Misinterpreted Persuasion

Mengxi Sun*

November 2024

link to current version

Abstract

This paper studies a behavioral model of persuasion, where the receiver mistakes one signal realization for another with positive probability. By solving a binary pure persuasion model, we find that the sender bears the entire cost of informational loss due to misinterpretation. The receiver is unaffected as long as he is Bayesian about the interpretive errors. If the receiver is unaware of these interpretive errors, the naïveté hinders the receiver's optimal decision-making. The sender benefits from the suboptimal choice because the receiver demands too little information in equilibrium. Lastly, we extend the model to confirmation bias where the direction of misinterpretation is endogenous.

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

1 Introduction

People may confuse messages from other people when communicating with each other. Such friction can arise from exchanging complex ideas that are difficult to abstract into simple labels. For example, juries might misunderstand forensic testimonies, voters may misinterpret policy effects, or consumers may react unexpectedly to advertisements. Beyond complexity, cognitive limitations can also bias our information assessment. For instance, individuals with stereotypes might skew personal evaluations based on a group attribute; motivated reasoners allow preferences to distort their judgment on new information; those with confirmation bias obtain different information depending on their prior beliefs. This paper provides a framework to analyze misinterpretations arising from various sources of communication frictions.

Building on Kamenica and Gentzkow (2011) (henceforth, KG), we consider a scenario in which a sender (she) persuades a receiver (he) by strategically choosing an information policy that specifies how likely different messages are sent in each state. While the classic model assumes that the sender and receiver get the same message generated according to the sender-designed information policy, our receiver may confuse one message for another with a positive probability. We define a sophisticated receiver as one who understands this likelihood of misinterpretation and updates his beliefs accordingly, factoring in both the sender's strategy and the exogenously given probability of misinterpretation. A naïve receiver, by contrast, not only misinterprets messages but also fails to account for this noise in his interpretation.

We analyze the baseline model with each receiver type—sophisticated and naïve. Our findings reveal that misinterpretation decreases the sender's probability of successful persuasion and does not affect the receiver's welfare. Naïveté benefits the sender at the expense of the receiver due to suboptimal decision-making. In addition to decomposing the effects of the receiver's misinterpretation and naïveté, we further break down the effects by examining different directions of misinterpretation and demonstrate how these effects shift with parameters. Finally, we extend the baseline model to consider confirmation bias, allowing the direction of misinterpretation to depend on the sender's strategy. The baseline model's insights on misinterpretation and naïveté also apply under confirmation bias.

To illustrate, consider an example of a lawyer persuading a jury to acquit her client. Both the lawyer and the jury begin with a common prior on the likelihood of the suspect being innocent. While the lawyer always prefers acquittal, the jury wants to acquit only if the suspect is innocent. Information is revealed by a forensic expert's testimony. In KG, the lawyer can drastically improve the probability of acquittal through persuasion. For example, the rational jury optimally acquits 60% of suspects despite the suspect being innocent with only 30% probability before persuasion.

In our model, however, the jury may misinterpret testimonies. For simplicity, we restrict our attention to 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 misinterprets a testimony suggesting guilt as evidence of innocence, and unfavorable misinterpretation as the probability of interpreting a testimony suggesting innocence as evidence of guilt. Both directions reduce the lawyer's expected probability of winning the case but through different mechanisms: (1) favorable misinterpretation limits the lawyer's ability to push beliefs toward innocence, and (2) unfavorable misinterpretation limits the ability to push beliefs toward guilty state.

Favorable misinterpretation shuts down the persuasion channel when initial belief in innocence is so low that the lawyer cannot convince the jury to acquit even by revealing full information. Since a sophisticated jury aware of favorable noise requires more information to acquit, he remains skeptical of forensic evidence that suggests innocence. In equilibrium, the lawyer has to reveal more information to persuade a sophisticated jury to acquit than engaging with a naïve or rational jury. For example, given a prior belief of 30% probability being innocent, if out of all jury perceived innocent testimony, less than 30% comes from misinterpreting a guilty testimony, then the lawyer can still persuade the sophisticated jury to acquit her client 60% of the time, same as in the KG benchmark. However, if the jury misinterprets guilty testimony as innocent with higher probability, the lawyer can no longer persuade the sophisticated jury to acquit at all, resulting in 0% chance of winning the case.

In contrast, unfavorable misinterpretation compresses the lawyer's profit from persuasion– the ex-ante probability of winning the case and getting her client acquitted. The lawyer can increase the ex-ante probability of winning by revealing more information in the innocent state. Intuitively, telling more truth in the innocent state allows the lawyer to lie more in the guilty state. Thus, without misinterpretation, the guilty forensic outcome fully reveals the guilty state in equilibrium (Kamenica and Gentzkow, 2011). However, with unfavorable misinterpretation, the guilty testimony can no longer fully reveal the guilty state because, in the jury's view, it can come from either the information policy or misinterpreting innocent testimony as guilty. This unfavorable noise erodes the ex-ante probability that the lawyer successfully defends her client for acquittal. In equilibrium, the lawyer reveals the same amount of information but (objectively) less get transmitted to the jury who misinterprets. For example, suppose the jury misinterprets unfavorably with a 20% probability. Assuming that the favorable misinterpretation is infrequent so that the lawyer can still successfully persuade the jury to acquit given prior belief of 30%. At optimal, the lawyer gets her clients acquitted with only 48% probability, 20% less than in the KG benchmark.

From the perspective of verdict quality, our jury makes correct decisions as often as a rational jury except when unaware of favorable misinterpretation, leading to overly lenient acquittals. Consequently, naïveté benefits the sender only through favorable misinterpretation but not unfavorable misinterpretation. This is due to the differential effects of favorable and unfavorable misinterpretation on persuasion as we talked about before. The suboptimal verdict stems from demanding insufficient information to acquit, which the lawyer exploits. However, unfavorable misinterpretation does not translate to the same decision-making errors. Thus, the naïveté of the jury who unfavorably misinterprets offers the lawyer no advantage.

After solving the model with each type of receiver, we discuss the implications of comparative statics results. Suppose the jury misinterprets the testimony indicating innocence based on the minority identity of the suspect. Following the above analysis of misinterpreted persuasion, at each prior that needs persuasion, the acquittal rate of minority suspects by a misinterpreting jury is lower than that by a rational jury. We want to make a change and care about both equality and fairness. Given a distribution of prior beliefs, let us consider bringing up the average conviction/acquittal rate of the minority group as improving inequality; consider closing the gap of minority acquittal rates between misinterpreted persuasion and the Bayesian persuasion at each prior as achieving fairness.

How do we improve inequality? There are three channels. First, if we are able to decrease the unfavorable misinterpretation directly, then we not only bring up the average minority acquittal rate but also narrow the gap between misinterpreted persuasion outcomes and the Bayesian benchmark. Equality and fairness goals are reached at the same time. Secondly, how about we lower the belief threshold of acquittal? This is an easy route to take in practice. It can also increase the average minority acquittal rate, but it doesn't help in closing the fairness gap. Last but not least, what happens if the jury starts to misinterpret favorably? With a naïve jury, the average minority acquittal rate increases, but it drags down the average quality of verdict, which feeds more severe statistical discrimination. The average quality of verdict decreases because a naïve jury acquits disproportionally more minority suspects with little prior chance of acquittal. Can a sophisticated jury do better? Unfortunately, favorable misinterpretation is even more detrimental with a sophisticated jury. The average minority acquittal rate decreases further. It is particularly unfair to the most disadvantaged suspects in the minority group, who always get convicted with no chance to be exonerated.

So far, the misinterpretation is entirely exogenous. What happens to persuasion with endogenous misinterpretation? We are interested in a particular form–the confirmation bias. This behavior got attention firstly from a group of psychologists (Lord et al., 1979; Plous, 1991; Darley and Gross, 1983) and later from many more 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 matters in various situations like individual learning, political accountability, selective exposure, opinion polarization, etc. Our findings in this simple communication setting–persuasion–serves as a foundation to understand confirmation bias in more complicated settings involving inter-temporal dynamics or competition.

In section 3, we formally define confirmation bias and then extend the baseline models. A jury with confirmation bias misinterprets a disconfirming testimony as a confirming testimony with some probability. The direction of misinterpretation depends on the lawyer's strategy. In equilibrium, the confirmation bias of the sophisticated jury only adversely affects the lawyer when the jury needs a lot of persuasion (low prior) and has no impact on the lawyer when little persuasion is needed (high prior). It doesn't influence the probability that the jury makes an optimal verdict. However, naïve confirmation bias benefits the lawyer and hurts the jury when little persuasion is needed (high prior) because the naively confirmatory biased jury acquits more easily in equilibrium. Moreover, naïvete further benefits the sender by enlarging the range of priors where the favorable misinterpretation takes effect under an equilibrium strategy.

Related Literature

Before laying out the model, we highlight our contribution along with the related literature.

We contribute to the persuasion literature by considering behaviors that introduces interdependence among the beliefs in the support of posterior distribution. In (de Clippel and Zhang, 2022) and (Alonso and Câmara, 2016), the posterior beliefs of the sender and the receiver don't agree, but they can rewrite the receiver's posterior belief as a function of the sender's posterior belief induced by the same realization. KG's concavification technique is robust to these behavioral deviations. One of the key features is that the payoff for the sender from inducing each posterior belief is independent of irrelevant posterior beliefs. Thus, we can find the (distorted) indirect utility of the sender at least for each given prior. With Bayes-plausible posterior distribution that average back to prior, concavification of this indirect utility gives us the optimal value.

However, with misinterpretation, the neat concavification characterization fails even at the smallest perturbations. The value of a posterior belief for the sender can depend on any realizations that could be misinterpreted as the one that inducing it. For example, suppose the guilty testimony fully reveals guilty state for the lawyer, but the jury misinterprets guilty testimony as innocent with a positive chance. Then, knowing the client being guilty for sure after the seeing a guilty testimony, the lawyer has either 0 probability of winning the case if the jury convict upon perceiving innocent testimony, or some positive probability of winning if the jury acquit upon perceiving innocent testimony. Despite this, we can still use the belief approach by establishing bijection between supports of the sender's and the receiver's posterior distributions through the invertibility of the interpretative error matrix. 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 "Noisy Persuasion" by Tsakas and Tsakas (2021). Both mine and their paper study noise in the persuasion model with different motivations and emphases. Tsakas and Tsakas (2021) motivates from implementation errors. If we think of the data-generating process committed by the sender as a machine, they focus on a broken machine that adds symmetric noise when spitting out the information. If we cannot repair the machine such that the symmetric noise is inevitable, then the sender benefits from complicating the signal. This is because by increasing the number of realizations, the symmetric noise gets diluted within a posterior belief induced by copies of realziations. This paper is motivated by misinterpretation. We treat all the synonyms as one realization and focus on misinterpretation that only happens across different meanings. Our Sender doesn't benefit from complicating the signal. But both models follow the intuition that noise hurts the sender.

Relatedly, Eliaz et al. (2021) studying a multidimensional model of persuasion is motivated by the complexity of real-world communication. Unlike our model, their sender has an additional tool to influence the receiver's beliefs by choosing a decipher. Our sender is weakened by the complexity of communication which leaves room for flexible interpretation of information. Misinterpretation reduces the sender's ability to induce the receiver's posterior beliefs. In addition, this paper contributes to many behavioral models of persuasion that focus on a specific behavior, such as correlation neglect (Levy et al., 2022), base-rate neglect (Benjamin et al., 2019), wishful thinking (Augias and Barreto, 2023). Finaly, the analysis of this model proposes an alternative explanation for why agents sometimes don't respond to generic debiasing methods in the field (Alesina et al., 2024). Misinterpretation personalizes individual information environment. Without individualized feedback, it is hard to correct suboptimal behavior from naive misinterpretation.

In the following sections, we first introduce the baseline model and analyze the welfare effects. Then, we extend the baseline model to confirmation bias. Lastly, we discuss behavioral decomposition and welfare effects in the context of persuasion.

2 Model

This section focuses on the canonical Prosecutor-Judge example in KG. Suppose a lawyer (she, the sender) defending a suspect who is either guilty (L) or innocent (H) tries to persuade a jury (he, the receiver) for acquittal, $\omega \in \Omega = \{L, H\}$. 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 an information policy. She invites a forensic expert to testify, generating either a guilty (l) or innocent (h) testimony/signal realization, $s \in \mathcal{S} = \{l, h\}$. 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 decision: 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 infers information from the testimony and is different between conviction and acquittal if he believes the probability of innocence reaching $\bar{\mu} := \frac{\left(u(a_l,L)-u(a_h,L)\right)}{\left(u(a_h,H)-u(a_l,H)\right)+\left(u(a_l,L)-u(a_h,L)\right)} \in (0, 1]$.

In the KG benchmark of persuading a rational receiver who doesn't misinterpret, the sender's optimal strategy is characterized by the concavification of the sender's indirect utility function. Given a prior $\mu_0 < \bar{\mu}$, the sender's best value from persuasion is $\frac{\mu_0}{\bar{\mu}}$ by inducing both the sender's and the receiver's posterior beliefs to $(0, \bar{\mu})$.

Setup

Let π_ω represent the probability of sending h realization in state $\omega \in \{L, H\}$. The information matrix $\Pi = \begin{bmatrix} 1-\pi_L & \pi_L \\ 1-\pi_H & \pi_H \end{bmatrix}$ represents the sender-designed information policy. The sender commits to an information policy Π before Nature chooses a state. Then, a realization $s \in \{h, l\}$ is generated according to Π. Everything follows from the KG model up until now.

Miscommunication comes from the possible confusion about the realizations sent by design. The sender still receives the realization as designed, $s \in \{h, l\}$, but the receiver may perceive the realization differently, $\tilde{s} \in \{h, l\}$. The probability $\gamma(s \mid \tilde{s})$ that the receiver interprets s as $\tilde{s}$ is exogenously given. We parameterize the probability of misinterpretation as $\Gamma = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ with γ_h being the probability of misinterpreting the realization (h) intended to induce higher posterior belief as the realization (l) intended to induce lower posterior belief and γ_l being the probability of misinterpreting the realization (l) intended to induce lower posterior belief as the realization (h) intended to induce the higher posterior belief. We call $\gamma_h > 0$ the unfavorable noise and $\gamma_l > 0$ the favorable noise.

Implementability

We focus on Γ that satisfies the following two assumptions: (1) *Invertibility*: $1 - \gamma_h - \gamma_h \neq 0$ so that the sender can influence the receiver's belief. (2) *Error of meaning*: the probability of misinterpretation attaches to the movement of beliefs rather than the label of realizations. Consequently, the sender cannot choose the probability of misinterpretation between the two directions by flipping the realization labels. Additionally, any uninformative information policy doesn't noise the receiver's posterior beliefs away from the prior belief.

The receiver's effective information policy, denoted as Φ, is less informative than the sender-designed policy Π. Γ captures the correlation between the sender's effective policy Π

and the receiver's effective policy $\Phi \coloneqq \Pi\Gamma = \begin{bmatrix} 1-\phi_L & \phi_L \\ 1-\phi_H & \phi_H \end{bmatrix}$, where $\phi_\omega \coloneqq \pi_\omega(1-\gamma_h-\gamma_l)+\gamma_l$ is the probability that the receiver gets realization $\tilde{h}$ when the state is ω. We say the receiver is sophisticated if he knows Φ determined by both Π and Γ.

Both the sender and the sophisticated receiver update beliefs using Bayes' rule with respect to their effective information environment, (Π, s) and $(\Phi, \tilde{s})$, respectively. The sender arrives at her Bayesian posterior beliefs $\mu = (\mu_l, \mu_h)$, where

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

The sophisticated receiver who misinterprets $\mathcal{S}$ with probability Γ also arrives at his Bayesian posterior beliefs $\tilde{\mu} = (\tilde{\mu}_l, \tilde{\mu}_h)$, where

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

Given the vector of prior beliefs $P = \begin{bmatrix} 1-\mu_0 & \mu_0 \end{bmatrix}$, the sender sends realizations $s \in (l, h)$ and arrives at her Bayesian posterior beliefs μ with probability

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

The sophisticated receiver perceives realizations $\tilde{s} \in (l, h)$ and arrives at his Bayesian posterior beliefs $\tilde{\mu}$ with probability

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

Since both players are correctly specified and update beliefs according to Bayes' rule, their posterior distributions τ are Bayes plausible:

$$\tau_1^l\mu_l + \tau_1^h\mu_h = \mu_0 \text{ and } \tau_2^l\tilde{\mu}_l + \tau_2^h\tilde{\mu}_h = \mu_0$$

Now, we establish the sender's *Bounded Implementability* of the receiver's posterior distribution. Invertible Γ gives us bijection between Π and Φ. Bayes-plausibility gives us a bijection between information policies and posterior distributions for each player. With both, for any Bayes-plausible receiver's posterior distribution $\tau_2(\tilde{\mu})$, if the corresponding sender's posterior distribution $\tau_1(\mu)$ is also a valid probability distribution, we can say that there exist a pair of information policies (Π, Φ) that implements the posterior distribution $\tau(\mu, \tilde{\mu})$, where τ_1 and τ_2 are the marginal probabilities with respect to the first and the second component and Γ captures the partial correlation between the two marginals. In addition to Bayes-plausibility, the receiver's posterior beliefs have to satisfy an additional condition that the corresponding sender's posteriors need to be valid beliefs. So, given a prior, the set of the receiver's posterior beliefs that the sender can induce is weakly smaller than that without misinterpretation.

Optimality

With perturbations of the realizations, we cannot find solutions via the concavification technique as featured in many persuasion models (Kamenica and Gentzkow, 2011; de Clippel and Zhang, 2022; Alonso and Câmara, 2016). By assuming invertible misinterpretations, we have established the bijection between the pair of effective information policies and the pair of posterior belief distributions of the sender and the receiver. Thus, we are still able to reduce the ex-ante problem of choosing an optimal information policy pair to the ex-post problem of choosing an optimal posterior distribution pair. However, since the sender's and the receiver's posterior distributions are imperfectly correlated by Γ, each posterior belief $\tilde{\mu}_s$ in the support of posterior distribution can be affected by not only the realization $\tilde{s}$ that inducing it, but also any other realization s that could be misinterpreted as $\tilde{s}$. The concavification technique requires the independence of irrelevant realizations for each posterior belief. Misinterpretation introduces imperfect correlation among the posterior beliefs, hence violating the independence requirement. So, concavification is not helpful because we need to determine the entire support of an optimal posterior distribution simultaneously.

We can still use the belief approach and write the sender's problem in posterior beliefs. Without loss of generality, we show results for infrequent misinterpretations $\frac{\gamma_l}{1-\gamma_h} < 1$ that don't flip the meaning of the realizations between the sender and the receiver[1]. In a perfect Bayesian equilibrium, the receiver takes the sender-preferred action a_h with the ex-ante probability of τ_2^h if persuasion is possible. The sender wants to maximize τ_2^h subject to the receiver taking action a_h when perceiving $\tilde{h}$, $\tilde{\mu}_h \geq \bar{\mu}$.

In the next subsection, we solve the sender's problem when the receiver's only mistake is misinterpretation. In the subsection after that, we consider the sender's problem when the receiver makes two types of mistakes: misinterpreting the realizations and naively ignoring the noise introduced by misinterpretation. In the last subsection, we analyze the impact of the parameters through each type of mistake.

2.1 Persuading a Sophisticated Receiver

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

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

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

[1] Frequent misinterpretation produces qualitatively similar results. Results under $\frac{\gamma_l}{1-\gamma_h} > 1$ is shown in Appendix B

where

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

2.1.1 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 noise reduces the sender's ability to raise the receiver's posterior beliefs and hence narrows the range of prior where she benefits from persuasion. We save space with sketch proof after all formal results and the mathematical proofs are in Appendix A.

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

The intuition is that for low priors where a lot of information is needed to switch action, the favorable noise gets so much impact that even full information revelation by the sender 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 to a higher action by revealing full information. Conversely, if the sender cannot persuade the receiver even with full information revelation, then no strategy can.

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.

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

Proposition 2. *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*[2] *if and only if there is favorable misinterpretation: $\mu_h^* = 0$ and $\mu_h^* = \frac{\bar{\mu}\mu_0(1-\gamma_h-\gamma_l)}{\mu_0(1-\gamma_h-\gamma_l)-\gamma_l(\bar{\mu}-\mu_0)} > \bar{\mu} \Leftrightarrow \gamma_l > 0$*

A direct welfare implication of Proposition 2 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 the KG benchmark, misinterpretation reduces her payoff through reduced implementability. The figure below shows the value function with and without misinterpretation. 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 threshold belief $\bar{\mu}$. 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)*

1. *Misinterpretation has no welfare effect on the receiver.*

   The receiver who misinterprets but correctly accounts for the error is always indifferent in equilibrium, the same as in the KG benchmark without misinterpretation.

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

2. *Misinterpretation 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:* $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \frac{\mu_0}{\bar{\mu}} \Leftrightarrow \gamma_h > 0$.

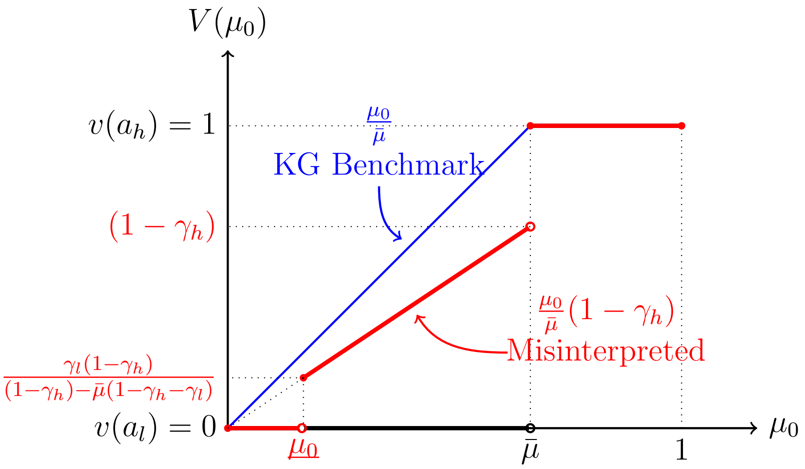


Figure 1: Value function comparison
— with *infrequent* misinterpretation
— without misinterpretation [3]

2.1.2 Comparative Statics

So far, we have solved and analyzed the effects of misinterpretation in persuasion. We attribute the effect of each direction to different channels. A natural next 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.

[3] Results under *frequent* misinterpretation is shown in the Appendix B.1.

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.

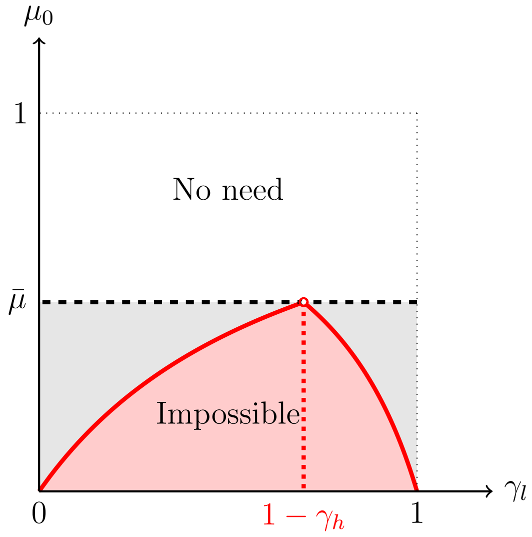


Figure 2: γ_l's Impact on Range of Prior that Sender Benefits

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[4], 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: $\mu_0 \geq \underline{\mu_0^f} \Leftrightarrow \gamma_h \geq (1 - \gamma_l)\frac{\bar{\mu}(1-\mu_0)}{\mu_0(1-\bar{\mu})} =: \bar{\gamma}_h^f$.

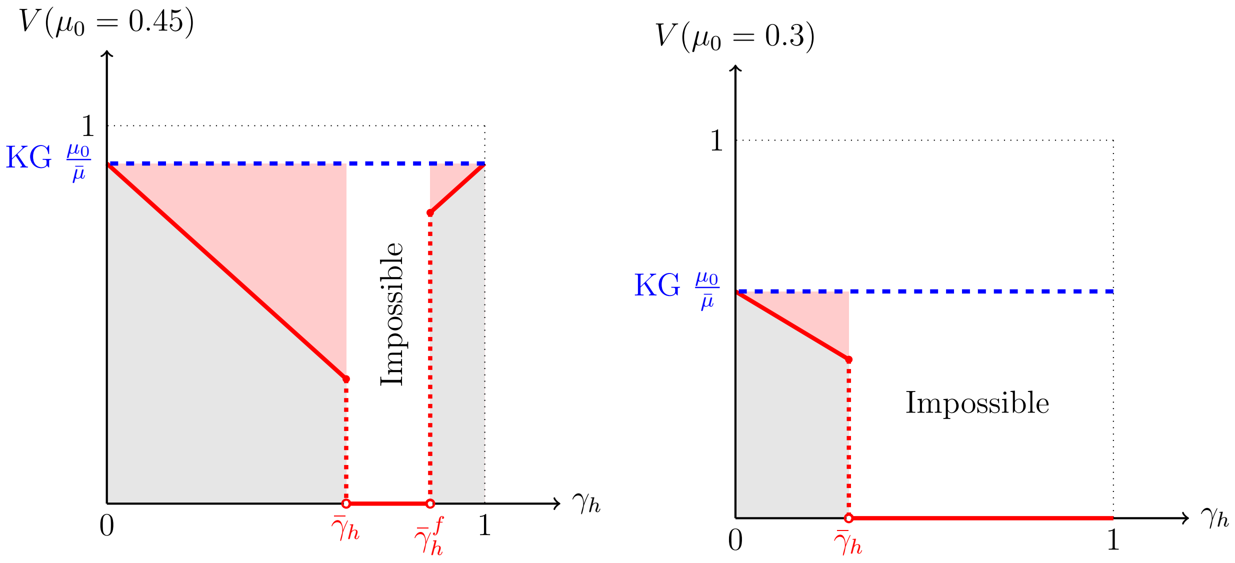


Figure 3: γ_h's Impact on Sender Profit from Misinterpreted Persuasion

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

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

2.2 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 realizations. 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. Since the naïveté is associated with misinterpretation, the composite behavior of the naïve receiver makes the sender's problem easier to solve than the previous problem with the only mistake of misinterpretation. The sender still maximizes the probability of the receiver taking the sender-preferred action a_h but is subject to a different constraint since the naïve receiver's subjective posteriors coincide with the sender's Bayesian posteriors μ instead of his own Bayesian posteriors $\tilde{\mu}$.

With infrequent misinterpretations ($\frac{\gamma_l}{1-\gamma_h} < 1$), the sender's problem has the same solution as the KG benchmark since the difference is a positive linear transformation.

$$\begin{aligned} \max_{\substack{\mu_l \in [0,\mu_0), \\ \mu_h \in (\mu_0, 1]}} \ & \tau_2^h(\mu_l, \mu_h) = \tau_1^h(\mu_l, \mu_h)(1 - \gamma_h - \gamma_l) + \gamma_l \\ \text{s.t. } & \mu_h \geq \bar{\mu} \end{aligned} \tag{O^N}$$

2.2.1 Solution and Welfare Analysis

Because different directions of misinterpretation differ in how they affect the persuasion outcome, the associated naiveté also has distinct channels of impact.

Proposition 3. *The sender has the same implemetability as in KG with the receiver fully naïve about his misinterpretations. When the sender benefits from naively misinterpreted persuasion ($\mu_0 \in (0, \bar{\mu})$), an optimal information policy induces the naïve receiver's misspec-*

ified posterior (μ_h) to the indifference threshold ($\bar{\mu}$). For $\gamma_l > 0$, the naïve receiver demands strictly less information to be persuaded than he should have.

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

.

Unlike the sophisticated receiver, the naïve receiver switches to sender-preferred action sub-optimally. He should take a_h when his Bayesian posterior is equal or above the indifference threshold $\bar{\mu}$. But in equilibrium, the sender only needs to bring the naïve Receiver's subjective posterior μ_h to $\bar{\mu}$, which is weakly easier since $\tilde{\mu}_h \leq \mu_h$ (with equality if no favorable misinterpretation $\gamma_l = 0$). Compared to the sophisticated benchmark, the receiver's naïveté has a zero-sum welfare shift 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.*

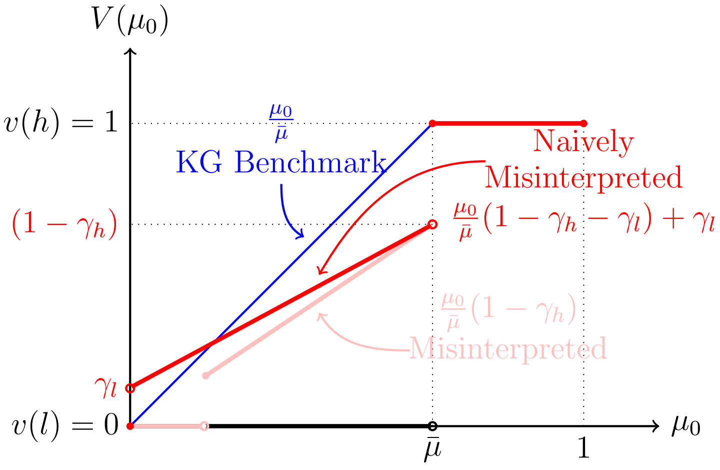


Figure 4: Value function comparison

— with *infrequent* misinterpretation and naïveté
— with *infrequent* misinterpretation and sophistication
— without misinterpretation [5]

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 persuasion (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.*

[5] Results under *frequent* misinterpretation is shown in the Appendix B.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.*

2.2.2 Comparative Statics

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.

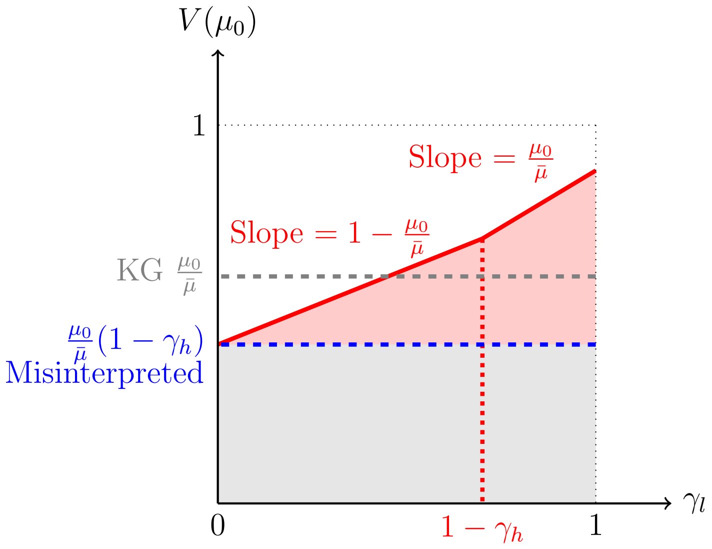


Figure 5: γ_l's Impact on Sender Profit from Naively Misinterpreted Persuasion

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.

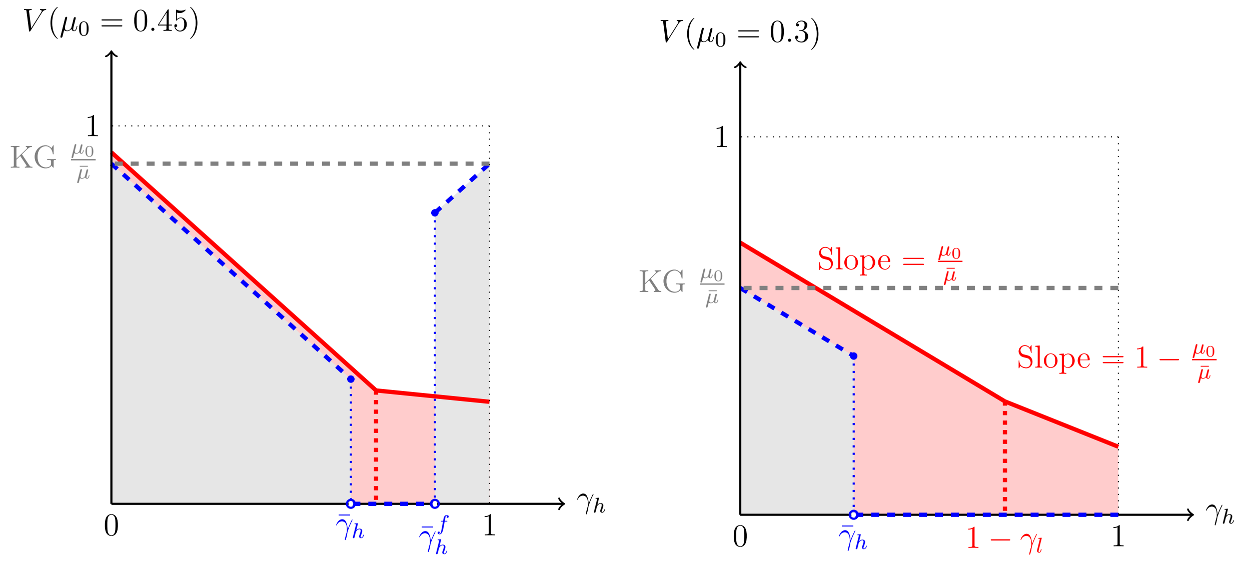


Figure 6: γ_h's Impact on Sender Profit from Naively Misinterpreted Persuasion

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

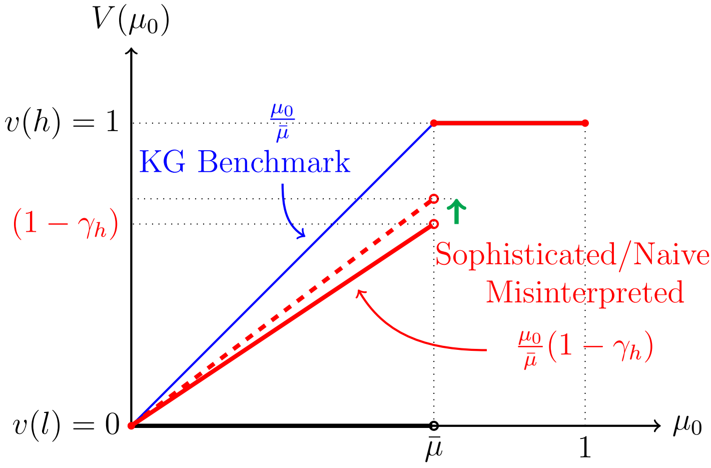
2.3 Policy Implications of (*Infrequent*) Misinterpretation

Let us return to the lawyer-jury example and think about what comparative statics means from the perspective of the f 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?

2.3.1 Discriminatory 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.

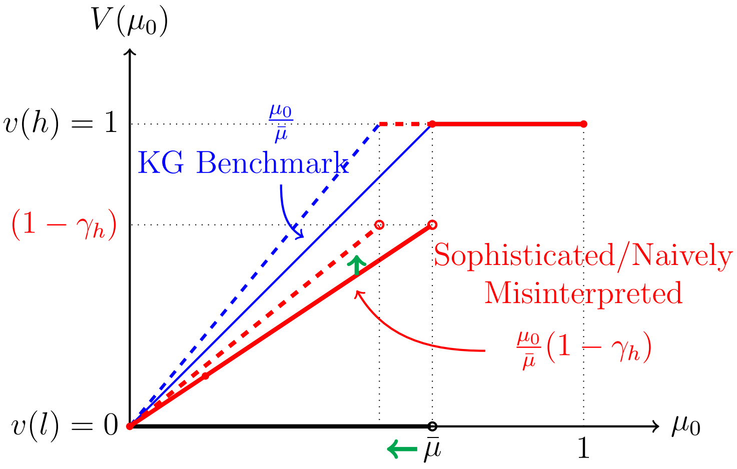
Figure 7A: reducing discrimination $\gamma_h \downarrow$



2.3.2 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 7B: relaxing standard $\bar{\mu} \downarrow$

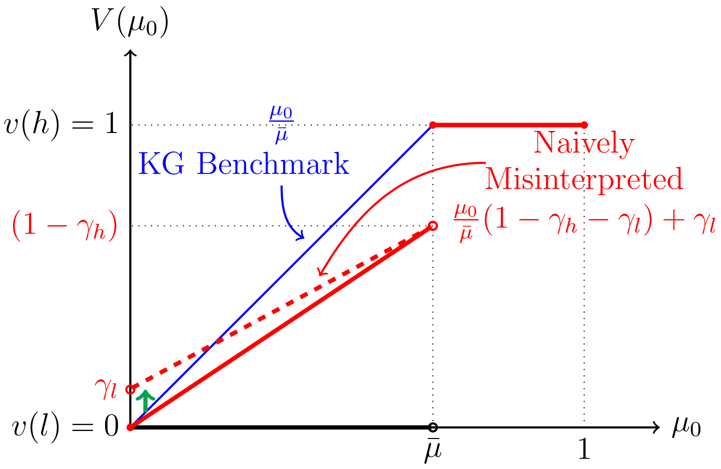


2.3.3 Favoritism noise γ_l

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.

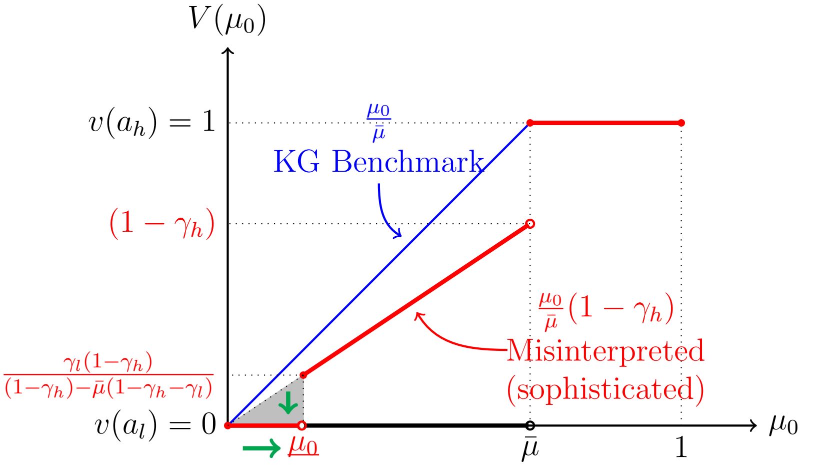
If the jury is a naïve receiver, then having favorable misinterpretation increases the minority acquittal rate on average. It helps the most disadvantaged (low priors) in the group more and helps the more fortunate (high priors) in the group less. However, the misinterpreted outcome gets distorted away 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 7C: introducing favoritism $\gamma_l \uparrow$ 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 7D: introducing favoritism $\gamma_l \uparrow$ with Sophisticated Receiver



3 Extension to Confirmation Bias

This section extends the baseline model to include confirmation bias. The setup is the same as the baseline model, but the specific behavior endogenizes the direction of misinterpretation. Instead of making errors represented by a single misinterpretation matrix, a receiver with confirmation bias misinterprets depending on the information policy. The receiver is more likely to perceive whichever realization confirming his prior. The persuasion outcome is almost a combination of two special cases of the baseline model in the previous section.

Specifically, a receiver with confirmation bias makes mistakes in two separate cases. On the one hand, when the prior belief is closer to his subjective high posterior, 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 prior belief is closer to his subjective low posterior, the jury may misinterpret innocent testimony (h) as guilty testimony (l) but never misinterpret guilty (l) as innocent (h). Figure 5 illustrates confirmation bias visually.

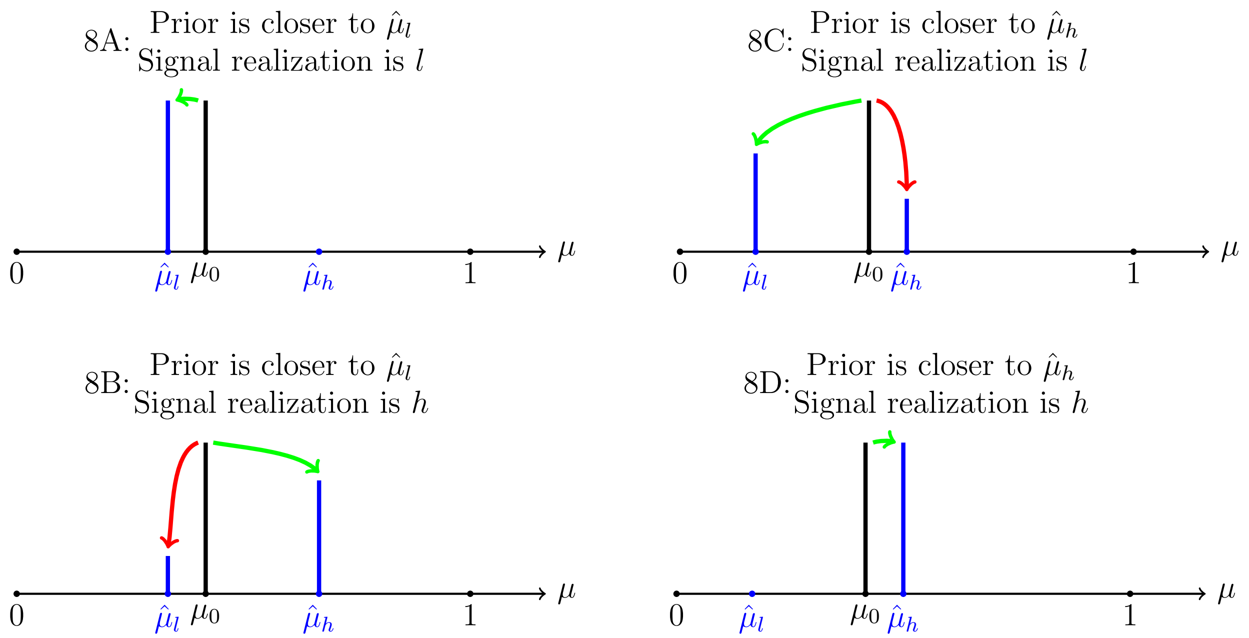

8A:
Prior is closer to $\hat{\mu}_l$
Signal realization is l
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ
8C:
Prior is closer to $\hat{\mu}_h$
Signal realization is l
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ
8B:
Prior is closer to $\hat{\mu}_l$
Signal realization is h
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ
8D:
Prior is closer to $\hat{\mu}_h$
Signal realization is h
0
$\hat{\mu}_l$
μ_0
$\hat{\mu}_h$
1
μ

Figure 8: Direction of Misinterpretation

$\longrightarrow$ Interpret as designed w.p. $1-\gamma_s$
$\longrightarrow$ Misinterpret w.p. γ_s

Like before, denote the probability of misinterpreting l realization as γ_l and the probability of misinterpreting h realization as γ_h. We can write the error matrices as $\Gamma_h = \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ for the case on the left (Figure 8A and 8B) and $\Gamma_l = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix}$ for the case on the right (Figure 8C and 8D). To formalize confirmation bias, we made a few choices that unsubstantially 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}.*

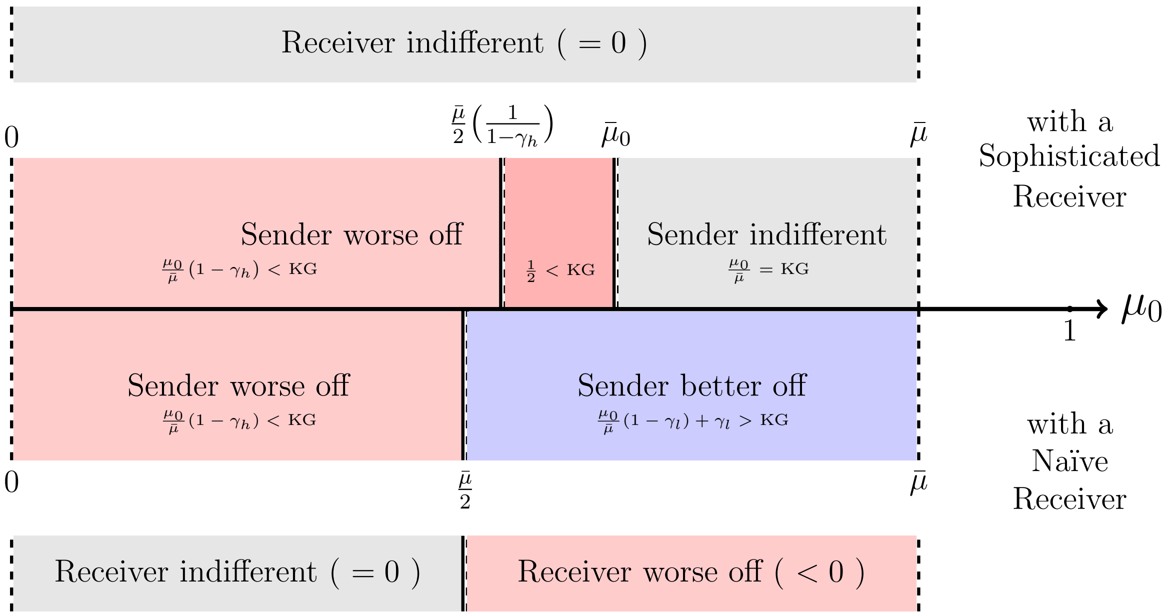
$$\Gamma^{NCB} = \begin{cases} \Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & , \textit{for} \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h+\mu_l}{2}\right\}; \\ \Gamma_l := \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & , \textit{for} \left\{(\mu_l, \mu_h)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

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.

Figure 9 overviews the welfare effects of confirmation bias compared to the KG benchmark. 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 9: 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
$\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \text{KG}$
$\frac{1}{2} < \text{KG}$
Sender indifferent
$\frac{\mu_0}{\bar{\mu}} = \text{KG}$
1
μ_0
Sender worse off
$\frac{\mu_0}{\bar{\mu}}(1-\gamma_h) < \text{KG}$
Sender better off
$\frac{\mu_0}{\bar{\mu}}(1-\gamma_l) + \gamma_l > \text{KG}$
with a Naïve Receiver
0
$\frac{\bar{\mu}}{2}$
$\bar{\mu}$
Receiver indifferent (= 0)
Receiver worse off (< 0)

In the following subsections, we find the equilibrium strategy, solve for the cutoffs, and state the equilibrium payoffs formally. The procedure 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 of the optimization problem.

3.1 Persuading a Sophisticated Confirmatory Biased Receiver

Proposition 4. *(Persuasion with Sophisticated Confirmation Bias)*

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

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

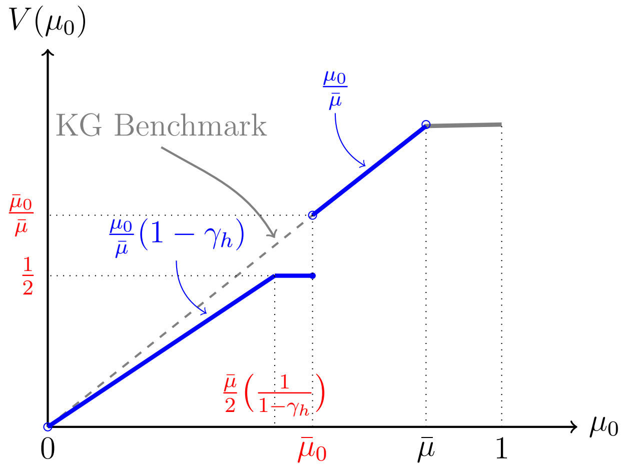
such that in equilibrium

- *For low priors ($\mu_0 \leq \bar{\mu}_0$), the receiver misinterprets against the sender (that is, the effective error matrix is Γ_h). Compared to the KG benchmark, 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 the KG benchmark. 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 the KG benchmark, 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 the KG benchmark, 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 10[6]. 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 10: Value Function with Sophisticated Confirmation Bias



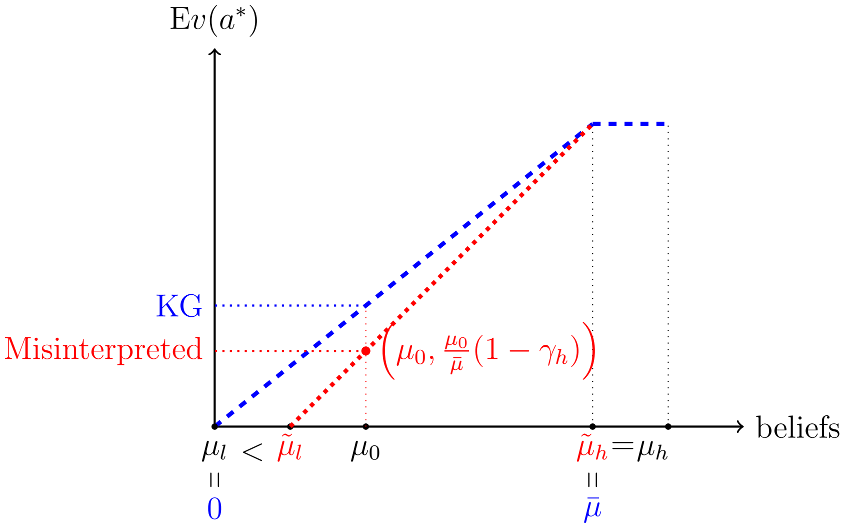
[6] The sender's problem is just a combination of two special cases of the baseline model with a sophisticated receiver, with an additional constraint on the posterior beliefs. The posterior condition restricts the solution to a half-space that doesn't change the nature of the convex optimization. We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

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 $\mu_0 \in (0, \frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big)]$, 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 $(\mu_l^*, \mu_h^*) = (0, \bar{\mu})$;
- The receiver updates to the receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \Big(\frac{\gamma_h \mu_0 \bar{\mu}}{\gamma_h \mu_0 + \bar{\mu} - \mu_0}, \bar{\mu}\Big)$;
- The sender gets $\frac{\mu_0}{\bar{\mu}}(1 - \gamma_h)$.

Figure 11A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big)]$

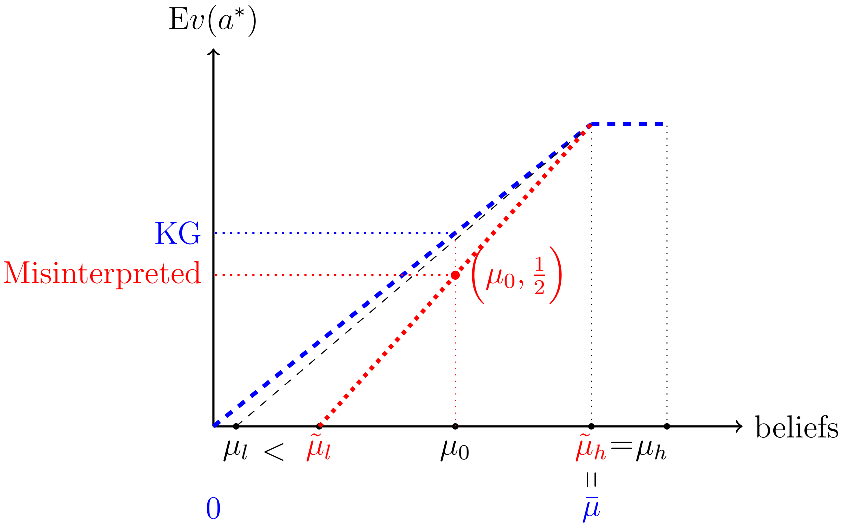


(2) For $\mu_0 \in \Big(\frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big), \bar{\mu}_0\Big]$, the receiver misinterprets under Γ_h in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- 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 $\frac{1}{2}$.

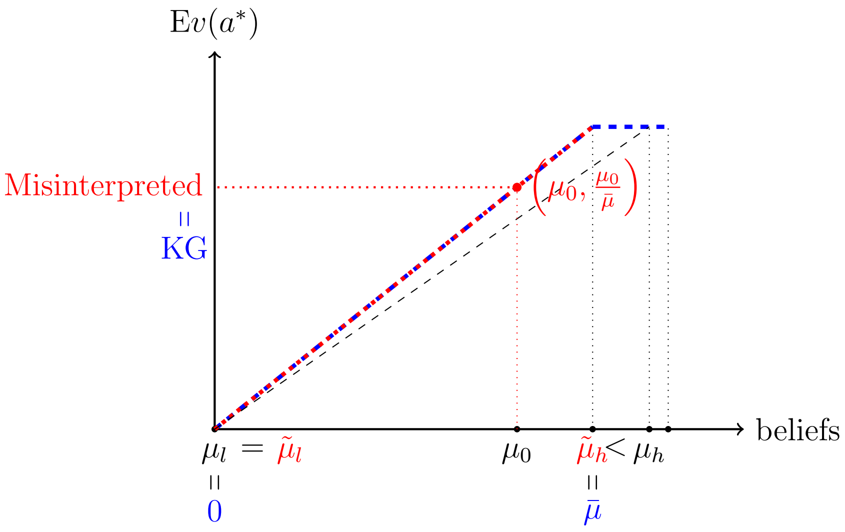
Figure 11B: solution at $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}_0\right]$



(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 11C: solution at $\mu_0 \in (\bar{\mu}_0, \bar{\mu})$



3.2 Persuading a Naïve Confirmatory Biased Receiver

This subsection states and proves the naïve equivalent of Proposition 4 in the previous subsection. The steps are the same and we are solving a simpler optimization problem since the naïve receiver thinks that he is rational.

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

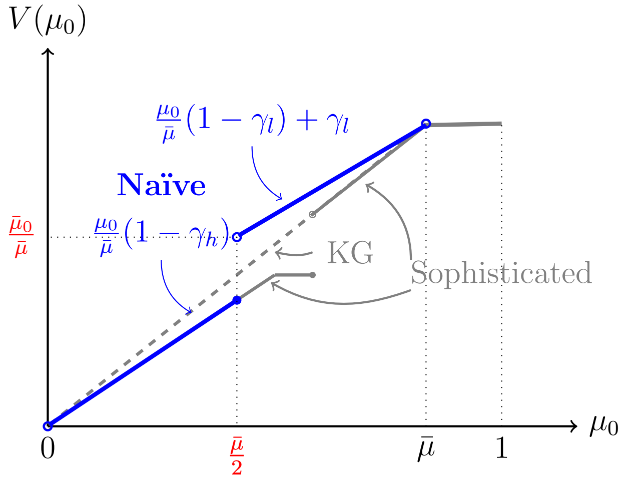
Suppose a confirmatory biased receiver is fully naïve and misinterprets according to Γ^{NCB}. Fixing an indifference threshold $\bar{\mu}$, there exists a prior belief threshold $\frac{\bar{\mu}}{2}$ such that in equilibrium

- *For low priors ($\mu_0 \leq \frac{\bar{\mu}}{2}$), the receiver misinterprets against the render (that is, the effective error matrix is Γ_h). Compared to the KG benchmark 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 3 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 12[7].

Figure 12: Value Function with Naïve Confirmation Bias



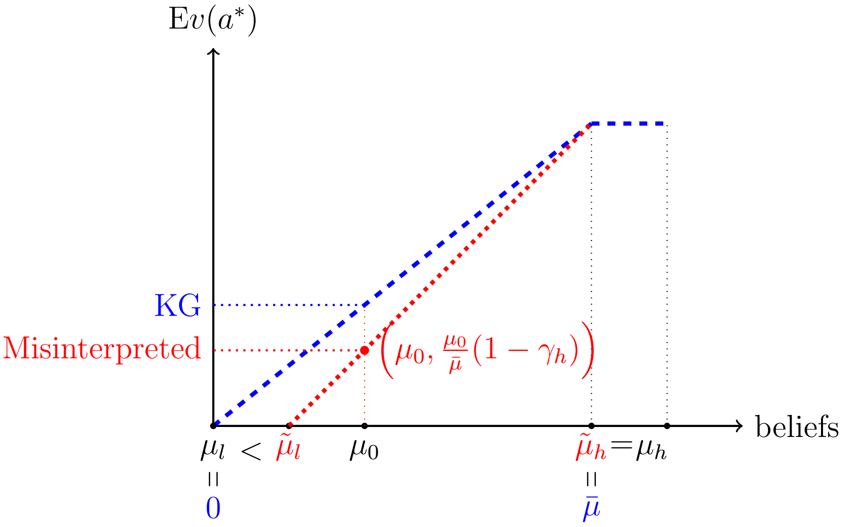
[7] Similarly, the sender's problem is just a combination of two special cases of the baseline model with a naïve receiver, with an additional constraint on the posterior beliefs. The posterior condition still restricts the solution to a half-space that doesn't change the nature of the convex optimization.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

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 13A: solution at $\mu_0 \in (0, \frac{\bar{\mu}}{2}]$



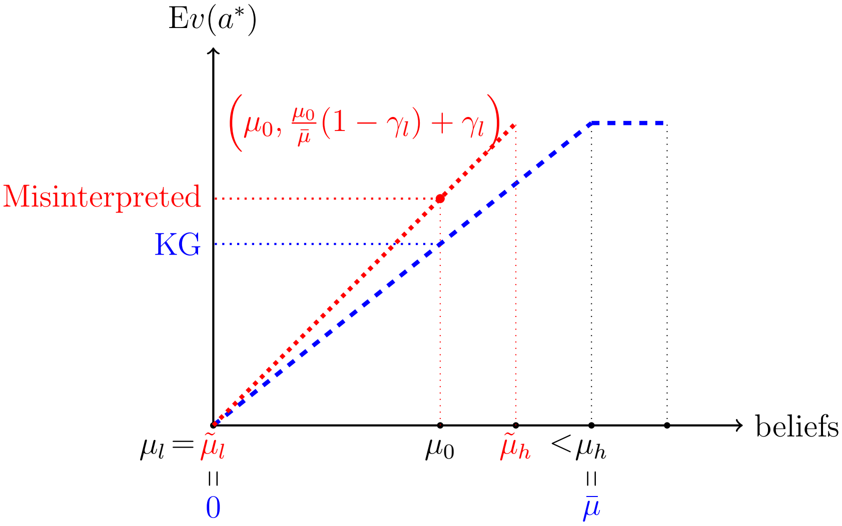
(2) For $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$, the receiver misinterprets under Γ_l in equilibrium and makes suboptimal decision ($\tilde{\mu}_h^* < \bar{\mu}$).

- 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 13B: solution at $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$



4 Conclusion

In this paper, we investigated how different forms of misinterpretation impact strategic communication. Misinterpretations—whether from language complexity, cognitive biases, or other communication frictions—affect persuasion by influencing the receiver's response to information. Using a Bayesian persuasion model, we examined both sophisticated and naïve receivers, showing that the sender's persuasive power is shaped by how the receiver understands his information environment.

Our findings reveal that the sophisticated receiver, aware of potential misinterpretation, requires more information to be persuaded than the naïve or rational receiver, which mitigates the sender's advantage. Conversely, naïveté benefits the sender at the receiver's expense, particularly through the channel of favorable misinterpretation.

We further extended the model to include confirmation bias, demonstrating how a receiver's pre-existing beliefs can skew information interpretation and thus affect equilibrium outcomes. Importantly, 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 matrix.

Through this framework, we show how misinterpretation and naïveté alter the dynamics of persuasion, with important implications for settings where both the endogenous information strategy and the exogenous probability of misinterpretation play a role, such as judicial processes, political campaigns, consumer marketing, etc. By outlining both theoretical insights and practical implications, our analysis offers a foundation for understanding how to navigate and, where possible, mitigate the effects of communication noise in real-world persuasion contexts.

Further research could relax the restriction on the realization space. The difficulty lies in specifying misinterpretation matrices for each number of distinct realizations and how the misinterpretation matrix contracts as the number of realizations decreases. Extending beyond binary state space posts a more dire challenge. To focus on the error of meaning consistent with the misinterpretation motivation, we need to order the posterior beliefs in the higher dimensional state space. With proper setting and assumptions, we conjecture that the insights generalize that misinterpretation hurts the sender and has no impact on the receiver but naïve misspecification shifts welfare surplus from the receiver to the sender.

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.

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 Baseline Model

A.1.1 Sophisticated receiver

Proof. of Proposition 1

"⇒" 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 $\tilde{\mu}_h(0, 1) \geq \bar{\mu}$.

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

"⇐" WTS 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,

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

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

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

■

Proof. of Proposition 2

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

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 favoritism noise $\gamma_l > 0$ hurts the sender by enlarging the region of prior that renders persuasion useless; the discriminatory noise $\gamma_h > 0$ hurts the sender by shrinking the profit from persuasion.

■

A.1.2 Naïve receiver

Proof. of Proposition 3

With naïveté misspecification, the sender's problem with *infrequent* misinterpretation is a *positive* linear transformation of the KG problem[8]. 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

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

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 3, 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 $\mu^* = (\mu_l^*, \mu_h^*) = (0, \bar{\mu})$. 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 favoritism noise ($\gamma_l > 0$) nor discriminatory 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

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

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.2 Confirmation Bias

A.2.1 Sophisticated Confirmation Bias

Proof. of Proposition 4

1. **Step 1 Case 1:**

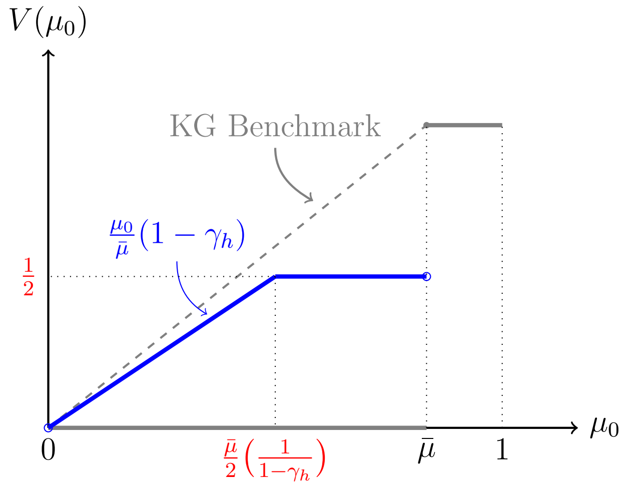
First, we search for a solution in $\left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}$ under $\Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint of the posterior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

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

Without the confirmation bias constraint on the posterior beliefs (CB_1^S), an optimal information policy induces Sender's posterior to $(0, \bar{\mu})$ by Corollary 1 and Sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$. For $\mu_0 \in \left(0, \frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right)\right]$, the CB_1^S constraint doesn't bind at the optimal Sender posterior $(0, \bar{\mu})$. For $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}\right)$, to satisfy the optimality (O_1^S) and the posterior (CB_1^S) constraints simultaneously, Sender can still induce $\hat{\mu}_h = \bar{\mu}$ by increasing μ_l so that CB_1^S is exactly satisfied. Then, Sender gets $\frac{1}{2}$. Figure 7A depicts the sender's value function with the Sophisticated Receiver in Case 1.

Figure 7A: Case 1 Value Function with Sophisticated Receiver



2. **Step 1 Case 2:**

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

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

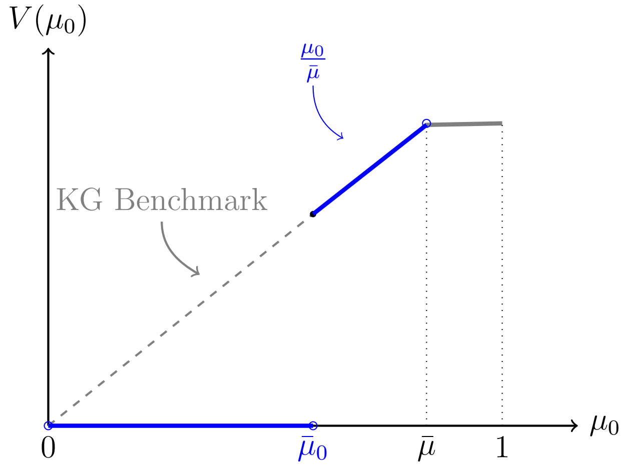
$$\max_{\mu_l,\mu_h} \tau_2(\mu_l, \mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\gamma_l)+\gamma_l$$

$$\text{s.t. } \tilde{\mu}_h(\mu_l, \mu_h) = \frac{(\mu_0-\mu_l)\mu_h+\gamma_l(\mu_h-\mu_0)\mu_l}{(\mu_0-\mu_l)+\gamma_l(\mu_h-\mu_0)} \geq \bar{\mu} \qquad (O_2^S)$$

$$\mu_0 > \frac{\mu_h+(1-2\gamma_h)\mu_l}{2(1-\gamma_h)} \qquad (CB_2^S)$$

   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 $\bar{\mu}_0^{O_2^S} := \frac{\gamma_l \bar{\mu}}{\gamma_l \bar{\mu} + 1 - \bar{\mu}}$[9] respectively. If either is violated, no strategy can induce the receiver to take the a_h action and the sender always gets 0. Therefore the cutoff belief $\bar{\mu}_0$ of the value function is just the larger of $\bar{\mu}_0^{CB_2^S}$ and $\bar{\mu}_0^{O_2^S}$.

Figure 7B: Case 2 Value Function with Sophisticated Receiver

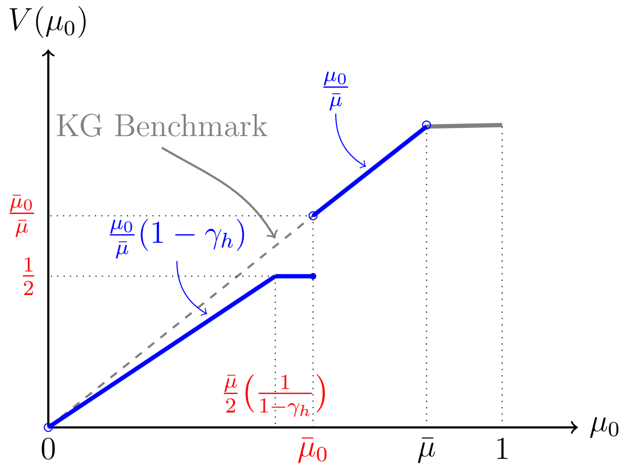


3. **Step 2 best of the two cases:**

   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 4.

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

Figure 7: Value Function with Sophisticated Confirmation Bias



■

A.2.2 Naïve Confirmation Bias

Proof. of Proposition 5

1. **Step 1 Case 1:**

   First, we search for a solution in $\left\{(\mu_l,\mu_h) \mid \mu_0 \le \frac{\mu_h+\mu_l}{2}\right\}$ 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:

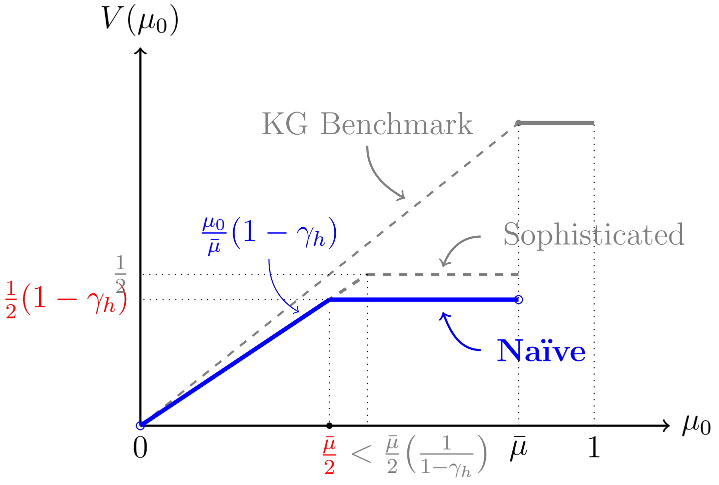
$$\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\tau_h)$$

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

$$\mu_0 \leq \frac{\mu_h+\mu_l}{2} \tag{CB_1^N}$$

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.

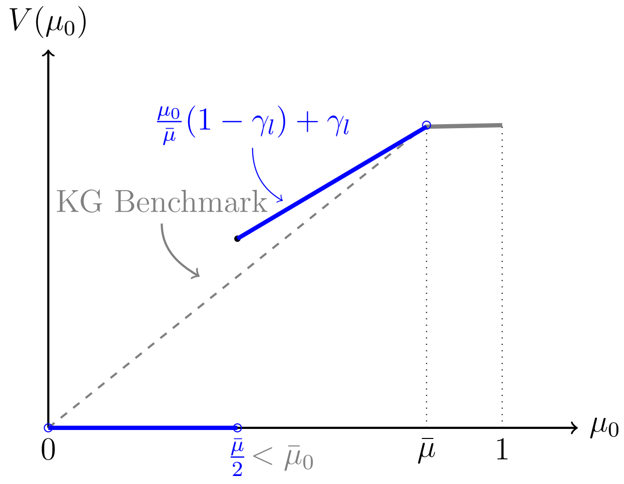
With Naïveté misspecification, the receiver updates to misspecified posterior coinciding with the sender's posterior μ like in Case 1. But the receiver's Bayesian high posterior $\tilde{\mu}_h$ is strictly less than his misspecified high posterior μ_h under Γ_l. Thus, the receiver makes a sub-optimal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

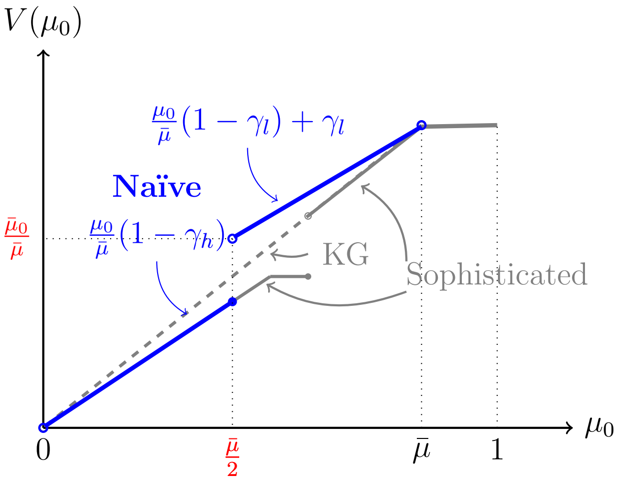
Figure 9B: Case 2 Value Function with Naïve Receiver



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 5.

Figure 9: Value Function with Naïve Confirmation Bias



■

B Results for Frequent Misinterpretation

B.1 Frequent Misinterpreted Receiver with Sophistication

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

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

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

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

where

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

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

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

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 $\gamma_l < 1$.

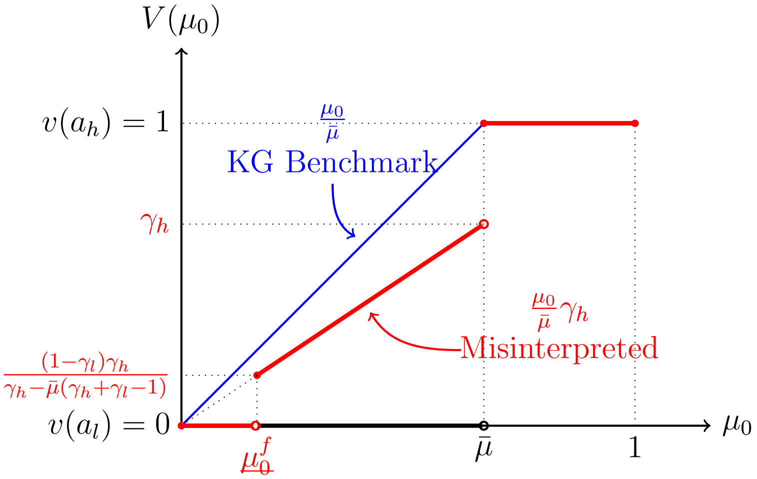


Figure 1^f: Value function comparison
— with *frequent* misinterpretation
— without misinterpretation

B.2 Frequent Naively Misinterpreted Receiver

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[10]. 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

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

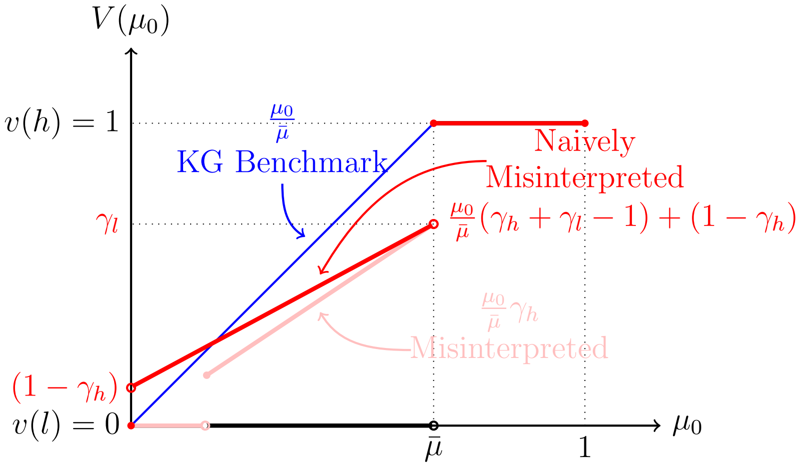


Figure 4^f: Value function comparison

— with *frequent* misinterpretation and naïveté

— with *frequent* misinterpretation and sophistication

— without misinterpretation

[10] 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 $\mu^* = (\mu_l^*, \mu_h^*) = (0, \bar{\mu})$.