Human-AI Interactions and Societal Pitfalls

Francisco Castro, Jian Gao
Anderson School of Management, UC Los Angeles, Los Angeles, CA 90024, francisco.castro@anderson.ucla.edu, jian.gao.phd@anderson.ucla.edu

Sébastien Martin
Kellogg School of Management, Northwestern University, Evanston, IL 60208, sebastien.martin@kellogg.northwestern.edu

When working with generative artificial intelligence (AI), users may see productivity gains, but the AI-generated content may not match their preferences exactly. To study this effect, we introduce a Bayesian framework in which heterogeneous users choose how much information to share with the AI, facing a trade-off between output fidelity and communication cost. We show that the interplay between these individual-level decisions and AI training may lead to societal challenges. Outputs may become more homogenized, especially when the AI is trained on AI-generated content. And any AI bias may become societal bias. A solution to the homogenization and bias issues is to improve human-AI interactions, enabling personalized outputs without sacrificing productivity.

1. Introduction

Generative artificial intelligence (AI) systems, particularly large language models (LLMs), have improved at a rapid pace. For example, ChatGPT recently showcased its advanced capacity to perform complex tasks and human-like behaviors (OpenAI 2023), reaching 100 million users within two months of its 2022 launch (Hu 2023). This progress is not limited to text generation, as demonstrated by other recent generative AI systems such as Midjourney (Midjourney 2023) (a text-to-image generative AI) and GitHub Copilot (Github 2023) (an AI pair programmer that can autocomplete code). Eloundou et al. (2023) estimated that about 80% of the U.S. workforce could be affected by the introduction of LLMs, and 19% of the workers may have at least 50% of their tasks impacted. In particular, AI can make users more productive by generating complex content in seconds, while users can simply communicate their preferences. For example, Noy and Zhang (2023) highlighted that ChatGPT can substantially improve productivity in writing tasks, and GitHub claims that Copilot increases developer productivity by up to 55% (Kalliamvakou 2023).

However, content generated with the help of AI is not exactly the same as content generated without AI. The boost in productivity may come at the expense of users' idiosyncrasies, such as personal style and tastes, preferences we would naturally express without AI. To let users express their preferences, many AI systems let users edit their prompt (e.g., Midjourney) or allow more natural interactions (e.g., ChatGPT), and users can always review and edit the AI-generated output themselves (Vaithilingam et al. 2022). However, aligning a user's intentions with an AI's output can take time and may not always be worth it if the AI's first or default output "does the job." Consider a simple example where we use Copilot to code a Python function that calculates the sum of numbers in a nested list. Figure 1 shows that Copilot's default output (the first to the left) was correct and functional. However, it did not correspond to our own way of writing the same function given enough time (at the bottom of the figure). To push Copilot to better match our style, we could provide more information by articulating a more detailed prompt. However, the figure shows this may require many steps, which goes against the goal of being more productive. Similarly, Lingard (2023) described guiding ChatGPT through incremental prompting. In essence, users' time and effort to convey information about their desired outcome to an AI can enhance the output's alignment with their preferences, albeit at the expense of an increased communication cost. In short, users face a trade-off between AI *output fidelity* and *communication cost*.[1]

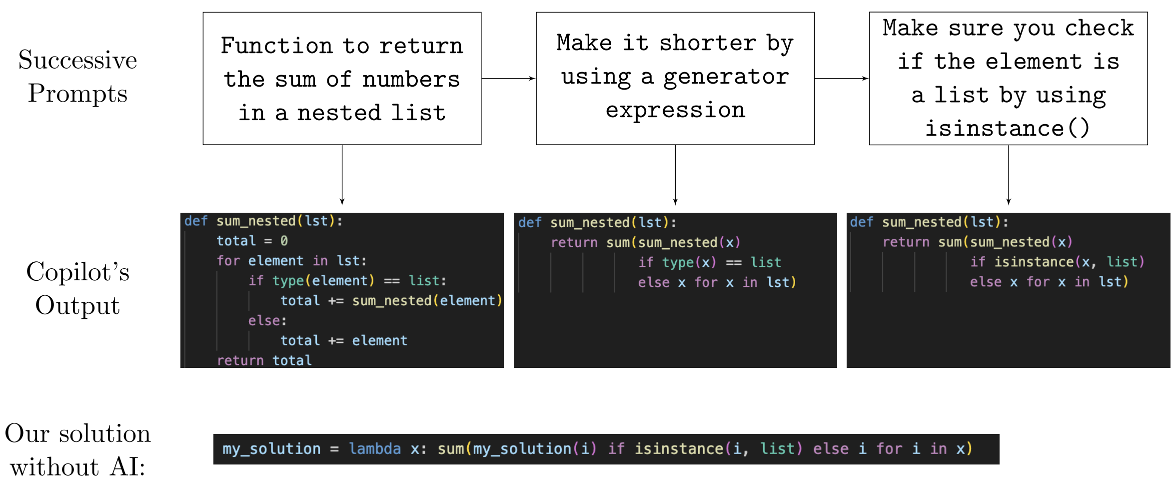

Figure 1 **The incremental information provided to align GitHub Copilot's Python code output with our preference. While the initial output in Attempt 1 is functional, it significantly differs from our solution without AI. Bridging the gap requires several iterations.**

Different users may respond to this trade-off differently, but those who value productivity more than fidelity will rely on AI more and willingly let go of their own preferences. We are interested in the potential societal consequences of these choices. First, working with AI may be more beneficial for some users than others: in the Copilot example, users who prefer the default output would not even need to communicate with the AI to have high fidelity to their preferences. Second, as users

[1] In fact, the importance and the associated costs of communicating with AIs have given rise to a new profession called prompt engineering (Mok 2023), and spurred the creation of novel marketplaces like PromptBase (URL: https://promptbase.com/).

do not share complete preferences with the AI and let it "choose" for them, the produced content may be, on average, *homogenized* towards the AI's default choices. For example, ChatGPT has been trained with reinforcement learning from human feedback (RLHF) (Kinsella 2023) to have a specific tone and language. If students use ChatGPT's help for their homework, their writing style may be influenced by ChatGPT's. More generally, AIs are built by a few but used by many, and there is a risk any AI bias could turn into a *societal bias*. The AI training process may involve censoring (e.g., the choice of the dataset) and human input (e.g., RLHF), which could intentionally or unintentionally lead to bias. For example, some studies discuss ChatGPT's inclination towards left-leaning political stances (Hartmann et al. 2023, Rozado 2023, Motoki et al. 2023). All in all, because of the benefits of increased productivity and the balance between output fidelity and communication costs, users could willingly produce less diverse content that is vulnerable to potential AI biases.

We propose a Bayesian model to study the societal consequences of human-AI interactions. For a given task, rational users can exchange information with the AI to align its output with their heterogeneous preferences. The AI has a knowledge of the distribution of preferences in the population and uses a Bayesian update to create an output with maximal expected fidelity given the information shared by the user. Users choose the amount of information they share to maximize their utility, balancing the cost of communication with the fidelity of the output. In this setting, we aim to formalize and evaluate the societal risks of homogenization and AI bias and how they could be mitigated.

When solving for each user's optimal decision, we find that their use of AI depends on how "unique" they are. Users with more common preferences simply accept the default output, avoiding any communication costs at the expense of a small fidelity mismatch. In contrast, users with more unique preferences share information with the AI to reduce fidelity errors, albeit with higher communication costs. And the most unique users do not benefit from the AI and simply perform the task themselves. To establish the homogenization effect, we prove that any output resulting from human-AI interactions isless unique than what a user would have done without AI. This is confirmed at the population level, where the AI-generated output distribution has lower variance than the users' preference distribution. This phenomenon is exacerbated when AI-generated content is used to train the next generation of AI: we show numerically that the users' rational decisions and the AI's training process can mutually reinforce each other, leading to a homogenization "death spiral". We also study the effects of AI bias, identifying who benefits or loses when using an AI model that does not accurately reflect the population preference distribution. At the population level, the censoring type of bias (e.g., biasing against the more unique preferences) negatively impacts the population utility, especially users with uncommon preferences who rely on AI interactivity the most. Directional biases (e.g., a slightly left-leaning AI) are not as harmful in terms of utility but any directional bias will influence the users' chosen output, leading to a societal bias. On the positive side, the user interactions with the AI partially counter the effects of AI bias, highlighting the need to consider human decisions to fully understand the impact of generative AI.

We show that tasks that are either hard to do without AI (e.g., image generation) or for which speed is particularly important (e.g., grammar correction) are especially sensitive to the risks of homogenization and bias. However, our research demonstrates that creating models that facilitate human-AI interactions can significantly limit these risks and preserve the population preference diversity.

The way we model the human-AI interaction shares similarities with the frameworks of information design (Kamenica and Gentzkow 2011), costly persuasion (Gentzkow and Kamenica 2014), the theory of rational inattention (Sims 2003), as well as the interpretation of LLMs with Bayesian inference (Wei et al. 2021, Xie et al. 2022) (see Section EC.3 for an extensive literature review). This study is also related to the recent modeling studies about human-AI interaction in operations management (e.g., Ibrahim et al. (2021), Boyacı et al. (2023), de Véricourt and Gurkan (2023)). They focus on decision-making when human and AI options exist separately, unlike the interactive setting we consider. Furthermore, while empirical research has recently proliferated around generative AIs (e.g., Binz and Schulz (2023), Noy and Zhang (2023), Webb et al. (2023)), our paper, to the best of our knowledge, is the first modeling study that employs a human-centric perspective to understand the societal consequences of generative AIs.

2. Model Setup

We develop a Bayesian model to represent the process of working with generative AI on a given task. Users have preferences on how to complete the task, and the AI knows the population's distribution of these preferences (through its training). Each user can also interact with the AI to share information about her specific preferences. This interaction will help the AI produce an output closer to what a user would have done without AI, leading to a better output *fidelity*. However, sharing information requires effort, which entails a *communication cost*. Users must choose how much information they share to balance the benefits of fidelity and the cost of communication.

User preferences and Fidelity We use $\theta \in \mathbb{R}$ to denote a user's specific *preference*, and we assume that θ is normally distributed in the population: $\theta \sim N(\mu_\theta, \sigma_\theta^2)$. Here μ_θ represents the average population preference and σ_θ the diversity of the population's preferences. In practice, the users' preferences should be represented by a high-dimensional space, but we will interpret θ as a specific *feature* of a user's preferences, as illustrated in the following example.

Example 1 (News Article). A journalist would like to write an article about a piece of breaking news and wants to use ChatGPT to write faster. θ measures the political orientation of the article this journalist would have written without AI. If $\theta > \mu_\theta$, the journalist is more right-leaning than the average journalist. When using AI, the journalist may be able to write faster but may not meet her exact political orientation θ (low fidelity).

We will refer to a "user θ" to describe a user with preference θ, and to an "output θ_A" to refer to an AI's output matching some preference θ_A. We define the output's *fidelity loss* as $(\theta-\theta_A)^2$, and we interpret it as a loss in utility for a user θ receiving an output θ_A (users prefer an output matching their preference). For example, if $\theta_A = \theta - 1$, the AI of Example 1 outputs a more left-leaning article than the journalist's preferred political orientation, and the fidelity loss is 1.

AI Bayesian Inference We model the interaction between a user and AI as an exchange of information about θ. The AI has a prior belief of the population distribution of θ. This belief corresponds to a normal distribution $N(\mu_A, \sigma_A^2)$ with density $\pi_A(\cdot)$. To capture that the AI has been trained on a representative dataset, we assume that the AI's prior is exactly the population distribution, $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$ (this assumption is relaxed in Section 4 to study the effects of a biased AI). We model the exchange of information between a user θ and the AI as a Bayesian update with normal distributions: the user shares a noisy signal $q = \theta + \epsilon_q$ where $\epsilon_q \sim N(0, \sigma_q^2)$, and the AI refines its belief using Bayes' rule: $\theta | q \sim \pi_A(\cdot|q)$. It then returns the output with the maximum expected fidelity:

$$\theta_A \triangleq \arg\min_{\hat{\theta}} E[(\hat{\theta}-\theta)^2|q] = E[\theta|q] = \frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\cdot q + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}\cdot \mu_A. \tag{1}$$

Note that θ_A is a weighted average between q and the prior mean (Berger 1985), which is a random variable since q is noisy. Additionally, a lower value of σ_q corresponds to more information shared with the AI: if $\sigma_q = 0$, then the user θ shares her exact preference, and the AI returns $\theta_A = \theta$. In the limit $\sigma_q \to +\infty$, the signal is uninformative, and the AI outputs the mean of its prior, $\theta_A = \mu_A$.

The *fidelity error* of a user θ given σ_q is the expected output fidelity, denoted by:

$$e(\theta, \sigma_q) \triangleq E[(\theta_A - \theta)^2|\theta]. \tag{2}$$

We can decompose it into two terms,

$$e(\theta, \sigma_q) = Var(\theta_A|\theta) + [E(\theta_A|\theta) - \theta]^2.$$

For a user θ, the first term corresponds to the variability in the AI's output that stems from the information exchange, while the second term is the impact of the bias in the AI response, which will be a focus of the paper. In the context of Example 1, the bias is high if the AI-written article consistently leans more to the left than the journalist orientation.

Information and Communication Cost Given an exchange of information parametrized by σ_q, we measure the "communication cost" of the user to share this information with the AI. Following standard assumptions in the rational inattention (Sims 2003) and costly persuasion (Gentzkow and Kamenica 2014) literature, we assume the communication cost to be proportional to the expected reduction in the AI uncertainty of θ relative to the population distribution of θ given σ_q:

$$\gamma I(\sigma_q) \triangleq \gamma[H(\theta) - E[H(\theta|q)]] = \gamma \left[\ln(\sigma_\theta \sqrt{2\pi e}) - \ln\left(\sqrt{\frac{\sigma_\theta^2 \sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}} \sqrt{2\pi e} \right) \right] = -\frac{\gamma}{2} \ln\left(\frac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2} \right),$$

where $\gamma > 0$ is the marginal cost of communication, $I(\sigma_q)$ is the mutual information, and $H(\cdot)$ denotes the differential entropy. Intuitively, $I(\sigma_q)$ corresponds to the "amount of information" the user shares about her preference θ. Note that sharing the exact value of θ ($\sigma_q = 0$) requires an *infinite* amount of information $I = +\infty$. Conversely, providing an uninformative signal about θ ($\sigma_q \to +\infty$) requires no information $I = 0$. To interpret this situation, remember that our model assumes that the AI knows the task description and that the human-AI interaction is about sharing user preferences. Without preference information, the AI uses its knowledge of the preference distribution to return a "default" output for the task, $\mu_A = \mu_\theta$ (e.g., the first answer of ChatGPT) . In Example 1, ChatGPT would write an article that expresses an "average" political orientation that does not necessarily reflect the journalist's views. And, in the GitHub Copilot coding example of Figure 1, the initial Copilot's output is a default function that does not consider the user's specific preference. In both cases, the AI requires more information to deliver better fidelity.

User's decision Each user θ chooses the information I they share with the AI (parametrized by σ_q) to minimize their *utility loss* l given by the sum of the fidelity error and the communication cost

$$l(\theta, \sigma_q) \triangleq e(\theta, \sigma_q) + \gamma I(\sigma_q). \tag{3}$$

That is, a user θ chooses an optimal $\sigma_q^\star(\theta)$ that minimizes her utility loss,

$$\sigma_q^\star(\theta) \triangleq \arg\min_{\sigma_q \geq 0} l(\theta, \sigma_q). \tag{4}$$

Importantly, γ controls the tradeoff between fidelity error and communication cost, and we will refer to it as the *cost of human-AI interactions.* A task has a low γ if it is particularly easy to interact with the AI and share preferences (e.g., a chat interface like ChatGPT or voice-based interactions) and/or if users care a lot about fidelity. γ will be high if users care more about minimizing effort than matching their specific preferences. Because the task of Example 1 is about breaking news, the journalist may be in a hurry and have a high γ.

Choosing to work with AI If the cost of human-AI interaction is high and fidelity is important, a user might be better off not using the AI and doing the work herself. In this case, the output would have no fidelity error by definition. However, manual work takes time, which we model as a fixed utility cost $\Gamma > 0$ that depends on the task but is the same for everyone.

If Γ is smaller than the expected utility loss $l(\theta, \sigma_q^\star)$, then a user θ will not use the AI. We define the output $\theta^\star$ chosen by a user θ and the corresponding expected utility loss l^* as:

$$\theta^\star \triangleq \begin{cases} \theta_A|(\theta, \sigma_q^\star) & \text{if } l(\theta, \sigma_q^\star) \leq \Gamma \\ \theta & \text{otherwise} \end{cases}, \qquad l^* \triangleq \min(l(\theta, \sigma_q^\star), \Gamma). \tag{5}$$

3. Human-AI Interactions and Homogenization

A consequence of our model is that different users may interact with the AI differently, sharing varying amounts of information about their preferences or even choosing not to use the AI. We first describe these individual-level choices and then study their implied societal consequences and how to mitigate them.

3.1. Individual Level: Heterogeneous Use of AI

Analyzing the optimal decision of each user θ requires solving Problem (4), and the results are presented in Proposition 1. Users' choices depend on their *uniqueness*, the distance of their preference θ to the population mean μ_θ, $d(\theta) \triangleq |\theta - \mu_\theta|$. We note that the derivation of $\sigma_q^\star(\theta)$, presented in Section EC.1, is non-trivial as Equation (3) is neither concave nor convex.[2]

Proposition 1. *Optimal user strategies solving Equation* (5) *have the following characteristics:*

1. *More unique users have a higher utility loss: $l^\star$ increases*[3] *in $d(\theta)$.*
2. *More unique users interact more with the AI (if they choose to use it): $\gamma I^\star$ increases in $d(\theta)$.*
3. *Users use the AI if they are below a uniqueness threshold $\tau_a > 0$: $d(\theta) \leq \tau_a \iff l(\theta, \sigma_q) \leq \Gamma$.*
4. *Users that use AI are characterized by another uniqueness threshold τ_d such that:*
   (a) *If $d(\theta) \leq \tau_d$, users choose the default AI output ($I^\star = 0$) and their fidelity error $e(\theta, \sigma_q^\star)$ increases with their uniqueness $d(\theta)$.*
   (b) *If $d(\theta) > \tau_d$, users interact with AI ($I^\star > 0$) and their fidelity error* decreases *with their uniqueness.*

The main takeaway from Proposition 1 is that users with more "common" preferences have a utility advantage (Item 1) and choose to interact less with the AI (Item 2). The fundamental driver of this, crucial throughout the paper, is that more common users can have a small fidelity error with limited shared information. There are three types of users: users who use AI but do

[2] We present the proofs for all the statements in Section EC.2.

[3] All references to "increasing" or "decreasing" functions are meant in a weak sense (i.e., "non-decreasing").

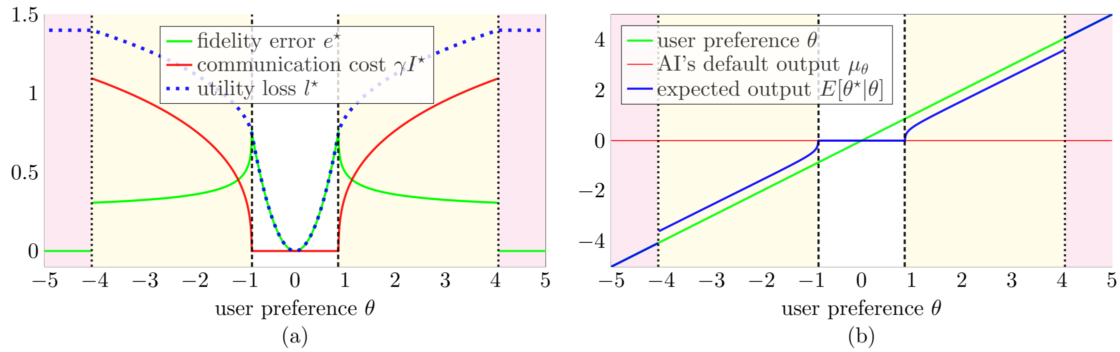

Figure 2 **The black dashed vertical lines are at $d(\theta) = \tau_d$, and the black dotted vertical lines are at $d(\theta) = \tau_a$. The white region indicates the users who choose the default output; the yellow region indicates those who send information to the AI; the red region indicates those not using AI. We use $\mu_\theta = 0, \sigma_\theta = 1, \gamma = 1, \Gamma = 1.4$.**

not share information, those who share information, and users who do not use AI. The most common users, with $d(\theta) \le \tau_d$ as described in Item 4a, accept the default output of the AI, zero communication cost but rapidly increasing fidelity error as these users become more unique, and the default output $\theta_A = \mu_\theta$ becomes worse (see the region in the center of Figure 2. Users with $\theta > \tau_d$ then choose to interact with the AI (Item 4b), which will reduce their fidelity error at the expense of communication cost (Item 2). Indeed, as illustrated in Figure 2 (a), while the fidelity error (green curve) dominates the utility loss of users with common preferences, more unique users prefer to pay an increasing communication cost (red curve) that dominates a decreasing fidelity error.[4] Interacting with the AI eventually reaches such high communication costs for the most unique users ($d(\theta) > \tau_a$) that the no-AI option becomes preferable (Item 3) as shown in the red area of Figure 2 (a).

Many users have a positive fidelity error, so the AI's output may not always align perfectly with a user's preference. The next proposition shows that this output is misaligned in a specific way: on average, a user's chosen output $\theta^\star$ tends to revert toward the population's mean preference.

Proposition 2. *The expected chosen output $E[\theta^\star|\theta]$ of any user θ is closer to the population's mean than their actual preference: $|\mathbb{E}[\theta^\star|\theta] - \mu_\theta| \le |\theta - \mu_\theta|$. Moreover, the inequality is strict for almost all users that use the AI: when $d(\theta) < \tau_a$ and $\theta \ne \mu_\theta$.*

We illustrate this result in Figure 2 (b). The most common users (with $d(\theta) \le \tau_d$) provide an uninformative signal to the AI ($I^\star = 0$) and accept the AI's default output $\theta_A = \mu_\theta$, which is a direct revert to the mean. As users become more unique, they interact with the AI ($I^\star > 0$), which mitigates the mean reversion in the AI's output. However, it doesn't completely vanish due to the high communication cost. The mean reversion disappears only for those very unique users who

[4] A more detailed description of effects at play can be found in Section EC.5.

chose to work themselves and not to use the AI. In Example 1, a journalist with $\theta > \mu_\theta$ would write a (slightly) more left-leaning article than her preference. As discussed in the next section, this can be an issue at the population level.

3.2. Societal Level: Homogenization

In a world without AI, the distribution of people's output would exactly match the distribution of their preference $\theta \sim N(\mu_\theta, \sigma_\theta^2)$. However, with AI, the output is $\theta^\star$, which does not have the same distribution, as we saw that users of AI tend to choose an output closer to the mean μ_θ. At the population level, this leads to *homogenization*, where the output distribution has a lower variance than the distribution of preferences.

THEOREM 1. *When everyone uses AI ($\Gamma \to +\infty$), the variance of the population output is lower than the variance of the population preferences, $Var(\theta^\star) < Var(\theta)$, and strictly decreases in the cost of human-AI interactions γ. In the general case ($\Gamma < +\infty$), $\lim_{\gamma \to 0} Var(\theta^\star) = Var(\theta)$ and $\lim_{\gamma \to +\infty} Var(\theta^\star) < Var(\theta)$.*

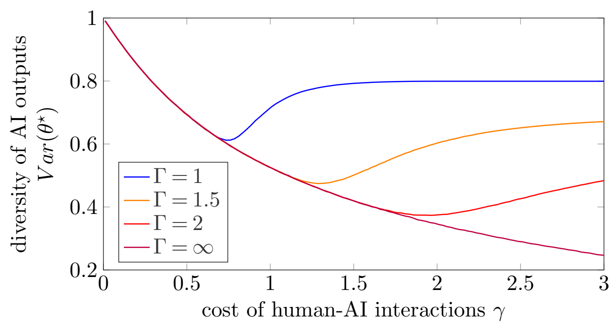

Figure 3 **We use $\mu_\theta = 0$, $\sigma_\theta = 1$.**

Theorem 1 formalizes the risk of homogenization and points to its solution. When everyone uses AI ($\Gamma \to +\infty$), reducing the cost of human-AI interactions γ encourages users to interact more with the AI and share their specific preferences more accurately, limiting homogenization and helping to preserve the population's diversity. When $\gamma \to 0$, there is no more homogenization as users can share their precise preference for free. The case $\Gamma < +\infty$ is more involved, as some users renounce the AI when the cost of human-AI interactions is high, partially improving the chosen output's diversity. We illustrate it in Figure 3 and present a more in-depth analysis in Section EC.6. An interesting special case is when $\Gamma < +\infty$ and $\gamma \to +\infty$. Only two types of users remain: those who complete the task themselves and those who accept the default AI output, leading to homogenization on average. In all cases, Theorem 1 underscores that enhancing the interactivity of AI tools (e.g., through better interfaces, multi-modal inputs, or real-time feedback mechanisms) to achieve a sufficiently low γ is an effective strategy to encourage users toward higher fidelity, reduce homogenization, and ultimately, preserve population preference diversity.

3.3. AI-generated content and the "Death Spiral" of Homogenization

While homogenization can easily be perceived as a negative societal outcome, we argue it may also have concerning long-term consequences. As more and more content becomes AI-generated, this content could be used to train the next generation of AI. Because of the homogenization issue, this would lead to an incorrect AI distribution of human preference (the AI's prior). The next AI generation would be even more likely to return homogenized outputs, resulting in a "death spiral" of homogenization, a dreadful outcome for human preference diversity. We study this phenomenon within our model, considering a *self-training loop* where the population's output distribution at time t becomes the AI prior at time $t+1$ (as illustrated in Figure 4).

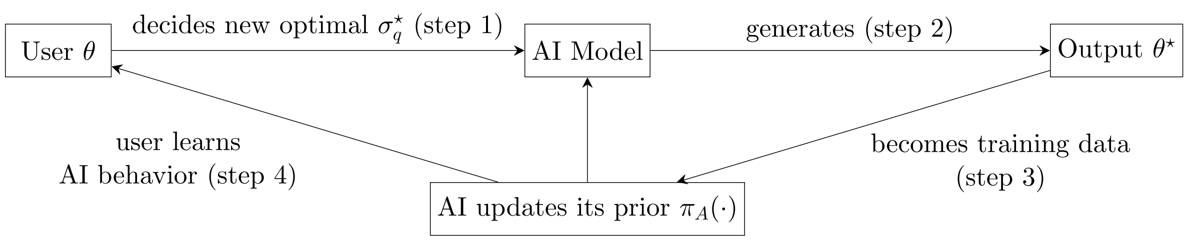

Figure 4 Steps in each iteration of the self-training loop.

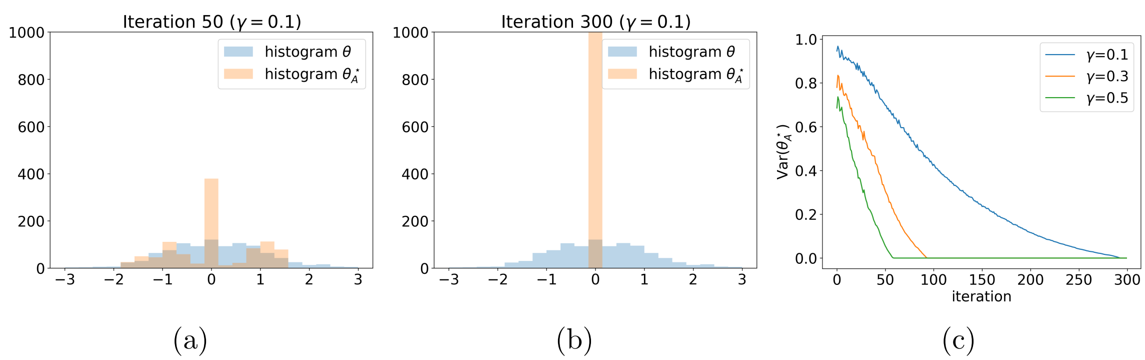

Figure 5 The iterative convergence of the distribution of $\theta_A^\star$. We use $\mu_\theta = 0$, $\sigma_\theta = 1$, $\Gamma = +\infty$. A full simulation description is provided in Section EC.4.

This model is not tractable, as our analysis does not apply when the AI has a non-normal prior, which happens after the first iteration of the self-training loop. However, we use simulations, as shown in Figure 5, to illustrate its consequences for homogenization. For simplicity, we choose a setting with $\Gamma = +\infty$, and everyone uses AI. Note that while the human distribution of preference θ (in blue) remains unchanged, the distribution of the outputs $\theta^\star = \theta_A$ (in orange) becomes more and more homogeneous. As the AI's prior becomes increasingly erroneous, the communication cost necessary to reduce the fidelity error becomes unmanageable, and more and more users start to accept the default output until we converge to a completely homogenized world.

These findings offer a different perspective than the technical explanation for the homogenization problem in Shumailov et al. (2023) (they call it model collapse), where the authors suggest the problem is caused by sampling and approximation errors. Our model also indicates that human and technical factors may mutually reinforce each other, potentially leading to an exacerbated homogenization problem. In our model, the homogenization loop is due to technical factors, the misspecified AI prior, and human behavior, who maximize their utility and are willing to let go of their specificity to limit the communication costs. We also offer a solution: creating models facilitating human-AI interactions (i.e., low γ) can significantly slow down the homogenization process (see Figure 5 (c)).

4. Human-AI Interactions and AI Bias

The homogenization phenomenon shows that the use of AI "influences" the user outputs, in the sense that $\theta^\star \neq \theta$ for many users θ. This is potentially concerning, as any choices made in the AI training, any bias it might have, would then influence the users' choice of output. Indeed, generative AIs are not necessarily trained to reflect the population's preferences exactly. For example, the AI's training data may be censored to avoid illegal or dangerous behavior (Thompson 2023). Moreover, the training of LLMs uses Reinforcement Learning from Human Feedback (Ziegler et al. 2020), in which a small group of humans "teach" the model what output is preferable. These training choices of a few can then influence the output of the entire population interacting with AI. We model this potential AI "bias" via an AI prior that does not exactly reflect the population's preference distribution (i.e., $\mu_A \neq \mu_\theta$ or $\sigma_A \neq \sigma_\theta$), leaving the true user preference distribution and the rest of the Bayesian inference unchanged. We refer to $\mu_A \neq \mu_\theta$ as a *directional bias* and to $\sigma_A < \sigma_\theta$ as a *censoring bias*. In Example 1, the AI may have a slight bias towards a political side (directional biased) or it may avoid extreme political views (censoring bias). We first discuss how the two types of AI bias affect users and the effectiveness of human-AI interactions. We then evaluate how much influence a biased AI can have on society and ways to mitigate this influence.

4.1. AI Bias and User Utility

A biased AI may be less useful for some users but may also help others, as summarized below.

PROPOSITION 3. *The utility loss $l^\star$ of a user θ changes when with a biased AI as follows:*

1. *the directional bias favors users the AI is biased towards: $l^\star$ increases with $|\mu_A - \theta|$;*
2. *the censoring bias benefits users with common preferences: $l^\star$ is increasing in σ_A when $\sigma_A \geq |\mu_A - \theta|$, and decreasing in σ_A when $\sigma_A < |\mu_A - \theta|$.*

Item 1 of the proposition states that directional bias is detrimental to users of the "opposite" direction. In Example 1, if the AI is slightly right-leaning, a left-leaning journalist may need more communication cost to obtain an article they will be satisfied with. However, a right-leaning journalist could be directly satisfied with the default output. The ideal case for user θ is $\mu_A = \theta$, as the default AI output would correspond to a perfect utility $l^\star = 0$. Item 2 states a similar result for the censoring bias. To clarify it, suppose $\mu_A = \mu_\theta$, and consider a user with "common" preferences less than a standard deviation away from the mean, i.e., $|\mu_\theta - \theta| < \sigma_\theta$. Then she would be better off if a slight censoring is used, with σ_A such that $|\mu_\theta - \theta| < \sigma_A < \sigma_\theta$. When reducing σ_A, the AI is more likely to return outputs closer to the mean, benefiting this user. However, this hurts users with more unique preferences, who will need more communication costs to maintain a reasonable fidelity or will stop using the AI. Therefore, both types of bias can increase some users' utility loss and decrease others'. The next results consider the aggregate-level consequences of bias and its effect on the population utility, defined as the expected utility loss $E[l^\star]$ taken across the users θ.

Proposition 4. *Directional and censoring bias have contrasting effects on the population utility:*

1. *A small directional bias has a limited negative effect on the population utility:* $\left.\frac{\partial E[l^\star]}{\partial \mu_A}\right|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = 0$ *and* $E[l^\star]$ *is minimized at* $\mu_A = \mu_\theta$.
2. *A small censoring bias can have a stronger negative impact: for example,* $\left.\frac{\partial E[l^\star]}{\partial \sigma_A}\right|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} < 0$ *when* $\gamma \geq 2\sigma_\theta^2$ *and* $\Gamma \to \infty$.

The proposition first shows that, while any directional bias hurts the population utility, a small directional bias has a negligible effect. Intuitively, if $\mu_A = \mu_\theta + \varepsilon$ for $\varepsilon > 0$ small, slightly less than half of the users (above $\mu_A + \epsilon/2$) benefit from the bias because they have a lower communication cost for the same fidelity, while the other half (below μ_θ) is hurt because of an increased communication cost for the same fidelity. These two populations balance each other, which limits the total loss of utility.

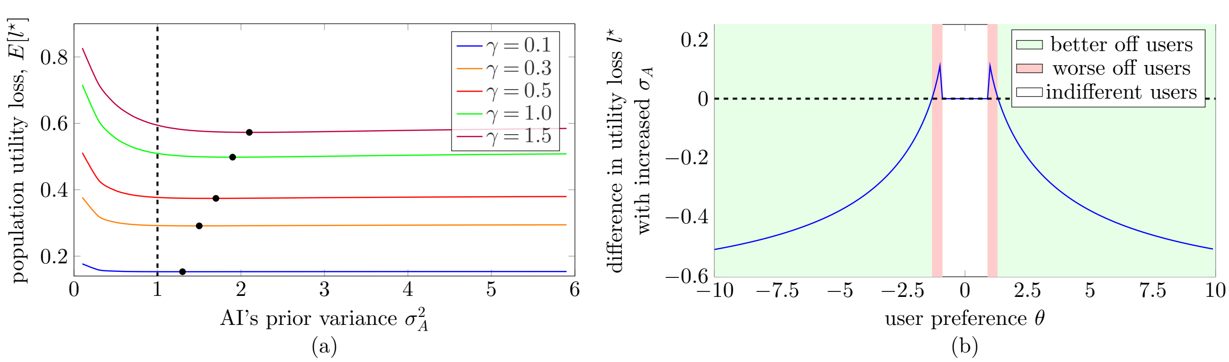

Figure 6 **(a) the black circles indicate the AI prior variance σ_A^2 that would minimize the population utility loss. (b) the utility loss $l^\star$ with $\sigma_A^2 = 2$ minus those with $\sigma_A^2 = 1$, when $\gamma = 1$. In both panels, we use $\mu_A = \mu_\theta = 0, \sigma_\theta^2 = 1, \Gamma = +\infty$.**

The case of censoring bias (Item 2 of Proposition 4) is maybe more surprising. Unlike the effect directional bias, setting $\sigma_A = \sigma_\theta$ (an unbiased prior) does not generally minimize the population utility loss $E[l^\star]$. Both the proposition and Figure 6 (a) show that when $\Gamma \to +\infty$, it is preferable to have $\sigma_A > \sigma_\theta$ (the opposite of censoring). Remember from Section 3.1 that there are two types of user behavior when $\Gamma \to +\infty$: using the default AI output or interacting with the AI, and the choice of σ_A only influences the utility of the interacting users. Therefore, the AI Bayesian update is more accurate when the choice of σ_A better represents the interacting users, who are the ones with more unique preferences (Proposition 1). This is why choosing $\sigma_A > \sigma_\theta$ improves the population utility. This effect is illustrated in Figure 6 (b): when increasing σ_A, common-preference users do not lose utility, but more unique users see a large improvement in utility loss. While this result may have implications for the design of interactive AI, it also warns against the potential negative effects of censoring bias. Decreasing σ_A is particularly hurtful to the most unique users, *who rely on human-AI interactions the most*. While censoring can be useful in preventing dangerous or illegal uses of AI, our results also highlight the importance of training AI on datasets that reflect a wide range of preferences.

4.2. AI Bias Becomes Societal Bias

Another interpretation of Item 1 of Proposition 4 is that a small directional bias $|\mu_A - \mu_\theta| > 0$ (referred to as *AI bias* in this section) may be hard to detect in practice, as it does not strongly affect the population's utility. However, it may still significantly influence the user output $\theta^\star$. For example, users who accept the default output ($I^\star = 0$) have $\theta^\star = \mu_A$, directly inheriting the AI bias. On the other hand, users may choose to share more information to correct this bias and maintain a high-fidelity output. To study which effect dominates, we analyze the consequences of the AI bias on the *societal bias*, defined as the bias of the output distribution: $|E[\theta^\star] - \mu_\theta|$.

THEOREM 2. *Given the AI bias $|\mu_A - \mu_\theta|$ and the societal bias $|E[\theta^\star] - \mu_\theta|$,*

1. *the societal bias is lower than the AI bias,*
2. *the societal bias is minimized when $\gamma \to 0$ or $\Gamma \to 0$: $|E[\theta^\star] - \mu_\theta| = 0$,*
3. *the societal bias is maximized when $\gamma \to +\infty$ and $\Gamma \to +\infty$: $|E[\theta^\star] - \mu_\theta| = |\mu_A - \mu_\theta|$,*
4. *if everyone uses AI ($\Gamma \to +\infty$), the societal bias increases with the cost of human-AI interactions γ.*

This theorem is illustrated in Figure 7 and shows an encouraging result: human-AI interactions can partially prevent AI bias from becoming societal bias. For example, a left-wing journalist in Example 1 may increase their interactions with the AI to correct the output if the AI is biased to the right. This is particularly true when either the cost of human-AI interactions, γ, or the cost of not using AI, Γ, is low. It is much easier for users to correct bias if they can easily interact or simply

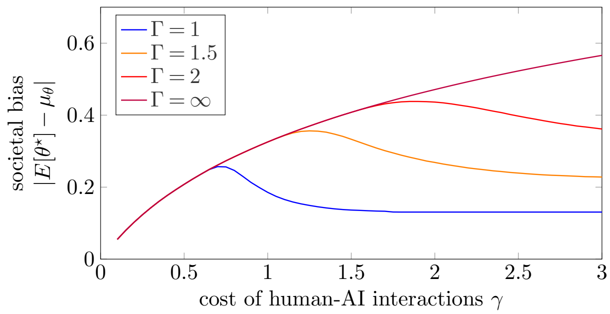

Figure 7 **We use** $\mu_\theta = 0$, $\mu_A = 1$, $\sigma_\theta = \sigma_A = 1$ **(e.g., the AI bias is** $|\mu_A - \mu_\theta| = 1$**).**

stop using AI. However, Theorem 2 also comes with a warning. For larger tasks that are painful to do without AI (high Γ), and if the human-AI interactions are not efficient (high γ), rational users will simply accept the AI bias, which will be fully converted into a societal bias. For example, generative AI systems that favor speed over interactivity (e.g., the AI writing assistant Grammarly) or tackle complex tasks with limited interactivity (e.g., the image generator Midjourney) may fall into this category. Any bias they introduce may have a stronger influence on societal output than systems designed for interactivity (e.g., ChatGPT).

5. Conclusions

The widespread introduction of generative AI enables significant productivity gains. However, we show that the power of these tools may lead users to accept homogenized or biased outputs and abandon their idiosyncrasies, even when given the possibility to communicate their preferences. At the societal level, this can lead to homogenization (reinforced by training loop effects) and the potential influence of AI training choices on the societal output. These risks are particularly strong for labor-intensive tasks (e.g., image/sound generation) or with AI tools that favor speed over preference-sharing (e.g., grammar correction). Nonetheless, we also show that enabling easier human-AI communication and training the AI on diverse data can significantly limit these negative effects, allowing the best of both worlds: high productivity and human diversity.

The topic studied in this work combines technical and behavioral complexity, as we need to capture how the AI tool works and how users interact with it. Our Bayesian framework is a simplified representation of this interaction that still enables nontrivial insights, but there are effects we do not capture. For example, it is a simplification to assume that a one-dimensional normal distribution can represent the vast space of human preferences and outputs and that the complexity of human-AI communication can be represented as a simple normal signal and Bayesian inference. We also assume all users have the same no-AI utility loss Γ, and the same human-AI interaction cost γ for a given task. Nonetheless, we believe our framework is versatile enough to study deeper variants and is a first step towards understanding the societal consequences of human-AI interactions.

Recent empirical studies examine the multifaceted implications of generative AIs across various domains, such as education (Baidoo-Anu and Owusu Ansah 2023), labor markets (Eloundou et al. 2023), and marketing (Brand et al. 2023). Understanding the general effects of user behaviors while interfacing with an AI remains an open question that is difficult to study empirically. We hope our analytical approach highlights the importance of adopting a human-centric perspective rather than solely focusing on AI technology. Indeed, while AIs could surpass human abilities in various aspects (Binz and Schulz 2023, Webb et al. 2023), their impact may largely depend on how we employ them. The interaction with AIs could offer a novel medium for production and creation but also introduce an extra risk: AIs may filter and even replace our original preferences, styles, and tastes, thereby leading to a world dictated by the AI creators' perspective — a potentially homogenized and biased world. Improving human-AI interactions and encouraging users to authentically voice their unique views is crucial to avoid these societal pitfalls.

References

Baidoo-Anu D, Owusu Ansah L (2023) Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484* .

Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer New York).

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120(6).

Boyacı T, Canyakmaz C, de Véricourt F (2023) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* .

Brand J, Israeli A, Ngwe D (2023) Using GPT for market research. *Available at SSRN 4395751* .

Chaney AJB, Stewart BM, Engelhardt BE (2018) How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *Proceedings of the 12th ACM Conference on Recommender Systems*, 224–232, RecSys '18.

de Véricourt F, Gurkan H (2023) Is your machine better than you? you may never know. *Management Science* .

Eloundou T, Manning S, Mishkin P, Rock D (2023) GPTs are GPTs: An early look at the labor market impact potential of large language models. *Available at arXiv:2303.10130* .

Gentzkow M, Kamenica E (2014) Costly persuasion. *American Economic Review* 104(5):457–62.

Github (2023) Github copilot · your AI pair programmer. URL `https://github.com/features/copilot`, Last accessed: 2023-08-13.

Hartmann J, Schwenzow J, Witte M (2023) The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *Available at arXiv:2301.01768* .

Hu K (2023) ChatGPT sets record for fastest-growing user base. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, Last accessed: 2023-09-18.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Kalliamvakou E (2023) Research: quantifying github copilot's impact on developer productivity and happiness. URL https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/, Last accessed: 2023-09-15.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *American Economic Review* 101(6):2590–2615.

Kinsella B (2023) OpenAI to offer ChatGPT customization and shares bias guidelines. URL https://synthedia.substack.com/p/openai-to-offer-chatgpt-customization, Last accessed: 2023-09-14.

Lingard L (2023) Writing with chatgpt: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education* 12(1):261–270.

Midjourney (2023) Midjourney. URL www.midjourney.com, Last accessed: 2023-08-14.

Mok A (2023) 'Prompt engineering' is one of the hottest jobs in generative AI. here's how it works. URL https://www.businessinsider.com/prompt-engineering-ai-chatgpt-jobs-explained-2023-3, Last accessed: 2023-09-18.

Motoki F, Neto VP, Rodrigues V (2023) More human than human: measuring chatgpt political bias. *Public Choice* .

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

OpenAI (2023) Introducing ChatGPT. URL https://openai.com/blog/chatgpt, Last accessed: 2023-09-18.

Rozado D (2023) The political biases of ChatGPT. *Social Sciences* 12(3).

Shumailov I, Shumaylov Z, Zhao Y, Gal Y, Papernot N, Anderson R (2023) The curse of recursion: Training on generated data makes models forget. *Available at arXiv:2305.17493* .

Sims CA (2003) Implications of rational inattention. *Journal of Monetary Economics* 50(3):665–690.

Thompson SA (2023) Uncensored chatbots provoke a fracas over free speech. URL https://www.nytimes.com/2023/07/02/technology/ai-chatbots-misinformation-free-speech.html, Last accessed: 2023-09-14.

Vaithilingam P, Zhang T, Glassman EL (2022) Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22 (Association for Computing Machinery).

Webb T, Holyoak KJ, Lu H (2023) Emergent analogical reasoning in large language models. *Nature Human Behaviour* .

Wei C, Xie SM, Ma T (2021) Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, eds., *Advances in Neural Information Processing Systems*.

Xie SM, Raghunathan A, Liang P, Ma T (2022) An explanation of in-context learning as implicit bayesian inference. *International Conference on Learning Representations*.

Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, Irving G (2020) Fine-tuning language models from human preferences. *Available at arXiv:1909.08593* .

Appendix

EC.1. Characterization of Optimal Decision

In this section, we characterize the user's optimal effort decision. We first show the closed form of the fidelity error $e(\theta, \sigma_q)$ and illustrate how the user's decision may impact the error. Then, the optimal solution to Problem (4) is derived. As in Section 3, we assume $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$.

Let's first explore how the fidelity error $e(\theta, \sigma_q)$ varies with respect to σ_q.

PROPOSITION EC.1. *For any θ, σ_q, the fidelity error is*

$$e(\theta, \sigma_q) = \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2} \tag{EC.1}$$

Furthermore,

- *$e(\theta, \sigma_q)$ increase in $d(\theta)$.*
- $\lim_{\sigma_q^2 \to 0} e(\theta, \sigma_q) = 0$, $\lim_{\sigma_q^2 \to \infty} e(\theta, \sigma_q) = d(\theta)^2$
- *If $d(\theta) \geq \sigma_\theta/\sqrt{2}$, $e(\theta, \sigma_q)$ is monotonically increasing in σ_q; if $d(\theta) < \sigma_\theta/\sqrt{2}$, there exists a threshold $t > 0$ such that $e(\theta, \sigma_q)$ increases in $1/\sigma_q \in (0, t)$ and decreases in $1/\sigma_q \in (t, \infty)$.*

Proposition EC.1 reveals that for any given σ_q, the uniqueness of a user's preference results in a larger error. If the user makes no effort and simply accepts the default output, the fidelity error increases the uniqueness metric, $d(\theta)$. Conversely, if the user offers substantial information, the fidelity error approaches zero. More intriguingly, the third part of Proposition EC.1 suggests that the fidelity error is monotonically decreasing in $1/\sigma_q$ only when the uniqueness $d(\theta)$ is sufficiently large (as demonstrated in the left panel of Figure EC.1). In other words, if a user's preference markedly deviates from the average, any increase in effort level to articulate their preference can yield a reduction in the AI's fidelity error. However, when a user's preference aligns closely with the majority (i.e., $2(\mu_\theta - \theta)^2 < \sigma_\theta^2$), there exists a threshold, t such that if the user is reluctant to exert sufficient effort such that $1/\sigma_q > t$, providing more information may backfire, causing the user to be worse off than if they provided less information.

To further investigate the cause of the non-monotonic fidelity error, as introduced in Section 2, we can decompose the fidelity error into a bias and a variance term,

$$e(\theta, \sigma_q) = Var(\theta_A|\theta) + [E(\theta_A|\theta) - \theta]^2,$$

We call $Var(\theta_A|\theta)$ the *fidelity uncertainty error* denoted by $e_u(\theta, \sigma_q)$, and $[E(\theta_A|\theta) - \theta]^2$ the *fidelity bias error* denoted by $e_b(\theta, \sigma_q)$. This decomposition is depicted in the right panel of Figure EC.1, highlighting that the non-monotonic fidelity error is primarily driven by the variance component.

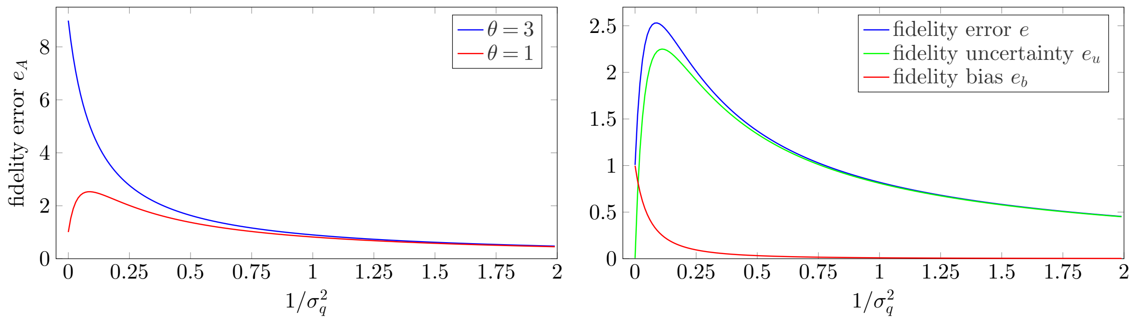

Figure EC.1 **Left panel: the fidelity error with respect to $1/\sigma_q^2$. Right panel: the decomposition of fidelity error for $\theta = 1$. In both panels, We use $\mu_\theta = 0$ and $\sigma_\theta^2 = 9$.**

Intuitively, when a user knows that the AI's default output μ_θ is closely aligned with their preference without the need for further information, any vague information could introduce ambiguity and cause the AI to deviate from the user's true preference. For example, in Example 1, users with a neutral opinion may find it advantageous to accept the AI's default output (suppose $\mu_\theta = 0$). If they were to loosely explain their reasoning without detailing specifics, they risk introducing noisy information and receiving a less desirable result. Hence, if you're not inclined to invest enough effort in providing precise information and you're aware your preference aligns closely with the AI's default output, it may be beneficial to exert less effort or allow the AI to make decisions on your behalf. In other words, offering nothing may be preferable to providing ambiguous information.

By Proposition EC.1, we can solve Problem (4) and derive the following Lemma EC.1.

Lemma EC.1. *The optimal solution to Problem* (4) *is*

$$\sigma_q^\star = \begin{cases} \sqrt{\dfrac{w^\star \sigma_\theta^2}{1 - w^\star}} & d(\theta) \geq \tau_d \\ \infty & \text{otherwise} \end{cases} \tag{EC.2}$$

where $w^\star = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\theta - \mu_\theta)^2 - \sigma_\theta^2)}}{4((\theta - \mu_\theta)^2 - \sigma_\theta^2)}$, *and* $\tau_d > 0$ *is a threshold that strictly increases in* γ *and is not less than* $\sqrt{\max\{0, \sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\gamma}\}}$. *In particular,* $\tau_d = \dfrac{1}{2}\sigma_\theta^2 + \dfrac{1}{4}\gamma$ *when* $\gamma > \sigma_\theta^2$.

It is not trivial to solve Problem (4), since the objective function is neither concave nor convex when θ is small. This non-convexity emerges from the non-monotonicity of the fidelity error, as outlined in Proposition EC.1. Lemma EC.1 implies that the users with common preferences are best suited to send no information. As discussed previously, these users may find it advantageous to rely on the AI's default output instead of introducing ambiguity. Furthermore, it can be observed that the more unique a user's preference is, the more effort she should spend.

EC.1.1. Proof of the Results in Section EC.1.

Proof of Proposition EC.1. By the definition of $e(\theta,\sigma_q)$,

$$\begin{aligned} e(\theta,\sigma_q) = E_{q\sim f_Q}[(\theta_A(\sigma_q^2)-\theta)^2|\theta] &= E\left[\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_q^2}(\theta+\epsilon_q)+\frac{\sigma_q^2}{\sigma_\theta^2+\sigma_q^2}\cdot\mu_\theta-\theta\right)^2|\theta\right] \\ &= E\left[\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_q^2}\epsilon_q+\frac{\sigma_q^2}{\sigma_\theta^2+\sigma_q^2}(\mu_\theta-\theta)\right)^2|\theta\right] \\ &= \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_q^2}\right)^2 E[\epsilon_q^2]+\left(\frac{\sigma_q^2}{\sigma_\theta^2+\sigma_q^2}(\mu_\theta-\theta)\right)^2 \\ &= \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_q^2}\right)^2 \sigma_q^2+\left(\frac{\sigma_q^2}{\sigma_\theta^2+\sigma_q^2}(\mu_\theta-\theta)\right)^2 \\ &= \frac{\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^2} \end{aligned}$$

- $\lim_{\sigma_q^2\to 0} e(\theta,\sigma_q) = \lim_{\sigma_q^2\to 0}\frac{\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^2} = 0$
  $\lim_{\sigma_q^2\to\infty} e(\theta,\sigma_q) = \lim_{\sigma_q^2\to\infty}\frac{\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^2} = \lim_{\sigma_q^2\to\infty}\frac{\sigma_\theta^4+2\sigma_q^2(\mu_\theta-\theta)^2}{2(\sigma_\theta^2+\sigma_q^2)} = \lim_{\sigma_q^2\to\infty}\frac{2(\mu_\theta-\theta)^2}{2} = (\mu_\theta-\theta)^2$
- Take the derivative of $e(\theta,\sigma_q)$ with respect to σ_q^2:

$$\begin{aligned} \frac{\partial e(\theta,\sigma_q)}{\partial\sigma_q^2} &= \frac{\partial\frac{\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^2}}{\partial\sigma_q^2} \\ &= \frac{(\sigma_\theta^4+2\sigma_q^2(\mu_\theta-\theta)^2)(\sigma_\theta^2+\sigma_q^2)^2-2(\sigma_\theta^2+\sigma_q^2)\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^4} \\ &= \frac{(\sigma_\theta^4+2\sigma_q^2(\mu_\theta-\theta)^2)(\sigma_\theta^2+\sigma_q^2)-2\sigma_q^2(\sigma_\theta^4+\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^3} \\ &= \frac{\sigma_\theta^6+\sigma_\theta^4\sigma_q^2+2\sigma_q^2(\mu_\theta-\theta)^2\sigma_\theta^2+2\sigma_q^4(\mu_\theta-\theta)^2-2\sigma_q^2\sigma_\theta^4-2\sigma_q^4(\mu_\theta-\theta)^2}{(\sigma_\theta^2+\sigma_q^2)^3} \\ &= \frac{\sigma_\theta^6-\sigma_\theta^4\sigma_q^2+2\sigma_q^2(\mu_\theta-\theta)^2\sigma_\theta^2}{(\sigma_\theta^2+\sigma_q^2)^3} \\ &= \frac{\sigma_\theta^2(\sigma_\theta^4-\sigma_\theta^2\sigma_q^2+2\sigma_q^2(\mu_\theta-\theta)^2)}{(\sigma_\theta^2+\sigma_q^2)^3} \\ &= \frac{\sigma_\theta^2(\sigma_\theta^4+\sigma_q^2(2(\mu_\theta-\theta)^2-\sigma_\theta^2))}{(\sigma_\theta^2+\sigma_q^2)^3} \end{aligned}$$

  which is positive for all $\sigma_q^2\geq 0$ if and only if $2(\mu_\theta-\theta)^2\geq\sigma_\theta^2$. And when $2(\mu_\theta-\theta)^2<\sigma_\theta^2$, $e(\theta,\sigma_q)$ increases for $\sigma_q\in\left(0,\sqrt{\frac{\sigma_\theta^4}{\sigma_\theta^2-2(\mu_\theta-\theta)^2}}\right)$, and decreases for $\sigma_q\in\left(\sqrt{\frac{\sigma_\theta^4}{\sigma_\theta^2-2(\mu_\theta-\theta)^2}},\infty\right)$, so $t=\sqrt{\frac{\sigma_\theta^2-2(\mu_\theta-\theta)^2}{\sigma_\theta^4}}$.

□

Proof of Lemma EC.1. Let $w \triangleq \dfrac{\sigma_q^2}{\sigma_q^2+\sigma_\theta^2}$, and by Proposition EC.1, we can rewrite Problem (4) as:

$$w^\star(\theta) \triangleq \arg\min_{w\in[0,1]} w(1-w)\sigma_\theta^2 + w^2(\mu_\theta-\theta)^2 - \frac{\gamma}{2}\ln w \tag{EC.3}$$

Let $f(w) \triangleq w(1-w)\sigma_\theta^2 + w^2(\mu_\theta-\theta)^2 - \dfrac{\gamma}{2}\ln w$. On the boundary, we have $f(0)=\infty$ and $f(1)=(\mu_\theta-\theta)^2$.

Take the first-order condition:

$$\frac{\partial f}{\partial w} = 2((\mu_\theta-\theta)^2-\sigma_\theta^2)w + \sigma_\theta^2 - \frac{\gamma}{2w} = 0$$

we can get $w^\star(\theta) = \dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ or $\dfrac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$.

Moreover, we have to make sure $w^\star(\theta)\in[0,1]$ and $f(w^\star(\theta)) \le (\mu_\theta-\theta)^2$ because Problem (4) is non-convex.

1. $(\mu_\theta-\theta)^2 \ge \sigma_\theta^2$

   Because $w^\star(\theta)\ge 0$, it is only possible to have $w^\star(\theta) = \dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$. Also,

$$\begin{aligned}
& w \le 1 \\
\iff & -\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)} \le 4((\mu_\theta-\theta)^2-\sigma_\theta^2) \\
\iff & \sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) \le (4(\mu_\theta-\theta)^2-3\sigma_\theta^2)^2 \\
\iff & 4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) \le (4(\mu_\theta-\theta)^2-4\sigma_\theta^2)(4(\mu_\theta-\theta)^2-2\sigma_\theta^2) \\
\iff & \gamma \le 2(2(\mu_\theta-\theta)^2-\sigma_\theta^2) \\
\iff & (\mu_\theta-\theta)^2 \ge \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma
\end{aligned}$$

   Additionally, when $(\mu_\theta-\theta)^2\ge\sigma_\theta^2$, $\dfrac{\partial f}{\partial w}$ is negative for $w<w^\star(\theta)$ and positive for $w>w^\star(\theta)$, so $f(w^\star(\theta)) \le f(1) = (\mu_\theta-\theta)^2$.

   Therefore, when $(\mu_\theta-\theta)^2\ge\sigma_\theta^2$, $w^\star(\theta) = \dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is optimal if $(\mu_\theta-\theta)^2 \ge \dfrac{1}{2}\sigma_\theta^2+\dfrac{1}{4}\gamma$; otherwise, $w^\star(\theta)=1$ is optimal.

2. $(\mu_\theta-\theta)^2 < \sigma_\theta^2$

   To make sure the first condition is satisfied (otherwise, $w^\star(\theta)=1$ is optimal), we need $\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)\ge 0$. That is, $(\mu_\theta-\theta)^2 \ge \sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\gamma}$.

   In addition, we can see that $\dfrac{\partial f}{\partial w}$ is negative for $w < \dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ or $w > \dfrac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$. And $\dfrac{\partial f}{\partial w}$ is positive for $w \in$

$\left(\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}, \dfrac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\right)$. Thus, the local minimum is at $w=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$, and the local maximum is at $w=\dfrac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$.

For $w\le 1$,

$$\begin{aligned}\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}&\le 1\\ \Longleftrightarrow \sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}&\ge 4(\mu_\theta-\theta)^2-3\sigma_\theta^2\end{aligned}$$

The above inequality always holds if $(\mu_\theta-\theta)^2\le\frac{3}{4}\sigma_\theta^2$, otherwise

$$\begin{aligned}&\Longleftrightarrow \sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)\ge(4(\mu_\theta-\theta)^2-3\sigma_\theta^2)^2\\ &\Longleftrightarrow \gamma\le 2(2(\mu_\theta-\theta)^2-\sigma_\theta^2)\\ &\Longleftrightarrow (\mu_\theta-\theta)^2\ge\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma\end{aligned}$$

Thus, if $\gamma\le\sigma_\theta^2$, we only need $(\mu_\theta-\theta)^2\ge\sigma_\theta^2-\dfrac{\sigma_\theta^4}{4\gamma}$ such that $w^\star(\theta)=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\in[0,1]$. Otherwise, $w^\star(\theta)=1$ is optimal.

If $\gamma>\sigma_\theta^2$, we need $(\mu_\theta-\theta)^2\ge\max\{\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma,\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma}\}$. However, notice that $\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma\ge\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma}$ because $\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma-(\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma})=(\gamma-\sigma_\theta^2)^2/(4\gamma)\ge 0$. So we need $(\mu_\theta-\theta)^2\ge\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma$ such that $w^\star(\theta)=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\in[0,1]$. Otherwise, $w^\star(\theta)=1$ is optimal.

Now we need to show the conditions when $w^\star(\theta)=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is optimal.

First, notice that $w^\star(\theta)=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is the global minimum if $\dfrac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\ge 1$. And,

$$\begin{aligned}&\frac{-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\ge 1\\ \Longleftrightarrow\ &-\sigma_\theta^2-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}\le 4((\mu_\theta-\theta)^2-\sigma_\theta^2)\quad\text{since } (\mu_\theta-\theta)^2<\sigma_\theta^2\\ \Longleftrightarrow\ &-\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}\le 4((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2\end{aligned}$$

The above inequality always holds if $(\mu_\theta-\theta)^2 \geq \frac{3}{4}\sigma_\theta^2$, otherwise

$$
\begin{aligned}
&\Longleftrightarrow \sigma_\theta^4 + 4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) \geq (4((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2)^2 \\
&\Longleftrightarrow \sigma_\theta^4 + 4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) \geq 16((\mu_\theta-\theta)^2-\sigma_\theta^2)^2 + 8\sigma_\theta^2((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^4 \\
&\Longleftrightarrow \gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) \geq 4((\mu_\theta-\theta)^2-\sigma_\theta^2)^2 + 2\sigma_\theta^2((\mu_\theta-\theta)^2-\sigma_\theta^2) \\
&\Longleftrightarrow \gamma \leq 4((\mu_\theta-\theta)^2-\sigma_\theta^2)+2\sigma_\theta^2 \quad \text{since } (\mu_\theta-\theta)^2<\sigma_\theta^2 \\
&\Longleftrightarrow \gamma \leq 4(\mu_\theta-\theta)^2-2\sigma_\theta^2 \\
&\Longleftrightarrow (\mu_\theta-\theta)^2 \geq \frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma
\end{aligned}
$$

Thus, if $\gamma>\sigma_\theta^2$ and $\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma \leq (\mu_\theta-\theta)^2 \leq \sigma_\theta^2$, $w^\star(\theta)=\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is optimal.

Let's see what we have shown now. We have shown that if $(\mu_\theta-\theta)^2 \geq \frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma$, $w^\star(\theta)=\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is feasible and optimal; if $(\mu_\theta-\theta)^2<\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma$ and $\gamma>\sigma_\theta^2$, $w^\star(\theta)=1$ is optimal; if $\gamma\leq\sigma_\theta^2$ and $(\mu_\theta-\theta)^2\leq\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma}$, $w^\star(\theta)=1$ is optimal. In addition, we have shown that if $\gamma\leq\sigma_\theta^2$ and $(\mu_\theta-\theta)^2\in(\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma},\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma)$, $w=\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\in[0,1]$ is feasible, but we need to show whether it is optimal since $w=1$ is another local minimum.

We want to show that if $\gamma\leq\sigma_\theta^2$, there exists a threshold $t\geq\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma}$ such that when $(\mu_\theta-\theta)^2>t$, $w^\star(\theta)=\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$ is optimal; otherwise, $w^\star(\theta)=1$ is optimal. Since $l(1)=(\mu_\theta-\theta)^2$, the former statement is equivalent to showing that if $\gamma\leq\sigma_\theta^2$ and $(\mu_\theta-\theta)^2\in(\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma},\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma)$,

$$
g((\mu_\theta-\theta)^2) \triangleq (\mu_\theta-\theta)^2 - l\left(\frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}\right)
$$

has at most one zero point. In particular, we want to show that $g((\mu_\theta-\theta)^2)$ is monotonically increasing for any $(\mu_\theta-\theta)^2\in\left(\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma},\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma\right)$.

By Lemma EC.5,

$$
\begin{aligned}
\frac{\partial g}{\partial(\mu_\theta-\theta)^2} &= 1-\frac{\sigma_\theta^2\sqrt{\Delta}-\sigma_\theta^4-2\gamma((\mu_\theta-\theta)^2-\sigma_\theta)}{8((\mu_\theta-\theta)^2-\sigma_\theta)^2} \\
&= \frac{8((\mu_\theta-\theta)^2-\sigma_\theta)^2-\sigma_\theta^2\sqrt{\Delta}+\sigma_\theta^4+2\gamma((\mu_\theta-\theta)^2-\sigma_\theta)}{8((\mu_\theta-\theta)^2-\sigma_\theta)^2}
\end{aligned}
$$

Let $h(\gamma) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 - \sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta)$ represents the numerator of $\frac{\partial g}{\partial(\mu_\theta - \theta)^2}$. We have $h(0) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 \geq 0$ and

$$h(\gamma = \sigma_\theta^2) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 + \sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta)$$

$$\text{Since } \gamma \leq \sigma_\theta^2 \text{ and } (\mu_\theta - \theta)^2 \leq \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma \implies 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 \geq 2(\gamma - 2\sigma_\theta^2)((\mu_\theta - \theta)^2 - \sigma_\theta)$$

$$\begin{aligned} h(\gamma = \sigma_\theta^2) &\geq 2(\sigma_\theta^2 - 2\sigma_\theta^2)((\mu_\theta - \theta)^2 - \sigma_\theta) + \sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta) \\ &= \sigma_\theta^2\sqrt{\Delta} - \Delta \\ &= \sqrt{\Delta}(\sigma_\theta^2 - \sqrt{\Delta}) \\ &\geq 0 \quad \text{since } \gamma \leq \sigma_\theta^2 \text{ and } (\mu_\theta - \theta)^2 \leq \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma \implies (\mu_\theta - \theta)^2 \leq \sigma_\theta^2 \end{aligned}$$

In addition,

$$\begin{aligned} \frac{\partial h}{\partial \gamma} &= \frac{\sigma_\theta^2}{2\sqrt{\Delta}}4((\mu_\theta - \theta)^2 - \sigma_\theta) - 2((\mu_\theta - \theta)^2 - \sigma_\theta) \\ &= 2((\mu_\theta - \theta)^2 - \sigma_\theta)(\frac{\sigma_\theta^2}{\sqrt{\Delta}} - 1) \leq 0 \quad \text{since } (\mu_\theta - \theta)^2 \leq \sigma_\theta^2 \end{aligned}$$

Thus, $h(\gamma) \geq 0$ for any $\gamma \leq \sigma_\theta^2$, which implies that $\frac{\partial g}{\partial(\mu_\theta - \theta)^2} \geq 0$.

Therefore, $g((\mu_\theta - \theta)^2)$ is monotonically increasing for any $(\mu_\theta - \theta)^2 \in (\sigma_\theta^2 - \frac{\sigma_\theta^4}{4\gamma}, \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma)$, which implies that if $\gamma \leq \sigma_\theta^2$, there exists a threshold $\eta \geq \sigma_\theta^2 - \frac{\sigma_\theta^4}{4\gamma}$ such that when $(\mu_\theta - \theta)^2 > \eta$, $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ is optimal.

In summary, when $\gamma > \sigma_\theta^2$, then $\tau_d(\gamma) \triangleq \sqrt{\frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma}$ is a threshold such that $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ is optimal if and only if $|\mu_\theta - \theta| \geq \tau_d(\gamma)$; and when $\gamma \leq \sigma_\theta^2$, then $\tau_d(\gamma) \triangleq \sqrt{\eta}$ is a threshold such that $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ is optimal if and only if $|\mu_\theta - \theta| \geq \tau_d(\gamma)$. Additionally, it is clear that $\frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\gamma$ strictly increases in γ; and by Equation (EC.16), $\frac{\partial f(w^\star(\theta))}{\partial \gamma} = \frac{3\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} = \frac{3\gamma}{8\sqrt{\Delta}w^\star(\theta)} > 0$, so η strictly increases in γ. These imply that $\tau_d(\gamma)$ strictly increases in γ.

Hence, the optimal solution to Problem (4) is

$$\sigma_q^\star = \begin{cases} \sqrt{\frac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}} & |\mu_\theta - \theta| \geq \tau_d(\gamma) \\ \infty & \text{otherwise} \end{cases} \tag{EC.2}$$

where $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$, and $\tau_d(\gamma) > 0$ is a threshold that increases in γ and is not less than $\sqrt{\max\{0, \sigma_\theta^2 - \frac{\sigma_\theta^4}{4\gamma}\}}$. $\square$

EC.2. Proof of the Main Results

EC.2.1. Proof of the Results in Section 3.

EC.2.1.1. Auxiliary lemmas

LEMMA EC.2. *Let* $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$, *then*

1.
$$\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} = \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \tag{EC.4}$$

2.
$$\frac{\partial w^\star(\theta)}{\partial\gamma} = \frac{1}{2\sqrt{\Delta}} \tag{EC.5}$$

3.
$$\frac{\partial w^\star(\theta)}{\partial\sigma_\theta^2} = \frac{(\sigma_\theta^2 - \sqrt{\Delta})(\mu_\theta - \theta)^2 + 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \tag{EC.6}$$

where $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof of Lemma EC.2. Let $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Since $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} = \dfrac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.
$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} &= \frac{\frac{1}{2\sqrt{\Delta}}4\gamma \cdot 4((\mu_\theta - \theta)^2 - \sigma_\theta^2) - 4(-\sigma_\theta^2 + \sqrt{\Delta})}{16((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{2\gamma \cdot ((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2\sqrt{\Delta} - \Delta}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}
\end{aligned}$$

2.
$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial\gamma} &= \frac{\frac{1}{2\sqrt{\Delta}} \cdot 4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \\
&= \frac{1}{2\sqrt{\Delta}}
\end{aligned}$$

3.
$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial\sigma_\theta^2} &= \frac{\left(-1 + \frac{2\sigma_\theta^2 - 4\gamma}{2\sqrt{\Delta}}\right)4((\mu_\theta - \theta)^2 - \sigma_\theta^2) + 4(-\sigma_\theta^2 + \sqrt{\Delta})}{16((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(-\sqrt{\Delta} + \sigma_\theta^2 - 2\gamma)((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} + \Delta}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})((\mu_\theta - \theta)^2 - \sigma_\theta^2) - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(\sigma_\theta^2 - \sqrt{\Delta}) + 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})(\mu_\theta - \theta)^2 + 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}
\end{aligned}$$

□

Lemma EC.3. *Let* $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, *then we can rewrite* $e(\theta, \sigma_q)$ *as*

$$e(\theta, w) = w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 \tag{EC.7}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta-\theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta-\theta)^2} = \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2} \tag{EC.8}$$

2.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \gamma} = \frac{\sigma_\theta^2 + \sqrt{\Delta}}{4\sqrt{\Delta}} \tag{EC.9}$$

3.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \sigma_\theta^2} = \frac{[2(\mu_\theta-\theta)^2 - \sigma_\theta^2](-\sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta^2)^2} \tag{EC.10}$$

where $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)$.

Proof of Lemma EC.3. Let $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, then $\sigma_q^2 = \dfrac{w\sigma_\theta^2}{1-w}$. Substitute $\sigma_q^{2\star} = \dfrac{w\sigma_\theta^2}{1-w}$ into Equation (EC.1), then we have Equation (EC.7).

Let $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)$,

1.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta-\theta)^2}$$

$$= \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial(\mu_\theta-\theta)^2} + 2(\mu_\theta-\theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial(\mu_\theta-\theta)^2} + w^{\star 2}$$

Substitute Equation (EC.4) into the above equation

$$= \sigma_\theta^2 \cdot \frac{\sigma_\theta^2 + 2((\mu_\theta-\theta)^2 - \sigma_\theta^2) - \sqrt{\Delta}}{2((\mu_\theta-\theta)^2 - \sigma_\theta^2)} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2}$$
$$+ 2(\mu_\theta-\theta)^2 \cdot \frac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta-\theta)^2 - \sigma_\theta^2)} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2}$$
$$+ \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{8((\mu_\theta-\theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{\sigma_\theta^4 + 2\sigma_\theta^2((\mu_\theta-\theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} - (\mu_\theta-\theta)^2\sigma_\theta^2 + (\mu_\theta-\theta)^2\sqrt{\Delta}}{2((\mu_\theta-\theta)^2 - \sigma_\theta^2)} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2}$$
$$+ \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{8((\mu_\theta-\theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{\sigma_\theta^2 + \sqrt{\Delta}}{2} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2} + \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{8((\mu_\theta-\theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{\sigma_\theta^2(\Delta - \sigma_\theta^4) - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2)(\sigma_\theta^2 + \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2} + \frac{\sigma_\theta^4\sqrt{\Delta} - \sigma_\theta^2\Delta + 2\gamma\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{8\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta-\theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2 - \sigma_\theta)^2}$$

2.

$$
\begin{aligned}
&\frac{\partial e(\theta, w^\star(\theta))}{\partial \gamma} \\
&= \sigma_\theta^2(1-2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial \gamma} + 2(\mu_\theta-\theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial \gamma} \\
&\quad \text{Substitute Equation (EC.5) into the above equation} \\
&= \sigma_\theta^2 \cdot \frac{\sigma_\theta^2 + 2((\mu_\theta-\theta)^2-\sigma_\theta^2) - \sqrt{\Delta}}{2((\mu_\theta-\theta)^2-\sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}} \\
&+ 2(\mu_\theta-\theta)^2 \cdot \frac{-\sigma_\theta^2+\sqrt{\Delta}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}} \\
&= \frac{\sigma_\theta^4 + 2\sigma_\theta^2((\mu_\theta-\theta)^2-\sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} - (\mu_\theta-\theta)^2\sigma_\theta^2 + (\mu_\theta-\theta)^2\sqrt{\Delta}}{2((\mu_\theta-\theta)^2-\sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}} \\
&= \frac{\sigma_\theta^2+\sqrt{\Delta}}{4\sqrt{\Delta}}
\end{aligned}
$$

3.

$$
\begin{aligned}
&\frac{\partial e(\theta, w^\star(\theta))}{\partial \sigma_\theta^2} \\
&= w^\star(\theta)(1-w^\star(\theta)) + \sigma_\theta^2(1-2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2} + 2(\mu_\theta-\theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2} \\
&\quad \text{Substitute Equation (EC.6) into the above equation} \\
&= \frac{-\sigma_\theta^2+\sqrt{\Delta}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)} \cdot \frac{4((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-\sqrt{\Delta}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)} + \frac{\sigma_\theta^2+\sqrt{\Delta}}{2} \cdot \frac{(\sigma_\theta^2-\sqrt{\Delta})(\mu_\theta-\theta)^2 + 2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{4((\mu_\theta-\theta)^2-\sigma_\theta^2)(-\sigma_\theta^2+\sqrt{\Delta}) - (\sigma_\theta^2-\sqrt{\Delta})^2}{16((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&+ \frac{-4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)(\mu_\theta-\theta)^2 + 2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)(\sigma_\theta^2+\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{2((\mu_\theta-\theta)^2-\sigma_\theta^2)(-\sigma_\theta^2+\sqrt{\Delta})\sqrt{\Delta} - \sqrt{\Delta}(\sigma_\theta^4 + 2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&+ \frac{-4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)(\mu_\theta-\theta)^2 + 2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)(\sigma_\theta^2+\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{2((\mu_\theta-\theta)^2-\sigma_\theta^2)[(-\sigma_\theta^2+\sqrt{\Delta})\sqrt{\Delta} - \gamma\sqrt{\Delta} - 2\gamma(\mu_\theta-\theta)^2 + \gamma(\sigma_\theta^2+\sqrt{\Delta})] - \sqrt{\Delta}\sigma_\theta^2(\sigma_\theta^2-\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{2((\mu_\theta-\theta)^2-\sigma_\theta^2)\sqrt{\Delta}(-\sigma_\theta^2+\sqrt{\Delta}) - 2((\mu_\theta-\theta)^2-\sigma_\theta^2)\gamma[2(\mu_\theta-\theta)^2-\sigma_\theta^2] + \sqrt{\Delta}\sigma_\theta^2(-\sigma_\theta^2+\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{[2(\mu_\theta-\theta)^2-\sigma_\theta^2]\sqrt{\Delta}(-\sigma_\theta^2+\sqrt{\Delta}) - 2((\mu_\theta-\theta)^2-\sigma_\theta^2)\gamma[2(\mu_\theta-\theta)^2-\sigma_\theta^2]}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \\
&= \frac{[2(\mu_\theta-\theta)^2-\sigma_\theta^2](-\sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2}
\end{aligned}
$$

□

Lemma EC.4. *Let* $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, *then we can rewrite* $I(\sigma_q)$ *as*

$$I(w) = -\frac{\gamma}{2}\ln w \tag{EC.11}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{\gamma}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})} \tag{EC.12}$$

2.
$$\frac{\partial I(w^\star(\theta))}{\partial\gamma} = -\frac{\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} \tag{EC.13}$$

where $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof of Lemma EC.4. Let $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$,

1.
$$\begin{aligned}\frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} &= -\frac{\gamma}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2}\\ &\text{Substitute Equation (EC.4) into the above equation}\\ &= -\frac{\gamma}{2}\cdot\frac{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{-\sigma_\theta^2 + \sqrt{\Delta}}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}\\ &= -\frac{\gamma}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})}\end{aligned}$$

2.
$$\begin{aligned}\frac{\partial I(w^\star(\theta))}{\partial\gamma} &= -\frac{\gamma}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial\gamma} - \frac{\ln w^\star(\theta)}{2}\\ &\text{Substitute Equation (EC.5) into the above equation}\\ &= -\frac{\gamma}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{1}{2\sqrt{\Delta}} - \frac{\ln w^\star(\theta)}{2}\\ &= -\frac{\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} - \frac{\ln w^\star(\theta)}{2}\end{aligned}$$

□

Lemma EC.5. *Let* $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, *then we can rewrite Equation* (3) *as*

$$l(\theta, w) = w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 - \frac{\gamma}{2}\ln w \tag{EC.14}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial l(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta)^2} \tag{EC.15}$$

2.
$$\frac{\partial l(\theta, w^\star(\theta))}{\partial\gamma} = \frac{3\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} \tag{EC.16}$$

where $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof of Lemma EC.5. By Lemma EC.3 and Lemma EC.4, it is clear that $l(\theta, w^\star(\theta)) = e(\theta, w^\star(\theta)) + I(w^\star(\theta)) = w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 - \frac{\gamma}{2}\ln w$. In addition,

1.
$$
\begin{aligned}
&\frac{\partial l(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} \\
&= \frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} + \frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} \\
&= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2} - \frac{\gamma}{2} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))(-\sigma_\theta^2 + \sqrt{\Delta}) - 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta)(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= \frac{(-\sigma_\theta^4 + \sigma_\theta^2\sqrt{\Delta} - 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta))(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= \frac{\sqrt{\Delta}(-\sqrt{\Delta} + \sigma_\theta^2)(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= -\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta)^2}
\end{aligned}
$$

2.
$$
\begin{aligned}
\frac{\partial l(\theta, w^\star(\theta))}{\partial\gamma} &= \frac{\partial e(\theta, w^\star(\theta))}{\partial\gamma} + \frac{\partial I(w^\star(\theta))}{\partial\gamma} \\
&= \frac{\sigma_\theta^2 + \sqrt{\Delta}}{4\sqrt{\Delta}} + \frac{\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= \frac{\Delta - \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} \\
&= \frac{3\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})}
\end{aligned}
$$

□

EC.2.1.2. Proof of the results.

Proof of Proposition 1. Because $d(\theta) = |\mu_\theta - \theta|$ by definition and $|\mu_\theta - \theta|$ increases with $(\mu_\theta - \theta)^2$, we only have to show the change of $l(\theta, \sigma_q^\star(\theta)), I(\sigma_q^\star(\theta))$ and $e(\theta, \sigma_q^\star(\theta))$ with respect to $(\mu_\theta - \theta)^2$. By Lemma EC.1,

$$
l(\theta, \sigma_q^\star(\theta)) = \begin{cases} l\left(\theta, \sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}}\right) & |\mu_\theta - \theta| \geq \tau_d(\gamma) \\ (\mu_\theta - \theta)^2 & \text{otherwise} \end{cases}
$$

$$
I(\sigma_q^\star(\theta)) = \begin{cases} I\left(\sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}}\right) & |\mu_\theta - \theta| \geq \tau_d(\gamma) \\ 0 & \text{otherwise} \end{cases}
$$

$$e(\theta,\sigma_q^\star(\theta))=\begin{cases} e\left(\theta,\sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1-w^\star(\theta)}}\right) & |\mu_\theta-\theta|\geq\tau_d(\gamma)\\ (\mu_\theta-\theta)^2 & \text{otherwise}\end{cases}$$

where $w^\star(\theta)=\dfrac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$, and $\tau_d(\gamma)>0$ is a threshold that increases in γ and is not less than $\sigma_\theta^2-\dfrac{\sigma_\theta^4}{4\gamma}$. Now, let's apply the results of Lemma EC.3, Lemma EC.4, and Lemma EC.5.

1. When $|\mu_\theta-\theta|\geq\tau_d(\gamma)$, by Lemma EC.5,

$$\frac{\partial l(\theta,\sigma_q^\star(\theta))}{\partial(\mu_\theta-\theta)^2}=-\frac{\sigma_\theta^2\sqrt{\Delta}-\sigma_\theta^4-2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}{8((\mu_\theta-\theta)^2-\sigma_\theta)^2}$$

where $\Delta=\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)$.

We only need to show $\sigma_\theta^2\sqrt{\Delta}-\sigma_\theta^4-2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)\leq 0$. When $|\mu_\theta-\theta|\geq\tau_d(\gamma)$, by the proof of Lemma EC.1, we know $\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta)\geq 0$, so $\sigma_\theta^4+2\gamma((\mu_\theta-\theta)^2-\sigma_\theta)\geq 0$. Thus,

$$\begin{aligned}&\sigma_\theta^2\sqrt{\Delta}-\sigma_\theta^4-2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)\leq 0\\ \iff\;&\sigma_\theta^4\Delta\leq[\sigma_\theta^4+2\gamma((\mu_\theta-\theta)^2-\sigma_\theta)]^2\\ \iff\;&\sigma_\theta^4(\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2))\leq\sigma_\theta^8+4\gamma\sigma_\theta^4((\mu_\theta-\theta)^2-\sigma_\theta)+4\gamma^2((\mu_\theta-\theta)^2-\sigma_\theta)^2\\ \iff\;&4\gamma^2((\mu_\theta-\theta)^2-\sigma_\theta)^2\geq 0\end{aligned}\tag{EC.17}$$

When $|\mu_\theta-\theta|<\tau_d(\gamma)$, $\dfrac{\partial l(\theta,\sigma_q^\star(\theta))}{\partial(\mu_\theta-\theta)^2}=1$. And $l(\theta,\sigma_q^\star(\theta))$ is continuous at $|\mu_\theta-\theta|=\tau_d(\gamma)$. Thus, $l(\theta,\sigma_q^\star(\theta))$ increases in $|\mu_\theta-\theta|$. By definition, $l^\star=\min(\Gamma,l(\theta,\sigma_q^\star(\theta))$, so $l^\star$ increases in $|\mu_\theta-\theta|$.

2. When $|\mu_\theta-\theta|\geq\tau_d(\gamma)$, by Lemma EC.4,

$$\frac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta-\theta)^2}=-\frac{\gamma}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta}-\sigma_\theta^4-2\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)(-\sigma_\theta^2+\sqrt{\Delta})}$$

By the proof of Lemma EC.1, we know $w^\star(\theta)\geq 0$ when $|\mu_\theta-\theta|\geq\tau_d(\gamma)$, so $((\mu_\theta-\theta)^2-\sigma_\theta)(-\sigma_\theta^2+\sqrt{\Delta})\geq 0$. Because of the above Inequality (EC.17), we conclude that $\dfrac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta-\theta)^2}\geq 0$.

When $|\mu_\theta-\theta|<\tau_d(\gamma)$, $\dfrac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta-\theta)^2}=0$. We conclude that $I(\sigma_q^\star(\theta))$ increases in $|\mu_\theta-\theta|$.

3. Firstly, notice that $l(\theta,\sigma_q^\star)=0$ for $d(\theta)=0$ and we have shown that $l(\theta,\sigma_q^\star(\theta))$ monotonically increases in $d(\theta)$. In addition, we can see that $w^\star(\theta)\to 0$ as $d(\theta)\to\infty$, which leads to $I(\sigma_q^\star(\theta))\to\infty$ and $l(\theta,\sigma_q^\star)\to\infty$ as $d(\theta)\to\infty$. These imply that for any $\Gamma>0$, there must exist a threshold $\tau_a>0$ such that $d(\theta)\geq\tau_a\iff l(\theta,\sigma_q^\star)\leq\Gamma$.

4. When $|\mu_\theta - \theta| < \tau_d(\gamma)$, by Lemma EC.1, $\sigma_q^\star(\theta) = \infty$, thereby $e(\theta, \sigma_q^\star(\theta)) = (\mu_\theta - \theta)^2$ and $\dfrac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = 1 > 0$.

   When $|\mu_\theta - \theta| \geq \tau_d(\gamma)$, by Lemma EC.3,

$$\frac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

   Because of the above Inequality (EC.17), we have $\dfrac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} \leq 0$.

   We conclude that if $|\mu_\theta - \theta| < \tau_d(\gamma)$, $e(\theta, \sigma_q^\star(\theta))$ increases in $(\mu_\theta - \theta)^2$; if $|\mu_\theta - \theta| \geq \tau_d(\gamma)$, $e(\theta, \sigma_q^\star(\theta))$ decreases in $|\mu_\theta - \theta|$.

□

Proof of Proposition 2. By definition, if $d(\theta) \geq \tau_a$, users will work on their own and $\theta^\star = \theta$, so $|E[\theta^\star|\theta] - \mu_\theta| = |\theta - \mu_\theta|$.

If $d(\theta) < \tau_a$, $\theta^\star = \theta_A^\star$. By Equation (1), we know $E[\theta_A^\star|\theta] = \dfrac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^2} \cdot \theta + \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2} \cdot \mu_\theta$, so $|E[\theta_A^\star|\theta] - \mu_\theta| = \dfrac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^{\star 2}}|\theta - \mu_\theta|$. Since $l(\theta, \sigma_q) \to \infty$ as $\sigma_q \to 0$ and $\sigma_q^\star = \infty$ is feasible, we must have $\sigma_q^\star > 0$. Thus, $\dfrac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^{\star 2}} < 1$, implying $|E[\theta_A^\star|\theta] - \mu_\theta| < |\theta - \mu_\theta|$ when $\theta \neq \mu_\theta$. □

Proof of Theorem 1. By Lemma EC.1, the AI's estimator $\theta_A(\sigma_q^\star(\sigma_\theta))$ is

$$\theta_A(\sigma_q^\star(\theta)) = \begin{cases} (1 - w^\star(\theta))q + w^\star(\theta)\mu_\theta & |\mu_\theta - \theta| \geq \tau_d(\gamma) \\ \mu_\theta & \text{otherwise} \end{cases}$$

where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$, and $\tau_d(\gamma) > 0$ is a threshold that increases in γ and is not less than $\sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\gamma}$.

By definition, the unconditional variance of $\theta^\star$ is

$$Var(\theta^\star) = E[(\theta^\star - E[\theta^\star])^2]$$

Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative density function of $N(0,1)$, respectively. We know

$$E[\theta^\star] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \theta^\star \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

First, when $\tau_d > \tau_a$, we know that for any $\theta < \tau_a < \tau_d$, $w^\star(\theta) = 1$ and $\theta^\star = \mu_\theta$; for any $\theta > \tau_a$, $\theta^\star = \theta$, so

$$
\begin{aligned}
E[\theta^\star] &= \int_{d(\theta)<\tau_a} \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} \theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)<\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&\text{Because } \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta = \int_{d(\theta)>\tau_a} (\theta-\mu_\theta) \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \mu_\theta
\end{aligned}
$$

When $\tau_d \le \tau_a$,

$$
\begin{aligned}
&E[\theta^\star] \\
= &\int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [(1-w^\star(\theta))q + w^\star(\theta)\mu_\theta] \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)<\tau_d} \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} \theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= &\int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))\theta + w^\star(\theta)\mu_\theta] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)<\tau_d} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&\text{Because } \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta = \int_{d(\theta)>\tau_a} (\theta-\mu_\theta) \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta, \\
= &\int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))(\theta-\mu_\theta) + \mu_\theta] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)<\tau_d} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= &\int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))(\theta-\mu_\theta) \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&\text{Notice that } \int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))(\theta-\mu_\theta) \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta = 0, \\
&\text{because } (1-w^\star(\theta))(\theta-\mu_\theta) \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) \text{ is symmetric with respect to } \theta = \mu_\theta. \\
= &\int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= &\mu_\theta
\end{aligned}
$$

Thus, when $\tau_d > \tau_a$,

$$
\begin{aligned}
Var(\theta^\star) &= \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta
\end{aligned} \tag{EC.18}
$$

When $\tau_d \leq \tau_a$

$$
\begin{aligned}
& Var(\theta^\star) \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [\theta_A(\sigma_q^\star(\theta)) - \mu_\theta]^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [(1-w^\star(\theta))q + w^\star(\theta)\mu_\theta - \mu_\theta]^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 (q-\mu_\theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 (\theta+\epsilon_q-\mu_\theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 [\epsilon_q^2 - 2\epsilon_q(\mu_\theta-\theta) + (\mu_\theta-\theta)^2] \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))^2 [\sigma_q^\star(\theta)^2 + (\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
+ & \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta
\end{aligned}
$$

Thus,

$$
\begin{aligned}
Var(\theta^\star) = 2\Big[& \int_{\mu_\theta+\tau_d}^{\tau_a} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 \\
& + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{\mu_\theta+\tau_a}^{\infty} (\mu_\theta-\theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \Big]
\end{aligned}
\tag{EC.19}
$$

1. Now, let us first show that when $\Gamma \to \infty$, $Var(\theta^\star)$ is strictly decreasing in γ. In this case, $Var(\theta^\star) = 2\int_{\mu_\theta+\tau_d}^{\infty} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta$. Let $h(\theta) \triangleq [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2]$, then

$$
Var(\theta_A(\sigma_q^\star(\theta))) = 2\int_{\mu_\theta+\tau_d(\gamma)}^{\infty} h(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta
$$

By the Leibniz integral rule,

$$\frac{\partial Var(\theta_A(\sigma_q^\star(\theta)))}{\partial\gamma} = -2h(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)|_{\theta=\mu_\theta+\tau_d(\gamma)}\cdot\frac{\partial\sqrt{\tau_d(\gamma)}}{\partial\gamma}$$
$$+2\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}\frac{\partial h(\theta)}{\partial\gamma}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$

Since $\frac{\partial\sqrt{\tau_d(\gamma)}}{\partial\gamma}\geq 0$ by Lemma EC.1, we only need to show $\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}\frac{\partial h(\theta)}{\partial\gamma}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\leq 0$.

$$\begin{aligned}
&2\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}\frac{\partial h(\theta)}{\partial\gamma}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\\
=&\,2\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}\frac{\partial h(\theta)}{\partial w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial\gamma}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\\
=&\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}[(1-2w^\star(\theta))\sigma_\theta^2+2(w^\star(\theta)-1)(\mu_\theta-\theta)^2]\frac{1}{\sqrt{\Delta}}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\\
&\text{where }\Delta=\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)\\
=&\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}[2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-2(\mu_\theta-\theta)^2]\frac{1}{\sqrt{\Delta}}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta
\end{aligned}$$

Let $g(\theta)\triangleq[2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-2(\mu_\theta-\theta)^2]\frac{1}{\sqrt{\Delta}}$, we want to show $\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta<0$.

(a) First, when $\gamma>\frac{\sigma_\theta^2}{2}$, we want to show $g(\theta)\leq 0$ for any $\theta\geq\mu_\theta+\tau_d(\gamma)$.

By Lemma EC.1, $\tau_d(\gamma)>\sqrt{\sigma_\theta^2-\frac{\sigma_\theta^4}{4\gamma}}$, so $\tau_d(\gamma)>\frac{\sigma_\theta}{\sqrt{2}}$. This implies that for any $\theta\geq\mu_\theta+\tau_d(\gamma)$, $(\mu_\theta-\theta)^2>\frac{\sigma_\theta^2}{2}$.

If $(\mu_\theta-\theta)^2>\sigma_\theta^2$, $2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-2(\mu_\theta-\theta)^2\leq-\sigma_\theta^2<0$, because $w^\star(\theta)\leq 1$. And if $\frac{\sigma_\theta^2}{2}<(\mu_\theta-\theta)^2\leq\sigma_\theta^2$, $2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-2(\mu_\theta-\theta)^2\leq\sigma_\theta^2-2(\mu_\theta-\theta)^2<0$, because $w^\star(\theta)>0$. Thus, $(\mu_\theta-\theta)^2>\frac{\sigma_\theta^2}{2}$ implies $2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2)+\sigma_\theta^2-2(\mu_\theta-\theta)^2<0$, which further implies $g(\theta)<0$.

Therefore, when $\gamma>\frac{\sigma_\theta^2}{2}$, $\int_{\mu_\theta+\tau_d(\gamma)}^{\infty}g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta<0$.

(b) When $\gamma\leq\frac{\sigma_\theta^2}{2}$:

Let $\alpha=\frac{\gamma}{\sigma_\theta^2}$ (so $\gamma\leq\frac{\sigma_\theta^2}{2}$ implies $\alpha\leq\frac{1}{2}$).

$\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2) = \sigma_\theta^4\left[1+\frac{4\gamma}{\sigma_\theta^2}\left(\frac{(\mu_\theta-\theta)^2}{\sigma_\theta^2}-1\right)\right] = \sigma_\theta^4\left[1+4\alpha\left(\left(\frac{\mu_\theta-\theta}{\sigma_\theta}\right)^2-1\right)\right]$. Similarly, we can get

$$w^\star(\theta) = \frac{-\sigma_\theta^2+\sqrt{\sigma_\theta^4+4\gamma((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)} = \frac{-1+\sqrt{1+4\alpha\left(\left(\frac{\mu_\theta-\theta}{\sigma_\theta}\right)^2-1\right)}}{4\left[\left(\frac{\mu_\theta-\theta}{\sigma_\theta}\right)^2-1\right]}$$

So the substitution $x \triangleq \frac{\theta-\mu_\theta}{\sigma_\theta}$ yields

$$\begin{aligned}\int_{\mu_\theta+\tau_d(\gamma)}^{\infty} g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta &= \int_{\frac{\tau_d(\gamma)}{\sigma_\theta}}^{\infty}[(1-2\hat{w}(x,\alpha))\sigma_\theta^2+2(\hat{w}(x,\alpha)-1)\sigma_\theta^2x^2]\frac{1}{\sigma_\theta^2\sqrt{\hat{\Delta}(x,\alpha)}}\phi(x)\sigma_\theta dx\\ &= \int_{\frac{\tau_d(\gamma)}{\sigma_\theta}}^{\infty}[(1-2\hat{w}(x,\alpha))+2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}}\phi(x)\sigma_\theta dx\\ &= \frac{1}{\sqrt{2\pi}}\int_{\hat{\tau_d}(\alpha)}^{\infty}[(1-2\hat{w}(x,\alpha))+2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}}\exp\left(-\frac{x^2}{2}\right)dx\end{aligned}$$

where $\hat{\tau_d}(\alpha) = \frac{\tau_d(\gamma)}{\sigma_\theta}$,$\hat{w}(x,\alpha) = \frac{-1+\sqrt{1+4\alpha(x^2-1)}}{4(x^2-1)}$ and $\hat{\Delta}(x,\alpha) = 1+4\alpha(x^2-1)$.

Note that $[(1-2\hat{w}(x,\alpha))+2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}} = \frac{1}{2}\left[1+\frac{1-4x^2}{\sqrt{1+4\alpha(x^2-1)}}\right]$

Define $G(\alpha) \triangleq \int_{\hat{\tau_d}(\alpha)}^{\infty}\left[1+\frac{1-4x^2}{\sqrt{1+4\alpha(x^2-1)}}\right]\exp\left(-\frac{x^2}{2}\right)dx$. We want to show $\forall\alpha\in[0,1/2]$, $G(\alpha)<0$.

Let's do another change of variables: $y \triangleq x^2-1$, which implies $dy = 2xdx$ and $x=\sqrt{y+1}$. This yields

$$G(\alpha) = \int_{\hat{\tau_d}^2(\alpha)-1}^{\infty}\left[1-\frac{3+4y}{\sqrt{1+4\alpha y}}\right]\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$$

Let $\omega(y,\alpha) \triangleq 1-\frac{3+4y}{\sqrt{1+4\alpha y}}$. Note that

i. If $y \geq 0$, $\omega(y,\alpha)$ is increasing α.

ii. If $y \in [-3/4, 0)$, $\omega(y,\alpha)$ is decreasing α.

iii. If $y \in [-1,-3/4)$, $\omega(y,\alpha)$ is increasing α.

Correspondingly,

i. Let $G_0(\alpha) \triangleq \int_0^\infty \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$, we have $G_0(\alpha) \leq G_0(1/2) \leq G_0(1) < -0.96$.

ii. $\hat{\tau_d}^2(\alpha) - 1 \geq -3/4 \Longleftrightarrow \hat{\tau_d}^2(\alpha) \geq 1/4$

Note that $\hat{\tau_d}^2(\alpha) = \dfrac{\tau_d(\gamma)}{\sigma_\theta}$, and by the definition of $\tau_d(\gamma)$ in the proof of Lemma EC.1, $\tau_d(\gamma)$ solves

$$(\tau_d^2(\gamma,\sigma_\theta) - \sigma_\theta^2)m^2 + \sigma_\theta^2 m - \frac{\gamma}{2}\ln(m) = \tau_d^2(\gamma,\sigma_\theta) - \sigma_\theta^2)$$

where $m = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma(\tau_d^2(\gamma,\sigma_\theta) - \sigma_\theta^2)}}{4(\tau_d^2(\gamma,\sigma_\theta) - \sigma_\theta^2)}$ This is equivalent to that $\hat{\tau}_d(\alpha)$ solves

$$(\hat{\tau_d}^2(\alpha) - 1)m^2 + m - \frac{\alpha}{2}\ln(m) = \hat{\tau_d}^2(\alpha)$$

where $m = \dfrac{-1 + \sqrt{1 + 4\alpha(\hat{\tau_d}^2(\alpha) - 1)}}{4(\hat{\tau_d}^2(\alpha) - 1)}$

Thus, we can get there exists $\alpha^\star$ such that $\hat{\tau_d}^2(\alpha) \geq 1/4 \Longleftrightarrow \alpha \geq \alpha^\star$. And we can numerically compute $\alpha^\star \approx 0.13845$.

Let $G_1(\alpha) \triangleq \int_{\hat{\tau_d}^2(\alpha)-1}^{0} \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$. Since $\omega(y,\alpha)$ is decreasing in α, we have

$$G_1(\alpha) \leq \int_{\hat{\tau_d}^2(\alpha)-1}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \leq \int_{-3/4}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$$

We can numerically find $\int_{-3/4}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy < 0$.

Thus, $G(\alpha) = G_0(\alpha) + G_1(\alpha) < 0$.

iii. $\hat{\tau_d}^2(\alpha) - 1 < -3/4 \Longleftrightarrow \alpha < \alpha^\star$

$$\begin{aligned}
G_1(\alpha) &= \int_{\hat{\tau_d}^2(\alpha)-1}^{\hat{\tau_d}^2(\alpha^\star)-1} \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\
&\quad + \int_{\hat{\tau_d}^2(\alpha^\star)-1}^{0} \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\
&\leq \int_{\hat{\tau_d}^2(\alpha)-1}^{\hat{\tau_d}^2(\alpha^\star)-1} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\
&\quad + \int_{\hat{\tau_d}^2(\alpha^\star)-1}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\
&= \int_{\hat{\tau_d}^2(\alpha)-1}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\
&\leq \int_{-1}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy
\end{aligned}$$

We can numerically find $\int_{-1}^{0} \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy < 0.817$.

Thus, $G(\alpha) = G_0(\alpha) + G_1(\alpha) < -0.96 + 0.817 \leq 0$.

We conclude that $\forall \alpha \in [0,1/2],\ G(\alpha) < 0$.

Hence, $Var(\theta_A(\sigma_q^{2\star}))$ strictly decreases in γ.

2. When $\gamma = 0$, we know $\forall\theta, w^\star(\theta)$. By Equation (EC.19), $Var(\theta^\star) = \int_{-\infty}^{\infty}(\mu_\theta - \theta)^2\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta = Var(\theta) = \sigma_\theta^2$. Thus, $\lim_{\gamma\to 0} Var(\theta^\star) = Var(\theta)$.

   When $\gamma \to \infty$, by definition, for any θ, $l \to \infty$ if σ_q is finite, so the optimal decision is $\sigma_q^\star = +\infty$ with $l^\star = (\theta - \mu_\theta)^2$. Thus, by Equation (EC.19), $\lim_{\gamma\to\infty}, Var(\theta^\star) = 2\int_{\mu_\theta+\tau_a}^{\infty}(\mu_\theta - \theta)^2\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$. And for any $\Gamma > 0$, there exists $\tau_a > 0$ such that $\forall\theta[0,\tau_a]$, $(\theta-\mu_\theta)^2 \le \Gamma$. Hence, $Var(\theta^\star) < Var(\theta)$.

3. (see Section EC.6) Next, we want to show $Var(\theta^\star) < Var(\theta)$ if $\gamma \ge \sigma_q^2/2$ or $\Gamma \le \hat\Gamma$ or $\Gamma \ge \tilde\Gamma$ for some $\hat\Gamma > 0, \tilde\Gamma > 0$. From Equation (EC.19), we need to show $D \triangleq \int_{\mu_\theta}^{\mu_\theta+\tau_a}(\mu_\theta-\theta)^2\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - \int_{\mu_\theta+\tau_d}^{\mu_\theta+\tau_a}[(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta > 0$.

   We can do the same change of variables as the above steps. In particular, let $y = ((\theta - \mu_\theta)/\sigma_\theta)^2 - 1$, then we have

$$D = \frac{\sigma_\theta}{\sqrt{2\pi}}\left[\int_{-1}^{\hat{\tau_a}^2-1}(y+1)\frac{\exp(-(y+1)/2)}{\sqrt{y+1}}d\theta - \int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1}(1-\hat w)(1+(1-\hat w)y)\frac{\exp(-(y+1)/2)}{\sqrt{y+1}}d\theta\right]$$

   where $\hat{\tau_a} = \tau_a/\sigma_\theta, \hat{\tau_d} = \tau_d/\sigma_\theta$, $\hat w = \frac{-1+\sqrt{1+4\alpha y}}{4y}$ and $\alpha = \gamma/\sigma_\theta^2$. Let $\omega = (1-\hat w)(1+(1-\hat w)y)$.

   (a) When $\gamma \ge \frac{\sigma_\theta^2}{2}$, by Lemma EC.1, $\tau_d \ge \sqrt{\sigma_\theta^2 - \frac{\sigma_\theta^4}{4\gamma}}$, so $\hat{\tau_d} \ge \frac{1}{\sqrt 2}$.

   $\frac{\partial\omega}{\partial\hat w} = -1 - 2(1-\hat w)y$, which is non-positive if and only if $(1-\hat w)y \ge -1/2$. Because $y \ge \hat{\tau_d} - 1 > -1/2$ and $\hat w \in [0,1]$, this implies that $(1-\hat w)y \ge -1/2$ and $\frac{\partial\omega}{\partial\hat w} \le 0$. Thus, $\int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1}(1-w)(1+(1-w)y)\exp\left(-\frac{y+1}{2}\right)\frac{1}{\sqrt{y+1}}d\theta$ is maximized at $w=0$, which is equal to $\int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1}(1+y)\exp\left(-\frac{y+1}{2}\right)\frac{1}{\sqrt{y+1}}d\theta$. This means $D = \frac{\sigma_\theta}{\sqrt{2\pi}}\int_{-1}^{\hat{\tau_d}^2-1}(y+1)\frac{\exp(-(y+1)/2)}{\sqrt{y+1}}d\theta$. And by Lemma EC.1, we know $\forall\gamma > 0$, we must have $\tau_d > 0$. Thus, $D > 0$.

   (b) Let $\hat\Gamma \triangleq l^\star(\theta = \mu_\theta + \tau_d) > 0$

   First, when $0 < \Gamma \le \hat\Gamma$, this means $\tau_a < \tau_d$, by Equation (EC.18), $Var(\theta^\star) = \int_{d(\theta)>\tau_a}(\mu_\theta - \theta)^2\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$, which is less than $Var(\theta)$, since $\tau_a > 0$ whenever $\Gamma > 0$.

   (c) Let $\acute\Gamma \triangleq l^\star(\theta = \mu_\theta + \sigma_\theta^2/2) > 0$

   In Item (a), we have seen that if $y \ge -1/2$, $\frac{\partial\omega}{\partial\hat w} \le 0$. This means that if $y \ge -1/2$, $\omega \le (1+y)$, which implies D increases in τ_a if $\tau_a \ge \acute\Gamma$.

   We already prove that $D > 0$ when $\Gamma \to \infty$, meaning that $D > 0$ when $\tau_a \to \infty$. Because D is continuous in τ_a, we either have $D > 0$ for $\Gamma \ge \acute\Gamma$ or there exists another threshold $\tilde\Gamma > \acute\Gamma$ such that $D > 0$ whenever $\Gamma \ge \tilde\Gamma$.

□

EC.2.2. Proof of the Results in Section 4.

EC.2.2.1. Auxiliary lemmas

LEMMA EC.6. *For any* θ, σ_q^2,

$$e(\theta,\sigma_q) = \frac{\sigma_q^2(\sigma_A^4 + \sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^2} \tag{EC.20}$$

In addition,

- *Both* $l(\theta,\sigma_q^2)$ *and* $e(\theta,\sigma_q)$ *strictly increase in* $(\mu_A-\theta)^2$.
- *Both* $l(\theta,\sigma_q^2)$ *and* $e(\theta,\sigma_q)$ *strictly decrease in* σ_A^2 *for* $\sigma_A^2 < (\mu_A-\theta)^2$ *and increase in* σ_A^2 *for* $\sigma_A^2 \geq (\mu_A-\theta)^2$.

Proof of Lemma EC.6. By Equation (1), $\theta_A = \frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}q + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}\mu_A$. Then,

$$\begin{aligned}
e(\theta,\sigma_q) &= E\left[\left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}(\theta+\epsilon_q) + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}\mu_A - \theta\right)^2 |\theta\right] \\
&= E\left[\left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\epsilon_q + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 |\theta\right] \\
&= \left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\right)^2 E[\epsilon_q^2] + \left(\frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 \\
&= \left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\right)^2 \sigma_q^2 + \left(\frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 \\
&= \frac{\sigma_q^2(\sigma_A^4+\sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^2}
\end{aligned}$$

It is clear that $e(\theta,\sigma_q)$ strictly increases in $(\mu_A-\theta)^2$, and $l(\theta,\sigma_q^2)$ strictly increases in $(\mu_A-\theta)^2$ (Note that $I(\sigma_q^2)$ does not depend on either μ_A or σ_A).

Take the derivative of $e(\theta,\sigma_q)$ with respect to σ_A^2:

$$\begin{aligned}
\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} &= \frac{2\sigma_q^2\sigma_A^2(\sigma_A^2+\sigma_q^2)^2 - 2(\sigma_A^2+\sigma_q^2)\sigma_q^2(\sigma_A^4+\sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^4} \\
&= \frac{2\sigma_q^2\sigma_A^2(\sigma_A^2+\sigma_q^2) - 2\sigma_q^2(\sigma_A^4+\sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^3} \\
&= \frac{2\sigma_q^4(\sigma_A^2-(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^3}
\end{aligned}$$

Thus, $\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} < 0$ if $\sigma_A^2 < (\mu_A-\theta)^2$, and $\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} \geq 0$ if $\sigma_A^2 \geq (\mu_A-\theta)^2$. This implies that both $l(\theta,\sigma_q^2)$ and $e(\theta,\sigma_q)$ strictly decrease in σ_A^2 for $\sigma_A^2 < (\mu_A-\theta)^2$ and increase in σ_A^2 for $\sigma_A^2 \geq (\mu_A-\theta)^2$.
□

LEMMA EC.7. *Let* $w^\star(\theta,\gamma) = \frac{\sigma_q^{\star 2}(\theta,\gamma)}{\sigma_A^2+\sigma_q^{\star 2}(\theta,\gamma)}$. $\forall \theta,\mu_A,\sigma_A,\gamma_1>\gamma_2,\ w^\star(\theta,\gamma_1) \geq w^\star(\theta,\gamma_2)$.

Proof of Lemma EC.7. For the sake of contradiction, assume $w^\star(\theta,\gamma_1) < w^\star(\theta,\gamma_2)$ for some θ. Since $I(w)$ strictly decreases in w, we have $\delta_I \triangleq I(w^\star(\theta,\gamma_1)) - I(w^\star(\theta,\gamma_2)) > 0$. Let $\delta_e \triangleq e(\theta,w^\star(\theta,\gamma_1)) - e(\theta,w^\star(\theta,\gamma_2))$. And because $w^\star(\theta,\gamma_1)$ is optimal when $\gamma=\gamma_1$, we must have $l(\theta,w^\star(\theta,\gamma_1),\gamma_1) - l(\theta,w^\star(\theta,\gamma_2),\gamma_1) < 0$. This implies $\delta_e < 0$ and $\delta_e + \gamma_1\delta_I < 0$. However, because $\gamma_1 > \gamma_2$ and $\delta_I > 0$, $\delta_e + \gamma_2\delta_I < \delta_e + \gamma_1\delta_I < 0$, meaning that $l(\theta,w^\star(\theta,\gamma_1),\gamma_2) - l(\theta,w^\star(\theta,\gamma_2),\gamma_2) < 0$. This contradicts the assumption that $w^\star(\theta,\gamma_2)$ is optimal when $\gamma=\gamma_2$. Therefore $\forall\theta$, $w^\star(\theta,\gamma_1) \geq w^\star(\theta,\gamma_2)$ whenever $\gamma_1 > \gamma_2$. □

EC.2.2.2. Proofs of the results

Proof of Proposition 3.

- Suppose $|\mu_{A_1} - \theta| > |\mu_{A_2} - \theta|$ for some $\mu_{A_1}, \mu_{A_2}, \theta$. Let $\sigma^\star_{q_1}$ and $\sigma^\star_{q_1}$ denote the optimal decision for user θ in Problem (4) when $\mu_A = \mu_{A_1}$ and $\mu_A = \mu_{A_2}$, respectively. By definition of l in Equation (3), let $l_1^\star = l(\theta, \sigma^\star_{q_1}, \mu_{A_1})$ and $l_2^\star = l(\theta, \sigma^\star_{q2}, \mu_{A_2})$.

  We want to show $l_1^\star > l_2^\star$. Let us prove this by contradiction. Suppose $l_1^\star \leq l_2^\star$. By Lemma EC.6, $l_1^\star = l(\theta,\sigma^\star_{q_1},\mu_{A_1}) > l(\theta,\sigma^\star_{q_1},\mu_{A_2})$. This implies $l(\theta,\sigma^\star_{q_1},\mu_{A_2}) < l_2^\star = l(\theta,\sigma^\star_{q_2},\mu_{A_2})$. This contradicts the assumption that $\sigma^\star_{q_2}$ minimizes $l(\theta,\sigma^\star_q,\mu_{A_2})$. Therefore, $l_1^\star > l_2^\star$. We conclude that $l^\star$ increases in $|\mu_A - \theta|$.

- Suppose $\sigma_{A_1} < \sigma_{A_2} < |\mu_A - \theta|$ for some $\sigma_{A_1}, \sigma_{A_2}, \mu_A, \theta$. Let $\sigma^\star_{q_1}$ and $\sigma^\star_{q_1}$ denote the optimal decision for user θ in Problem (4) when $\sigma_A = \sigma_{A_1}$ and $\sigma_A = \sigma_{A_2}$, respectively. By definition of l in Equation (3), let $l_1^\star = l(\theta,\sigma^\star_{q_1},\sigma_{A_1})$ and $l_2^\star = l(\theta,\sigma^\star_{q2},\sigma_{A_2})$.

  We want to show $l_1^\star > l_2^\star$. Let us prove this by contradiction. Suppose $l_1^\star \leq l_2^\star$. By Lemma EC.6, $l_1^\star = l(\theta,\sigma^\star_{q_1},\sigma_{A_1}) > l(\theta,\sigma^\star_{q_1},\sigma_{A_2})$. This implies $l(\theta,\sigma^\star_{q_1},\sigma_{A_2}) < l_2^\star = l(\theta,\sigma^\star_{q_2},\sigma_{A_2})$. This contradicts the assumption that $\sigma^\star_{q_2}$ minimizes $l(\theta,\sigma^\star_q,\sigma_{A_2})$. Therefore, $l_1^\star > l_2^\star$. We conclude that $l^\star$ decreases in σ_A when $\sigma_A < |\mu_A - \theta|$.

  Similarly, when $|\mu_A - \theta| \leq \sigma_{A_1} < \sigma_{A_2}$, we want to show $l_1^\star \leq l_2^\star$. Suppose $l_1^\star > l_2^\star$. By Lemma EC.6, $l_2^\star = l(\theta,\sigma^\star_{q_2},\sigma_{A_2}) > l(\theta,\sigma^\star_{q_2},\sigma_{A_1})$. This implies $l(\theta,\sigma^\star_{q_2},\sigma_{A_1}) < l_1^\star = l(\theta,\sigma^\star_{q_1},\sigma_{A_1})$. This contradicts the assumption that $\sigma^\star_{q_1}$ minimizes $l(\theta,\sigma^\star_q,\sigma_{A_1})$. Therefore, $l_1^\star \leq l_2^\star$. We conclude that $l^\star$ increases in σ_A when $\sigma_A \geq |\mu_A - \theta|$.

  □

Proof of Proposition 4. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative density function of $N(0,1)$, respectively. And let $w(\theta) = \dfrac{\sigma_q^2(\theta)}{\sigma_A^2 + \sigma_q^2(\theta)}$.

- Let us first show $E[l^\star(\theta,\mu_A)]$ is minimized at $\mu_A = \mu_\theta$. $\forall \mu_{A1} \neq \mu_\theta$, we want to show $E[l^\star(\theta,\mu_{A1})] > E[l^\star(\theta,\mu_\theta)]$. Without loss of generality, suppose $\mu_{A1} > \mu_\theta$.

  By definition, $E[l^\star(\theta,\mu_A)] = \int_{-\infty}^{\infty} l^\star(\theta,\mu_A)\phi\left(\dfrac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta$. We want to show $\int_{-\infty}^{\infty} [l^\star(\theta,\mu_{A1}) - l^\star(\theta,\mu_\theta)]\phi\left(\dfrac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta > 0$.

By Equation (EC.20), $\forall \sigma_q, \theta_1, \theta_2, \theta_1 - \mu_A = \mu_A - \theta_2 \implies e(\theta_1, \sigma_q) = e(\theta_2, \sigma_q)$, so $w^\star(\theta_1) = w^\star(\theta_2)$, meaning that $w^\star(\theta)$ and $l^\star(\theta, \mu_A)$ are axisymmetric with respect to $\theta = \mu_A$. Also, $\forall \theta, \mu_A$, $w^\star(\theta)$ and $l^\star(\theta, \mu_A)$ are constant as long as $|\mu_A - \theta|$ is constant. This implies $[l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]$ is centrosymmetric with respect to the point $\left(\frac{\mu_{A1} + \mu_\theta}{2}, 0\right)$. That is, $\forall \theta_1 > \theta_2$, $\theta_1 - (\mu_{A1} + \mu_\theta)/2 = (\mu_{A1} + \mu_\theta)/2 - \theta_2 \implies -[l^\star(\theta_1, \mu_{A1}) - l^\star(\theta_1, \mu_\theta)] = [l^\star(\theta_2, \mu_{A1}) - l^\star(\theta_2, \mu_\theta)] > 0$, where the positivity is because $l^\star(\theta, \mu_A)$ increases in $|\mu_A - \theta|$ by Proposition 3.

Let $\bar{\mu}$ denote $(\mu_{A1} + \mu_\theta)/2$. Because $\mu_A > \mu_\theta \implies \bar{\mu} > \mu_\theta$, we have $Pr(\theta \le \bar{\mu}) > Pr(\theta > \bar{\mu})$, and $\forall \theta_1 > \theta_2$, $\theta_1 - \bar{\mu} = \bar{\mu} - \theta_2 \implies \phi\left(\frac{\theta_1 - \mu_\theta}{\sigma_\theta}\right) < \phi\left(\frac{\theta_2 - \mu_\theta}{\sigma_\theta}\right)$. Because $[l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]$ is centrosymmetric with respect to the point $(\bar{\mu}, 0)$, these imply $0 < -[l^\star(\theta_1, \mu_{A1}) - l^\star(\theta_1, \mu_\theta)]\phi\left(\frac{\theta_1 - \mu_\theta}{\sigma_\theta}\right) < [l^\star(\theta_2, \mu_{A1}) - l^\star(\theta_2, \mu_\theta)]\phi\left(\frac{\theta_2 - \mu_\theta}{\sigma_\theta}\right)$.

Hence, $\int_{-\infty}^{\infty} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta = \int_{-\infty}^{\bar{\mu}} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_{\bar{\mu}}^{\infty} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta > 0$. This implies $E[l^\star(\theta, \mu_A)]$ is minimized at $\mu_A = \mu_\theta$.

And because $l^\star(\theta, \mu_A)$ is continuous for any θ and μ_A, by the fundamental theorem of calculus, we know $E[l^\star(\theta, \mu_A)]$ is differentiable, thereby $\frac{\partial E[l^\star]}{\partial \mu_A}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = 0$.

- By Equation (3) and Equation (EC.20), $l^\star(\theta) = \frac{\sigma_q^{\star 2}(\theta)(\sigma_A^4 + \sigma_q^{\star 2}(\theta)(\mu_A - \theta)^2)}{(\sigma_A^2 + \sigma_q^{\star 2}(\theta))^2} - \frac{\gamma}{2}\ln\left(\frac{\sigma_q^{\star 2}(\theta)}{\sigma_q^{\star 2}(\theta) + \sigma_\theta^2}\right)$.

By the chain rule, $\frac{\partial l^\star}{\partial \sigma_A^2} = \frac{dl^\star}{d\sigma_q^{\star 2}} \cdot \frac{d\sigma_q^{\star 2}}{d\sigma_A^2} + \frac{dl^\star}{d\sigma_A^2}$. Because $\sigma_q^{\star 2}$ is optimal, $\frac{dl^\star}{d\sigma_q^{\star 2}} = 0$. This implies $\frac{\partial l^\star}{\partial \sigma_A^2} = \frac{dl^\star}{d\sigma_A^2}$. With some algebra, we can get $\frac{dl^\star}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{2\sigma_q^{\star 4}(\sigma_\theta^2 - (\mu_\theta - \theta)^2)}{(\sigma_q^{\star 2} + \sigma_\theta^2)^3}$. Since $w(\theta) = \frac{\sigma_q^2(\theta)}{\sigma_A^2 + \sigma_q^2(\theta)}$, we can rewrite it as

$$\frac{dl^\star}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{dl^\star(\theta)}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{2}{\sigma_\theta^2} w^\star(\theta)^2 (1 - w^\star(\theta))(\sigma_A^2 - (\mu_\theta - \theta)^2)$$

where $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ and $\Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)$ by Lemma EC.1.

$E[l^\star]\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \int_{-\infty}^{\infty} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta = \int_{|\mu_\theta - \theta| \ge \tau_d} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_{|\mu_\theta - \theta| < \tau_d} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$, where τ_d is defined in Lemma EC.1.

When $\mu_A = \mu_\theta$, $l(\theta)$ is symmetric with respect to $\theta = \mu_\theta$, so $E[l^\star]\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = 2\left[\int_{\mu_\theta + \tau_d}^{\infty} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_0^{\mu_\theta + \tau_d} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right]$. And when $w = 1$ we know $l = (\mu_\theta - \theta)^2$, so $E[l^\star]\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = 2\left[\int_{\mu_\theta + \tau_d}^{\infty} l^\star(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_0^{\mu_\theta + \tau_d} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right]$.

Thus, by the Leibniz integral rule,

$$
\begin{aligned}
\frac{\partial E[l^\star]}{\partial \sigma_A^2}\Big|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} &= 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{\partial l^\star(\theta)}{\partial \sigma_A^2}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - (\mu_\theta-\tau_d)^2\cdot\frac{\partial \tau_d}{\sigma_A^2} + (\mu_\theta-\tau_d)^2\cdot\frac{\partial\tau_d}{\sigma_A^2}\right] \\
&= 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{\partial l^\star(\theta)}{\partial \sigma_A^2}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right] \\
&= 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{2}{\sigma_\theta^2}w^\star(\theta)^2(1-w^\star(\theta))(\sigma_A^2-(\mu_\theta-\theta)^2)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right] \\
&= \frac{1}{\sigma_\theta^2}\left[\int_{\mu_\theta+\tau_d}^{\infty} w^\star(\theta)^2(1-w^\star(\theta))(\sigma_A^2-(\mu_\theta-\theta)^2)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]
\end{aligned}
$$

Let $g(\theta)=w^\star(\theta)^2(1-w^\star(\theta))(\sigma_A^2-(\mu_\theta-\theta)^2)$

When $\gamma\geq 2\sigma_\theta^2$, by Lemma EC.1, $\gamma>2\sigma_\theta^2\geq\sigma_\theta^2 \implies \tau_d=\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\gamma>\frac{1}{2}\sigma_\theta^2+\frac{1}{4}\cdot 2\sigma_\theta^2=\sigma_\theta^2$. So $g(\theta)$ is always negative for any $\theta>\mu_\theta+\tau_d$. Thus, $\int_{\mu_\theta+\tau_d}^{\infty} g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta<0$.

□

Proof of Theorem 2. Let $w=\frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}$. By Equation (1), $\theta_A=(1-w)q+2\mu_A$, where $q=\theta+\epsilon_q$, $\epsilon_q\sim N(0,\sigma_q^2)$ and $\theta\sim N(\mu_\theta,\sigma_\theta^2)$. We further define $w^\star(\theta)=\frac{\sigma_q^{\star 2}(\theta)}{\sigma_A^2+\sigma_q^{\star 2}(\theta)}$. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative density function of $N(0,1)$, respectively.

$$
\begin{aligned}
E[\theta^\star] &= \int_{|\mu_A-\theta|\leq\tau_a}\int_{-\infty}^{\infty}\theta_A^\star\phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right)d\epsilon_q\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta+\int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\leq\tau_a}\int_{-\infty}^{\infty}[(1-w^\star(\theta))q+w^\star(\theta)\mu_A]\phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right)d\epsilon_q\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&+\int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\leq\tau_a}[(1-w^\star(\theta))\theta+w^\star(\theta)\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta+\int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\leq\tau_a}[(1-w^\star(\theta))(\theta-\mu_A)+\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta+\int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\leq\tau_a}[(1-w^\star(\theta))(\theta-\mu_A)+\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta+\mu_\theta-\int_{|\mu_A-\theta|\leq\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\leq\tau_a}w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta+\mu_\theta
\end{aligned}
$$

This implies that

$$
|E[\theta^\star]-\mu_\theta|=\left|\int_{|\mu_A-\theta|\leq\tau_a}w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| \tag{EC.21}
$$

1. Without loss of generality, suppose $\mu_A\geq\mu_\theta$. Then, $Pr(\theta\leq\mu_A)\geq Pr(\theta>\mu_A)$, and $\forall\theta_1>\theta_2$, $\theta_1-\mu_A=\mu_A-\theta_2 \implies \phi\left(\frac{\theta_1-\mu_\theta}{\sigma_\theta}\right)<\phi\left(\frac{\theta_2-\mu_\theta}{\sigma_\theta}\right)$. Because $w^\star(\theta)$ is symmetric with respect to $\theta=\mu_A$, we have $w^\star(\theta_1)=w^\star(\theta_2)$. These imply $0\leq w^\star(\theta_1)|\theta_1-\mu_A|\phi\left(\frac{\theta_1-\mu_\theta}{\sigma_\theta}\right)\leq$

$w^\star(\theta_2)|\theta_2 - \mu_A|\phi\left(\frac{\theta_2-\mu_\theta}{\sigma_\theta}\right)$. Thus, $|E[\theta^\star]-\mu_\theta| = \left|\int_{|\mu_A-\theta|\le\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| \le \left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right|$ and $\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \ge 0$

Let $\gamma_1 > \gamma_2$. By Lemma EC.7, $\forall\theta$, $w^\star(\theta,\gamma_1) \ge w^\star(\theta,\gamma_2)$. Because $w^\star(\theta)$ is symmetric with respect to $\theta=\mu_A$, $\forall\theta_1>\theta_2$, $\theta_1-\mu_A=\mu_A-\theta_2 \implies 0 \ge (w^\star(\theta_1,\gamma_2) - w^\star(\theta_1,\gamma_1))(\theta_1-\mu_A)\phi\left(\frac{\theta_1-\mu_\theta}{\sigma_\theta}\right) \ge (w^\star(\theta_2,\gamma_2)-w^\star(\theta_1,\gamma_1))(\mu_A-\theta_2)\phi\left(\frac{\theta_2-\mu_\theta}{\sigma_\theta}\right)$. This implies $\int_{\theta\le\mu_A} w^\star(\theta,\gamma_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - \int_{\theta\le\mu_A} w^\star(\theta,\gamma_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \ge -\left[\int_{\theta>\mu_A} w^\star(\theta,\gamma_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - \int_{\theta>\mu_A} w^\star(\theta,\gamma_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right] \ge 0$. Rearrange the inequality, we can get $\int_{-\infty}^{\infty} w^\star(\theta,\gamma_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \ge \int_{-\infty}^{\infty} w^\star(\theta,\gamma_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$. Thus, $\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$ increases in γ.

And because $w^\star(\theta,\gamma)\to 1$ as $\gamma\to\infty$, by the monotone convergence theorem (Pugh (2015)), we get the upper bound $\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \le \mu_A-\mu_\theta$. Hence, $|E[\theta^\star]-\mu_\theta| \le |\mu_A-\mu_\theta|$.

2. When $\gamma=0$, for any θ, $w^\star(\theta)=0$, by Equation (EC.21), we have $|E[\theta^\star]-\mu_\theta|=0$. And when $\Gamma=0$, $\tau_a=0$, $|E[\theta^\star]-\mu_\theta| = \left|\int_{|\mu_A-\theta|=0} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| = 0$
3. When $\Gamma\to\infty$, by Equation (EC.21), $|E[\theta^\star]-\mu_\theta| = \left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right|$. And when $\gamma\to\infty$, for any θ, $w^\star(\theta)\to 1$. Without loss of generality, suppose $\mu_A\ge\mu_\theta$. In Item 1, we have shown $\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$ is non-negative and increases in γ. By the monotone convergence theorem (Pugh (2015)), we have
$$\lim_{\gamma\to\infty}\left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| = \left|\int_{-\infty}^{\infty}(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| = |\mu_A-\mu_\theta|$$
Thus, when $\Gamma\to\infty$ and $\gamma\to\infty$, $|E[\theta^\star]-\mu_\theta| = |\mu_A-\mu_\theta|$.
4. When $\Gamma\to\infty$, by Equation (EC.21), $|E[\theta^\star]-\mu_\theta| = \left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right|$. Without loss of generality, suppose $\mu_A\ge\mu_\theta$. In Item 1, we have shown $\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$ is non-negative and increases in γ. Hence, when $\Gamma\to\infty$, $|E[\theta_A^\star]-\mu_\theta|$ increases in γ.

□

EC.3. Extensive literature review

Related studies on human-AI interactions. Our paper is related to some recent modeling studies about human-AI interaction in operations management (e.g., Agrawal et al. (2018), Ibrahim et al. (2021), Chen et al. (2022), Dai and Singh (2023), Boyacı et al. (2023), de Véricourt and Gurkan (2023)). Essentially, their primary focus lies in examining the potential impact of the coexistence of humans and an AI on performance, such as accuracy, and exploring how the predictive performance can be enhanced or hindered compared to decisions made solely by humans or AI. For example, de Véricourt and Gurkan (2023) consider the human-AI interactions in which a human agent supervises an AI to make some high-stakes decisions. They show that the agent may be subject to a verification bias and hesitates forever whether the AI performs better than the agent because the agent can overrule the AI before observing the correctness of the AI's predictions. Ibrahim et al. (2021) build a stylized model to analyze how human judgments can improve AI predictions. In the paper of Boyacı et al. (2023), the authors consider a situation in which a human agent has to spend a cognitive cost collecting information in a decision process, whereas an AI can provide him with some additional information without cognitive cost. They show that the AI input can improve the overall accuracy of human decisions, but may incur a higher propensity for certain types of errors. In contrast, in our paper, users can provide an AI with more information to improve the AI's outputs but have to spend a communication cost.

In these papers, the authors typically evaluate whether the coexistence of humans and AI can outperform an AI or a human solely. Rather than comparing performance, our study delves into the rational decision-making process of individuals when interacting with AIs and the subsequent individual and societal influence of these interactions with AIs. Ultimately, we aim to understand the broader implications for our lives in terms of the issues of homogenization and bias.

Related studies on generative AIs. With the popularity of ChatGPT, many scholars have engaged in research on its impact on people's lives and in their respective fields, such as labor markets (Eloundou et al. (2023)), marketing (Brand et al. (2023)), healthcare (Sallam (2023)), and so on. Most of the research uses empirical analysis to investigate whether generative AI, represented by ChatGPT, can truly bring us more benefits and conveniences. For instance, Noy and Zhang (2023) show that ChatGPT can substantially improve productivity in mid-level professional writing tasks. Binz and Schulz (2023) tested GPT-3 with some experiments from the cognitive psychology literature. They find that GPT-3 can solve many of those tasks well and even sometimes outperform humans' performance. Our study approaches this question from a different angle. Through a modeling method, we attempt to foreshadow how our lives may change under the widespread application of AIs in the future. We place particular emphasis on the societal impact of people's rational decision-making on the issues of homogenization and bias during the process of interacting with AI.

We assume that the output of AIs depends on the information provided by users. In fact, many empirical studies have observed that AIs are quite sensitive to users' inputs (Liu et al. (2023), Brand et al. (2023), Binz and Schulz (2023)). For example, Brand et al. (2023), who adopted ChatGPT to conduct marketing research, found that GPT is sensitive to the phrasing of queries in their empirical work. When querying GPT with a list of options, they found that GPT is more likely to choose the first option. Denny et al. (2023) also indicated that sending proper prompts is critical for the performance of Copilot.

Related studies on the issues of homogenization and bias A few studies have a focus related to the homogenization issue (Saatci and Wilson (2017), Chaney et al. (2018), Shumailov et al. (2023)). They mainly focus on how the technical algorithm design of an AI may reduce the diversity of AI-generated content. For example, Shumailov et al. (2023) observed that the tails of the original content distribution disappear when AIs are successively trained from AI-generated content (they call it model collapse). They analyze how the functional approximation error and sampling error can contribute to the process of model collapse. In addition, a similar homogenization issue has also been found in the literature of recommendation systems. For instance, by using a simulation, Chaney et al. (2018) demonstrate that a feedback loop, where a recommendation system is trained on data from users already exposed to AI recommendations, may homogenize user behavior. On the other hand, the bias issue of AI has also been shown (Hartmann et al. (2023), Rozado (2023), Motoki et al. (2023)). For example, Rozado (2023) implemented 15 different political orientation tests to ChatGPT. The author found that ChatGPT's answers manifested a preference for left-leaning opinions in 14 of the 15 tests.

This paper considers these societal issues from a human-centric perspective. Instead of analyzing AI's inherent technical design or training process, we concentrate on how the issues of homogenization and bias may be influenced by human-AI interactions. In our setting, the issues would disappear if every user is able to provide infinite information. Our results highlight the importance of improving the usability of AIs to mitigate the issues of homogenization and bias.

Related studies on the modeling approach. Regarding the modeling approach, our model is primarily related to the framework of information design (Kamenica and Gentzkow (2011)) and costly persuasion (Gentzkow and Kamenica (2014)), the theory of rational inattention (Sims (2003)), as well as the interpretation of LLMs with Bayesian inference (Wei et al. (2021), Xie et al. (2022)). The user's decision is modeled as an information design process. The sender (i.e., the user) sends a signal to the receiver (i.e., the AI) to inform the receiver about a true state (i.e., the user's preference). The utility of the sender is determined by the receiver's decision (i.e., the AI's output). Additionally, we employ the framework of costly persuasion (Gentzkow and Kamenica (2014)) and the theory of rational inattention (Sims (2003)) to model the user's communication cost when sending the signal. In particular, we follow the standard way in the literature to model the cost of information as the expected reduction in entropy. This assumption can also be found in other modeling papers, such as the cognitive cost defined in Boyacı et al. (2023). Note that we assume the reduction in entropy is relative to the population distribution of users' preferences (see Section 2) instead of AI's prior. As Gentzkow and Kamenica (2014) suggested, the reduction in entropy can be defined relative to any proper fixed reference belief. So we use the population distribution of users' preferences as the fixed reference belief to indicate the communication cost is independent of AI's prior but relevant to how difficult to distinguish her preference from the others. Furthermore, we model the AI's behavior as a Bayesian inference (Wei et al. (2021), Xie et al. (2022)). For instance, Xie et al. (2022) interpret that the learning of an LLM can be viewed as an implicit Bayesian inference. The prior of the LLM is formulated during pretraining. Conditional on a prompt, the LLM characterizes a posterior distribution to make an output.

EC.4. The description of the simulation for the self-training loop.

In this section, we explain the simulation process of the self-training loop depicted in Section 3.3. The detailed pseudo code is provided in Algorithm 1, Algorithm 2, Algorithm 3, and Algorithm 4. Let $\phi(\cdot)$ denote the probability density function of $N(0,1)$. To construct Figure 4, we use $\mu_\theta = 0, \sigma_\theta = 1, N = M = 1000, n = 400$.

Algorithm 1 is the main algorithm that runs the simulation. There are two points we want to highlight. First, for computational tractability, we discretize the population distribution of θ. To this end, we evenly select M points from the range $[\mu_\theta - 3\sigma_\theta, \mu_\theta + 3\sigma_\theta]$ which covers 99.7% of the sample space of $N(\mu_\theta, \sigma_\theta^2)$. These selected points constitute the support for the population distribution of θ, denoted by S. And the probability mass function is given by $\pi_\theta(\theta) = \phi(\theta)/\sum_{x\in S}\phi(x), \ \forall \theta \in S$. Second, at the end of each iteration, the AI's prior is updated by the AI outputs. That is, the AI's prior is replaced by the distribution of θ_A. This corresponds to the self-training loop in which the AI perfectly learns the AI content generated in the previous iteration, so that the distribution of the AI outputs overrides the AI's prior.

Algorithm 2 is used to produce the AI output given the information sent by a user, as depicted in Section 2.

Algorithm 3 is used to compute the posterior distribution with respect to the population distribution, π_θ, given q. It helps us to compute the mutual information e given σ_q in Algorithm 4.

Algorithm 4 is used to compute the utility loss given σ_q and θ. Note that we compute e by its definition $e \triangleq H(\theta) - E[H(\theta|q)]$.

Algorithm 1 The steps of the simulation model

Input: $\mu_\theta, \sigma_\theta$, sample size N, iteration number n, support size M.

Output: $\theta^{\star,i}_{A,j}$, $\forall i \in \{1,2,\ldots,n\}$, $\forall j \in \{1,2,\ldots,N\}$.

Discretize Normal distribution: Evenly select M points from $[\mu_\theta - 3\sigma_\theta, \mu_\theta + 3\sigma_\theta]$ as the support of θ, denoted by S. And compute $\pi_\theta(\theta) = \phi(\theta)/\sum_{x\in S}\phi(x),\ \forall \theta \in S$ as the population distribution of θ.

Generate users' preferences: Randomly sample $\{\theta_1, \ldots, \theta_N\}$ from S with replacement as the set of users' preferences.

Initialize the AI's prior: $\pi^1_A(\theta) = \pi_\theta(\theta),\ \forall \theta \in S$

for $i = 1,2,\ldots,n$ **do**

  for $j = 1,2,\ldots,N$ **do**

    Find the optimal $\sigma^{\star,i}_{q,j} = \arg\min_{\sigma_q} l(\sigma_q|\theta_j, \pi^i_A, \pi_\theta, S, \gamma)$ (Algorithm 4)

    Sample q^i_j from $N(\theta_j, \sigma^{\star,i}_{q,j})$

    Output $\theta^{\star,i}_{A,j} = \theta_A(\pi^i_A, q^i_j, \sigma^{\star,i}_{q,j}, S)$ (Algorithm 2)

  end for

  Compute the distribution of $\theta^{\star,i}_{A,j}$ and use it as the new AI prior to the next iteration.

  for $m = 1,2,\ldots,M$ **do**

    $\forall \theta \in S, \pi^{i+1}_A(\theta) = \sum_{j=1}^N \mathbf{1}_{\theta^{\star,i}_{A,j}=\theta}/N.$

  end for

end for

Algorithm 2 Output θ_A

Input: π_A, q, σ_q, S

Output: θ_A

Compute the likelihood: $\forall \theta \in S, p(q|\theta) = \phi((q-\theta)/\sigma_q)$

Compute the posterior with respect to π_A: $\forall \theta \in S, p(\theta|q) = \dfrac{p(q|\theta)\pi_A(\theta)}{\sum_{x\in S} p(q|x)\pi_A(x)}$.

Compute the posterior mean: $\theta_A = \sum_{\theta\in S} \theta \cdot p(\theta|q)$

Algorithm 3 Posterior with respect to π_θ

Input: q, π_θ, σ_q, S

Output: $p_\theta(\cdot|q)$

Compute the likelihood: $\forall \theta \in S, p(q|\theta) = \phi((q-\theta)/\sigma_q)$

Compute the posterior with respect to π_θ: $\forall \theta \in S, p_\theta(\theta|q) = \dfrac{p(q|\theta)\pi_\theta(\theta)}{\sum_{x\in S} p(q|x)\pi_\theta(x)}$.

Algorithm 4 Compute the utility loss l

1: **Input**: σ_q, θ, π_A, π_θ, S, γ

2: **Output**: l

3: Compute the fidelity error $e = \int [\theta_A(\pi_A, q, \sigma_q, S) - \theta]^2 \phi((q-\theta)/\sigma_q) dq$.

4: Compute the mutual information where $p_\theta(\cdot|q)$ is given by Algorithm 3

$$I = -\sum_{\theta \in S} \pi_\theta(\theta) \log(\pi_\theta(\theta)) + \int \sum_{\theta \in S} p_\theta(\theta|q) \log(p_\theta(\theta|q)) \phi((q-\theta)/\sigma_q) dq$$

5: Compute $l = e + \gamma I$

EC.5. Extensive explanation of Proposition 1: Decomposition of the fidelity error

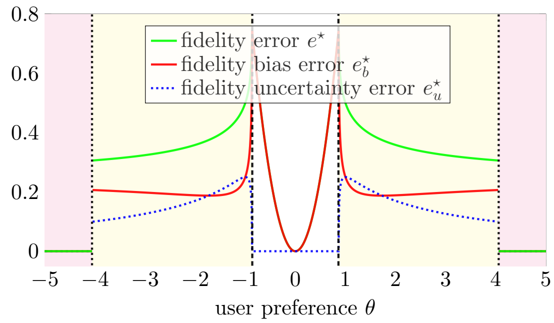

Figure EC.2 The black dashed vertical lines are at $d(\theta) = \tau_d$, and the black dotted vertical lines are at $d(\theta) = \tau_a$. The white region indicates the users who simply accept the default output; the yellow region indicates the users interacting with the AI by sending information; the red region indicates the users without using AI. We use $\mu_\theta = 0, \sigma_\theta = 1, \gamma = 1, \Gamma = 1.4$.

To further understand the variation of fidelity error shown in Proposition 1 for the users with $d(\theta) < \tau_a$, we decompose their the fidelity error into a bias and a variance term, $e^\star = Var(\theta^\star|\theta) + [E(\theta^\star|\theta) - \theta]^2$, as introduced in Section 2. Again, we call $Var(\theta^\star|\theta)$ the *fidelity uncertainty error* denoted by $e_u^\star$, and $[E(\theta^\star|\theta) - \theta]^2$ the *fidelity bias error* denoted by $e_b^\star$. This decomposition is depicted in Figure EC.2. We can see that for the users with $d(\theta) < \tau_d$, the fidelity bias error e_b largely contributes to the fidelity error since they accept the AI's default output without sending any informative signal. At the point of $d(\theta) = \tau_d$, the user starts providing information, leading to a decrease in e_b but an increase in e_u. As the uniqueness further grows, they share more information, resulting in lower fidelity errors. This reduction is primarily driven by the decrease in e_u, since providing more information effectively reduces the noise of the communication but hardly eliminates the inherent difference between the mean and their actual preferences. We formalize this observation in the following Proposition EC.2.

Proposition EC.2. *For users with $d(\theta) < \tau_a$,*

1. *The fidelity uncertainty error $e_u^\star$ is zero when $d(\theta) \le \tau_d$, then increases, and finally decreases in $d(\theta)$.*
2. *The fidelity bias error $e_b^\star$ increases when $d(\theta) \le \tau_d$, then decreases and finally increases in $d(\theta)$.*

Proof of Proposition EC.2. For users with $d(\theta) < \tau_a$, by definition, $e_u(\theta, \sigma_q) = Var(\theta_A|\theta)$ and $e_b(\theta, \sigma_q^2) = [E(\theta_A|\theta) - \theta]^2$. As what we did in the proof of Proposition EC.1 and Lemma EC.3, we can show $e_u(\theta, \sigma_q) = w(1-w)\sigma_\theta^2$ and $e_b(\theta, \sigma_q^2) = w^2(\mu_\theta - \theta)^2$, where $w = \frac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$. Thus, $e_u(\theta, \sigma_q^\star(\theta)) = w^\star(\theta)(1 - w^\star(\theta))\sigma_\theta^2$ and $e_b(\theta, \sigma_q^\star(\theta)) = w^{\star 2}(\theta)(\mu_\theta - \theta)^2$, where $w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ given by Lemma EC.1.

Then,

1. Fidelity uncertainty error:

   We know that $w^\star(\theta) = 1$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, and $w^\star(\theta) < 1$ for $|\mu_\theta - \theta| \ge \tau_d(\gamma, \sigma_\theta)$. Thus, $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, and $e_u(\theta, \sigma_q^\star(\theta)) > 0$ for $|\mu_\theta - \theta| \ge \tau_d(\gamma, \sigma_\theta)$.

   When $|\mu_\theta - \theta| \ge \tau_d(\gamma, \sigma_\theta)$,

   $$\frac{\partial e_u(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = \frac{\partial[w^\star(\theta)(1 - w^\star(\theta))\sigma_\theta^2]}{\partial(\mu_\theta - \theta)^2} = \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2}$$

   We know $\frac{\partial I^\star}{\partial(\mu_\theta - \theta)^2} \ge 0$ by Proposition 1 and $I^\star = -\frac{\gamma}{2}\ln w^\star(\theta)$ by Lemma EC.4. These imply $\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} \le 0$. Thus, when $|\mu_\theta - \theta| \ge \tau_d(\gamma, \sigma_\theta)$, the sign of $\frac{\partial e_u(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2}$ depends on $(1 - 2w^\star(\theta))$. If $(1 - 2w^\star(\theta)) < 0$ for small $|\mu_\theta - \theta|$, then $e_u(\theta, \sigma_q^\star(\theta))$ first increases and then decreases in $|\mu_\theta - \theta|$; if $(1 - 2w^\star(\theta)) \ge 0$ for any $|\mu_\theta - \theta|$, monotonically decreases in $|\mu_\theta - \theta|$. (Notice that $(1 - 2w^\star(\theta))$ is always positive for sufficiently large $|\mu_\theta - \theta|$, because $w^\star(\theta) \to 0$ as $|\mu_\theta - \theta| \to \infty$.)

   Hence, we either have $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, first increases and then decreases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| \ge \tau_d(\gamma, \sigma_\theta)$; or $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, then there is a jump at $|\mu_\theta - \theta| = \tau_d(\gamma, \sigma_\theta)$ ($e_u(\theta, \sigma_q^\star(\theta))$ jumps to a positive value), and then $e_u(\theta, \sigma_q^\star(\theta))$ monotonically decreases in $|\mu_\theta - \theta|$.

2. Fidelity bias error:

   We know that $w^\star(\theta) = 1$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, so $e_b(\theta, \sigma_q^\star(\theta)) = (\mu_\theta - \theta)^2$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, which is increasing in $|\mu_\theta - \theta|$.

   At $|\mu_\theta - \theta| = \tau_d(\gamma, \sigma_\theta)$, since $w^\star(\theta) < 1$ is optimal, we have $e_b(\theta, \sigma_q^\star(\theta)) = e(\theta, \sigma_q^{\star 2}(\theta)) - e_u(\theta, \sigma_q^\star(\theta)) < e(\theta, \sigma_q^{\star 2}(\theta)) < (\mu_\theta - \theta)^2$. Thus, $e_b(\theta, \sigma_q^\star(\theta))$ decreases at $|\mu_\theta - \theta| = \tau_d(\gamma, \sigma_\theta)$.

When $|\mu_\theta - \theta| > \tau_d(\gamma, \sigma_\theta)$,

$$
\begin{aligned}
\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} &= \frac{\partial[w^{\star 2}(\theta)(\mu_\theta - \theta)^2]}{\partial(\mu_\theta - \theta)^2} \\
&= 2(\mu_\theta - \theta)^2 w^\star(\theta) \frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + w^{\star 2}(\theta) \\
&= w^\star(\theta)\left[2(\mu_\theta - \theta)^2 \frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + w^\star(\theta)\right] \\
&\text{Substitute Equation (EC.4) into the above equation} \\
&= w^\star(\theta)\left[2(\mu_\theta - \theta)^2 \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} + w^\star(\theta)\right] \\
&\text{where } \Delta = \sigma_\theta^4 + 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2) \\
&\text{With some simplifications} \\
&= w^\star(\theta) \cdot \frac{\sigma_\theta^2[((\mu_\theta - \theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma((\mu_\theta - \theta)^2 - \sigma_\theta^2)]}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}
\end{aligned}
$$

Let $d(\theta) = |\mu_\theta - \theta|$. Then,

$$
\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2} = w^\star(\theta) \cdot \frac{\sigma_\theta^2[(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2)]}{4\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2}
$$

Now, we want to show that when $d(\theta) \geq \tau_d(\gamma, \sigma_\theta)$ and $d(\theta)$ is finite, $\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$. This is equivalent to showing $\frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2}$ has at most one zero point with respect to $d(\theta)$. Let $\widehat{d(\theta)}$ denote a solution of $d(\theta)$ such that $\frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2} = 0$.

First, let the nominator $(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2) = 0$, we get:

$$
\begin{aligned}
& (d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) = 4\gamma(d(\theta)^2 - \sigma_\theta^2) \\
\Longrightarrow\ & (d(\theta)^2 + \sigma_\theta^2)\sqrt{\Delta} = 4\gamma(d(\theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2) \\
\Longrightarrow\ & \sqrt{\Delta} = \frac{4\gamma(d(\theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2)}{d(\theta)^2 + \sigma_\theta^2} \\
\Longrightarrow\ & 1 = \frac{4\gamma(d(\theta)^2 - \sigma_\theta^2)}{(d(\theta)^2 + \sigma_\theta^2)^2} + \frac{2\sigma_\theta^2}{d(\theta)^2 + \sigma_\theta^2} \\
\Longrightarrow\ & (d(\theta)^2 + \sigma_\theta^2)^2 = 4\gamma(d(\theta)^2 - \sigma_\theta^2) + 2\sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2) \\
\Longrightarrow\ & (d(\theta)^2 - \sigma_\theta^2)(d(\theta)^2 + \sigma_\theta^2 - 4\gamma) = 0
\end{aligned}
$$

So the candidates of $\widehat{d(\theta)}$ are $\widehat{d(\theta)} = \sigma_\theta$, and $\widehat{d(\theta)} = \sqrt{4\gamma - \sigma_\theta^2}$ if $4\gamma \geq \sigma_\theta^2$

Furthermore, by using L'Hôpital's rule, one can get

$$
\lim_{\widehat{d(\theta)} \to \sigma_\theta} \frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2} = \frac{2\sigma_\theta^2\gamma - 4\gamma^2}{\sigma_\theta^6}
$$

which is zero if and only if $\sigma_\theta^2 = 2\gamma$. This means when $\sigma_\theta^2 \neq 2\gamma$, there is no real $\widehat{d(\theta)}$ if $4\gamma < \sigma_\theta^2$ or $\widehat{d(\theta)} = \sqrt{4\gamma - \sigma_\theta^2}$ if $4\gamma \geq \sigma_\theta^2$.

And when $\sigma_\theta^2 = 2\gamma$, $\widehat{d(\theta)} = \sqrt{4\gamma - \sigma_\theta^2} = \sigma_\theta$, so we also only have one solution for $\widehat{d(\theta)}$.

Thus, $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$.

In addition, if $d(\theta) > \sigma_\theta$, $(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2) > (d(\theta)^2 - \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\gamma(d(\theta)^2 - \sigma_\theta^2) = (d(\theta)^2 - \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2 - 4\gamma)$, which is positive if $d(\theta) > \max\{\sigma_\theta, \sqrt{|4\gamma - \sigma_\theta^2|}\}$. This means that for any $d(\theta) > \max\{\sigma_\theta, \sqrt{|4\gamma - \sigma_\theta^2|}\}$, $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ is positive. Because we have shown $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$, the intermediate value theorem implies that when $|\mu_\theta - \theta| > \tau_d(\gamma, \sigma_\theta)$, $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ is either always positive, or negative for small $|\mu_\theta - \theta|$ and then positive for large $|\mu_\theta - \theta|$.

Hence, we either have $e_b(\theta, \sigma_q^\star(\theta))$ first increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, then decreases and finally increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| \geq \tau_d(\gamma, \sigma_\theta)$; or $e_b(\theta, \sigma_q^\star(\theta))$ first increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| < \tau_d(\gamma, \sigma_\theta)$, then there is a jump at $|\mu_\theta - \theta| = \tau_d(\gamma, \sigma_\theta)$ ($e_b(\theta, \sigma_q^\star(\theta))$ jumps to a smaller value), and then $e_b(\theta, \sigma_q^\star(\theta))$ monotonically increases in $|\mu_\theta - \theta|$.

□

EC.6. More Detailed Version of Theorem 1.

In this section, we present a more detailed description of Theorem 1 :

THEOREM 1 *(Full version) When $\Gamma \to +\infty$, the variance of the population output is lower than the variance of the population preferences, $Var(\theta^\star) < Var(\theta)$, and strictly decreases in the cost of human-AI interactions γ. When $\Gamma < +\infty$, $\lim_{\gamma \to 0} Var(\theta^\star) = Var(\theta)$ and $\lim_{\gamma \to +\infty} Var(\theta^\star) < Var(\theta)$. In addition, $Var(\theta^\star) < Var(\theta)$ if $\gamma \geq \sigma_q^2/2$ or $\Gamma \leq \hat{\Gamma}$ or $\Gamma \geq \tilde{\Gamma}$ for some $\hat{\Gamma} > 0, \tilde{\Gamma} > 0$.*

The full proof is provided in Section EC.2.2. The last sentence in this detailed version is the additional part compared with the version that we presented in the main text. In particular, we show that the population variance of the output is strictly less than the population variance of the preferences if γ is sufficiently large or Γ is outside an interval $(\hat{\Gamma}, \tilde{\Gamma})$.

In Figure EC.3, we show why it is possible that the population variance of the output can be larger than the population variance of the preferences when $\gamma < \sigma_q^2/2$ and $\Gamma \in (\hat{\Gamma}, \tilde{\Gamma})$. By the tower property of conditional expectation, we know $Var(\theta^\star) = E[E[(\theta^\star - \mu_\theta)^2|\theta]]$ (Notice that $E[\theta^\star] = \mu_\theta$ is shown in the proof of Theorem 1). So if $E[(\theta^\star - \mu_\theta)^2|\theta] < (\theta - \mu_\theta)^2$, we must have $Var(\theta^\star) < Var(\theta)$. However, it is possible that $E[(\theta^\star - \mu_\theta)^2|\theta] \geq (\theta - \mu_\theta)^2$ for some θ whose $d(\theta)$ are close to τ_d. Since τ_d is the root of a transcendental equation, it is complicated to find the closed form of this region. Despite this possibility, it is actually hard to find a scenario such that $Var(\theta^\star) > Var(\theta)$ in our numerical tests.

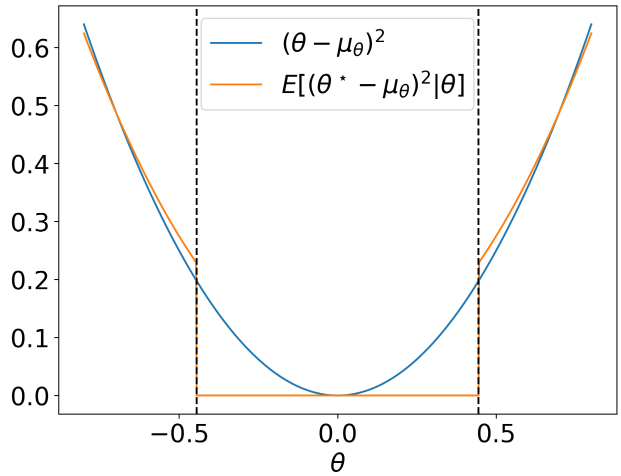

Figure EC.3 **We use $\mu_\theta = 0, \sigma_\theta = 1, \gamma = 0.1$. The black dashed vertical lines are at $d(\theta) = \tau_d$. The orange curve is above the blue curve for some θ with $d(\theta) > \tau_d$ but close to τ_d, showing that $E[(\theta^\star - \mu_\theta)^2|\theta] \geq (\theta - \mu_\theta)^2$ for these user θ.**

Intuitively, the users with $d(\theta) > \tau_d$ but close to τ_d will send a small amount of information. Since the information is always noisy, it will also add more randomness and uncertainty to the outputs. So these users have a higher $E[(\theta^\star - \mu_\theta)^2|\theta]$ and can "contribute" more to the population variance of outputs. However, there are also many users who simply accept the default outputs (i.e., $d(\theta) < \tau_d$) and users with preferences that are far from the mean. They have a lower $E[(\theta^\star - \mu_\theta)^2|\theta]$, thereby reducing the population variance of outputs. These two counter-forces interact with each other, leading to a change in the variance.

References for Appendix

Agrawal A, Gans J, Goldfarb A (2018) *Prediction, Judgment, and Complexity: A Theory of Decision-Making and Artificial Intelligence*, 89–110 (University of Chicago Press).

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120(6).

Boyacı T, Canyakmaz C, de Véricourt F (2023) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* .

Brand J, Israeli A, Ngwe D (2023) Using GPT for market research. *Available at SSRN 4395751* .

Chaney AJB, Stewart BM, Engelhardt BE (2018) How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *Proceedings of the 12th ACM Conference on Recommender Systems*, 224–232, RecSys '18.

Chen N, Hu M, Li W (2022) Algorithmic decision-making safeguarded by human knowledge. *Available at arXiv:2211.11028* .

Dai T, Singh S (2023) Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Available at SSRN 3987454* .

de Véricourt F, Gurkan H (2023) Is your machine better than you? you may never know. *Management Science* .

Denny P, Kumar V, Giacaman N (2023) Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 1136–1142, SIGCSE 2023.

Eloundou T, Manning S, Mishkin P, Rock D (2023) GPTs are GPTs: An early look at the labor market impact potential of large language models. *Available at arXiv:2303.10130* .

Gentzkow M, Kamenica E (2014) Costly persuasion. *American Economic Review* 104(5):457–62.

Hartmann J, Schwenzow J, Witte M (2023) The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *Available at arXiv:2301.01768* .

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *American Economic Review* 101(6):2590–2615.

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55(9).

Motoki F, Neto VP, Rodrigues V (2023) More human than human: measuring chatgpt political bias. *Public Choice* .

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

Pugh CC (2015) *Real Mathematical Analysis* (Springer Cham).

Rozado D (2023) The political biases of ChatGPT. *Social Sciences* 12(3).

Saatci Y, Wilson AG (2017) Bayesian gan. *Advances in Neural Information Processing Systems*, volume 30 (Curran Associates, Inc.).

Sallam M (2023) Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* 11(6).

Shumailov I, Shumaylov Z, Zhao Y, Gal Y, Papernot N, Anderson R (2023) The curse of recursion: Training on generated data makes models forget. *Available at arXiv:2305.17493* .

Sims CA (2003) Implications of rational inattention. *Journal of Monetary Economics* 50(3):665–690.

Wei C, Xie SM, Ma T (2021) Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, eds., *Advances in Neural Information Processing Systems*.

Xie SM, Raghunathan A, Liang P, Ma T (2022) An explanation of in-context learning as implicit bayesian inference. *International Conference on Learning Representations*.