# Misinterpreted Persuasion

Mengxi Sun[*]

October 2024

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 and the Receiver is unaffected as long as he is Bayesian about the interpretive errors. If the Receiver is unaware of these potential 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 showcase an application of the binary model in confirmation bias.

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

# 1 Introduction

Misinterpretation is common in communication. It could come from exchanging complex ideas that are hard to abstract into simple labels. We are also constrained by cognitive limitations that can easily bias our information assessment, such as stereotypes, prejudices, motivated reasoning, confirmation bias, etc. This paper provides a framework to analyze biased noise motivated by the above communication frictions of misinterpretation.

We investigate the signal realization perturbations in the Bayesian persuasion framework. Particularly, this paper focuses on the invertible error matrix that attaches the probability of mistakes to the information rather than the labels. We find that misinterpretation impairs the Sender's power to influence the Receiver's beliefs by introducing noise. The Receiver who misinterprets still makes the optimal decisions as long as he incorporates the potential errors in the Bayesian belief updating. We call the Receiver naïve if he doesn't know about his misinterpretation. Naïveté weakly hurts the naïve Receiver because he switches to the Sender-preferred action sub-optimally. The Sender gets the full benefits of the Receiver's sub-optimal decisions due to the commitment assumption in persuasion.

From the tractability perspective, misinterpretations that garble the information introduce interdependence among the meaning of realizations. Even a small perturbation of interpretation destroys the concavification characterization featured widely in the persuasion literature. Despite this technical difficulty it creates, misinterpretation opens up interesting channels of comparative statics that serve as cautionary tales for policy implementation.

To illustrate the intuitions, we want you to think about a persuasion scenario. Suppose that a politician attracts support from a voter by advocating a policy. The voter has a noisy and potentially biased assessment process of the policy outcomes. We say that the noisy assessment in favor of the politician is the favoritism noise and the noisy assessment biased against the politician is the discriminatory noise. A voter with discriminatory noise may misinterpret a good policy outcome as bad. Conversely, a voter with favoritism noise may misinterpret a bad policy outcome as a good one. I will use this election scenario as a running example, but you can think about other persuasion examples. We care about the effects of the voter's behaviors on both the politician's welfare (ex-ante probability of getting elected, minority representation, fairness, etc.) and the voter's welfare. The welfare of the politician is just the likelihood of getting into office if the politician is purely office-oriented and doesn't care about ideology. The voter's welfare depends on whether he makes the correct voting decision.

If misinterpretation is the only error that the voter makes, the likelihood of the politician getting elected is lower than when the voter has an accurate assessment of the policy outcomes. Both directions of noise, discriminatory and favoritism, hurt the politician's chance. Discriminatory noise hurts the politician because a bad policy outcome is no longer fully revealing for the sophisticated voter who correctly takes the noise into consideration. Favoritism noise hurts the politician because it completely closes down the persuasion channel for politicians with little chance to begin with. If the voter overlooks the noise in the assessment of policy outcome, this naïveté misspecification helps the politician and hurts the voter with favoritism because the voter is easily persuaded. The composite effect of (favoritism) misinterpretation and naïveté misspecification helps the politician the most with the voter who needs the most persuasion.

Consider a situation in which the voter has a noisy assessment of a policy advocated by a minority politician but accurately interprets policy outcome with a mainstream politician. This could be because the voter uses information–the minority group identity that shouldn't have affected his judgment. Then, the minority representation is lower than the mainstream politicians due to misinterpretative noise. We want to make change and care about both equality and fairness. Let us consider bringing up the average minority representation to the level of mainstream politicians as improving inequality; consider closing the gap between Misinterpreted Persuasion and the statistically correct Bayesian Persuasion as achieving fairness.

How do we improve inequality? There are three channels. First, if we are able to decrease the discriminatory noise directly, then we not only bring up the average minority representation but also narrow the gap between misinterpreted persuasion outcomes and the Bayesian benchmark. Equality and fairness goals are reached at the same time, yay!

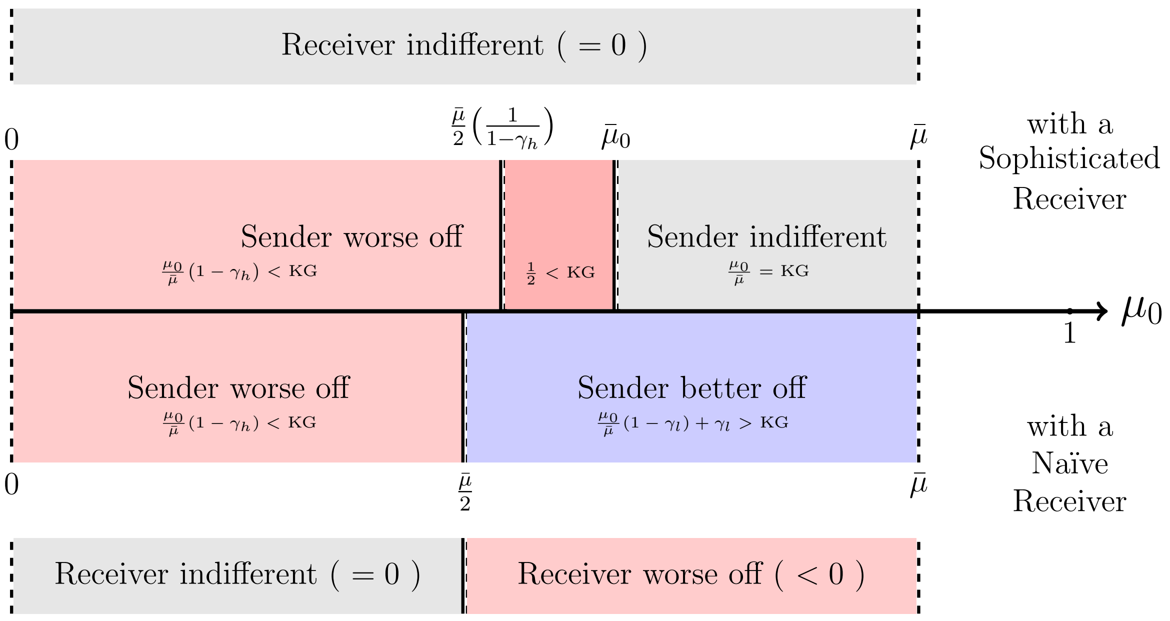
Secondly, how about we relax the standard for a minority candidate? This is an easy route to take in reality. It can also increase the average minority representation, but it disproportionally helps the more fortunate individual of the group who has a high chance of getting into the office to begin with. Dropping the bar indeed improves inequality but at the cost of deviating away from the Bayesian benchmark of fairness.

Last but not least, what if we increase favoritism noise? The minority representation increases if the voter is "naïve", but it drags down the average quality of minority elects, which fuels statistical discrimination. This is because a naive voter with favoritism votes disproportionally more for minority candidates with little prior chance. Can a sophisticated voter do better, who only misinterprets and correctly specifies his effective information environment? Unfortunately, favoritism manifests into a different detrimental consequence with a sophisticated voter. The minority representation decreases with an increase in favoritism. This is because the increasing favoritism noise of a sophisticated voter enlarges the region where the persuasion channel shuts down. Favoritism without naivety not only reduces the average minority representation but also is particularly unfair to the most disadvantaged candidates in the group with little prior chance of getting into office.

In essence, misinterpretation alone hurts the Sender and does not affect the Receiver in persuasion. Misinterpretatiive noise damages the Sender's ability to induce the Receiver's posterior beliefs. It has no impact on the Receiver because the Receiver uses Bayes' rule correctly with respect to his effect more noisy information environment, which guarantees optimal decision. On top of misinterpretation, naïveté misspecification has a zero-sum welfare effect: it weakly benefits the Sender and weakly hurts the Receiver. The Naïve Receiver with favoritism noise may make sub-optimal choices because he is too persuaded easily. The Sender takes advantage of this Receiver's mistake and she benefits the most from the Naive Receiver who needs the most persuasion (low prior).

These two key insights of misinterpretation and naivete misspecification carry beyond the binary model. We apply the binary results to demonstrate the welfare effects of a more complex behavior–confirmation bias–in persuasion. This is an important application of the binary misinterpreted persuasion model. A voter with confirmation bias misinterprets a disconfirming realization as a confirmation realization with some probability. Confirmation bias of the voter alone only adversely affects the politician when the voter needs a lot of persuasion (low prior) and has no impact on the politician when little persuasion is needed (high prior). It doesn't influence the probability that the voter makes correct vote decisions at all. However, naive confirmation bias benefits the politician and hurts the voter when little persuasion is needed because the voter's naive confirmation bias makes persuasion even more easily in equilibrium.

**Figure: 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
$\frac{\mu_0}{\bar{\mu}} = \text{KG}$
$\mu_0$
1
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 )

We are also working towards extending to the finite case. The invertibility of the misinterpretation matrix is sufficient for us to use the belief approach to solve the persuasion problem. But to generalize the welfare intuition, we want the finite model to be consistent with the Receiver's misinterpretation motivation in the binary model so that the Sender has no control over misinterpretation by manipulating the realization labels. The difficulty is that we define misinterpretation as perturbations of the realizations but need the error attached to the posterior beliefs rather than the labels. To confine misinterpretation to the meanings rather than the labels, we need to find a way to completely order all posterior beliefs, where we are currently stuck.

## Related Literature

This paper contributes to both persuasion and behavioral literature.

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 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. However, 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 follow from the intuition that noise hurts the Sender.

We also contribute to the persuasion literature by considering a behavior that introduces interdependence among the beliefs in the support of posterior distributions. As a result, the Sender's posterior beliefs are partially correlated to the Receiver's posterior beliefs. In (de Clippel and Zhang, 2022) and (Alonso and Câmara, 2016), the posterior beliefs of the Sender and the Receiver also don't agree, but we can rewrite the Receiver's posterior belief as a function of the Sender's posterior belief induced by the same realization. The concavification technique as prominently raised by Kamenica and Gentzkow (2011) is robust to these behaviors. However, with misinterpretation, the neat concavification characterization fails even at the smallest perturbations. Despite that, 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 naivete misspecification complements the welfare analysis of the system distortion as per de Clippel and Zhang (2022) by Bordoli (2024).

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 complex communication that leaves room for flexible interpretation of information because noise 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). We are particularly interested in confirmation bias, which is widely studied first by a group of psychologists (Lord et al., 1979; Plous, 1991; Darley and Gross, 1983) and later by 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). [Need to explain contribution on confirmation bias]

The analysis of the model corroborates experimental evidence and proposes an explanation for why agents sometimes don't respond to generic debiasing methods (Alesina et al., 2024; **?**). [Need to find more experimental refs]

In the following sections, we first introduce the binary model and analyze the welfare effects. The key insights of the binary model carry through the remainder of the paper. In section 3, we apply the results from the binary model to confirmation bias. Lastly, we discuss behavioral decomposition and welfare effects in the context of persuasion.

# 2 Binary Model

This section focuses on the canonical Prosecutor-Judge example in KG. Suppose a politician (she, the Sender) of ability $\omega \in \{L, H\}$ tries to persuade a voter (he, the Receiver) for support (donation, vote, etc.). The voter could either do nothing ($a_l$) or support ($a_h$), $\mathcal{A} = \{a_l, a_h\}$. The politician and the voter share a common prior belief at $\mu_0 := Prob.(H) \in (0, 1)$.

The politician can influence the voter's belief through an information policy generating either a bad ($l$) or good ($h$) outcome/signal realization, $\mathcal{S} = \{l, h\}$. She gets $v(a_h) = 1$ if the voter supports her and $v(a_l) = 0$ if the voter does nothing, regardless of the politician's ability. However, the voter only wants to support the politician if he believes that the politician is sufficiently likely to have $H$ ability. We denote the voter's indifference between doing nothing ($a_l$) and supporting ($a_h$) as $\bar{\mu} := \frac{\big(u(a_l,L)-u(a_h,L)\big)}{\big(u(a_h,H)-u(a_l,H)\big)+\big(u(a_l,L)-u(a_h,L)\big)} \in (0, 1]$[1].

Let $\pi_\omega$ represent and probability that the Sender sends $h$ realization in state $\omega \in \{L, H\}$ and the 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 $\Pi$ before Nature chooses a state. Then, a realization $s \in \{h, l\}$ is generated according to $\Pi$. Everything follows from the KG model up until now.

Here comes the miscommunication. The Receiver interprets $s$ as $\tilde{s}$ with probability $\gamma(s \mid \tilde{s})$. The Sender still receives the realization as designed, $s \in \{h, l\}$, but the Receiver may perceive the realization differently, $\tilde{s} \in \{h, l\}$. We parameterize the probability of the Receiver's misinterpretation as $\Gamma = \begin{bmatrix} 1-\gamma_l & \gamma_l \\ \gamma_h & 1-\gamma_h \end{bmatrix}$ with $\gamma_h$ being the probability of misinterpreting the realization ($h$) that induces higher posterior belief as the realization ($l$) that induces lower posterior belief and $\gamma_l$ being the probability of misinterpreting the realization ($l$) that induces lower posterior belief as the realization ($h$) that induces the

[1] In general, taking an action that matches the state is better than mismatching.

Doing nothing is strictly better than supporting a $L$ ability candidate: $u(a_l, L) > u(a_h, L)$;
Supporting a $H$ ability candidate is weakly better than doing nothing: $u(a_h, H) \geq u(a_l, H)$.

higher posterior belief. We call $\gamma_h > 0$ the discriminatory noise and $\gamma_l > 0$ the favoritism noise.

### Implementability

We focus on $\Gamma$ 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 with noise. (2) *Error of meaning*: the probability of misinterpretation attaches to the movement of beliefs rather than the label of realizations. As a consequence, the Sender cannot choose between the probability of misinterpretation between the two directions by flipping the realization labels. Additionally, any uninformative information policy doesn't noise the Receiver's posteriors away from the prior.

As a result, the Receiver's effective information policy, denoted as $\Phi$, is less informative than $\Pi$. $\Gamma$ captures the correlation between Sender's effective policy $\Pi$ and the Receiver's effective policy $\Phi := \Pi\Gamma = \begin{bmatrix} 1-\phi_L & \phi_L \\ 1-\phi_H & \phi_H \end{bmatrix}$, where $\phi_\omega := \pi_\omega(1-\gamma_h-\gamma_l)+\gamma_l$ is the probability that Receiver gets realization $\tilde{h}$ when the state is $\omega$.

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

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

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

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

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

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

Consequently, the Receiver perceives realizations $\tilde{s} \in (l, h)$ and arrives at his Bayesian posterior beliefs $\tilde{\mu}$ with probabilities

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

Since both players are correctly specified and update beliefs according to Bayes' rule, their posterior distributions $\tau$ 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 $\Gamma$ gives us bijection between $\Pi$ and $\Phi$. 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 valid probability distribution, we can say that there exist a pair of information policies $(\Pi, \Phi)$ that implements the posterior distribution $\tau(\mu, \tilde{\mu})$, where $\tau_1$ and $\tau_2$ are the marginal probabilities with respect to the first and the second component and $\Gamma$ captures the correlation between the two marginals. We know that, given a prior, the set of Receiver's posterior beliefs that the Sender can induce is weakly smaller than without misinterpretation. In addition to Bayes-plausibility, the Receiver's posterior beliefs have to satisfy another condition that the corresponding Sender's posterior needs to be valid beliefs.

### Optimality

With misinterpretations that perturb the realizations, we cannot find solutions through 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 Sender's and Receiver's posterior distributions are imperfectly correlated by $\Gamma$, 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 violates this independence requirement. So, concavification is not helpful because we need to determine the entire support of optimal posterior distribution simultaneously.

In this binary model, the Sender wants to maximize the probability of the Receiver taking the desirable action $a_h$. 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[2]. In a perfect Bayesian equilibrium, the Receiver takes the Sender-preferred action $a_h$ with the ex-ante probability of $\tau_2^h$ if possible.

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 mistake, misinterpretation and naïveté misspecification. In the last subsection, we analyze the impact of parameters through each type of mistake.

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

## 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

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

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

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

Full information revelation is always a strategy of 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. With misinterpretation, the noise weakly reduces the Sender's ability to induce the Receiver's posterior distributions, and hence weakly narrows the range of prior where she benefits from persuasion. We save space with sketch proof after all formal results and to see mathematical proofs, refer to Appendix A.

**Proposition 1.** *Given $\bar{\mu}$, $\gamma_l$, and $\gamma_h$, Sender benefits from persuasion with a Sophisticated Receiver if and only if the common prior is large enough so that the Sender can persuade the Receiver to switch to the desirable action, $\mu_0 \geq \frac{\gamma_l \bar{\mu}}{(1-\gamma_h)(1-\bar{\mu}) + \gamma_l \bar{\mu}} =: \underline{\mu_0}$.*

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. 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, $\tilde{\mu}_h(\mu_l, \mu_h) \leq \tilde{\mu}_h(0,1) \forall (\mu_l, \mu_h) \in [0, \mu_0) \times (\mu_0, 1]$. Consequently, the Sender cannot get a strictly better payoff through misinterpreted persuasion when even revealing full information is not enough to convince 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 at the prior and any posterior beliefs in the support of equilibrium posterior distribution.

**Proposition 2.** *When Sender benefits from persuasion with a Sophisticated Receiver, an optimal information policy induces the Receiver's Bayesian posterior to the indifference threshold, $\tilde{\mu}_h(\mu_l^*, \mu_h^*) = \bar{\mu}$.*

A direct welfare implication of Proposition 2 is that the Receiver with misinterpretation still makes the correct decisions as long as he is Bayesian w.r.t. his effective information policy $\Phi$. For the Sender, compared to the KG benchmark, misinterpretation reduces her payoff only 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 favoritism ($\gamma_l > 0$) by the Sophisticated Receiver reduces the Sender's ability to move the high posterior too far away from the prior to the action-switching threshold belief $\bar{\mu}$. For high priors ($\mu_0 > \underline{\mu_0}$), misinterpretation hurts the Sender because discrimination ($\gamma_h > 0$) by the Sophisticated Receiver reduces the Sender's ability to move the low posterior too far away from to prior to 0, which maximizes the ex-ante probability of sending $h$ realization. Formally,

**Corollary 1.** *(Welfare effects of misinterpretation)*

1. *Misinterpretation doesn't affect the Receiver's payoff.*

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

2. *Misinterpretation strictly hurt the Sender by reducing 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 favoritism* $\gamma_l > 0$.
   - *The Sender's gain from misinterpreted persuasion is strictly less than that in KG if and only if there is discrimination* $\gamma_h > 0$.

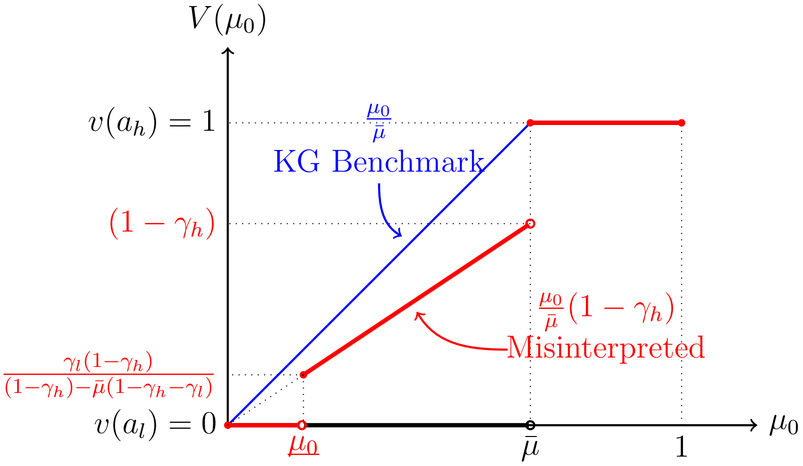


**Figure 1: Value function comparison**
🟥 with *infrequent* misinterpretation
🟦 without misinterpretation [3]

### 2.1.2 Comparative Statics

So far, we have solved and analyzed the effects of misinterpretation in persuasion. We attribute these effects to each parameter. A natural next question would be how the effects change with the 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 previous analysis, $\gamma_l$ negatively affects the Sender by limiting her ability to induce the Sophisticated Receiver's high posterior beliefs. 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 $\gamma_l$ increases but the total noise is infrequent such that the meaning of realizations doesn't flip between the Sender and the Receiver, more misinterpretation hinders information transmission: $\underline{\mu_0}$ increases in $\gamma_l$. However, as $\gamma_l$ continues to increase passing the point where the meaning of realizations flips, more misinterpretation starts to ameliorate the negative impact: $\underline{\mu_0}^f$ decreases in $\gamma_l$. The following graph plots the range of priors where the Sender can benefit from misinterpreted persuasion against $\gamma_l$.

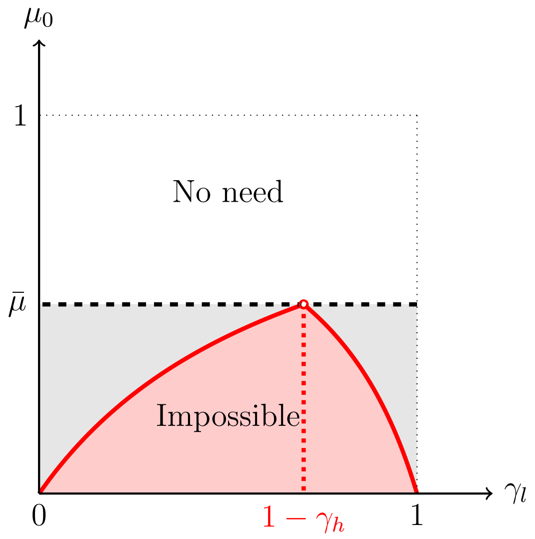


**Figure 2: $\gamma_l$'s Impact on Range of Prior that Sender Benefits**
The graph shown fixing $\gamma_h = 0.3$ and $\bar{\mu} = 0.5$
🟥 persuasion channel shuts down due to $\gamma_l$
▒ the range of priors where the Sender benefits
- - - the upper bound of persuasion: $\bar{\mu}$ exclusive
— the lower 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 persuasion[4], $\gamma_l$ has no impact and

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

$\gamma_h$ negatively affects the Sender by limiting her ability to induce Sophisticated Receiver's low posterior beliefs. This restriction manifests as an informational loss that reduces the Sender's profit from persuasion. The following graph depicts the effect of $\gamma_h$ on the Sender's persuasion profit at two examples of $\mu_0$.

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

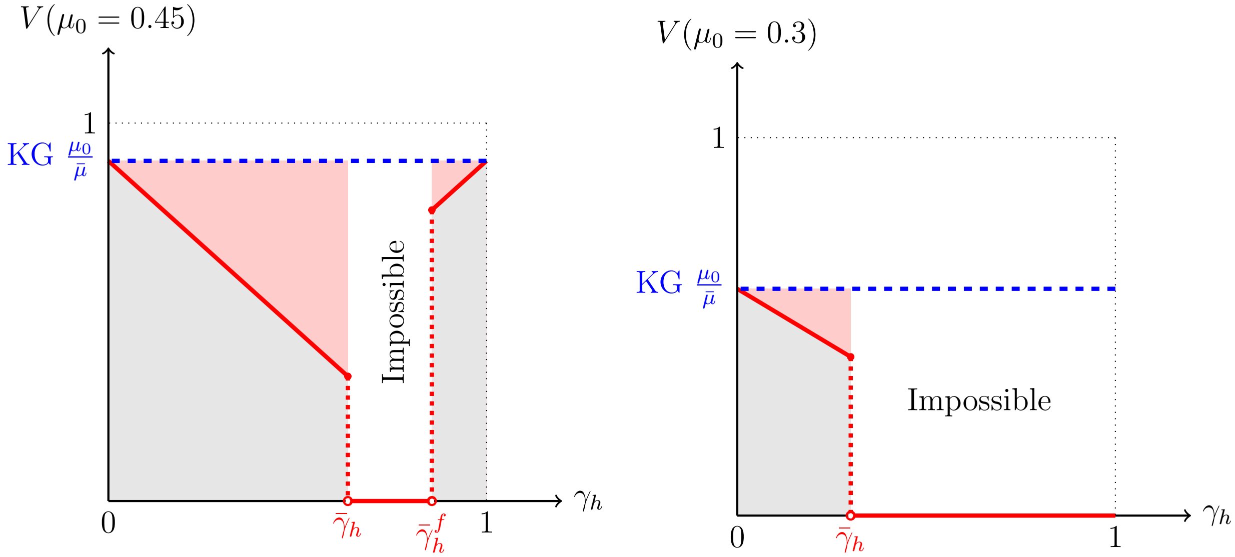


**Figure 3: $\gamma_h$'s Impact on Sender Profit from Misinterpreted Persuasion**
The graph shown fixing $\gamma_l = 0.3$ and $\bar{\mu} = 0.5$
🟥 informational loss due to $\gamma_h$
▒ Sender's profit from persuasion with misinterpretation
- - - the optimal value from Bayesian Persuasion
— the optimal value from Misinterpreted Persuasion

## 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 Naive Receiver who misspecifies his information environment to be what the Sender has announced, $(\Pi, \tilde{s})$, despite his true information environment being $(\Phi, \tilde{s})$. Hence, instead of Receiver's Bayesian posteriors $\tilde{\mu}$, the Naive Receiver arrives at misspecified posterior beliefs equal to the Sender's Bayesian posterior belief $\mu = (\mu_l, \mu_h)$, but still with probability $\tau_2$. Now, in addition to misinterpretation breaking the independence among posterior beliefs, we further lose Bayes-plausibility to naïveté misspecification. Since the naïveté misspecification is special to misinterpretation, the composite behavior of the Receiver makes the Sender's problem easier to solve than the previous one with only misinterpretation. The Sender still maximizes the probability of the Receiver taking the Sender-preferred action $a_h$, but subject to a different constraint since the Naive Receiver's subjective posteriors coincide with the Sender's Bayesian posteriors $\mu$ but not his Bayesian posteriors $\tilde{\mu}$ anymore.

### 2.2.1 Solution and Welfare Analysis

With infrequent misinterpretations ($\frac{\gamma_l}{1-\gamma_h} < 1$), the Sender's problem has the same solution as the KG benchmark.

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

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

**Proposition 3.** *The Sender has the same implemetability as in KG if the Receiver is fully naive about his misinterpretations. When Sender benefits $\mu_0 \in (0, \bar{\mu})$, an optimal information policy induces the Receiver's misspecified posterior ($\mu_h$) to the indifference threshold ($\bar{\mu}$), which is weakly (strictly if there exists favoritism $\gamma_l > 0$) higher than the Receiver's Bayesian posterior ($\tilde{\mu}_h$):*

$$\tilde{\mu}_h(\mu_l^*, \mu_h^*) \leq \mu_h^* = \bar{\mu}$$

.

Unlike the Sophisticated Receiver, the Naïve Receiver switches to Sender-preferred action sub-optimally. He should take $a_h$ when his Bayesian posterior is above the indifference threshold $\bar{\mu}$. But in equilibrium, the Sender only needs to bring the Naïve Receiver's subjective posterior $\mu_h$ to $\bar{\mu}$, which is weakly easier since $\tilde{\mu}_h \leq \mu_h$ (with equality if no favoritism $\gamma_l = 0$).

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

1. *Naïveté 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 favoritism.*

2. *Naïveté misspecification strictly 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 naive favoritism.*

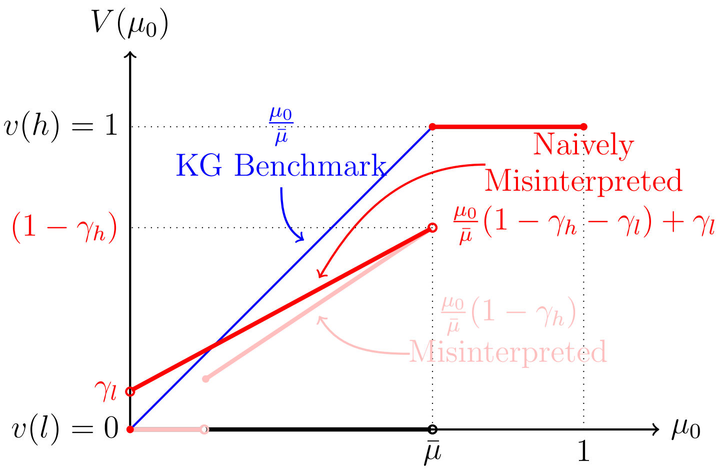


**Figure 4: Value function comparison**
— with *infrequent* misinterpretation and naïveté
— with *infrequent* misinterpretation and sophistication
— without misinterpretation [5]

[5] Results under *frequent* misinterpretation is shown in the Appendix B.2.

Combining the effects of both misinterpretation and naïveté 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 discrimination ($\gamma_h > 0$) restricting the Sender's ability to induce low posteriors and favoritism ($\gamma_l > 0$) restricting the Sender's ability to induce high posteriors. On the other hand, naïveté benefits the Sender only through favoritism ($\gamma_l > 0$) which magnifies the Naïve Receiver's easiness to be persuaded as more persuasion is needed (low prior). 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 gain from naïveté eventually trumps the cost of misinterpretation for the Sender.

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

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

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

### 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 Naïvely Misinterpreted Persuasion.

With naïveté misspecification, the Receiver doesn't respond to potential misinterpretations. Thus, $\gamma_l$ doesn't restrict the range of priors where the Sender can benefit from persuasion. However, it does affect how beneficial the naïveté misspecification is to the Sender because the larger $\gamma_l$ is, the more sub-optimal the Receiver's decision is.

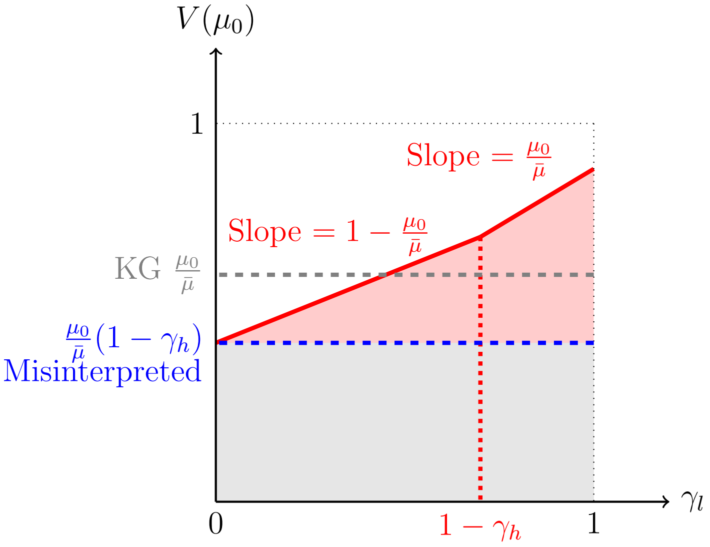


**Figure 5: $\gamma_l$'s Impact on Sender Profit from Naïvely Misinterpreted Persuasion**

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

▨ Sender's gain from naïveté due to $\gamma_l$

— the optimal value from Naïvely Misinterpreted Persuasion

▨ Sender's profit from persuasion with misinterpretation only

- - - the optimal value from Misinterpreted Persuasion

- - - the optimal value from Bayesian Persuasion

As $\gamma_h$ increases, the Sender's value from Naïvely Misinterpreted Persuasion decreases. When misinterpretation is *frequent*, the Sender may lose from naïveté when the prior is larger enough for some information to get through with Sophisticated misinterpretation. This is because if the Receiver is sophisticated, he should be able to infer the opposite meaning of the realizations and take the high action $a_h$ more often with high enough $\gamma_h$.

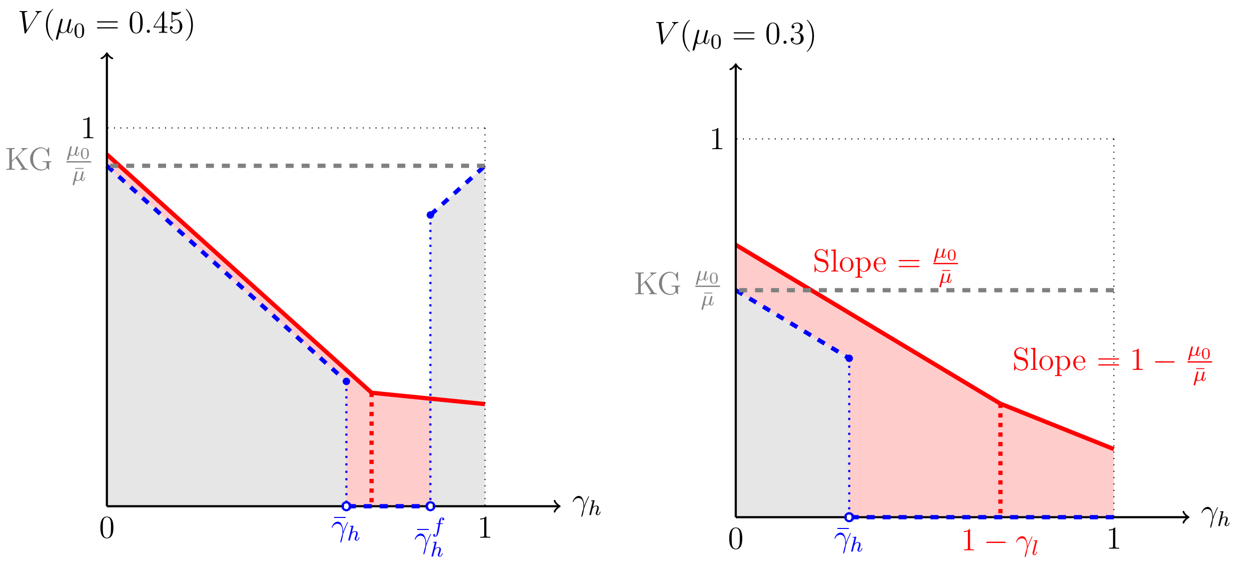


**Figure 6: $\gamma_h$'s Impact on Sender Profit from Naïvely Misinterpreted Persuasion**

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 Naïvely Misinterpreted Persuasion

⬜ Sender's profit from persuasion with misinterpretation only

- - - the optimal value from Misinterpreted Persuasion

- - - the optimal value from Bayesian Persuasion

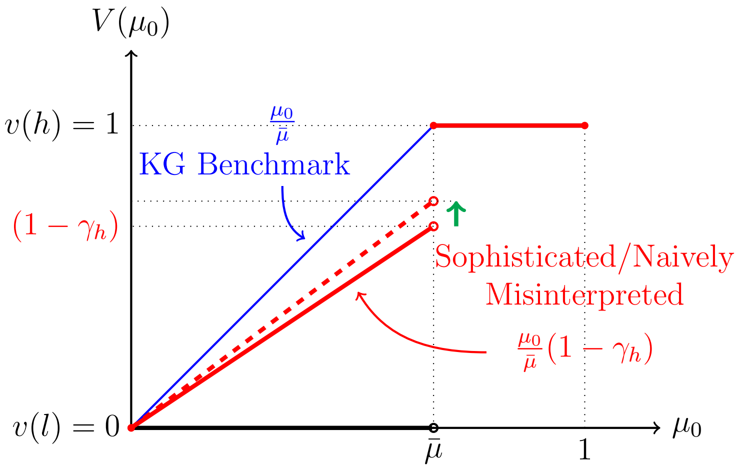
## 2.3 Policy Implications of (*Infrequent*) Misinterpretation

Let us return to the politician-voter example and think about what comparative statics means from the perspective of political representation. Suppose a voter discriminates against a minority politician by misinterpreting $h$ outcome as $l$ outcome with probability ($\gamma_h > 0$), and no favoritism ($\gamma_l = 0$). Then, the minority representation is lower than the KG rational benchmark across the board (for $\mu_0 \in (0, \bar{\mu})$) with either a Sophisticated or Naïve voter. Suppose we want to improve the average probability of a minority politician getting into office. How can we achieve this?

### 2.3.1 Discriminatory noise $\gamma_h$

If we were able to improve voter misinformation by decreasing discriminatory noise $\gamma_h$, then we not only increase the minority representation but also get closer to the KG benchmark across the board. If we take the KG benchmark as statistically fair, then reducing $\gamma_h$ achieves equality and fairness at the same time.

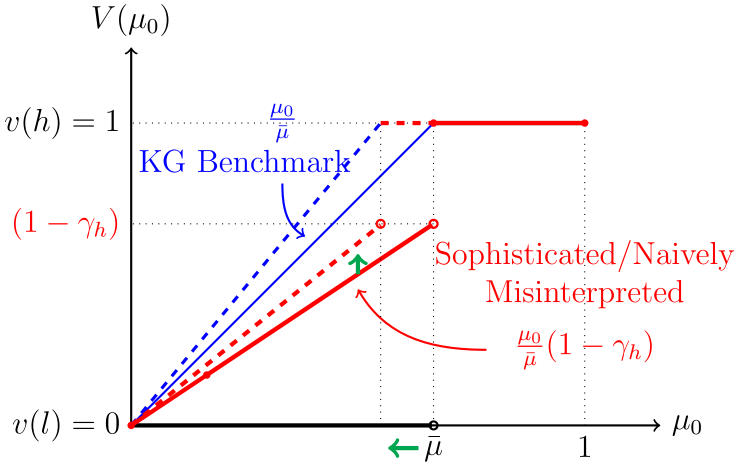
**Figure 7A: reducing discrimination $\gamma_h \downarrow$**



### 2.3.2 Belief threshold

Sometimes, it is difficult to directly improve discrimination ($\downarrow \gamma_h$). How about we drop the bar by reducing standards ($\downarrow \bar{\mu}$)? Minority representation increases on average. However, it doesn't help to close 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 $\bar{\mu} \downarrow$**

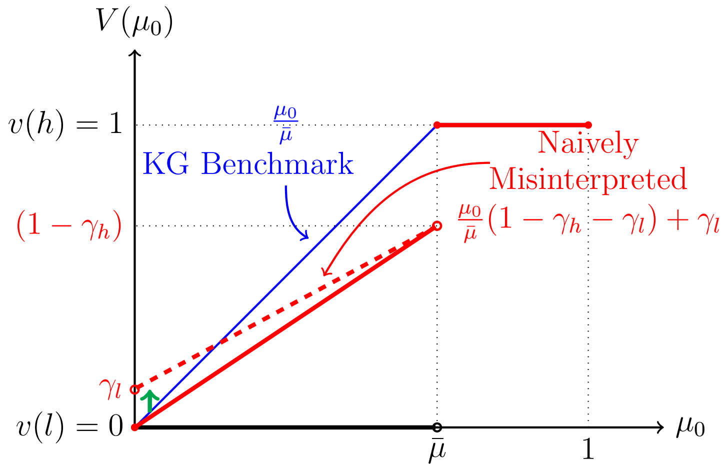


### 2.3.3 Favoritism noise $\gamma_l$

Lastly, we may (accidentally) introduce favoritism noise towards this minority candidate ($\uparrow \gamma_l$). The probability of misinterpreting $l$ outcome as $h$ outcome sounds favorable, but it is the most dangerous channel of the three.

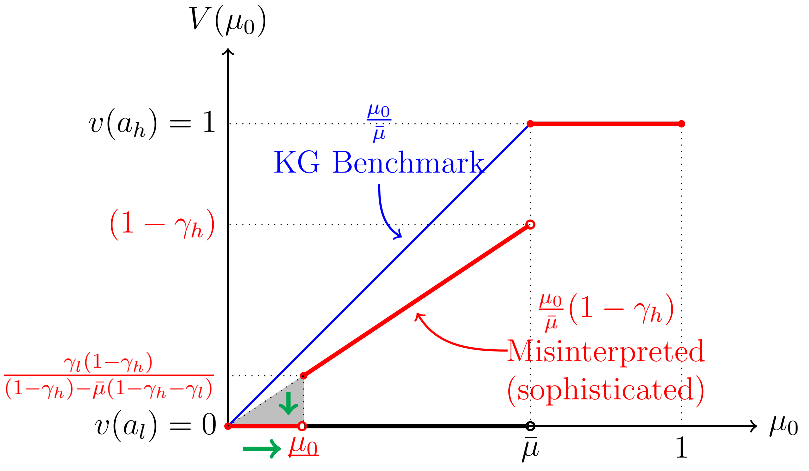
If the voter is a Naive Receiver, then having favoritism increases minority representation on average. It helps the most disadvantaged (low priors) in the group more and helps the more fortunate (high priors) in the group less. However, the misinterpreted outcome tilts away from the rational benchmark. The detrimental consequence is that the average quality of minority representation decreases, which fuels the statistical discrimination against minority politicians.

**Figure 7C: introducing favoritism $\gamma_l \uparrow$ with Naïve Receiver**



If the voter is a Sophisticated Receiver, then favoritism would hurt the most disadvantaged candidate in the group by shutting down the persuasion channel entirely. For low priors $\mu_0 \in (0, \underline{\mu_0})$ ($\neq \emptyset$ with $\gamma_l > 0$), it becomes impossible for these minority candidates to get elected.

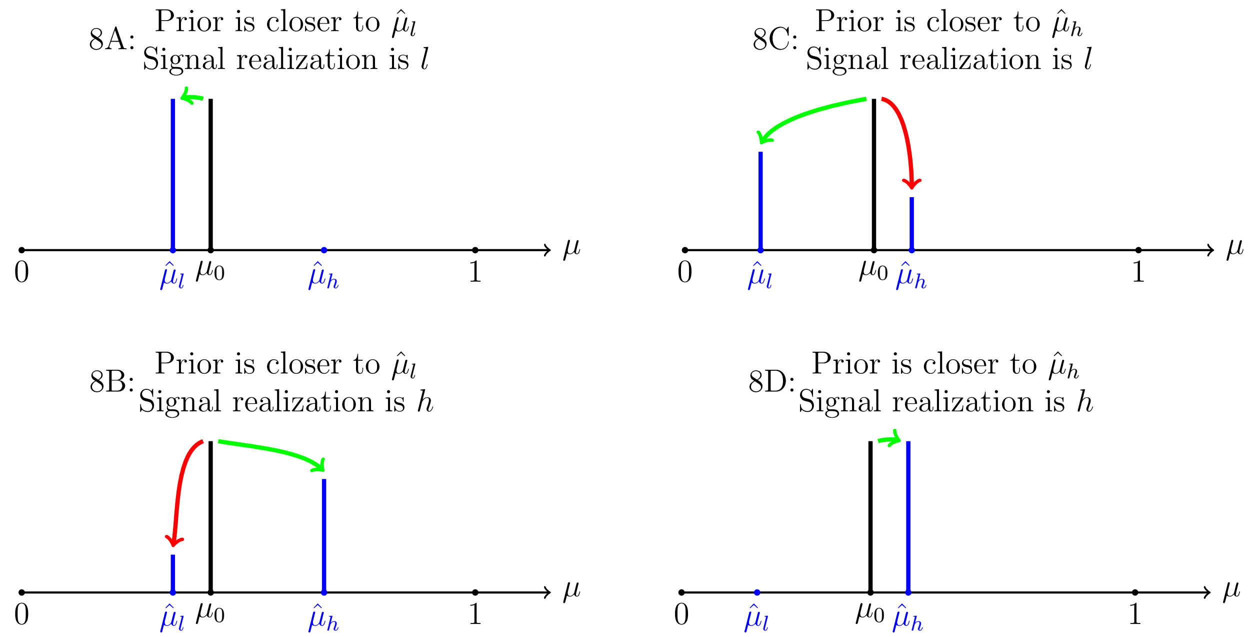
**Figure 7D: introducing favoritism $\gamma_l \uparrow$ with Sophisticated Receiver**



# 3 Application: Confirmation Bias

This section showcases an application–the confirmation bias. The setup is the same as the binary model. Instead of making errors represented by a single misinterpretation matrix, a Receiver with confirmation bias makes mistakes depending on the information policy. The Receiver is more likely to perceive whichever realization confirms his prior. In the binary setup, the results are almost a combination of two special cases of the binary 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 voter may misinterpret low realization ($l$) as high realization ($h$) but never misinterpret $h$ as $l$ realization. On the other hand, when the prior belief is closer to his subjective low posterior, the voter may misinterpret high realization $h$ as low realization $l$ but never misinterpret $l$ as $h$ realization. Figure 5 illustrates confirmation bias visually.



8A: Prior is closer to $\hat{\mu}_l$
Signal realization is $l$

8C: Prior is closer to $\hat{\mu}_h$
Signal realization is $l$

8B: Prior is closer to $\hat{\mu}_l$
Signal realization is $h$

8D: Prior is closer to $\hat{\mu}_h$
Signal realization is $h$

**Figure 8: Direction of Misinterpretation**

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

$\longrightarrow$ Misinterpret w.p. $\gamma_s$

Like before, denote the probability of misinterpreting $l$ as $\gamma_l$ and the probability of misinterpreting $h$ as $\gamma_h$. We can write the error matrice 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 $\mu_0$ and the Receiver's **subjective** posterior, which equates to Receiver's Bayesian posterior $\tilde{\mu}$ if he is sophisticated and coincides to Sender's Bayesian posterior $\mu$ if Receiver is naively misspecified. We also take the cutoff rule to be the one under $\Gamma_h$.

**Definition 1.** *(Confirmation Bias)*

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

1. *Sophisticated Receiver with confirmation bias exhibits errors represented by $\Gamma^{SCB}$.*

   - *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 $(\mu_l, \mu_h)$.*

2. *Naïve Receiver with confirmation bias exhibits errors represented by $\Gamma^{NCB}$.*

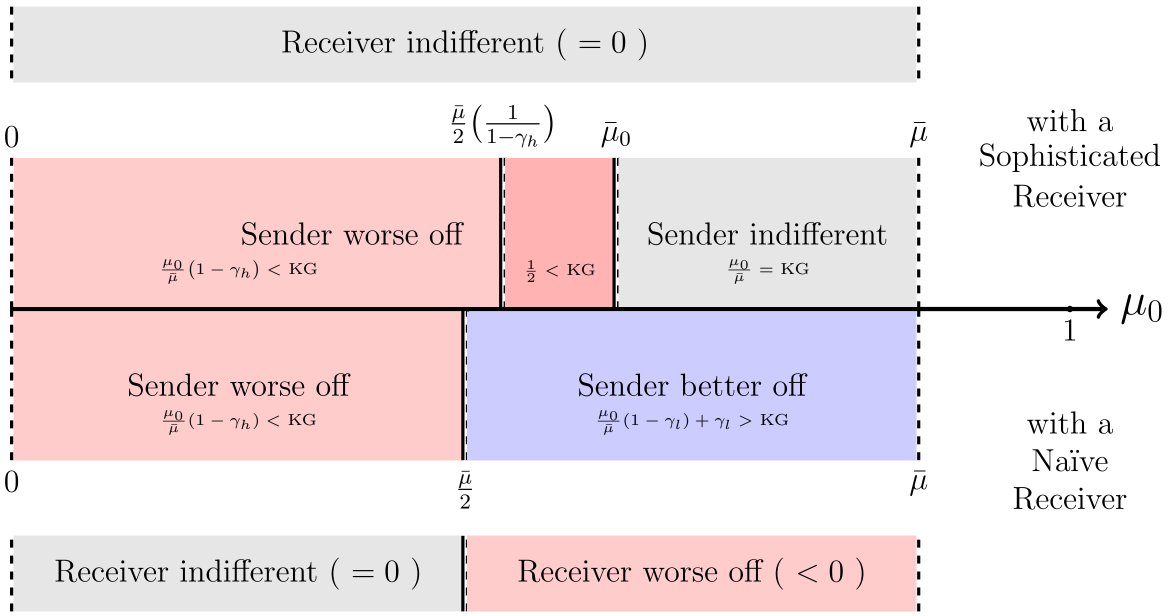
$$\Gamma^{NCB} = \begin{cases} \Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix} & , \textit{for} \left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h+\mu_l}{2}\right\}; \\ \Gamma_l := \begin{bmatrix} 1-\gamma_l & \gamma_l \\ 0 & 1 \end{bmatrix} & , \textit{for} \left\{(\mu_l, \mu_h)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

Given a problem with indifference threshold $\bar{\mu}$, prior $\mu_0$, and bias parameters $\gamma_l$ and $\gamma_h$, a Sender persuading a Receiver with confirmation bias solves the optimal strategy in two steps. First, she searches for a solution under each error matrix $\Gamma_h$ or $\Gamma_l$ 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 usually made indifferent in equilibrium. With an exception, when the Receiver is naive, he may make a sub-optimal decision by being over-precise/naïve.

Let us briefly return to the politician-voter example. Confirmation bias only hurts the voter when he is naïve and little persuasion is needed (high prior). In equilibrium, only high prior activates favoritism noise but low prior doesn't. As a consequence, the politician profits from confirmation bias more than in KG when the prior is high, which is opposite to Corollary 3. This is again due to the fact that confirmation bias is an interpretive error that varies with Sender's 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
$\frac{\mu_0}{\bar{\mu}} = \text{KG}$
$\mu_0$
1
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 welfare values formally. The method to find a solution with a Sophisticated or a Naive Receiver is the same. The difference is just in the 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 $\Gamma^{SCB}$. Fixing an indifference threshold $\bar{\mu}$, there exists a prior belief threshold*

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

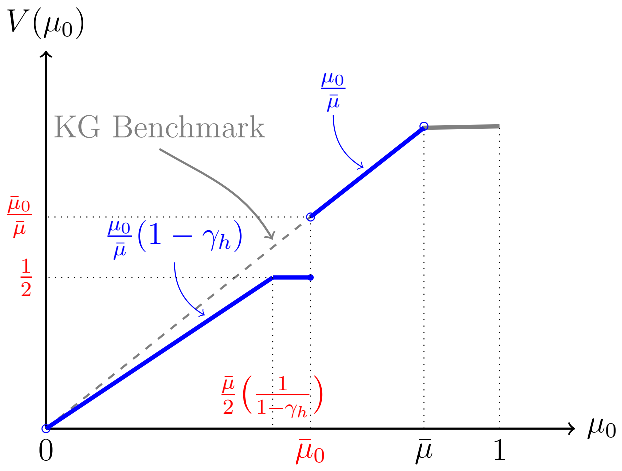
*such that in equilibrium*

- *For low priors ($\mu_0 \leq \bar{\mu}_0$), the Receiver misinterprets against the Sender (that is, the effective error matrix is $\Gamma_h$). 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 $\mu_0$ (above $\bar{\mu}_0$), the Receiver misinterprets in favor of the Sender (that is, the effective error matrix is $\Gamma_l$). Compared to the KG benchmark, 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 $\Gamma_h$ for low prior and $\Gamma_l$ 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 $(\mu_l, \mu_h)$, which generates the flat region in the middle.

**Figure 10: Value Function with Sophisticated Confirmation Bias**



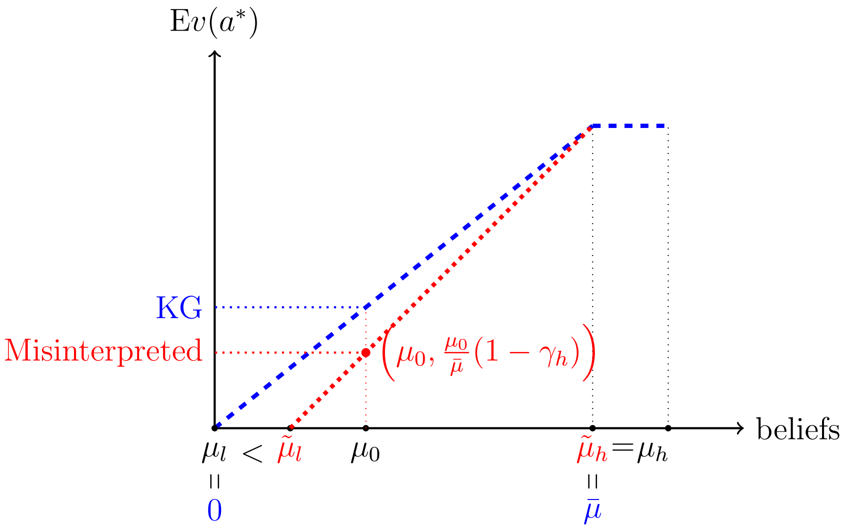
[6] The Sender's problem is just a combination of two special cases of the binary model with a Sophisticated Receiver, with an additional constraint on the posterior beliefs. The posterior constraint is a half-space that doesn't change the nature of the convex optimization. We show the detailed solution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

In the remainder of this subsection, we show a representative solution at $\mu_0$ in each of the three intervals. From these examples, we can see that the Receiver always makes the correct decisions by switching to higher action at the correct Bayesian belief, $\tilde{\mu}_h = \bar{\mu}$. If you are eager to learn the impact of Naïveté misspecification on top of confirmatory biased misinterpretation, skip to the next subsection.

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

- Sender updates to Sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = (0, \bar{\mu})$;
- Receiver updates to Receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \Big(\frac{\gamma_h \mu_0 \bar{\mu}}{\gamma_h \mu_0 + \bar{\mu} - \mu_0}, \bar{\mu}\Big)$;
- Sender gets $\frac{\mu_0}{\bar{\mu}}(1 - \gamma_h)$.

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

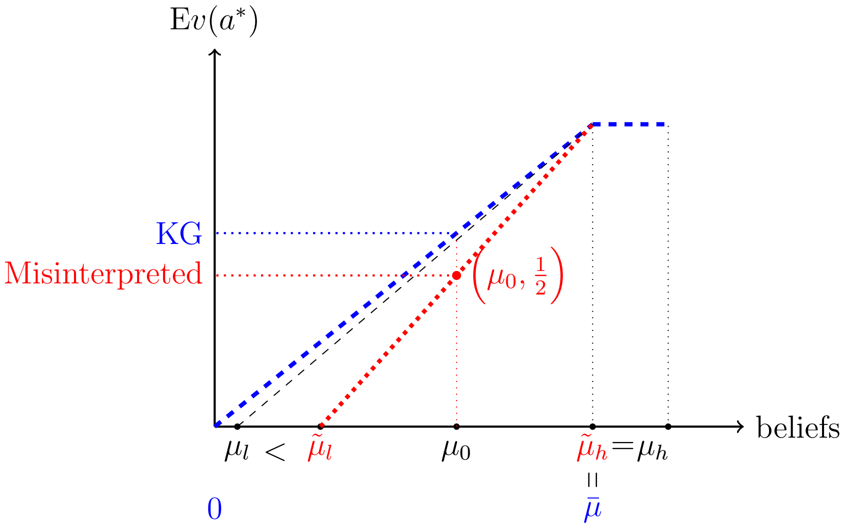


(2) For $\mu_0 \in \Big(\frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big), \bar{\mu}_0\Big]$, Receiver misinterprets under $\Gamma_h$ in equilibrium and always makes the optimal decision ($\tilde{\mu}_h^* = \bar{\mu}$).

- Sender updates to Sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = (\frac{2\mu_0 - \bar{\mu} + \gamma_h \bar{\mu}}{1+\gamma_h}, \bar{\mu})$;

- Receiver updates to Receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \Big(2\mu_0 - \bar{\mu}, \bar{\mu}\Big)$;
- Sender gets $\frac{1}{2}$.

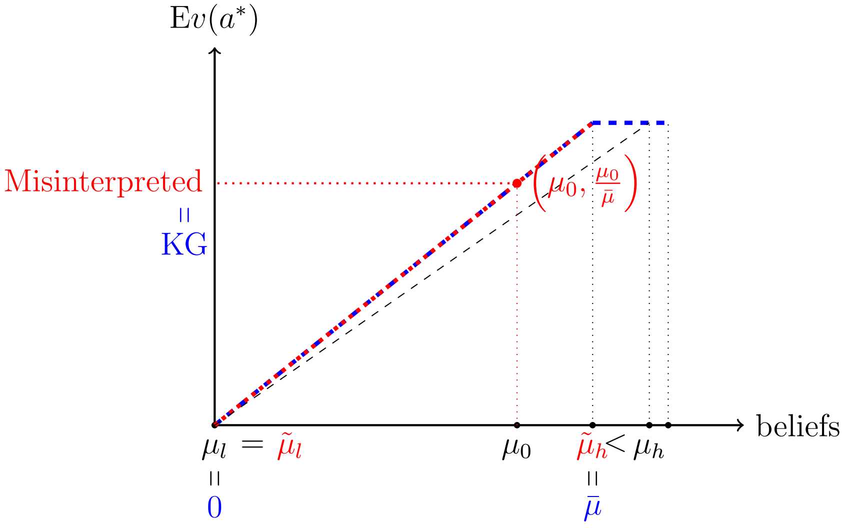
**Figure 11B: solution at** $\mu_0 \in \left(\frac{\bar{\mu}}{2}\left(\frac{1}{1-\gamma_h}\right), \bar{\mu}_0\right]$



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

- Sender updates to Sender's Bayesian posterior $(\mu_l^*, \mu_h^*) = \left(0, \frac{\bar{\mu}}{1-\frac{\gamma_l(\bar{\mu}-\mu_0)}{\mu_0(1-\gamma_l)}}\right)$;
- Receiver updates to Receiver's Bayesian posterior $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = (0, \bar{\mu})$;
- 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 Naive equivalent of Proposition 4 in the previous subsection. The steps are the same and we are solving a simpler optimization problem since Naive Receiver thinks that he is rational.

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

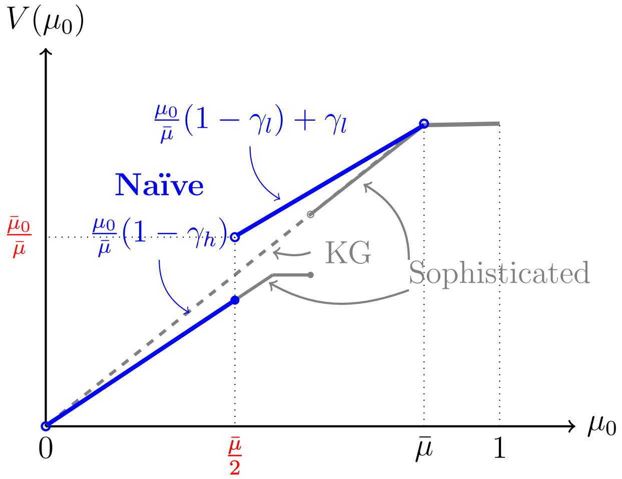
*Suppose a confirmatory biased Receiver is fully naïve and misinterprets according to* $\Gamma^{NCB}$*. Fixing an indifference threshold* $\bar{\mu}$*, there exists a prior belief threshold* $\frac{\bar{\mu}}{2}$ *such that in equilibrium*

- *For low priors* ($\mu_0 \leq \frac{\bar{\mu}}{2}$)*, the Receiver misinterprets against the Sender (that is, the effective error matrix is* $\Gamma_h$*). Compared to the KG benchmark and 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* $\Gamma_l$)*. Sender reveals the same amount of information compared to KG and less information compared to Sophisticated. The Receiver switches action before reaching* $\bar{\mu}$ *and thus gets strictly less payoff than in KG and Sophisticated benchmarks. However, the Sender gets a strictly higher payoff than in KG. Compared to Sophisticated, the Sender gains from naïveté; she profits the most for intermediate priors* $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu}_0]$.

Similarly, the outcome under naïve confirmation bias is also an almost direct application of Proposition 3 under $\Gamma_h$ for low prior and $\Gamma_l$ for high prior respectively. The seemingly contradictory result as opposed to Corollary 3 stems from the equilibrium strategy evoking different directions of misinterpretation (discriminatory with low prior and favoritism with high prior). With naïveté misspecification, confirmation bias also confines the solutions to half-spaces in $(\mu_l, \mu_h)$. 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**



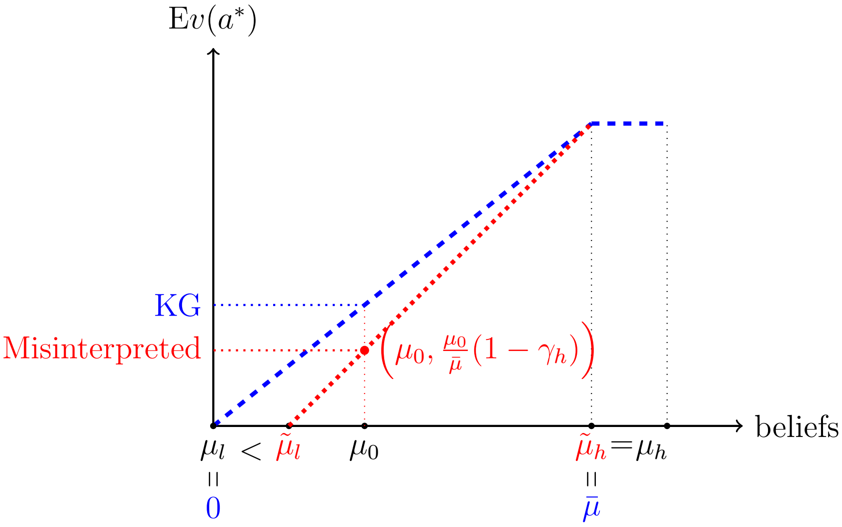
[7] Similarly, the Sender's problem is just a combination of two special cases of the binary model with Naïve Receiver, with an additional constraint on the posterior beliefs. The posterior constraint is still a half-space that doesn't change the nature of the convex optimization. We show the detailed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

Like in the Sophisticated case, the remainder of this subsection shows an example solution with a Naive Receiver for $\mu_0$ in each interval. These examples show that the naïve Receiver is worse off if and only if there is favoritism in equilibrium.

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

- Both Sender and (misspecified) Receiver updates to Sender's Bayesian posteriors at $(0, \bar{\mu})$.
- Receiver Bayesian posteriors are $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(\frac{\gamma_h \mu_0 \bar{\mu}}{\gamma_h \mu_0 + \bar{\mu} - \mu_0}, \bar{\mu}\right)$;
- Sender gets $\frac{\mu_0}{\bar{\mu}}(1 - \gamma_h)$.

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

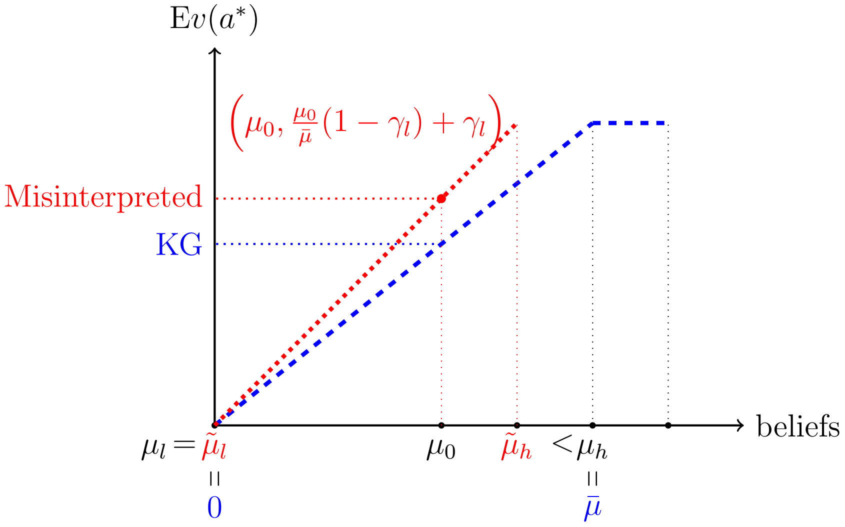


(2) For $\mu_0 \in (\frac{\bar{\mu}}{2}, \bar{\mu})$, Receiver misinterprets under $\Gamma_l$ in equilibrium and makes sub-optimal decision ($\tilde{\mu}_h^* < \bar{\mu}$).

- Both Sender and (misspecified) Receiver updates to Sender's Bayesian posteriors at $(0, \bar{\mu})$.

- Receiver Bayesian posteriors to $(\tilde{\mu}_l^*, \tilde{\mu}_h^*) = \left(0, \frac{\bar{\mu}}{1+\gamma_l(\frac{\bar{\mu}}{\mu_0}-1)}\right)$;
- 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 analyzed the impact of two types of Receiver errors in the persuasion model: misinterpretation and the related naïveté misspecification. We found that misinterpretation acts as a communication friction, imposing costs only on the Sender but not on the Receiver. Both discriminatory and favoritism noise limit the Sender's ability to guide the Receiver toward the most desirable beliefs. Both discriminatory and favoritism noise restrict the Sender's power to induce the most desirable Receiver's posterior beliefs. The Receiver can always make the correct decision as long as he is correctly specified of his effective information environment. On the contrary, naïveté misspecification creates a zero-sum welfare shift, favoring the Sender while disadvantaging the Receiver. Naïveté misspecification restores the Sender's loss of power to induce high posterior beliefs. The Sender gains from the Naïve

Receiver's loss by making a sub-optimal choice and requiring too little information from the Sender in equilibrium.

# 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 Binary Model

### A.1.1 Misinterpretation (only)

*Proof.* of Proposition 1

"$\Rightarrow$" Revealing full information to the Sender, $\mu = (\mu_l, \mu_h) = (0, 1)$, is always implementable as long as the posterior distribution $\tau_1$ over $\mu$ 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 $\tilde{\mu}_h(0, 1) \geq \bar{\mu}$.

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

"$\Leftarrow$" 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 $\mu_0$. 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,

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

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

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

■

*Proof.* of Proposition 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 $\mu_0$, 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 $\mu_h^*$ 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 Naïveté Misspecification on top of Misinterpretation

*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 $\tau_2^*$ 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 $\mu$ 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 $\tau_2^h$ 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 naïvely misinterpreted persuasion is

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

Compared to Misinterpreted only, the Sender is strictly better off for the range of prior that the Sender benefits from naïvely 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 naïvely 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:**

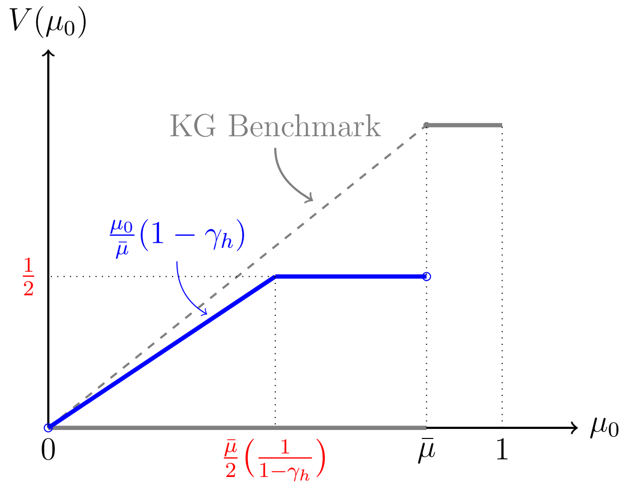
First, we search for a solution in $\left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h + (1-2\gamma_h)\mu_l}{2(1-\gamma_h)}\right\}$ under $\Gamma_h := \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint of the posterior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

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

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

Without the confirmation bias constraint on the posterior beliefs $(CB_1^S)$, an optimal information policy induces Sender's posterior to $(0, \bar{\mu})$ by Corollary 1 and Sender gets $\frac{\mu_0}{\bar{\mu}}(1-\gamma_h)$. For $\mu_0 \in \left(0, \frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big)\right]$, the $CB_1^S$ constraint doesn't bind at the optimal Sender posterior $(0, \bar{\mu})$. For $\mu_0 \in \left(\frac{\bar{\mu}}{2}\big(\frac{1}{1-\gamma_h}\big), \bar{\mu}\right)$, 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 $\mu_l$ 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 $(\mu_l,\mu_h)$, as opposed to $CB_1^S$ in Case 1.

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

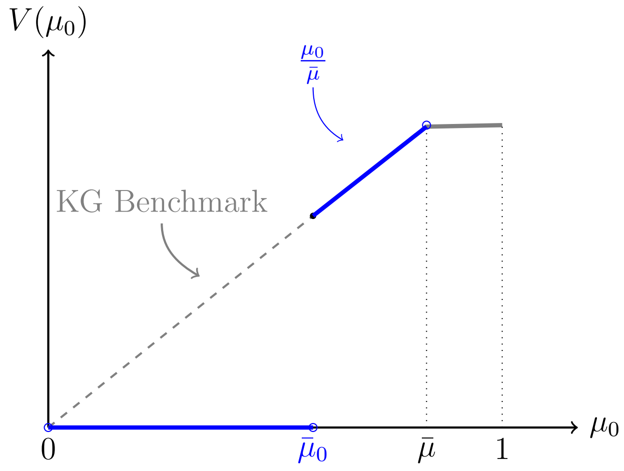
$$\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\gamma_l)+\gamma_l$$

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

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

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 $\mu_0$, and bias parameters $\gamma_l$ and $\gamma_h$, 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**

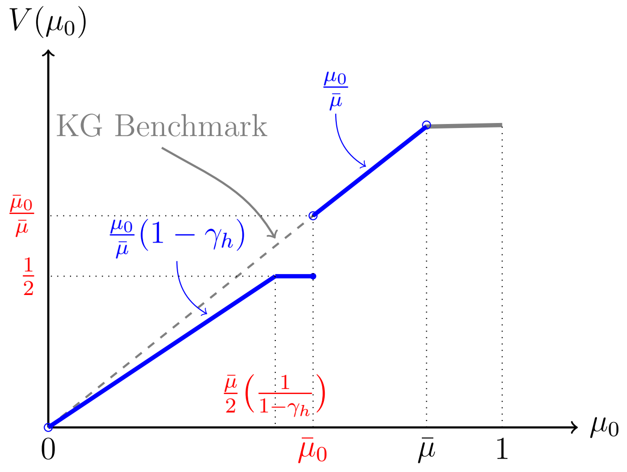


3. **Step 2 best of the two cases:**

   Now, we have solved the two cases separately. Given a prior $\mu_0$, the Sender can affect the effective direction of the bias by choosing different posterior pair $(\mu_l, \mu_h)$. So, she chooses the better between the two cases at each prior. For low priors below $\bar{\mu}_0$, $\Gamma_h$ takes effect and the Receiver misinterprets against the Sender in equilibrium; for high priors above $\bar{\mu}_0$, $\Gamma_l$ 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 $\left\{(\mu_l, \mu_h) \mid \mu_0 \leq \frac{\mu_h+\mu_l}{2}\right\}$ under $\Gamma_h = \begin{bmatrix} 1 & 0 \\ \gamma_h & 1-\gamma_h \end{bmatrix}$. This is a binary model in the previous section with an additional constraint on the posterior beliefs, which imposes solutions to a half-space in $(\mu_l, \mu_h)$.

   With Naïveté misspecification, the Receiver updates to a misspecified posterior coinciding with the Sender's Bayesian posterior $\mu$. Under $\Gamma_h$, the Receiver's Bayesian high posterior $\tilde{\mu}_h$ equals to the Sender's high posterior $\mu_h$. Thus, the Receiver makes optimal decisions in equilibrium even with misspecification.

The Sender solves the following problem:

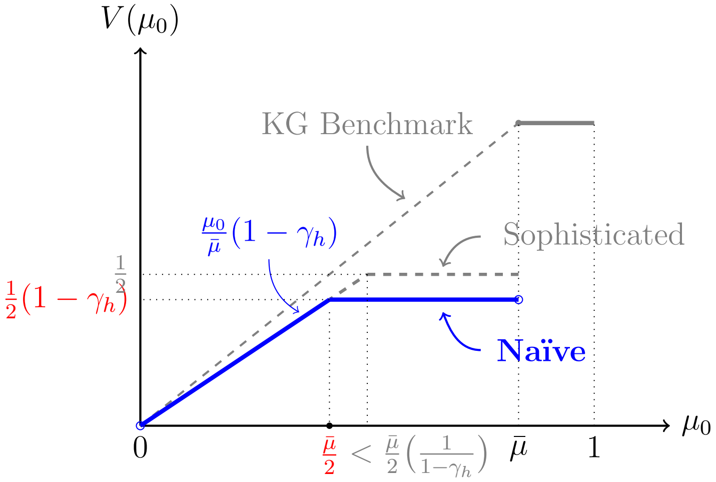
$$\max_{\mu_l,\mu_h} \tau_2(\mu_l,\mu_h) = \frac{\mu_0-\mu_l}{\mu_h-\mu_l}(1-\tau_h)$$

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

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

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 $\mu_h$ to $\bar{\mu}$ by increasing $\mu_l$ 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 $(\mu_l, \mu_h)$, as opposed to $CB_1^N$ in Case 1.

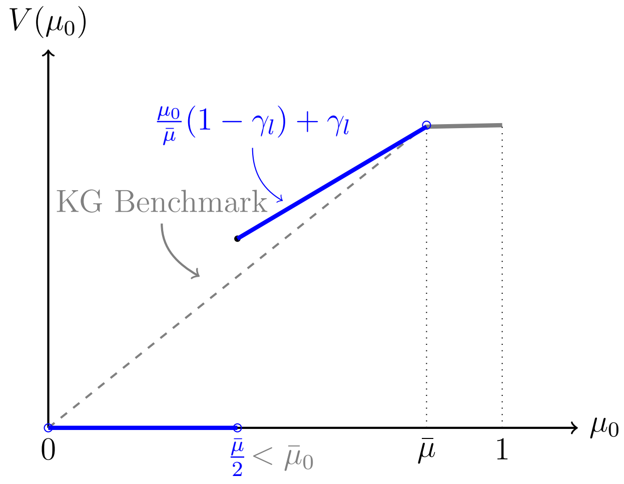
With Naïveté misspecification, the Receiver updates to misspecified posterior coinciding with the Sender's posterior $\mu$ like in Case 1. But the Receiver's Bayesian high posterior $\tilde{\mu}_h$ is strictly less than his misspecified high posterior $\mu_h$ under $\Gamma_l$. Thus, the Receiver makes 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.

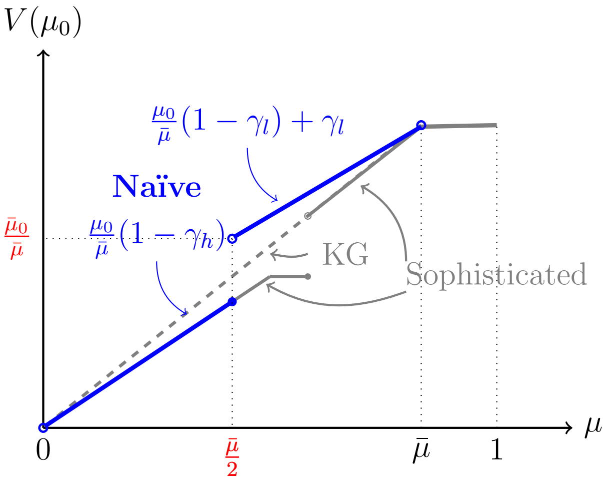
**Figure 9B: Case 2 Value Function with Naïve Receiver**



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

   Now, we have solved the two cases separately. Given a prior $\mu_0$, the Sender can decide the effective direction of the bias by choosing between the posterior pairs $(\mu_l, \mu_h)$. 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 $\gamma_l < 1$.

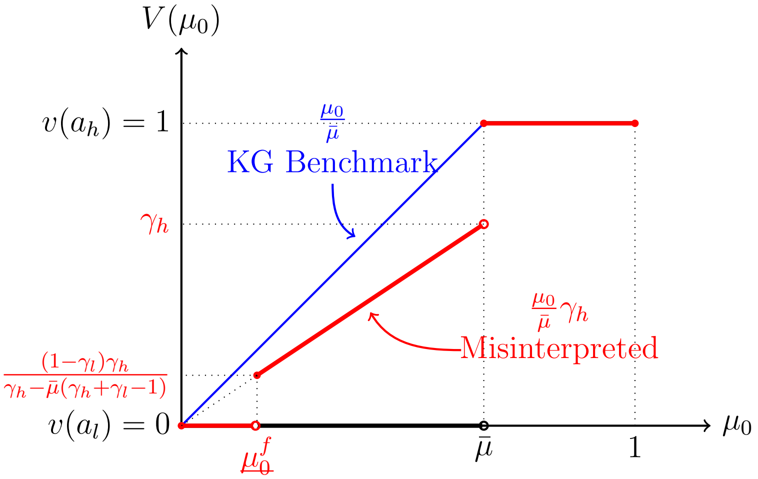


**Figure $1^f$: Value function comparison**
— with *frequent* misinterpretation
— without misinterpretation

## B.2 Frequent Naïvely 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* naive misinterpretation case under a different condition of the parameters.

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

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

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

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

The optimal strategy induces the posterior distribution to minimize $\tau_1^h(\mu_l, \mu_h) = \frac{\mu_0 - \mu_l}{\mu_h - \mu_l}$. So, the solution with *frequent* misinterpretation flips $\mu_l^*$ and $\mu_h^*$ of the solution with *infrequent* naïve misinterpretation[10]. Thus, for $\mu_0 \in (0, \bar{\mu})$, the Sender's optimal profit from *frequent* naïvely 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

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

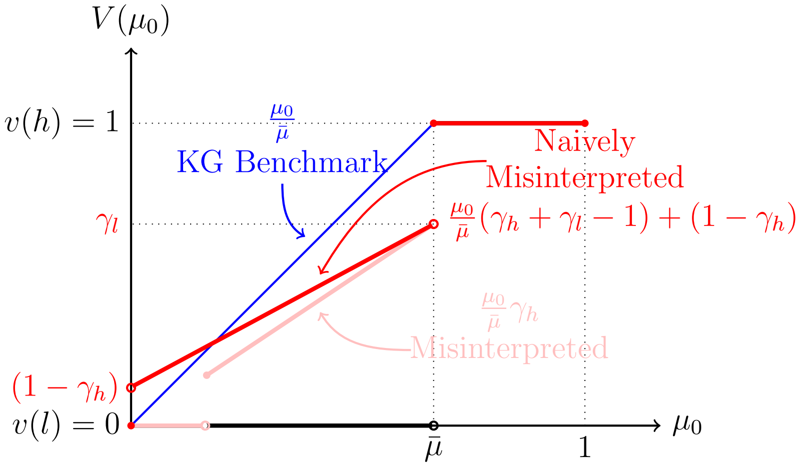


**Figure $4^f$: Value function comparison**

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