^f: Value function comparison

— 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.

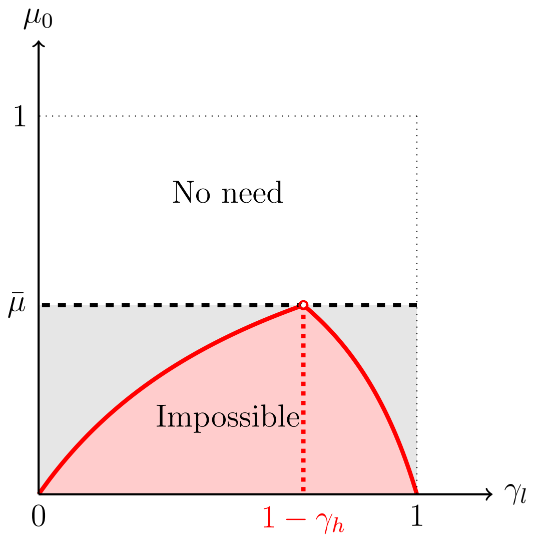


Figure 11: γ_l's Impact on Range of Prior that Sender Benefits
The graph shown fixing $\gamma_h = 0.3$ and $\bar{\mu} = 0.5$
🟥 persuasion channel shuts down due to γ_l
▨ the range of priors where the sender benefits
- - - the upper bound of persuasion: $\bar{\mu}$ exclusive
— the lower bound of persuasion: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

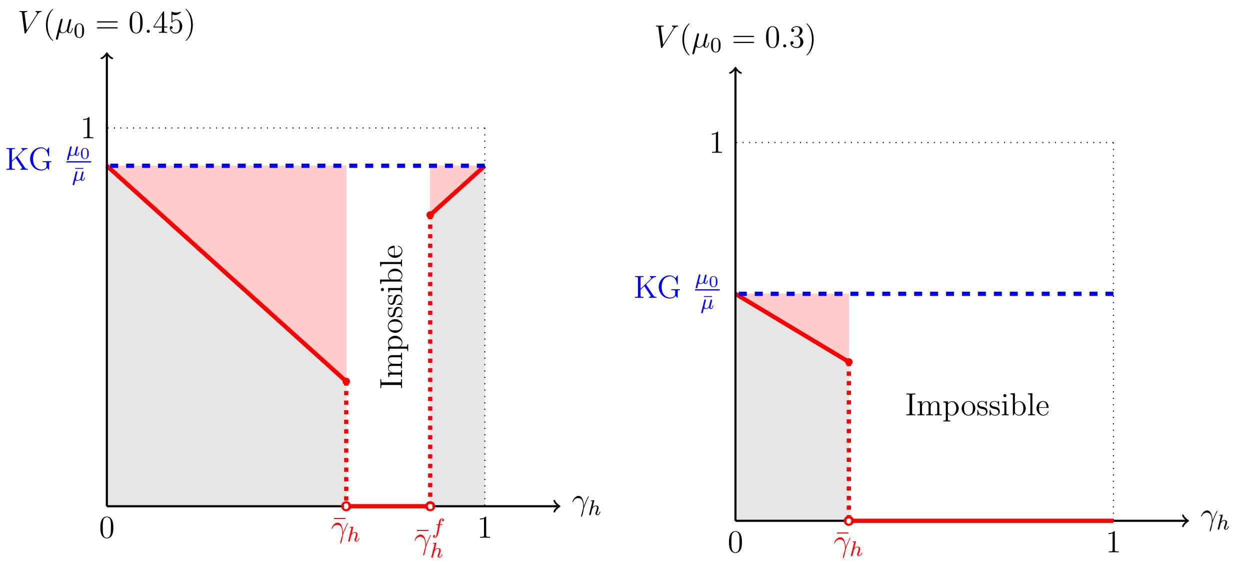


Figure 12: γ_h's Impact on Sender Profit from Misinterpreted Persuasion

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 Naive 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.

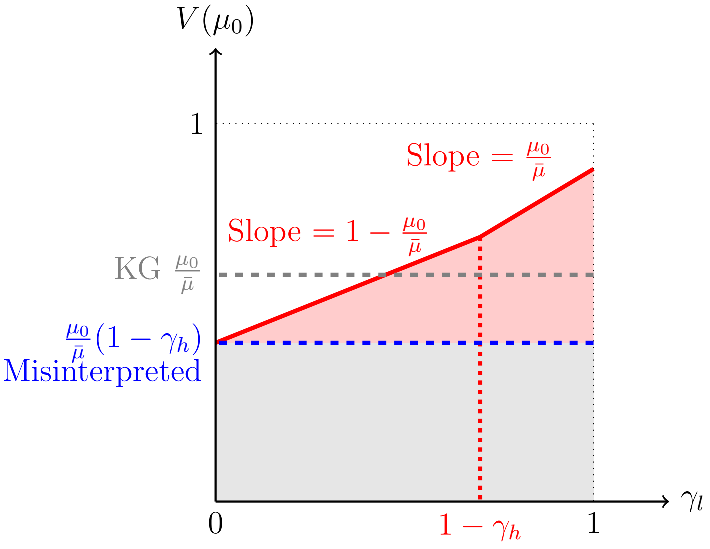


Figure 13: γ_l's Impact on Sender Profit from Naively Misinterpreted Persuasion

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.

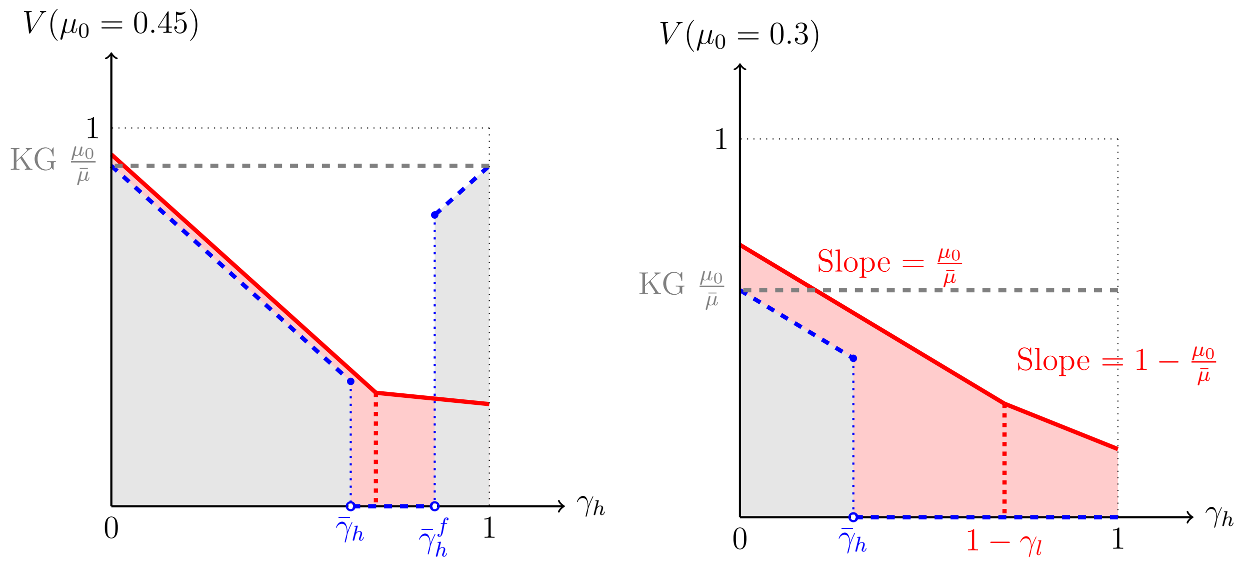


Figure 14: γ_h's Impact on Sender Profit from Naively Misinterpreted Persuasion

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

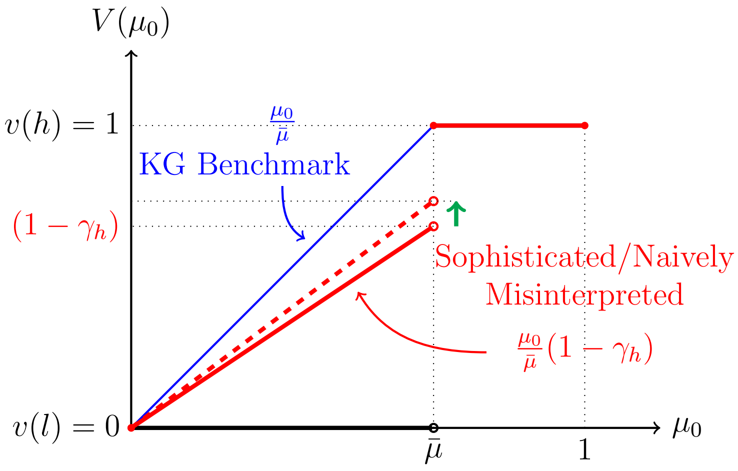
C.2 Policy Implications

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.

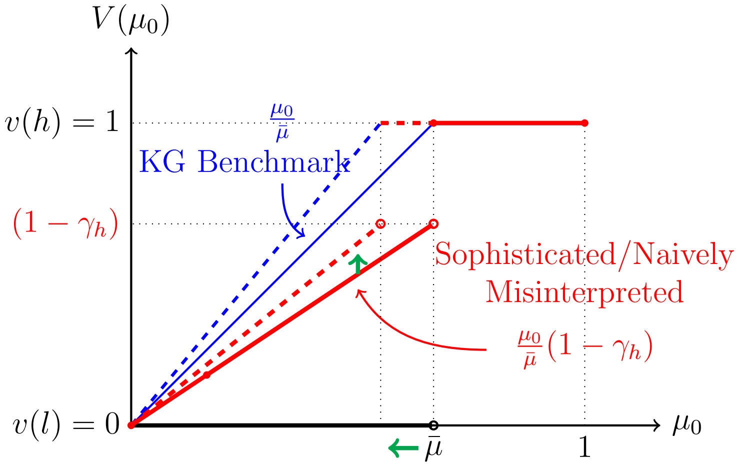
Figure 15A: reducing unfavorable noise $\gamma_h \downarrow$



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 $\bar{\mu} \downarrow$

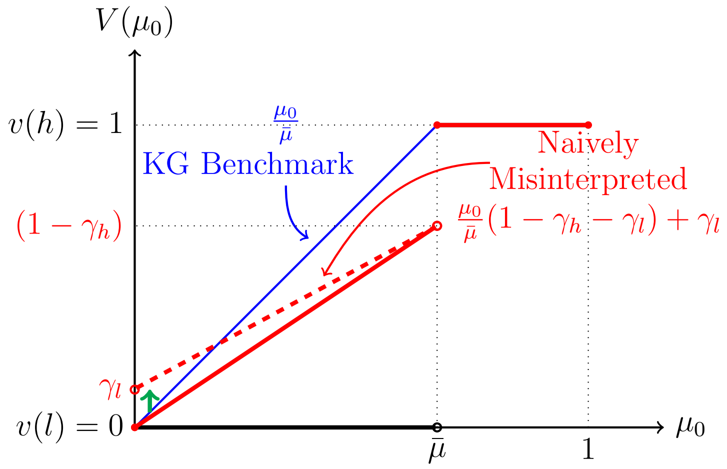


C.2.3 Favorable noise γ_l

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.

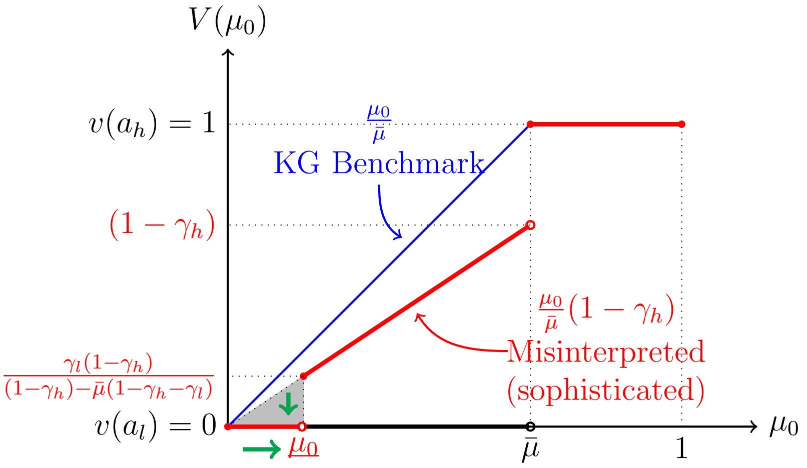
If the jury is a naïve receiver, then having favorable misinterpretation increases the minority acquittal rate on average. It helps the most disadvantaged (low priors) in the group more and helps the more fortunate (high priors) in the group less. However, the misinterpreted outcome gets distorted away 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


$V(\mu_0)$
$v(a_h) = 1$
$\frac{\mu_0}{\bar{\mu}}$
KG Benchmark
$(1 - \gamma_h)$
$\frac{\mu_0}{\bar{\mu}}(1 - \gamma_h)$
Misinterpreted
(sophisticated)
$\frac{\gamma_l(1-\gamma_h)}{(1-\gamma_h)-\bar{\mu}(1-\gamma_h-\gamma_l)}$
$v(a_l) = 0$
$\underline{\mu_0}$
$\bar{\mu}$
1
μ_0