DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 23-03

Over-Persuasion and Mechanism Design

Mengqi Zhang
University of Colorado Boulder

October 16, 2023

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

Over-Persuasion and Mechanism Design

Mengqi Zhang*

October 16, 2023
Latest Version

Abstract

The effect of Bayesian persuasion depends on the Receiver's prior belief. When the persuasive message must be broadcast, an optimal persuasion strategy chosen for one Receiver may over-persuade some other Receivers with different prior beliefs, resulting in a negative persuasion value. We design a mechanism in which contract transfer is contingent on signal realization. The agents have the option of accepting the contract and revealing the persuasion signal, or rejecting it and maintaining their prior beliefs. This mechanism, which also aims to improve persuasiveness, is designed to discourage certain Receivers from engaging in persuasion, as opposed to the conventional approach, which encourages Receivers to accept persuasive messages tailored to their types from a menu. By leveraging the conformation bias based on the Receiver's heterogeneous prior beliefs, incentive compatibility is easier to implement in this signal-contingent mechanism. Additionally, the signal-contingent mechanism allows a persuasive message be broadcast to all Receivers, making it more practical in the real world.

*Department of Economics, University of Colorado Boulder. Email: mengqi.zhang@colorado.edu

To gild refined gold, to paint the lily

.....

Is wasteful and ridiculous excess.

—Shakespeare, King John, Act 4, Scene 2

1 Introduction

The Backfire effect has been documented in the fields of marketing, politics, and healthcare, as attempts at persuasion can sometimes result in a counterproductive outcome (Bickart, 1993; Nyhan and Reifler, 2010; Lewandowsky et al., 2012). Typically, this effect is attributed to confirmation bias.[1] A study (Nyhan and Reifler, 2015) examining whether correcting the misconception that flu vaccines cause influenza could help improve individuals' willingness to accept the vaccines, however, may suggest otherwise. The experimental results demonstrate that, although a treatment group experienced a backfire effect and a significant decline in vaccine acceptance, the persuasion did effectively correct their misconceptions. Why would an effective persuasion result in a less desirable outcome?

This phenomenon, formally defined in this paper as *Over-Persuasion*, is not a concern within the canonical Bayesian persuasion framework. In the canonical model, the Receivers share a prior belief known to the Sender. Therefore, the Sender only persuades when it is expected to produce a desired outcome. When the Receivers possess heterogeneous prior beliefs, the optimal persuasion strategy for some may cause others to reverse their ex-ante decisions that the Sender desires. Providing more information about the actual side effects of the vaccine may be the most effective way to increase acceptance among those whose only concern about the vaccine is that it causes influenza. This persuasion strategy, which is broadcast to all Receivers, may exacerbate the concerns of those who are skeptical about other side effects, thereby contributing to a counterproductive outcome. When the

[1] A behavioral phenomenon in which agents tend to believe the signal that confirms their prior beliefs rather than one that goes against them. See Benjamin (2019) for a detailed review and discussion.

Receiver's prior belief is unknown, providing additional information may be gilding the lily.

In scenarios, mechanisms are designed to motivate Receivers to report their private information truthfully so that strategies can be tailored to best persuade Receivers of heterogeneous types (Kolotilin et al., 2017, Pham, 2023). However, it is unclear whether incentive compatibility will remain implementable when these scenarios are extended to a broader context that includes a discussion of the over-persuasion issue. Due to the nature of persuasion, in order for the Receivers to report their types and for the Sender to convey persuasive messages tailored to the Receiver's type, private channels must also be established, which is often impractical. This is especially true for political and health campaigns in which persuasive messages must be disseminated to an audience of millions to billions. In fact, even with a small audience base, preparing different pitch scripts for different audiences can be costly. Moreover, to ensure that these mechanisms are perfectly implemented, the persuasive message targeting specific Receivers must not reach other Receivers.

Given the challenges of tailoring the persuasive message to the Receiver's type, under what conditions does over-persuasion occur when the Sender must broadcast a persuasive message to a group of Receivers with heterogeneous prior beliefs? Are there any conditions that permit the design of a straightforward and practical mechanism to restore the persuasiveness loss caused by the over-persuasion issue while allowing the persuasive message to be broadcast?

In this study, we generalize analyses that were previously conducted only in specific contexts. These analyses indicate that if different prior beliefs have the same optimal support for the posterior belief distribution, they may require different information structures to achieve the same persuasion objective. Because of this incongruence among different prior beliefs, the persuasion strategy that is optimal for some beliefs may produce negative persuasion values for others. If the degree of incongruence is high, such as when some prior

beliefs support the Sender's most desired action, making persuasion unnecessary for them, it may be impossible to find a non-trivial persuasion strategy that produces non-negative persuasion values for them and other prior beliefs simultaneously.

In these situations where the over-persuasion issue arises, the Sender is constrained by the need to maintain certain prior beliefs that he desires the Receivers to hold. This constraint compromises the persuasiveness of the broadcast persuasive message when the Sender attempts to persuade some other Receivers to change their prior beliefs. Therefore, the Sender would benefit from the implementation of a mechanism capable of removing this constraint. Such a mechanism should effectively discourage Receivers from processing the broadcast persuasive messages if their prior beliefs lack persuasion values or are incongruent with those whose persuasion values are high.

To meet these requirements, we design a contract where the Receiver's transfer is contingent on the signal realization. In this mechanism, the Sender offers a same experiment and contract with transfers to all of the Receivers. The Receivers can either accept the contract and reveal the signal from the experiment, or reject the contract and maintain their prior beliefs. The fundamental concept of a signal-contingent mechanism is to exploit the varying degrees of confirmation biases exhibited by Receivers, which arise from their heterogeneous prior beliefs. According to Bayes' Rule, the Receivers with different prior beliefs will hold different perceptions regarding the probability of the occurrence of a given signal, even when presented with the same experiment. Specifically, a Receiver with a stronger belief in a particular state is less likely to believe the presence of a signal against it. Hence, the Receivers with different prior beliefs interpret the same reward or punishment in the transfers differently. By leveraging these rewards or punishments, the Sender may be able to manufacture differentiated motivations for Receivers to reveal the signals.

The implementation of the mechanism dependent on signal realization is widely used in practice. For example, to encourage certain potential buyers to skip the house inspec-

tion and make the purchase directly, the homeowner could opt to adjust the price of the home based on the inspection results. She charges the buyer more when the results are good and less when the results are bad. The flat information fee is a special and more common form of signal-contingent mechanism; the buyer could waive an inspection in exchange for a discount or to pay extra for permission to inspect before closing the deal. The signal-contingent mechanism can also be employed in an inter-temporal context to prevent information spillover. The Senders may, for instance, choose a time-sensitive pricing strategy, such as an early-bird discount. In such strategies, all customers are charged a premium once the information is revealed or when the product is proven to be successful on the market. In this context, the signal-contingent mechanism is analogous to the discount factor in the dynamic information disclosure design to address the heterogeneous prior belief issue (Au, 2015).

This study contributes to several topics, ranging from application to theory. When the Sender is uncertain or ambiguous about the Receiver's prior belief, she is more reluctant to disclose additional information out of concern for the possibility of over-persuasion. Many persuasion strategies can only be restored by the design of appropriate mechanisms. These mechanisms not only preserve the Sender's payoff but also, to some extent, align her interests with social benefits. More importantly, the mechanisms introduced in this study do not require a private communication channel, allowing for the effective dissemination of persuasive messages. This could make it easier to promote public campaigns such as vaccination acceptance.

In this study, we generalize analyses that were previously conducted only in specific contexts. Many previous studies assume the Receiver's full attention or acceptance of the persuasive messages, which downplays the importance of analyzing their motivation in persuasion games. We develop a generalized framework to analyze the Receiver's motivation to accept the belief-changing signal. This framework may provide methodological support for a growing body of research on inattention in persuasion games, where Receivers may not

be influenced by the persuasion signal if it is too costly to process. Furthermore, the implications of this study may contribute to other theoretical studies. In a robust persuasion design, a maxmin conjecture is commonly used to address the ambiguity of the Receiver's type in the persuasion game. This environment also poses a risk of over-persuasion, which could be substantiated by a maxmin approach. Consequently, a Sender may wish to design a mechanism that prevents certain Receivers from receiving additional information in order to improve efficacy in the context of robust persuasion, which may not only reduce the risk of over-persuasion but also lower the degree of ambiguity in the persuasion game.

1.1 Related Literature

This study is established on the canonical persuasion framework introduced by Kamenica and Gentzkow (2011). Alonso and Câmara (2016) extended the framework to include heterogeneous prior belief but the heterogeneity exists between the Sender and the Receiver. Laclau and Renou (2017) discussed the equilibrium in a persuasion game where audiences have heterogeneous prior beliefs. They also compared targeted and public persuasion but did not specify how mechanism design can granulate the public audience into targeted ones. Gitmez and Molavi (2022), and Boyaci et al. (2022) follow this framework to discuss the implications of the Receiver's belief being heterogeneous. Boyaci et al. (2022) pointed out that heterogeneous prior beliefs could modify the Sender's optimal persuasion strategy to be more conservative. Our paper complements their research by analyzing the persuasiveness loss due to this change and emphasizes the conditions that could lead to an over-persuasion issue.

The heterogeneity among players in the persuasion game, especially when their types are private information warrants the mechanism design as a resolution. The Sender may design an incentive-compatible persuasion mechanism where she picks the information structures according to the Receiver's self-report type and directly recommend actions to the Receivers (Kolotilin et al., 2017; Guo and Shimaya, 2019). As the closest study to

ours, Pham (2023) discusses a mechanism design based on the Receiver's heterogeneous prior beliefs in a specific context. Our research differs from these relevant studies in that we simply discourage certain Receivers from being persuaded, whereas these studies encourage all Receivers to choose various persuasion strategies. Due to this distinction, our mechanism does not require a private channel between the Sender and the Receivers to communicate types, signals, or recommendations. We emphasize the role of the information fee in establishing incentive compatibility, thereby making the contract contingent on the realized signal or the Receivers' action rather than their types. Its simplicity allows for theoretical analysis on a more generalized framework and makes it applicable in many occasions such as public health campaigns where persuasive messages need to be broadcast. To some extent, different mechanism designs necessitate certain predetermined conditions to ensure incentive compatibility, which are only satisfied on specific occasions. Therefore, the aforementioned distinctions establish our mechanisms to complement rather than challenge relevant prior research.

2 A Simple Example

This section gives a simple example that illustrate the main idea of this study. Consider a scenario where a homeowner sells a house of uncertain value for 1.2 (million dollars). The value of the house is either 1 or 1.3, corresponding to two possible states $\omega \in \{h, l\}$, respectively. The homeowner believes that h has a $\frac{7}{12}$ chance of being the actual state. Depending on whether he is of type a or b, a prospective buyer's prior belief that h is the actual state may be $p_a(h) = \frac{7}{12}$ or $p_b(h) = \frac{9}{12}$. If the buyer has a sufficiently high chance of being type b, the owner will prioritize preserving the buyer b's initial decision regardless of which signal is realized. To realize this design, the Sender should design an information structure such that states h and l send the good signal with probabilities of $\frac{5}{8}$ and $\frac{7}{16}$, respectively. With a posterior belief of $q_b(h) = \frac{2}{3}$, a type b buyer will always buy the house, even if a bad signal is revealed. Since $q_a(h) \geq \frac{2}{3}$ only when the good signal arrives, a type

a buyer has a chance of $\frac{5}{8} \times \frac{7}{12} + \frac{7}{16} \times \frac{5}{12} = \frac{35}{64}$ of purchasing the house from the seller's perspective.

Alternatively, if the buyer must pay a fee of $\frac{7}{120}$ to reveal the signal, the Sender can design a full disclosure information structure and outperform the aforementioned persuasion outcome. A full-disclosure information structure completely reveals the actual state. When the actual state is revealed to be l with a chance of $\frac{3}{12}$ from the perspective of type b buyer, he can avoid a 0.2 loss. On the other hand, the buyer of type a has a $\frac{7}{12}$ chance to avoid missing out on the 0.1 benefit. Consequently, their respective gains from the inspection are $\frac{7}{120}$ and 0.05, discouraging the type b buyer from conducting the inspection while encouraging the type a buyer to reconsider his initial decision following the home inspection. In this case, the buyer of type b maintains his initial decision to proceed with the purchase based on his prior belief. With a probability of $\frac{7}{12} > \frac{35}{64}$, the good signal arrives to advise the type a buyer to reverse his initial decision and purchase the home.[2] As a result, this mechanism design benefits the homeowner, even if the information fee is collected by third parties such as home inspectors or real estate agents.

When the buyer is highly likely to be of type b, a full disclosure information structure can only be implemented if a mechanism can deter the buyer of type b from inspecting the home. A flat-rate information fee can satisfy this requirement only if $0.2(1 - p_b) < 0.1p_a$, where $p_a < \frac{2}{3} < p_b$. A mechanism that allows for different transfer for different signals, on the other hand, is powerful enough to establish desired incentive compatibility and participation constraint for any $p_a < \frac{2}{3} < p_b$ under this circumstance. Let m_g and m_b represent the additional compensation transferred to the Receiver in the event of good and bad signals, respectively. To implement a signal-contingent mechanism on the full disclosure persuasion strategy, only $m_g - m_b > 0.1 - \frac{3p_b - 2}{10(p_b - p_a)}$ and $m_g \geq [(m_g - m_b) - 0.1]p_a$ are required. As long as m_g and m_b are unrestricted, the satisfaction of both conditions can be guaranteed.

[2] As we will demonstrate in the Application section, an optimal mechanism design coupled with a non-full-disclosure persuasion strategy may even outperform this result for this particular case.

3 Model

Our model is established on the canonical framework of a persuasion game. There are Receivers whose shared payoff $u(\alpha, \omega)$ is jointly determined by actions α selected from a compact set $\mathcal{A}$ and the actual state of the world ω selected by Nature from a finite state space Ω. Receivers are uncertain about the actual ω. They need to choose their actions based on their beliefs in $\Delta\Omega$. Each Receiver's own prior belief p is his private information. But the distribution of Receivers' heterogeneous prior beliefs $f(p)$ on a finite support $\mathcal{P} \subset \Delta\Omega$, where $\sum_{p\in\mathcal{P}} f(p) = 1$, is common knowledge among all players in the game.

Also uncertain about the actual ω is the only Sender in the game. Her payoff, denoted as $\sum_{p\in\mathcal{P}} f(p)v(\alpha_p)$, is exclusively determined by the actions taken by the Receivers, where α_p represents the action taken by Receivers with prior belief p. To influence the Receivers' beliefs and actions, the Sender can design a publicly known experiment $\pi : \Omega \to \Delta\mathcal{S}$ that maps the states to the probability distribution of a signal realization s belonging to a finite signal space $\mathcal{S}$. After observing signal s, the Receivers with prior belief p update their posterior belief to $q_p(s)$ based on Bayes' Rule.

The Receivers can choose whether to reveal the signal $\gamma \in \{0, 1\}$. They only observe the signal realization when $\gamma = 1$ is chosen. Assume that the Receivers' choices of γ are verifiable and contractible. Although we assume that sending and revealing the signal incur no cost, the Sender can design a mechanism $\theta : \mathcal{H} \to \Theta \subset \mathbb{R}$, where $\mathcal{H} = \{0, 1\}$ in signal-independent mechanism and $\mathcal{H} = \{0, 1\} \times \mathcal{S}$ in signal-contingent mechanism, to influence the Receivers' decision on $\gamma \in \{0, 1\}$. Without loss of generality, we suppose $\theta = 0$ when $\gamma = 0$. With this mechanism, the Receivers' payoffs are changed to $u(a, \omega, h) = u(a, \omega) + \theta(h)$. Accordingly, the Sender's payoff becomes $\sum_{p\in\mathcal{P}} f(p)\big[v(\alpha_p) - \beta\theta(h_p)\big]$, where $\beta \in [0, 1]$ reflects the loss when the contract value is being transferred between the Sender and the Receivers. If the context permits, for instance when $\theta < 0$, we assume $\beta = 0$. The purpose of this assumption is to minimize the value of the mechanism in order to highlight

the Sender's motivation to design the mechanism solely for persuasion purposes. [3]

In the persuasion game, the Receivers first choose γ based on the Sender's optimal mechanism design of θ and optimal information structure π. Based on their optimal decision of γ^* and the realized signal s, they then optimize the following objective.

$$\max_{\alpha \in \mathcal{A}} \sum_{\omega \in \Omega} \left[(1-\gamma^*) + \gamma^* \frac{\pi(s|\omega)}{\sum_{\omega \in \Omega} \pi(s|\omega) p(\omega)} \right] p(\omega) u(a, \omega) \tag{1}$$

The classical tie-breaking rule is adopted, wherein Receivers choose the actions optimal for the Sender from the set of actions that optimizes (1). If multiple qualified actions are remaining and $\gamma^* = 1$ is chosen, the Receiver will further narrow down the set by selecting the actions that would have optimized (1) had $\gamma^* = 0$ been chosen. Let $\mathcal{A}^*(q) \subset \mathcal{A}$ be the set containing the optimal options of α after the tie-breaking rule is applied.

When deciding the optimal γ^* at the first stage, the Receivers with prior belief p make rational predictions about their potential posterior beliefs, $q_p(\gamma, s) = \gamma q_p(s) + (1-\gamma)p$ and associated optimal action set, $\mathcal{A}^*\big(q_p(\gamma, s)\big)$. Denote $\mathcal{S}' = \{s \in \mathcal{S} | \mathcal{A}^*\big(q_p(\gamma, s)\big) \bigcap \mathcal{A}^*(p) = \emptyset\}$. Additionally, let $\alpha_p^{\gamma, s}$ represent an arbitrary element in $\mathcal{A}^*\big(q_p(\gamma, s)\big)$. The objective for the Receivers at the first stage is as follows.

$$\max_{\gamma \in \{0,1\}} \gamma \left\{ \sum_{s \in \mathcal{S}',\ \omega \in \Omega} \pi(s|\omega) p(\omega) \Big[u(a^{1,s}, \omega) - u(a^{0,s}, \omega) + \theta(h) \Big] \right\}, \tag{2}$$

where $h = \gamma$ in signal-independent mechanism, and $h = (\gamma, s)$ in signal-contingent mechanism.

The term $u(a^{1,s}, \omega) - u(a^{0,s}, \omega)$ in objective (2) suggests that a signal affects the Receiver's payoff from $u(\cdot)$ is solely through its influence on their decisions. If certain signals $s \in \mathcal{S} \backslash \mathcal{S}'$ do not change the Receiver's optimal choice of actions in the second stage, the possible occurrences of these signals should not encourage the Receivers to reveal these signals, as represented by $u(a^{1,s}, \omega)$ and $u(a^{0,s}, \omega)$ having the same weight. For example, a

[3] This setting also reflects the existence of transaction cost. For example, the seller does not receive the entire extra fee paid by the buyer for the home inspection.

signal in favor of the product's quality provides no actual benefit to a consumer who has already decided to buy that product, even if the signal increases the consumer's expected payoff. $\theta\big(h(\gamma)\big)$ in the objective demonstrates that the Sender can modify the Receiver's incentive to reveal the signal by implementing a mechanism. In the event of a tie, we assume that the Receiver always selects $\gamma = 1$.

According to (1) and (2), the choice of $\pi \in \Pi$ and $\theta \in \Theta$ have a direct impact on the Receivers' decisions on γ^*; γ^* along with s that is determined by π and the actual state ω determine the choice of $\alpha_p^{\gamma,s} \in \mathcal{A}^*\big(q_p(\gamma, s)$. Suppose p_v represents the Sender's prior belief. The objective of the Sender is as follows.

$$\max_{\pi\in\Pi,\ \theta\in\Theta} \sum_{p\in\mathcal{P},\ \omega\in\Omega,\ s\in\mathcal{S}} f(p)\pi(s|\omega)p_v(\omega)v\Big(\alpha_p^{\gamma(\theta,\pi),s}\Big) - \beta\theta(h), \tag{3}$$

where $h = \gamma$ in signal-independent mechanism, and $h = (\gamma, s)$ in signal-contingent mechanism.

The game unfolds in the following manner. In the beginning, Nature chooses the actual state, which remains unknown to all players unless the full disclosure information structure is selected. Subsequently, the Sender designed and made public both the mechanism and the information structure. Nature selects the signal based on the information structure and the actual state. The Receivers decide whether or not to reveal the signal. Nature then discloses the signal to those Receivers who have opted to reveal it. Following this, the Receivers' beliefs are updated based on the realized signal and their choice of whether to reveal it. Based on their current beliefs, all Receivers choose their actions optimally. Finally, the game concludes, and the payoff is realized in accordance with the game profile and history.

4 Over-Persuasion

This section focuses on the over-persuasion issue and assumes $\theta = 0$ to examine how this issue, if it exists, impacts the equilibrium in a persuasion game. The root cause of the over-persuasion issue is the heterogeneous beliefs of the Receivers. This, along with the disparity in prior beliefs between the Senders and Receivers, poses a significant challenge to the analysis of the model. Following Kamenica and Gentzkow (2011) and Alonso and Câmara (2016), we simplify and modify the Sender's problem in order to make the model tractable.

4.1 Simplifying the Sender's Problem

Let $\tau_p \in \Delta\Delta\Omega$ and $\tau_v \in \Delta\Delta\Omega$ be distributions over the Receivers' and the Sender's posterior beliefs of the states, respectively. According to Alonso and Câmara (2016), for each Receiver's posterior belief, $q_p \in \Delta\Omega$, there exists a bijection that maps the Sender's posterior $q_v \in \Delta\Omega$ to q_p such that $q_p(\omega) = \frac{q_v(\omega)\frac{p(\omega)}{p_v(\omega)}}{\sum_{\omega\in\Omega} q_v(\omega)\frac{p(\omega)}{p_v(\omega)}}$. For certain persuasion strategy π, if q_v is Bayes plausible, then it follows that q_p is also Bayes plausible. Assuming that reading the revealed signal incurs no cost for the Receivers in the absence of mechanism design, we set $\gamma = 1$ in this section, as we will further explain in Section 5. The Sender's problem of persuading the Receivers with prior belief p can be simplified to

$$\begin{aligned} &\max_{\tau_v} \; E_{\tau_v}\left[v_p(q)\right] \\ &s.t.\; E_{\tau_v}(q) = p_v, \end{aligned} \tag{4}$$

where

$$v_p(q) = v\left[\frac{q(\omega)\frac{p(\omega)}{p_v(\omega)}}{\sum_{\omega\in\Omega} q(\omega)\frac{p(\omega)}{p_v(\omega)}}\right].$$

This simplification reduces the Sender's problem to selecting a Bayes plausible distribution of her own posterior beliefs τ_v to maximize the value function $v_p(q)$. Since selecting τ_v is equivalent to selecting π, we will use τ_v to characterize the persuasion strategy hereafter.

For a certain signal, the Sender and Receivers can hold differing posterior beliefs as a result of their distinct prior beliefs. Nevertheless, their posterior beliefs exhibit some general connections. Specifically, if one state becomes more plausible for the Sender following the realization of a signal, the Receivers should exhibit a similar rationale. In the event that a signal confirms or negates a particular state, it is necessary for the Sender and the Receivers must be certain of the veracity or falsity of said state, respectively. In the absence of any new information, individuals should also maintain their prior beliefs. The connection between $v(q)$ and $v_p(q)$, as formalized by Lemma 1, demonstrates these relationships.

Lemma 1. *Let $\overline{\Delta\Omega}$ be the boundary of $\Delta\Omega$. For a given p, it follows that $v_p(q_p) = v(q_v)$ at $q_v \in \overline{\Delta\Omega}$, $v_p(p) = v(p_v)$ and that q_p is monotonically increasing in q_s.*

Lemma 1 provides a means to characterize the graph of $(q, v_p(q))$ based on the graph of $(q, v(q))$. The graph transformation from $v(q)$ to $v_p(q)$ can be described as follows. Initially, the rubber hypersurface of $v(q)$ is attached to the boundary of $\Delta\Omega$. Next, a fixed point $q = p$ is established on the given hypersurface, and the hypersurface is stretched until the fixed point reaches $q = p_v$. Figure 1 illustrates the relationship between $v(q)$ and $v_p(q)$ when there are only two states $\{\omega_1, \omega_2\}$ and $q(\omega_1)$ is the projection of a two-dimensional simplex q onto a one-dimensional simplex $[0, 1]$.

This transformation allows for the estimation of the shape of different $v_p(\cdot)$. When these graphs are aligned on the same simplex, it becomes straightforward to characterize the impact of a certain persuasion strategy on the Receivers who hold different prior beliefs.

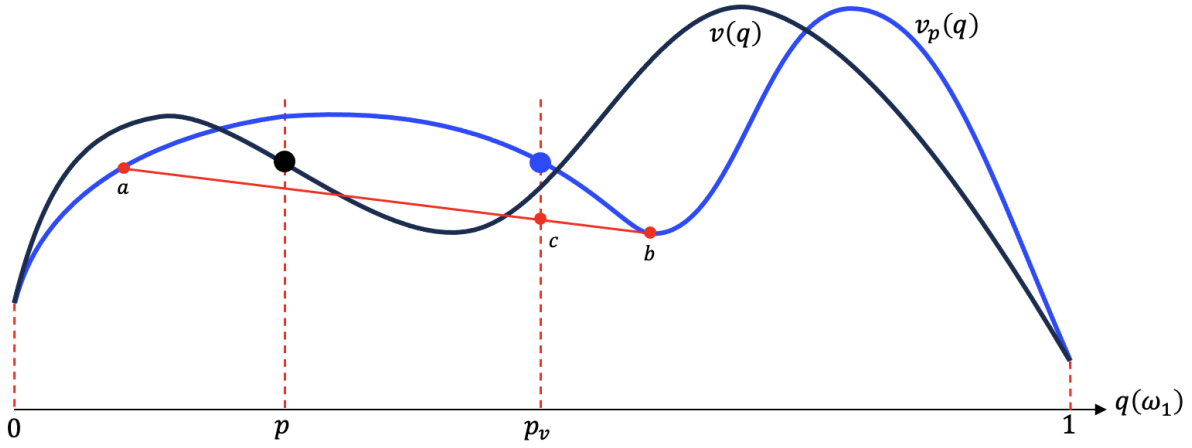

Figure 1: Sender's Problem Simplification

4.2 Heterogeneous Prior Beliefs and Over-persuasion

When the Sender persuades the Receivers, certain information structures may convey signals that cause some Receivers to change their prior choices, thereby changing the Sender's payoff. This change may bring positive persuasion value that benefits the Sender, but it could also result in negative persuasion value, which we formally define as over-persuasion.

Definition 1. *Receivers with prior beliefs p are over-persuaded by a persuasion strategy τ_v if it causes negative persuasion value or specifically,*

$$E_{\tau_v}\left[v_p(q)\right] < v_p(p_v). \tag{5}$$

Also, let $\mathcal{P}(\tau_v) = \left\{p \middle| E_{\tau_v}\left[v_p(q)\right] < v_p(p_v)\right\}$ denote the set of prior beliefs with which Receivers are over-persuaded by τ_v.

Condition (5) is absent in canonical persuasion games wherein the Receivers share a prior belief known to the Sender. It becomes problematic, however, if the Sender is uncertain about the Receivers' prior beliefs or if she must use a single strategy to persuade a group of Receivers with heterogeneous prior beliefs. If a persuasion strategy differs significantly from the one that maximizes the persuasion value of Receivers with prior

belief p, the LHS of inequality (5) is substantially discounted. This discounting effect may result in a negative persuasion value of the Receivers with prior belief p, provided that their optimal persuasion value is sufficiently low.

In Figure 1, any line segment connecting the end points of the curve $v_p(q)$ within an interval that contains p_v represents a Bayes plausible persuasion strategy τ_v. The distance between the line segment and the curve $v_p(q)$ at $q = p_v$ represents the persuasion value of the Receivers with prior belief p when the strategy is given as τ_v. The line segment connecting points a and b in Figure 1 represents a Bayes plausible persuasion strategy. Because point c on the line segment lies below the corresponding point on the curve, this strategy yields a negative persuasion value for Receivers holding p as prior belief. When p is the only belief held by Receivers, this strategy is dominated and will not be adopted. However, in cases where a substantial number of Receivers hold a prior belief p_v, it is possible that this strategy will be selected because it produces the highest persuasion value of Receivers with prior belief p_v. In these cases where τ_v is adopted, the Receivers with prior belief p are over-persuaded.

If some points on the graph of $(q, v_p(q))$ are also on its concave closure, the corresponding prior belief is associated with a zero optimal persuasion value. Some of these beliefs, if held by Receivers in the persuasion game, may make over-persuasion unavoidable under certain conditions. Proposition 1 discusses these conditions and delineates the specific set of prior beliefs with which the Receivers will be over-persuaded by any persuasion strategy.

Proposition 1. *Let* $\mathbf{q}_\tau$ *be the vector space spanned by* $\operatorname{supp}(\tau)$*,* $\tilde{v}(\cdot)$ *be the concave closure of* $v(\cdot)$*. Then given a value function* $v(q)$ *and a persuasion strategy characterized by* τ_v*,* $\mathcal{P}(\tau_v) \neq \emptyset$ *as long as there exists a* $p \in \operatorname{Int}(\Delta\Omega)$*:*

(a) There exists $q \in \Delta\Omega$ *such that* $\tilde{v}_p(q) < \sup_q v_p(q)$*;*

(b) $\mathbf{q}_{\tau_v}$ *has full rank;*

(c) $\left\{q \middle| v_p(q) = \sup_q v_p(q)\right\} \cap \operatorname{Int}(\Delta\Omega) \neq \emptyset$*.*

Let $p \in \hat{\mathcal{P}}$ *if and only if* $v_p(p) = \sup_q v_p(q)$ *and for any* $\epsilon > 0$*, there exist* $\hat{p} \in \Delta\Omega$ *such*

that $\|p - \hat{p}\|_2 < \epsilon$ *and* $\tilde{v}_p(\hat{p}) < \sup_q v_p(q)$. $\hat{\mathcal{P}} \subset \mathcal{P}(\tau_v)$, $\forall \tau_v \in \Delta\Delta\Omega$.

The effectiveness of persuasion is contingent on the presence of specific posterior beliefs that advise the Receivers to choose actions that are more advantageous for the Sender than their default choices. Therefore, any attempt to persuade the Receivers who have already chosen the action $\alpha \in A^*$ that maximizes benefits for the Sender based on their prior belief would be unnecessary or even counterproductive, like gilding the fine gold or painting the lily.

Conditions (a) and (c) in Proposition 1 ensure that if a set of prior beliefs, denoted as $\hat{\mathcal{P}}$, is located within $\text{Int}(\Delta\Omega)$, and induces $\alpha \in A^*$, then these beliefs are adjacent to beliefs that induce less advantageous actions for the Sender. If a persuasion strategy induces fully mixed signals, as specified by condition (b), the Receivers with prior beliefs $p \in \hat{\mathcal{P}}$ may change their beliefs following the persuasion. Some of these potential posterior beliefs will recommend less advantageous actions for the Sender, resulting in over-persuasion.

Multiple conditions must be met for Proposition 1 to hold. Condition (a) implies that there is no persuasion strategy that ensures certain Receivers will take the action that the Sender desires most. Simply put, the persuasion outcome depends not only on the Sender's strategy but also on the Receivers' prior beliefs. Condition (b) requires that the number of functioning signals in the persuasion should equal the number of the actual states. Condition (c) indicates the Receivers are able to choose the action that is most advantageous for the Sender in the face of uncertainty. These requirements are not stringent. In fact, their implications are common in the real world. For example, our application in Section 7 exemplifies a typical commodity transaction scenario in which all of the conditions specified in Proposition 1 are satisfied. Consequently, it is the Sender's natural conjecture that some Receivers could be over-persuaded by any fully-mixed information structure if she is ambiguous about the Receivers' types.

When all the conditions specified in Proposition 1 are met, the identification of $\hat{\mathcal{P}}$ becomes straightforward when $v(q)$ is given. They are associated with the maximum value

of $v(q)$ and are adjacent to those with lower values. This property suggests that $\hat{\mathcal{P}}$ either are or are in close proximity to the targeted posterior beliefs that are induced by the optimal persuasion strategy. This raises concerns if the Receivers have access to external information sources that the Sender is unaware of. In contrast to the concern in robust persuasion design (Dworczak and Pavan, 2022), where an adversarial external resource may threaten the effectiveness of persuasion, a collaborative external resource may cause an over-persuasion issue. For example, a customer entering a store may have already been persuaded and be ready to make a purchase; any further persuasion risks backfiring due to the over-persuasion issue.

The existence of $\hat{\mathcal{P}}$ in Receiver types is not a necessary condition for the over-persuasion to occur. Receivers with a given prior belief can be over-persuaded by certain persuasion strategies, even if they have a positive persuasion value at optimal, such as point c in Figure 1. Accordingly, the over-persuasion issue arises when these types of Receivers are in the support $\mathcal{P}$, and certain persuasion strategies are endogenous in the persuasion game, chosen optimally by the Sender to solve the following simplified persuasion problem.

$$\begin{aligned} \max_{\tau_v} \ & \sum_{p\in\mathcal{P}} f(p) E_{\tau_v}\left[v_p(q)\right] \\ & s.t.\ E_{\tau_v}(q) = p_v. \end{aligned} \tag{6}$$

Let $\tau^*(\cdot)$ denote a function defined on $2^{\mathcal{P}}$ that represents the optimal strategy of a subset of the support $\mathcal{P}$. The over-persuasion issue is formally defined based on the Sender's simplified persuasion problem (6).

Definition 2. *For a given persuasion game with $\mathcal{P}$ being the support of the Receivers' prior belief distribution, the over-persuasion issue arises when $\left\{p|p \in \mathcal{P}\left(\tau^*(\mathcal{P}\backslash p)\right)\right\} \neq \emptyset$.*

According to Definition 2, over-persuasion becomes problematic when the Receivers who could be over-persuaded force the Sender to change her otherwise optimal strategy

to limit the loss from over-persuasion, or when some Receivers are over-persuaded as the Sender maintains her initial optimal strategy. While the latter case unambiguously suggests a loss in the Sender's persuasiveness, it remains unclear whether the Sender also loses in the former scenario. To characterize the effect of the over-persuasion issue, we define $\mathcal{P}^*$ as the support of the Receivers' prior belief distribution that maximizes the total persuasion value, as indicated by (7).

$$\mathcal{P}^* = \arg\max_{\mathcal{P}'} \left\{ \max_{\tau_v} \sum_{p \in \mathcal{P}'} f(p) V_p(\tau_v) \right\}, \tag{7}$$

where $V_p(\tau_v) = E_{\tau_v}\left[v_p(q)\right] - v(p)$ represents the persuasion value of the Receiver with prior belief p under the strategy τ_v satisfying the constraint in (4).

If the condition $\mathcal{P}^* \subsetneq \mathcal{P}$ is satisfied, the over-persuasion issue results in a loss of persuasiveness. This is because the condition implies both $\left\{p | p \in \mathcal{P}\left(\tau^*(\mathcal{P} \backslash p)\right)\right\} \neq \emptyset$ and the depreciation of persuasion value $\sup_{\tau_v} \sum_{p \in \mathcal{P}} f(p)\left[E_{\tau_v}\left[v_p(q)\right] - v(p)\right]$ due to the presence of $\mathcal{P} \backslash \mathcal{P}^*$ in the support of $f(p)$. According to the definition of $\mathcal{P}^*$, the persuasion value for all Receivers who hold prior beliefs in $\mathcal{P}^*$ must be positive when employing the persuasion strategy $\tau^*(\mathcal{P}^*)$. This requirement implies Proposition 2.

Proposition 2. *If $\hat{\mathcal{P}} \subsetneq \mathcal{P}$, or there exists p_1 and p_2 such that $\{\tau_v | V_{p_1}(\tau_v) > 0, V_{p_2}(\tau_v) > 0\} = \emptyset$, then it follows that $\mathcal{P}^* \subsetneq \mathcal{P}$.*

If the conditions specified in Proposition 2 are satisfied, any persuasion strategy will result in negative persuasion values for a non-empty subset of Receivers who hold prior beliefs $\mathcal{P} \backslash \mathcal{P}^*$. When these conditions induce the over-persuasion issue, the Sender's persuasiveness will inevitably decline. The existence of $\hat{\mathcal{P}}$ is demonstrated by Proposition 1. Provided that $v(q)$ is not globally concave, it is also not hard to identify a pair of prior beliefs p_1 and p_2 in $\Delta\Omega$.

Panel (i) of Figure 2 demonstrates how to derive the phase graphs of the set of persuasion strategies represented by $q \in [0, p_v] \times [p_v, 1]$ that generates positive persuasion values,

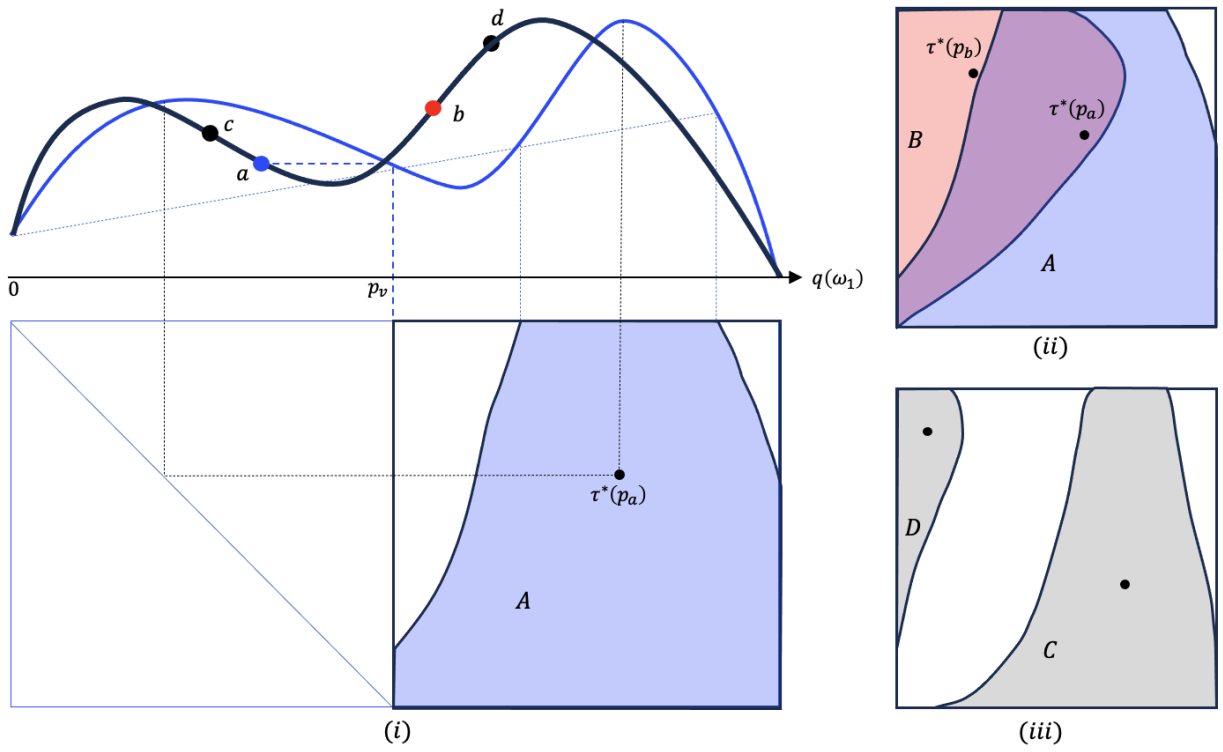

Figure 2: Heterogeneous Prior Beliefs and Over-persuasion Issue

when the transformed value function $v_p(q)$ is given in a persuasion game with two possible states ω_1 and ω_2. By merging these phase graphs as shown in Panel (ii) of Figure 2, we can find the overlapping region representing the feasible set of targeted posterior beliefs that ensures positive persuasion value for all Receivers with $p \in \mathcal{P} = \{p_a, p_b\}$.

While it is possible for prior beliefs $p = q$ within the convex region of the graph $\big(q, v(q)\big)$ to yield positive persuasion values at optimal, the persuasion strategies that induce these optimal persuasion values may emphasize different states. This distinction renders these heterogeneous prior beliefs incongruent. The change in information structure in favor of Receivers with a specific prior belief may reduce the persuasion values of other Receivers with incongruent prior beliefs. If these incongruent beliefs are located near the concave closure of the graph $\big(q, v(q)\big)$ and therefore associated with low optimal persuasion value, it may be difficult or even impossible to find a strategy that simultaneously induces

positive persuasion value for all of them. In Figure 2, the beliefs represented by a and c on the left slope of the convex segment of the graph are incongruent with the beliefs represented by b and d on the right slope. As p_a and p_b approach p_c and p_d, respectively, their optimal persuasion values decrease, ultimately resulting in detached phase graphs C and D. Therefore, p_c and p_d represent the pair of beliefs $\{p_1, p_2\}$ that we seek to identify.

As shown in Figure 2, the increase in heterogeneity leads to a rise in incongruence, causing the overlapping region to recede. Consequently, an over-persuasion issue arises as $\tau^*(p_b)$ becomes exposed by the region A. Additionally, the persuasiveness is discounted as the phase progressively retreats towards regions C and D. Heterogeneity among congruent prior beliefs also increases the risk of the Sender losing persuasiveness when over-persuasion occurs. When the Receivers exhibit significant heterogeneity, even if they all possess the prior belief $p \in \mathcal{P}^*$, once a Receiver with $p' \in \mathcal{P} \backslash \mathcal{P}^*$ is introduced, the heterogeneity makes it more challenging to find a persuasion strategy that yields positive persuasion values for both p' and all $p \in \mathcal{P}^*$. In Figure 2, p_c and p_a are distinctive but congruent. However, when p_b is included in $\mathcal{P}$, the heterogeneity between a and c produces a minimally overlapping region between A, B, and C. As a result, $\mathcal{P}^* = \mathcal{P}$ becomes highly improbable.

4.3 Distribution of Prior Beliefs and Over-persuasion

The presence of heterogeneity in the support of the Receivers' prior belief distribution causes the over-persuasion issue. Furthermore, when this support includes specific beliefs, it can even result in a loss of persuasiveness. But when these beliefs specified in Proposition 2 are absent from the support, the impact of over-persuasion on the Sender's persuasiveness is determined solely by the distribution.

Proposition 3. *Given the profile of a persuasion game, if there exists a finite $\mathcal{P} \subset \Delta\Omega$ that gives rise to the over-persuasion issue, then there exists a distribution $f(p)$ over $\mathcal{P}$ such that $\mathcal{P}^* \subsetneq \mathcal{P}$.*

When confronted with the over-persuasion issue, the Sender may endeavor to choose a

persuasion strategy that yields a positive persuasion value for $p \in \mathcal{P}\backslash\mathcal{P}^*$ while maintaining as high persuasion values as possible for $p \in \mathcal{P}^*$. However, if $f(p)$ is small enough for $p \in \mathcal{P}\backslash\mathcal{P}^*$ to make Receivers holding these prior beliefs less significant relative to the entire group, the cost outweighs the benefit of deviating from $\tau^*(\mathcal{P}^*)$, where $E_f V_p[\tau^*(\mathcal{P}^*)] < 0$ for $p \in \mathcal{P}\backslash\mathcal{P}^*$. As $\sum_{\mathcal{P}\backslash\mathcal{P}^*} f(p)$ increases, certain strategies that result in $E_f V_p[\tau^*(\mathcal{P}^*)] > 0$ for $p \in \mathcal{P}\backslash\mathcal{P}^*$ may become optimal as this group gains significance. However, these strategies may lead to $E_f V_p[\tau^*(\mathcal{P}^*)] < 0$ for $p \in \mathcal{P}^*$ when $\sum_{\mathcal{P}\backslash\mathcal{P}^*} f(p)$ is excessively great, rendering $p \in \mathcal{P}^*$ insignificant.

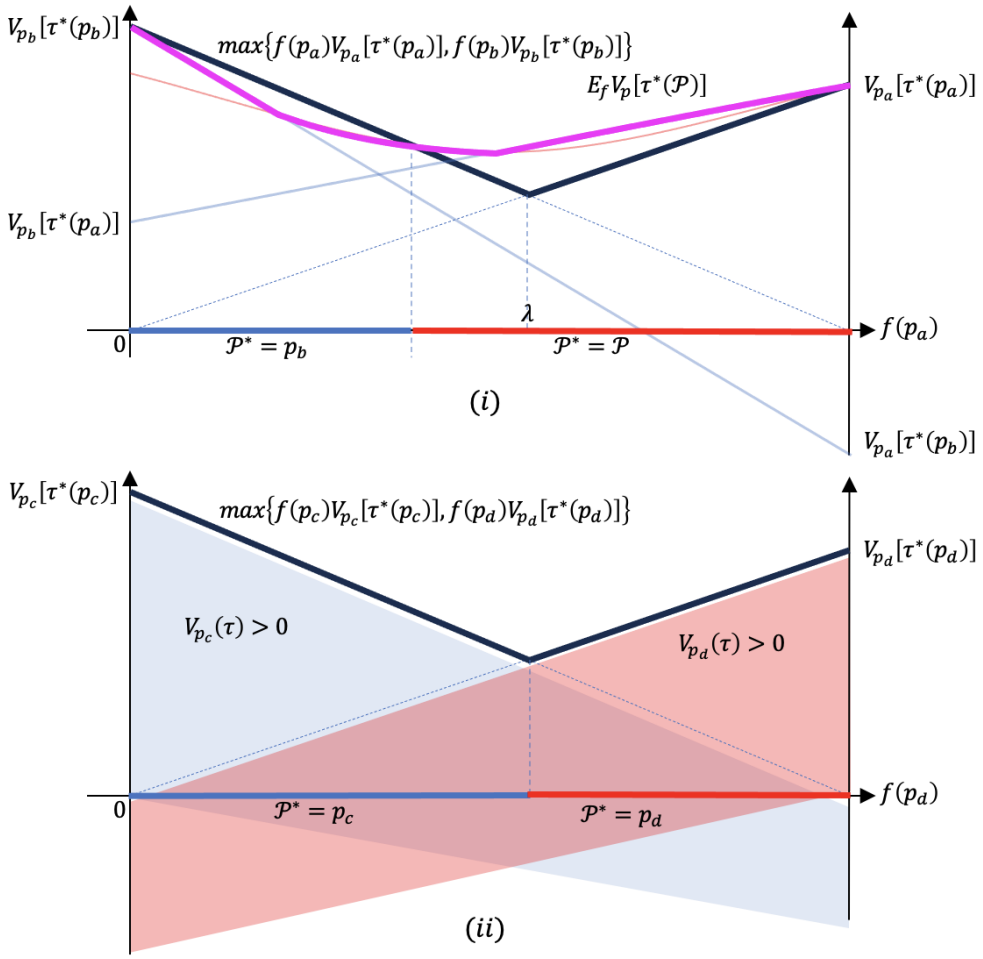

Figure 3: Over-persuasion Issue and Distribution $f(p)$

Panel (i) of Figure 3 illustrates an example based on p_a and p_b in Figure 2. This example demonstrates how the distribution $f(p)$ determines the loss of persuasiveness caused by the over-persuasion issue. When $f(p_a)$ is sufficiently small, the Sender's expected payoff is $E_f V_p[\tau^*(p_b)] = [1 - f(p_b)]V_{p_a}[\tau^*(p_b)] + f(p_b)V_{p_b}[\tau^*(p_b)]$. Since $E_f V_p[\tau^*(\mathcal{P})] = E_f V_p[\tau^*(p_b)]$ is smaller than $\max\{f(p_a)V_{p_a}[\tau^*(p_a)], f(p_b)V_{p_b}[\tau^*(p_b)]\}$, the Sender suffers persuasiveness loss, as indicated by $\mathcal{P}^* = \{p_b\} \subsetneq \mathcal{P}$. When $f(p_a)$ is large enough, there exist alternative persuasion strategies τ' such that $V_{p_a}(\tau') > 0$ and $E_f V_p(\tau') > E_f V_p[\tau^*(p_b)]$. $\tau^*(p_a)$ is such a strategy. Therefore, the purple curve representing $E_f V_p[\tau^*(\mathcal{P})]$ eventually exceeds $E_f V_p[\tau^*(p_b)]$ and even $\max\{f(p_a)V_{p_a}[\tau^*(p_a)], f(p_b)V_{p_b}[\tau^*(p_b)]\}$ as $f(p_a)$ increases, resulting in $\mathcal{P}^* = \mathcal{P}$.

In this example, $V_{p_b}[\tau^*(p_a)] > 0$ contributes to the transition from $\mathcal{P}^* \subsetneq \mathcal{P}$ to $\mathcal{P}^* = \mathcal{P}$ as $f(p_a)$ increases. When $f(p_a)$ exceeds λ, the prior belief p_a becomes the major type in $\mathcal{P}$. In this case, over-persuasion issue does not occur because $V_p(\tau) > 0$ for both p_a and p_b when $\tau^*(p_a)$ is adopted, as shown in the panel (ii) of Figure (2). Therefore, it follows that $E_f V_p[\tau^*(\mathcal{P}^*)] \geq E_f V_p[\tau^*(p_a)] > f(p_a)V_{p_a}[\tau^*(p_a)]$ when $f(p_a)$ is sufficiently large, indicating the absence of persuasion loss.

Panel (ii) of Figure 3 depicts an opposing example that demonstrates the validity of Proposition 2. According to Figure 2, there is no τ_v such that both $V_{p_c}(\tau_v) > 0$ and $V_{p_d}(\tau_v) > 0$ are simultaneously greater than 0. Therefore, all possible $E_f V_p(\tau_v)$ are within the shaded region and are therefore less than $\max\{f(p_c)V_{p_c}[\tau^*(p_c)], f(p_d)V_{p_d}[\tau^*(p_d)]\}$. Furthermore, if p_c or p_c belong to $\hat{\mathcal{P}}$, the shaded region is reduced to the area below the axis. In both cases, $\mathcal{P}^* \subsetneq \mathcal{P}$ is resulted.

The loss of persuasiveness due to over-persuasion, caused by the nature of heterogeneous Receivers, may explain the backfire effect in persuasion. The aforementioned vaccination example may be the result of the over-persuasion issue. Understanding that the flu vaccine does not cause flu may not have a monotonic effect on the likelihood that an individual will accept vaccination. For example, significant resistance and hesitancy to COVID-19

vaccines have been observed among healthcare workers and professionals (Khubchandani et al.,2022; Gu et al., 2022). For those who firmly believe that the flu vaccine causes the flu, more information about the vaccine's side effects may best clarify their myth and increase their vaccine acceptance. However, those who initially have a deeper understanding of the mechanism of the flu vaccine may be more concerned about the vaccine side effects with this information. As their beliefs shift away from the one that induces the highest likelihood of acceptance, they become increasingly hesitant about the vaccine.

5 Signal-Independent Mechanism

The over-persuasion issue does not necessarily result in a loss of persuasiveness. But when it does, as demonstrated in Section 4, it is indicated by $\mathcal{P}^* \subsetneq \mathcal{P}$. This indication implies that the persuasiveness loss can be mitigated or resolved through mechanism design, provided that it is possible to prevent only Receivers with $p \in \mathcal{P} \backslash \mathcal{P}^*$ from revealing the signal. In this section, we aim to explore the conditions under which it is possible to design a mechanism to address this problem, as well as how to determine the optimal design of such a mechanism.

In a persuasion game, the Sender possesses ownership of the signal or the subject that generates the signal. She may terminate the contract at any time if the Receivers fail to fulfill its terms. Once the Receivers fulfill the terms, sending the signal becomes advantageous for the Sender, and she has no incentive not to execute the contract. Consequently, a contract drafted by the Sender and contingent on the Receiver's decision to obtain the signal has inherent enforceability.

A contract contingent on the signal, on the other hand, is more challenging to enforce and may require external enforcement. Once the signal is revealed to the Receivers, they acquire the information, and the mechanism is incapable of ensuring their compliance with the contract terms. On the other hand, if the sender retains control of the signal prior to contract execution, a commitment problem arises. This section focuses on the

signal-independent mechanism, where $\theta(h) = \theta(\gamma)$. Under certain conditions, if a signal-independent mechanism is feasible for establishing incentive compatibility, the Sender may prioritize it over a signal-contingent mechanism, which is effective at generating feasibility despite its limited enforceability, as will be elaborated in Section 6.

5.1 Simplifying the Receiver's Problem

To design a mechanism that effectively discourages specific Receivers from revealing signals, it is crucial to understand their underlying motivations. Similar to the Sender's problem, The Receivers' problem can be simplified to choosing actions for different posterior beliefs ex-post and deciding whether to reveal the signal ex-ante based on their anticipation of optimal payoffs associated with potential posterior beliefs. Accordingly, the simplified problem faced by the receiver is

$$\max_{\gamma} \ \gamma E_{\tau_p}\Big[u(\alpha^{1,q}, q) - u(\alpha^{0,q}, q)\Big] - \theta(\gamma) \tag{8}$$

at the first stage, where $\alpha(\gamma, q)$ is the optimal reaction function determined at the second stage, and

$$\max_{\alpha} \ \gamma^* u(\alpha, q^*) + (1 - \gamma^*) u(\alpha, p) \tag{9}$$

at the second stage based on the optimal choice at the first stage γ^* and the realized posterior belief q^*.

Consider $u^*(q) = u(\alpha^{1,q}, q)$ to be an indirect utility function of posterior belief based on the optimal selection of α in (9), when $\gamma = 1$. Lemma 2 characterizes the basic property of $u^*(q)$ that facilitates its graphical representation and analysis.

Lemma 2. *$u^*(q)$ is continuous and convex in q and $u^*(q) = u(\alpha^{0,q}, q)$ at $q = p$, where p is the Receiver's prior belief.*

In Receivers' problems, they are confronted with several actions that determine $u(a, \omega)$ in different actual states ω. Once α is given, they expect a payoff of $u(\alpha, q) = E_{\tau_q} u(\alpha, \omega)$

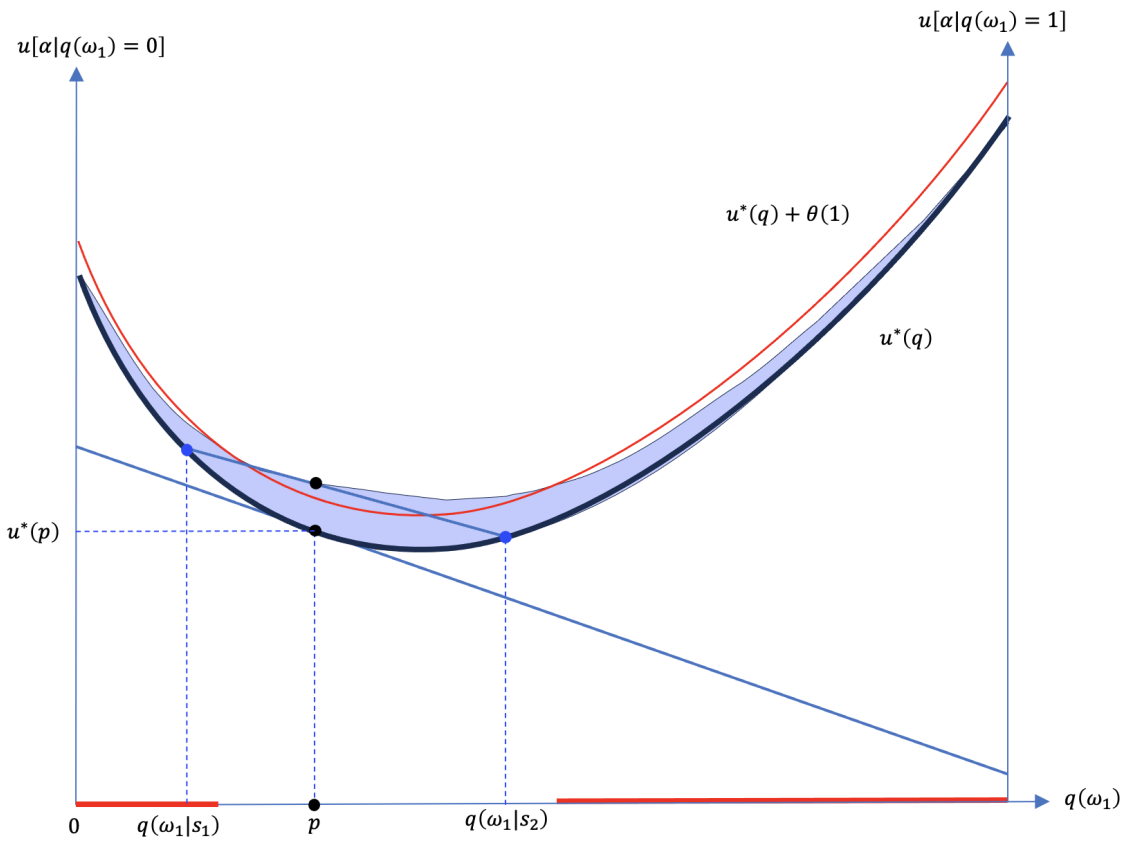

Figure 4: Receiver's Problem and the Feasibility of Mechanism Design

where $\tau_q \in \Delta\Omega$ is the distribution on the support ω such that $E_{\tau_q} e_i = q$, supposing $\{e_1, ...e_{n-1}\}$ is an orthonormal basis in $n-1$ dimensional vector space when $\dim(\omega) = n$. When choosing an action to maximize $u(\alpha, q)$, the Receivers are choosing a hyperplane with the highest value among a set of hyperplanes at q. Essentially, $u^*(q)$ is an envelope hypersurface for all hyperplanes $u(\alpha, q)$, $\alpha \in \mathcal{A}$, resulting in $u^*(q)$ being tangent to some $u(\alpha, q)$ graphically. In a two-state case, Figure 4 depicts the relationship between $u^*(q)$ and $u(\alpha, q)$. Occasionally, $u(a, q)$ and $u^*(q)$ could tangent at a region $\mathcal{Q}$ rather than a single point. In these cases, the same set of actions $\mathcal{A}^*$ may be optimal for different beliefs $q \in \mathcal{Q}$. According to our tie-breaking rule where the Receivers always select the action most advantageous for the Sender in cases of indifference, $v(q)$ should remain constant for

all $q \in \mathcal{Q}$. Since the tangent region appears at all $q \in \text{Int}(\Delta\Omega)$, $u^*(q)$ must be globally convex; otherwise, a contradiction arises as certain $u(a, q)$ could be greater than $u^*(q)$ at some q where $u(a, q)$ is not the optimal payoff the Receivers receive.

Since $\alpha^{0,q} = \arg u(\alpha, p)$ by definition, it follows that $u^*(p) = u(\alpha^{0,p}, p)$, and that $u^*(q)$ and $u(\alpha^{0,q}, q)$ are tangent at $q = p$. $u(\alpha^{0,q}, q)$ as a hyperplane is linear in q. Hence, $E_\tau u(\alpha^{0,q}, q) = u(\alpha^{0,p}, p)$ holds, provided that $E_\tau q = p$. In conjunction with the fact that $u^*(p) = u(\alpha^{0,p}, p)$, the Receiver's information value, defined as $U_p(\tau_p) = E_{\tau_p}[u^*(q) - u(\alpha^{0,q}, q)]$, where $E_{\tau_p} q = p$, is equivalent to $E_{\tau_p} u^*(q) - u^*(p)$. This equivalence implies that the information value is determined by the convexity of $u^*(q)$ within the domain established by the set of targeted posterior beliefs $\{q\}$ induced by the signals. In Figure 4, the shaded region represents the information value associated with prior beliefs varying within the entire interval $[0, 1]$, assuming the signals are $\{s_1, s_2\}$.

Corollary 1. *The information value $U_p(\tau_p)$ is non-negative for all $p \in \text{Int}(\Delta\Omega)$ and τ_p such that $E_{\tau_p} q = p$.*

This corollary follows directly from Lemma 2, which states that $u^*(q)$ is convex in all $q \in \text{Int}(\Delta\Omega)$. Therefore, when $\theta(\gamma) \equiv 0$, the Receivers will always choose $\gamma = 1$ to reveal the signal in a persuasion game in the absence of mechanism design. This clarifies the assumption we made without further explanation in Section 4, which essentially forms the basis for the over-persuasion issue. Within the framework of our model, this corollary also implies that any design of $u^*(q)$, such as pricing in business, is not possible to exclude only a subset of the Receivers from the persuasion game without the participation of intrinsic or imposed information fee. This implication emphasizes the significance of mechanism design as a solution to the over-persuasion issue.

In the persuasion game, when the information cost is given as $\theta(1) > 0$, the Receivers with prior beliefs yielding a lower information value than the curve representing $u^*(q) + \theta(1)$ will not reveal the signals. In the scenario illustrated in Figure 4, if the Sender designs a mechanism that imposes an information fee of $\theta(1)$ on the Receivers, only those whose

prior beliefs fall within the non-bolded segment on the horizontal axis will change their belief by the persuasion strategy that produces signals $\{s_1, s_2\}$.

5.2 Incentive Compatibility

For any given persuasion strategy τ, it is possible to rank the information value $U_p(\tau)$ associated with prior beliefs p. Accordingly, the Contour sets of $U_p(\tau)$ are a natural way to divide $\mathcal{P}$ into two different groups. In this persuasion game, incentive compatibility is defined as a feasible separating plan to single out a set of Receivers under the given persuasion strategy, as desired.

Definition 3. *Let $C[U_p(\tau)] = \bigcup_\theta \{p|U_p(\tau) \geq \theta\}$ be the union of upper Contour sets of U_p for given τ. Incentive Compatibility is established for a desired set $\mathcal{P}'$ under τ if $\mathcal{P}' \in \bigcap C[U_p(\tau)]$. Under the constraint of incentive compatibility, $\tilde{\mathcal{P}}^*(\tau)$ is defined as the set that maximizes persuasion value for given τ.*

$$\tilde{\mathcal{P}}^*(\tau) = \arg \max_{\mathcal{P}' \in C[U_p(\tau)]} \left[\sum_{p \in \mathcal{P}'} f(p) E_\tau v_p(q) + \sum_{p \in \mathcal{P} \setminus \mathcal{P}'} f(p) v_p(p) \right]$$

According to Definition 3, if the Sender can impose $\theta(\gamma)$ to single out a certain group to enhance persuasion value at optimal, the mechanism design can improve the persuasion outcome. Technically, we can find an $u^*(q)$ to satisfy this feasibility.

Proposition 4. *Given a finite set $\mathcal{P}'$ and a Bayes plausible persuasion strategy τ_v such that $\text{supp}(\tau_v) \setminus \text{Int}(\Delta\Omega) \neq \emptyset$, there exists an $u^*(q)$ that are not linear everywhere in such that incentive compatibility can be established.*

In order for mechanism design to be feasible, it is essential that $u^*(q)$ is not linear everywhere; otherwise, the information value is constantly zero for all Receivers, regardless of their prior beliefs, making it impossible to induce the Receivers' separating decisions to reveal the signals. Under this condition, Proposition 4 shows that when

$\text{supp}(\tau)\setminus \text{Int}(\Delta\Omega) \neq \emptyset$, so that posterior q varies in p for given information structures, the convexity of $u^*(q)$ may vary independently in $q = p$. It may allow $U_p(\tau)$ for all $p \in \mathcal{P}\setminus\mathcal{P}'$ to be smaller than $U_p(\tau) \geq \theta(1)$ for all $p \in \mathcal{P}'$, which induces the possibility of incentive compatibility.

Given that $\mathcal{P}^* \subsetneq \mathcal{P}$, there exists $\{\tilde{\tau}, \tilde{\mathcal{P}}\}$ that induces accumulated persuasion value $\sum_{p\in\tilde{\mathcal{P}}} f(p)V_p(\tilde{\tau})$ that is greater than the optimal total persuasion value when no Receiver is excluded, $\sum_{p\in\mathcal{P}} f(p)V[\tau^*(\mathcal{P})]$. According to Proposition 4, there exists $u^*(q)$ that ensures incentive compatibility for $\{\tilde{\tau}, \tilde{\mathcal{P}}\}$, which ensures that $\tilde{\mathcal{P}}$ can be separated from $\mathcal{P}$ by some $\theta(1) > 0$ to realize $\sum_{p\in\tilde{\mathcal{P}}} f(p)V_p(\tilde{\tau})$. Therefore, Proposition 4 implies that applying mechanism design to improve the persuasiveness in persuasion games with the over-persuasion issue is technically feasible with certain well-defined $u^*(q)$.

However, the existence of an ideal indirect utility function $u^*(q)$ as required in Proposition 4 cannot be guaranteed in persuasion games. When $u^*(q)$ is predetermined and the Sender and cannot be modified by the Sender, it is necessary to test whether the prior beliefs in the set $\tilde{\mathcal{P}}$ and its complement $\mathcal{P}\setminus\tilde{\mathcal{P}}$ are sufficiently distinctive to support incentive compatibility. For a given indirect utility function, the information value of Receivers is determined by the interval of targeted posteriors and the relative position of p within this interval. The Sender's persuasion strategy and the Receivers' prior beliefs uniquely determine the targeted posteriors, or the intervals on the same supersurface $u^*(q)$ for different Receivers. Similar prior beliefs result in similar intervals and similar positions within the interval. Since $u^*(q)$ is continuous and convex in q, significantly overlapping intervals imply similar information values for any information structure, which poses a challenge to incentive compatibility. In contrast, if $\tilde{\mathcal{P}}$ and $\mathcal{P}\setminus\tilde{\mathcal{P}}$ are sufficiently distinct, their information values are determined more independently, thereby facilitating the establishment of incentive compatibility. This is particularly true when there is significant variation in the convexity of $u^*(q)$.

In addition to the support of the Receiver's prior belief distribution, the persuasion

strategy also influences incentive compatibility. When a persuasion strategy discloses too much information or, in an extreme case, fully discloses the actual states, all Receivers, regardless of their prior beliefs, tend to have a similar or even identical interval on $u^*(q)$. Therefore, if the Receivers' prior beliefs are located distantly and exhibit non-symmetrical positioning with respect to the lowest point of the supersurface $u^*(q)$, they could have very different information values, even if the curvature of $u^*(q)$ varies only very slightly. When the distinction between $\tilde{\mathcal{P}}$ and $\mathcal{P}\backslash\tilde{\mathcal{P}}$ and the convexity of $u^*(q)$ provides a premise for incentive compatibility, the Sender may modify the optimal information structure to ensure the establishment of incentive compatibility for a certain group of Receivers. More information will be disclosed if the convexity of $u^*(q)$ exhibits minimal variation with respect to q; otherwise, the Sender may prefer to disclose less information.

5.3 Optimal Persuasion Strategy with Mechanism Design

When $\mathcal{P}^* \subsetneq \mathcal{P}$ and incentive compatibility is possible for a certain subset of Receivers, allowing the persuasiveness to be enhanced through mechanism design, we are able to find the optimal information fee to prevent the Receivers of some types from revealing the belief-changing signals and the optimal information structure to persuade those Receivers who remain in the persuasion game.

Proposition 5. *Provided that incentive compatibility can be established on* $(\tilde{\tau}, \tilde{\mathcal{P}})$*, there exists* $(\tilde{\tau}_v^*, \theta(1)^*)$ *where* $\theta(1) > 0$ *and* $\tilde{\mathcal{P}} \subsetneq \mathcal{P}$ *that maximize (6). The Sender's maximized expected payoff is greater than* $\sum_{p\in\mathcal{P}} f(p)v_p[\tau_v^*(\mathcal{P})]$.

The union $\bigcup(\tilde{\tau}, \tilde{\mathcal{P}})$ outlines all the possible combinations that could improve the Sender's expected payoff by preventing a subset of the Receivers from changing their initial decisions. When $\mathcal{P}$ is finite, $\bigcup(\tilde{\tau}, \tilde{\mathcal{P}})$ is a compact set, which ensures the existence of the pair $(\tilde{\tau}_v^*, \tilde{\mathcal{P}}^*) \in \bigcup(\tilde{\tau}, \tilde{\mathcal{P}})$ that optimizes (6). And the optimal $\theta(1)^*$ is to ensure Receivers are precisely separating into $\tilde{\mathcal{P}}^*$ and $\mathcal{P}\backslash\tilde{\mathcal{P}}^*$.

For the given value functions $v(q)$ and indirect utility function $u^*(q)$, the optimal $\{\tilde{\tau}^*, \theta(1)^*\}$ can be identified by the following procedure. First, we must identify $\tau_v^*(\mathcal{P})$ and determine $\sum_{p\in\mathcal{P}} f(p)E_{\tau_v^*(\mathcal{P})}V_p(q)$ as a benchmark persuasion value. Second, for all $\tau_v \in \Delta\Delta\Omega$, find the subset space of the full support of the Receiver's beliefs that generates greater a persuasion value than the benchmark, $\bigcup \tilde{\mathcal{P}}(\tau_v) = \{\mathcal{P}' | \sum_{p\in\mathcal{P}'} f(p)E_{\tau_v}V_p(q) > \sum_{p\in\mathcal{P}} f(p)E_{\tau_v^*(\mathcal{P})}V_p(q)\}$. Over-persuasion issue arises as long as $\bigcup_{\tau_v\in\Delta\Delta\Omega}\bigcup \tilde{\mathcal{P}}(\tau_v) \neq \emptyset$. If an over-persuasion issue exists, then check if incentive compatibility can be established on any pair of $\{\tau_v, \tilde{\mathcal{P}}(\tau_v)\}$. Specifically, for τ_v such that $\bigcup \tilde{\mathcal{P}}(\tau_v) \neq \emptyset$, find the contour set $C[U_p(\tau_v)] = \bigcup_\theta \{p | U_p(\tau_v) \geq \theta\}$. If $\left(\bigcup \tilde{\mathcal{P}}(\tau_v)\right) \bigcap C[U_p(\tau_v)] \neq \emptyset$, the incentive compatibility is established. Let $\tilde{\tau}_v$ be the persuasion strategy that passes the incentive compatibility test. Among the sets of prior beliefs associated with $\tilde{\tau}_v$ and pass the incentive compatibility test, let $\tilde{\mathcal{P}}^*(\tilde{\tau}_v)$ represent the one that maximizes the persuasion value. Finally, the Sender's simplified problem becomes

$$\max_{\tau_v} \sum_{p\in\tilde{\mathcal{P}}^*(\tilde{\tau}_v)} f(p)E_{\tau_v}V_p(q). \tag{10}$$

The optimal mechanism that realizes this outcome $\theta(1)$ is set to the minimum information value of the prior belief associated with the targeted group, $\inf_{p\in\tilde{\mathcal{P}}^*(\tilde{\tau}_v^*)} U_p(\tilde{\tau}_v^*)$, where $\tilde{\tau}_v^*$ is the persuasion strategy that optimizes the objective (10).

6 Signal-Contingent Mechanism

Information fee as a signal-independent mechanism may discourage some Receivers from obtaining belief-changing signals to maintain their initial decisions, while the other Receivers can be persuaded to change their mind and shift to a more advantageous action for the Sender. This contract is easier to enforce because it is executed before the signals are realized, but it has limitations. In this mechanism design, the incentive compatibility under the participation constraint is highly dependent on the predetermined indirect utility

function and is therefore not guaranteed to exist.

If this concern can be alleviated so that we can assume that the contract is enforceable even after the signal is realized, the sender may use a signal-contingent mechanism to address the over-persuasion issue in the persuasion game. Since the contract is signal-contingent, the Sender can even reward the Receivers upon the realization of a specific signal. This characteristic makes this mechanism more powerful at ensuring the simultaneous satisfaction of the participation constraint and incentive compatibility in the vast majority of situations.

6.1 Incentive Compatibility

To design a signal-contingent mechanism, the Sender designs $\theta(\gamma, q_v)$, which is dependent on whether the Receivers decide to obtain the signal and the signal realization. According to our previous analysis, q_v is uniquely associated with the Receiver's posterior beliefs given their prior beliefs. Therefore, once the signal is realized, the distribution of the posterior distribution τ_v is public information for contract enforcement. Without loss of generality, the Receivers always pay zero if not participating in the contract, implying $\theta(0, q_v) = 0$. In addition, we assume that $\beta = 1$ in this section. This assumption is particularly important in this section because, in order to achieve a desirable persuasion outcome, the Sender may need to pay information rent to the Receivers.

Proposition 6. *For any $\mathcal{P}$, $\tilde{\mathcal{P}} \subsetneq \mathcal{P}$, and τ_v, suppose that for all $p \in \tilde{\mathcal{P}}$ and all $p' \in \mathcal{P} \backslash \tilde{\mathcal{P}}$, there exists at least one $\omega \in \Omega$ such that $p(\omega) > p'(\omega)$, then there exists $\theta(\gamma, q_v)$ such that $U_p(\tau_v) > U_{p'}(\tau_v)$ for all $p \in \tilde{\mathcal{P}}$ and all $p' \in \mathcal{P} \backslash \tilde{\mathcal{P}}$.*

This proposition implies that mechanism design is applicable in any circumstance for addressing the over-persuasion issue when a signal-contingent contract is enforceable and a group of Receivers holds a stronger belief in one state than any other Receivers. Once the key assumptions, which are not uncommon in the real world, are satisfied, the Sender may

take advantage of this state to distinguish this group from the general public. Regardless of the information structure proposed by the Sender, this group of Receivers ex-ante believe that the signal favoring this particular state is more likely than any other signal in the population. Consequently, if the Sender rewards and punishes the Receivers upon the realization of the signal that favors or opposes this state, respectively, in the signal-contingent contract, the reward is amplified and the punishment is decreased for this targeted group, differentiating their incentive to obtain and process the persuasion signal from that of the others.

A graph analysis demonstrates how a signal contingent mechanism can address the over-persuasion issue, even if the indirect utility function is not ideal for signal-independent mechanism design. Consider a two-state scenario in which the Sender designs an arbitrary information structure that may produce signals s_1 and s_2 which relatively favor ω_2 and ω_1, respectively. The Receivers with p_a, who hold stronger beliefs in ω_2 prior to receiving information, are of the opinion that the occurrence of s_1 is more probable compared to the Receivers with p_b. If the Sender can only charge the Receiver an information fee based on their decision to obtain the signal, rather than the actual signal received, it is not possible to exclude Receivers with p_a from the persuasion game while keeping Receivers with p_b in the game. Given the information structure, it can be observed that the Receivers with p_a possess a greater natural informational value. Therefore, the signal-independent mechanism design may not effectively address the over-persuasion problem when $\mathcal{P}^* = \{p_a\}$.

Nonetheless, as shown in Figure 5, the signal-contingent mechanism design can circumvent this limitation. As depicted in panel (i) of the graph, when the Sender rewards the Receivers upon the arrival of signal s_2 and punishes them when signal s_1 is observed, there exist positive $\theta\big(1, q_v(s_2)\big)$ and negative $\theta\big(1, q_v(s_1)\big)$ such that the expected information value $E_{\tau_p}\left[u_p^*(p) + \theta(1, q)\right]$ generated by this mechanism is non-negative for the Receivers with p_b and negative for the Receivers with p_a. As a result, the establishment of incentive

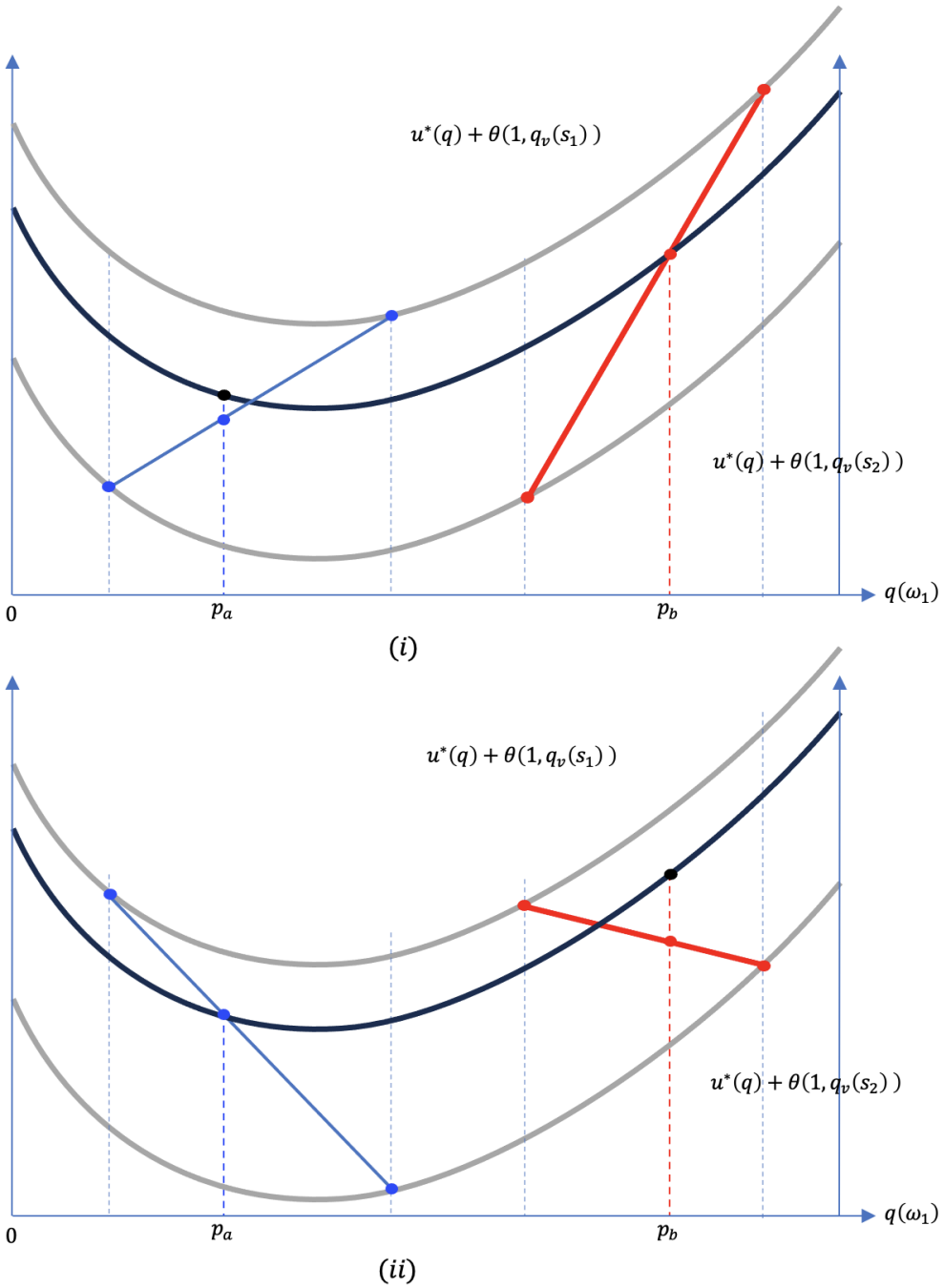

Figure 5: Incentive Compatibility in Signal-Contingent Mechanism Design

compatibility is facilitated by this mechanism. Also in Figure 5, panel (ii) of the graph indicates that the signal-contingent mechanism can also be applied to the opposite scenario where the Receivers with p_b should be excluded. By rewarding the realization of s_1 and punishing its absence, the Sender can simultaneously establish the required participation constraint and incentive compatibility to effectively address the over-persuasion issue.

The signal-contingent contract is a generalized version of the mechanism design employed in a persuasion game, ensuring incentive compatibility across different persuasion strategies. This guarantees that the persuasiveness of these strategies remains uncompromised, even when they are only optimal for a specific subset of Receivers. The signal-independent mechanism can be considered as a specific instance of a signal-contingent contract. To implement this mechanism design, the Sender needs only the upper and lower curves in Figure 6 to overlap beneath the utility curve. As indicated in Section 5, the incentive compatibility can be implemented only when the Sender intends to exclude the Receivers with p_a but eliminate those with p_b. Under certain conditions, this mechanism may emerge as the optimal mechanism and the Sender will still choose it even when it is possible to select a contract that is strictly signal-contingent.

In fact, if the Sender is free to choose between signal-contingent and signal-independent contracts, it may be observed that the Sender obtains a higher information value when she chooses the latter. This is not because the signal-independent contract generates higher information value. Signal-contingent mechanisms are typically employed as a final recourse in mechanism design when signal-independent mechanisms prove ineffective. Hence, it is always employed in certain scenarios where the Receivers who require motivation possess a lower natural information value compared to those who are intended to be discouraged. If the Sender generates motivation artificially rather than leveraging the natural information value, she may be responsible for the extra incentive costs. As a result, in cases where the persuasion values are not significantly different, the Sender may still give priority to $\tilde{\mathcal{P}}$ and τ_v that can be accommodated within a signal-independent mechanism.

6.2 Optimal Mechanism

Proposition 6 and Figure 5 suggest that incentive compatibility can be implemented with any τ_v and $\mathcal{P}'$, given that specific conditions are met. Therefore, if $\tilde{\mathcal{P}}$ satisfies this condition, the mechanism design can improve the persuasiveness. According to Proposition 5, there exists an optimal information structure $\tilde{\tau}_v^*$ to maximize the Sender's persuasion value at the second stage.

At the first stage of the game, the Sender needs to solve the following mechanism design problem.

$$\begin{aligned} \max_{\theta(1,q)\in\Theta} & \sum_{p\in\{p'\}} f(p) E_{\tilde{\tau}_v^*}\left[v_p(q) - \theta(1,q)\right] + \sum_{p\in\{p''\}} f(p) v_p(p) \\ s.t.\ & E_{\tau_{p'}}\left[u^*(q) - u^*(p') + \theta(\gamma, q)\right] \geq 0 \\ & E_{\tau_{p''}}\left[u^*(q) - u^*(p'') + \theta(\gamma, q)\right] \leq -\epsilon, \end{aligned} \tag{11}$$

where $\{p'\}\bigcap\{p''\} = \mathcal{P}$ and $\epsilon > 0$ is the smallest measure unit in the game.[4] In this Sender's problem, we need ϵ to ensure that Θ is compact when it is bounded. However, it is bounded only under some extra conditions.

Proposition 7. *Given $\tilde{\mathcal{P}}$ and τ_v, $\theta(1,q)^*$ is finite only if $p_v(\omega') > \inf_{p\in\tilde{\mathcal{P}}}\{p(\omega')\}$, where ω' is the state such that $p(\omega') > p'(\omega')$ for all $p \in \tilde{\mathcal{P}}$ and all $p' \in \mathcal{P}\backslash\tilde{\mathcal{P}}$. There exists a $\underline{p}(\omega')$ such that only when $p_v(\omega') \leq \underline{p}(\omega')$ does Sender have $E_{\tau_v}\left[\theta(1,q)\right] \geq 0$.*

The Receivers who hold prior beliefs $p = \inf_{p\in\tilde{\mathcal{P}}} p(\omega')$, where ω' is the state such that $p(\omega') > p'(\omega')$ for all $p \in \tilde{\mathcal{P}}$ and all $p' \in \mathcal{P}\backslash\tilde{\mathcal{P}}$ are willing to accept a contract with harsher punishments for the other signals, provided that there are greater rewards associated with signals that support $\{\omega'\}$. This modification of both higher rewards and punishments in

[4] To ensure the consistency of the models throughout this study, $\epsilon > 0$ is used when solving the problem, but $\epsilon = 0$ is assigned to the solution containing ϵ; this reflects the fact that ϵ is neglected by those who are not at the margin.

the contract ensures that the participation constraint is maintained for all Receivers with $p \in \tilde{\mathcal{P}}$, while still excluding the remaining Receivers with $p \in \mathcal{P} \backslash \tilde{\mathcal{P}}$. When the Sender holds a prior belief $p_v(\omega') < \inf_{p \in \tilde{\mathcal{P}}}\{p(\omega')\}$, she assigns a lower weight to potential rewards and a higher weight to potential punishments than the targeted Receivers. Therefore, provided that the constraints in (11) remain satisfied, the Sender can obtain a greater payoff from the Receiver's information value $\sum_{p \in \tilde{\mathcal{P}}} E_{\tau^*}\left[\theta\big(1, s(q)\big)\right]$ with these contract modifications. Since the modifications can be infinitely repeated, it is reasonable to anticipate infinitely positive and negative $\theta(1, q)$ values when all participants are risk neutral.

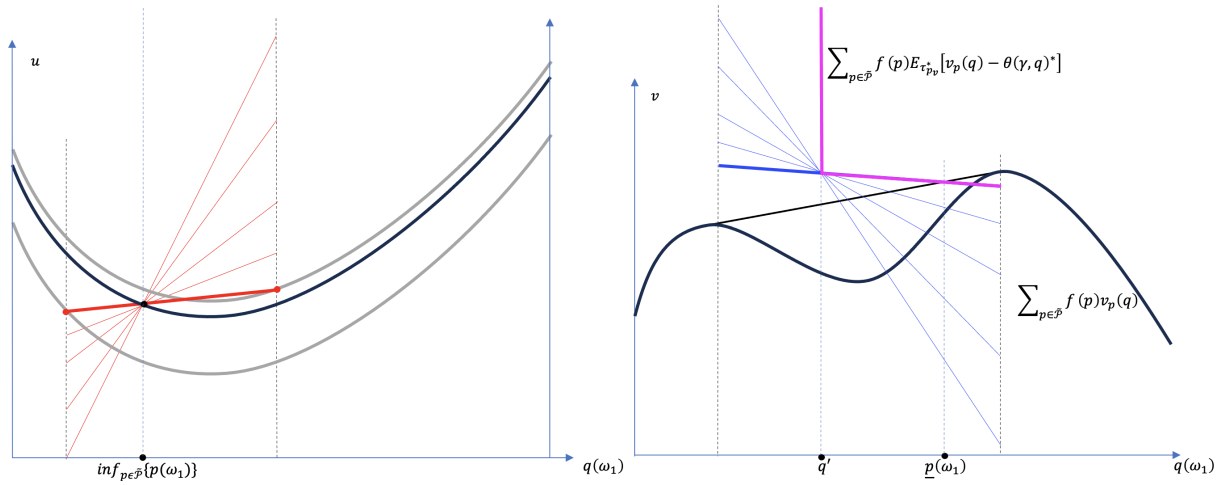

Figure 6: Optimal Signal-Contingent Mechanism

Suppose the thick linear curve on the left panel of Figure 6 represents the boundary mechanism. This mechanism establishes the minimum reward-to-punishment ratio necessary to satisfy both the participation constraint and incentive compatibility. For a fixed targeted posterior interval, these two constraints can continue to be satisfied as the Sender increases the gap $\theta\big(1, q_v(s_2)\big) - \theta\big(1, q_v(s_1)\big) > 0$ and steepens the linear curve.

The impact of the modifications on the Sender's payoff is illustrated in the right panel of Figure 6. The blue linear curve represents the boundary mechanism. When the curves are steepened, as indicated by the left panel of the graph, the Sender with $p < q'$ extracts

a higher information value from the Receiver. This leads to infinite optimal mechanisms. Conversely, the Sender with $p > q'$ incurs a loss due to this modification. It is easy to show that $p > \inf_{p\in\tilde{\mathcal{P}}} p(\omega')$ holds if and only if $p > q'$ holds. Therefore, a Sender with $p > \inf_{p\in\tilde{\mathcal{P}}} p(\omega')$ will optimally choose a finite boundary mechanism. As shown by the graph, $\inf_{p\in\tilde{\mathcal{P}}}\{p(\omega')\}$ and $\underline{p}(\omega')$ are independent. When the objective of mechanism design is to retain those Receivers with higher natural information value, the Sender may be able to collect this value as an extra benefit. Specifically, if $\inf_{p\in\tilde{\mathcal{P}}}\{p(\omega')\} < \underline{p}(\omega')$, the Sender with $p > q'$ where p is only slightly greater than $\inf_{p\in\tilde{\mathcal{P}}}\{p(\omega')\}$ can leverage this additional information benefit to offset the negative impact resulting from the disparity in beliefs with the targeted Receivers.

In a persuasion game, when the optimal mechanism possesses infinite characteristics, the information value produced by the mechanism may exceed the value attributed to the action itself. In the real world, such instances of putting the cart before the horse are not uncommon. For example, the bet-on (in Chinese, Duidu) agreement is common in the Chinese venture capital market today. When faced with high levels of risk in a financial market, employing state- or signal-contingent agreements can potentially facilitate the resolution of disputes between start-ups and venture capital firms regarding a company's future performance, thereby encouraging investment. However, it is not uncommon for the compensation specified in the contract to be disproportionately high compared to the potential value of the project.

The infinite characteristic of $\theta(\gamma, q)$ can essentially be attributed to the unbounded $\sup_{q\in\text{supp}(\tau_v)}\{\theta(1,q\} - \inf_{q\in\text{supp}(\tau_v)}\{\theta(1,q\}$, as illustrated in Figure 6. Hence, a proposed solution for ensuring the compactness of Θ is to impose a constraint $\sup_{q\in\text{supp}(\tau_v)}\{\theta(1,s\} - \inf_{q\in\text{supp}(\tau_v)}\{\theta(1,s\} \leq \iota$ where $\iota \in [0,\infty)$. This condition restricts the gap between the upper and lower curves depicted in Figure 5 and the left panel of Figure 6. As long as this gap is finite, both the optimal mechanism and the information value it generates will also be finite with the help of ϵ.

The cost of imposing such a constraint is the potential weakening of Proposition 6's validity. When the Sender intends to exclude a subset of Receivers with significantly higher natural information value than those the Sender desires to retain in the game, it is necessary to establish a substantial reward value in order to satisfy participation constraints. Simultaneously, it is crucial to set the punishment at a sufficiently high level to ensure the satisfaction of incentive compatibility. Consequently, a smaller gap between the reward and punishment may not simultaneously satisfy the participation constraint and incentive compatibility, causing the mechanism design to fail. The consideration of the trade-off between the feasibility of mechanism design and the realization of a reasonable optimal mechanism becomes important when the public seeks to regulate contracts pertaining to persuasion. However, the discussion of this topic is beyond the scope of this research, and therefore, we will defer it to future studies.

7 Application

To test the validity of the aforementioned generalized theory, we apply it to a stylized model representing a specific scenario in which the issue of over-persuasion arises.

7.1 Model

Consider a scenario in which there exists a seller and multiple buyers in the market. The seller acquires or produces the good at no cost and subsequently sells it at a price denoted as x. In order to simplify the problem and place greater emphasis on the persuasion aspect rather than the pricing analysis, we adopt the assumption that the seller acts as a sales representative of a brand who has no authority to change the price that has been set. The buyers' valuation of the good (such as an umbrella) can be either 0 or 1, contingent upon the actual state $\omega \in \{\omega_0, \omega_1\}$ (such as weather), which is unknown to all players. The buyers can be categorized into two groups based on their prior beliefs regarding state ω_1, denoted

as p_a and p_b, with a distribution of $f(p)$, $p = p_a, p_b$. It is assumed that $p_b > x > p_a$. To simplify the problem without loss of generality, we assume that the seller's prior belief p_v is equal to p_a. Before buyers choose a decision between purchasing ($\alpha = 1$) or not purchasing ($\alpha = 0$), the sender can design an information structure that reveals signals to influence buyers' beliefs and decisions. Let $q_a(s)$ and $q_b(s)$ represent the posterior beliefs about the state ω_1 after the persuasion, respectively. Additionally, let τ_p denote the distribution of these posterior beliefs when a player's prior belief is p. The seller's simplified problem is:

$$
\begin{aligned}
&\max_{\tau_{p_a}} E_{\tau_{p_a}}\left[f(p_a)v_{p_a}(q) + f(p_b)v_{p_b}(q)\right] \\
&s.t.\ E_{\tau_{p_a}} q = p_a \\
&\qquad v_{p_a}(q) = \begin{cases} 0 & \text{if } 0 < q < x \\ x & \text{if } x \le q \le 1 \end{cases} \\
&\qquad v_{p_b}(q) = \begin{cases} 0 & \text{if } 0 < q < \frac{xp_a(1-p_b)}{xp_a(1-p_b)+(1-x)(1-p_a)p_b} \\ x & \text{if } \frac{xp_a(1-p_b)}{xp_a(1-p_b)+(1-x)(1-p_a)p_b} \le q \le 1 \end{cases}
\end{aligned} \tag{12}
$$

7.2 Over-Persuasion

As a canonical persuasion example, when $f(p_b) = 0$, the optimal information structure generates $\tau^*_{p_a}(0) = \frac{x-p_a}{x}$ on s_1 and $\tau^*_{p_a}(x) = \frac{p_a}{x}$ on s_2, resulting in the expected payoff $E_{\tau^*_{p_a}} v_{p_a}(q) = p_a$.

This persuasion strategy over-persuades the buyers with $p_b > x$, who would have already made the purchase without the persuasion. With the baseline optimal persuasion strategy $\tau^*_{p_v}$, these buyers will update their posterior belief to $q_b(s_1) = 0$ and decide not to purchase when s_1 occurs with probability $\frac{x-p_a}{x} > 0$. According to our definition 2, the over-persuasion issue arises in this persuasion game because of $\{p | p \in \mathcal{P}(\tau^*_{p_a})\} = p_b$.

When $f(p_b) > 0$, the seller either maintains the baseline strategy or deviates to

$\tau^{**}_{p_a}\left(\frac{xp_a(1-p_b)}{xp_a(1-p_b)+(1-x)(1-p_a)p_b}\right) = 1-\frac{p_a(1-p_a)(p_b-x)}{x(1-x)(p_b-p_a)}$ and $\tau^{**}_{p_a}(x) = \frac{p_a(1-p_a)(p_b-x)}{x(1-x)(p_b-p_a)}$, resulting in an expected payoff of $f(p_a)\frac{p_a(1-p_a)(p_b-x)}{(1-x)(p_b-p_a)} + f(p_b)x$. Since $x > p_b$, the seller maintains the baseline strategy when $f(p_b)$ is relatively small and deviates to $\tau^{**}_{p_v}$ when $f(p_b)$ is sufficiently large. In Figure 7, Panel (i) displays the aggregated value function $f(p_a)v_{p_a}(q)+f(p_b)v_{p_b}(q)$. The middle section of the piecewise graph with the value of $f(p_b)$ is located at a higher position when $f(p_b)$ is larger, leading to the dominance of $\tau^{**}_{p_a}$ over $\tau^{*}_{p_a}$ as an optimal strategy.

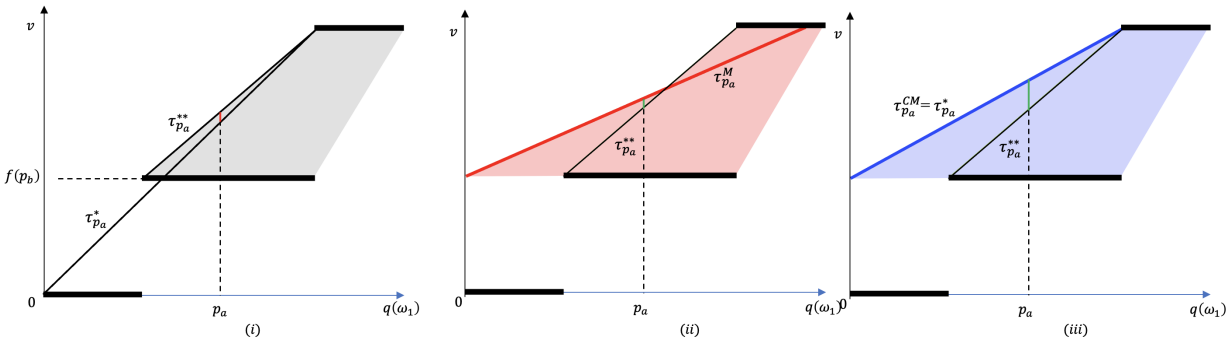

Figure 7: Value Functions with Different Mechanisms

p_b belongs to $\hat{\mathcal{P}}$ we defined in Proposition 1 because $q = p_b$ is associated with the maximum value of $v(q) = v_a(q)$ and is adjacent to a smaller value of its concave closure, $\tilde{v}(q)$. Particularly, any information structure that generates a positive persuasion value for p_a generates a negative value for p_b. In panel (i) of Figure 7, a strategy must be entirely contained within the shaded area to fully prevent buyers with p_b from reversing their initial purchase decision. The baseline strategy $\tau^{*}_{p_a}$ generates a negative persuasion value, $p_a - x < 0$, for buyers with p_b, resulting in a portion of $\tau^{*}_{p_a}$ falling outside the shaded region of the graph. In contrast, $\tau^{**}_{p_a}$ is entirely contained within the shaded region. When it is chosen to preserve a non-negative persuasion value for buyers with p_b, the persuasion value for buyers with p_a is $\frac{p_a(1-p_a)(p_b-x)}{(1-x)(p_b-p_a)} - p_a < 0$. Therefore, the over-persuasion issue in this example results in a loss of persuasiveness.

7.3 Signal-Independent Mechanism

To investigate the potential mechanisms that may enhance persuasiveness, it is first necessary to examine the decision-making motivations of buyers. The buyers possess the choice to either purchase the good or not, making their payoff function to be $u^*(q) = \max\{0, q-x\}$. Here, q is determined by whether or not the persuasion signal is obtained and what signal is realized. Accordingly, their information value is $U_p(\tau_p) = \left[E_{\tau_p} \max\{0, q-x\}\right] - \max\{0, p-x\}$.

Since the buyers with p_a will not buy the good without persuasion, the objective of designing a persuasiveness-enhancing mechanism should be to exclude buyers with p_b, resulting in $p_b = \mathcal{P}^*$. Consequently, $q_a(s_2)$ should always be greater than x; otherwise, buyers with p_a will consistently have zero information value and will be excluded as long as their peers with p_b are excluded by information fee. Given the price x, when the seller choose $\big(q_a(s_1), q_a(s_2)\big) = (q_1, q_2)\big)$, the buyers' information value are as follows:

$$\begin{aligned} U_{p_a} &= \frac{(p_a - q_1)(q_2 - x)}{q_2 - q_1} \\ U_{p_b} &= \frac{(p_a - q_2)\left[x p_a (q_1 + p_b - 1) - q_1 p_b (x + p_a - 1)\right]}{p_a(1-p_a)(q_2 - q_1)}. \end{aligned} \tag{13}$$

When charging buyers an information fee of $\theta > 0$ for a persuasion signal, the simultaneous satisfaction of participation constraint and incentive compatibility requires that $U_{p_a} - \theta \geq 0 > U_{p_b} - \theta$. Hence, the condition $U_{p_a}(\tau) > U_{p_b}(\tau)$ serves as a determinant for the feasibility of mechanism design when τ is chosen as the persuasion strategy. This condition is closely related to the support of $f(p)$ or, in this application, the values of p_a and p_b.

When the condition $U_{p_a} > U_{p_b}$ holds with $q_1 = 0$ fixed, the payoff associated with mechanism design exceeds that generated by $\tau^{**}_{p_v}$ if $(1-x)[p_a(1-p_a) - x(1-p_b)] > p_a(1-p_a)(p_b - x)$. Panel (iii) of Figure 8 shows the region that satisfies this condition

given that $p_a < x < p_b$. According to the graph, the optimal p_a should be located near 0.5 so that it is close to the bottom of the $u^*(q)$ curve. An excessively high value of p_a would necessitate a higher value of x that also improves the information value of p_b, whereas an excessively low value of p_a would cause a significant deviation from the curve bottom. On the other hand, it is desirable for the value of p_b to be as large as possible in order to approach the boundary of the curve, thereby minimizing its information value to the greatest extent possible. To simplify the analysis, the values of $p_a = 0.5$, $p_b = 0.85$, and $x = 0.75$ are employed in the subsequent optimization and analysis.

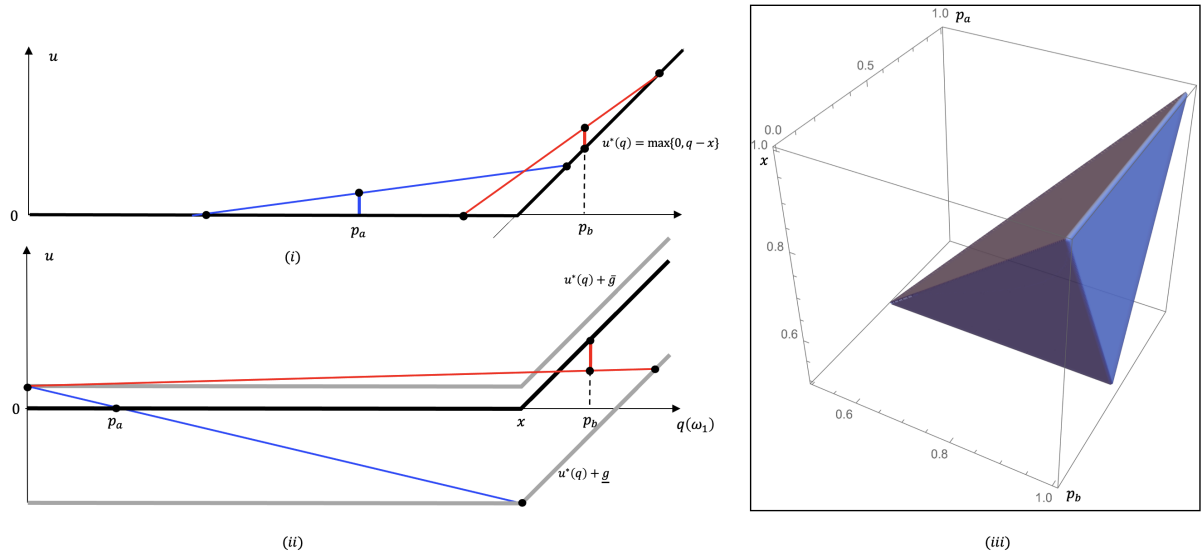

Figure 8: Incentive Compatibility and the Support of $f(p)$

If feasible, the (ii) panel of Figure 7 illustrates how the mechanism design could increase the seller's expected payoff. Starting from $\tau^{**}_{p_a}$, the boundary of the shaded region in panel (i) of the figure, simultaneously increasing q_2 and decreasing q_1 raises the information value for all buyers. Under certain conditions that make mechanism design feasible, the increase in information value for buyers with p_a exceeds that for buyers with p_b. As a result, the shaded region expands from the area depicted in panel (i) to the area depicted in panel (ii). With this expansion, the value function can attain p_a even if $q < \frac{xp_a(1-p_b)}{xp_a(1-p_b)+(1-x)(1-p_a)p_b}$,

given that buyers with p_b opt to maintain their initial decision. This is impossible without the mechanism design, which prevents buyers with p_b from receiving the potential signal that advises them against making the purchases. According to Figure 7, $\tau^M_{p_a}$ becomes the optimal information structure with the designed mechanism, partially recovering the persuasion value of buyers with p_a that was compromised by $\tau^{**}_{p_v}$.

With the mechanism design and the given values of the parameters, the Sender's problem becomes

$$\begin{aligned} \max_{\tau_{p_a}} \; & E_{\tau_{p_a}}\Big[f(p_a)v_{p_a}(q) + f(p_b)v_{p_b}(q)\Big] \\ s.t. \; & E_{\tau_{p_a}} q = 0.5 \\ & q_2 > \max\{0.75, \frac{0.066 - 0.1q_1}{0.069 - 0.088q_1}\} \\ & 0 \leq q_1 \leq 0.5 \end{aligned} \tag{14}$$

By substituting the given constraints into the objective function, $(q^M_1, q^M_2) = (0, 0.957)$ is the optimal solution to the seller's persuasion problem under mechanism design. Given that $\theta(0) = 0$, the information fee is set to $\theta(1) = 0.108$, which is equal to the information value of the buyers with p_a. Even if $\beta = 0$ so that the seller does not collect the information fee, this mechanism generates seller's expected payoff of $0.75 \times \frac{0.5}{0.975} f(p_a) + 0.75 f(p_b) = 0.385 f(p_a) + 0.75 f(p_b)$, which is greater than $0.286 f(p_a) + 0.75 f(p_b)$, the expected payoff generated by $\tau^{**}_{p_a}$.

7.4 Signal-Contingent Mechanism

When the indirect utility function $u^*(q)$ does not possess the desired characteristics for ensuring incentive compatibility, the seller can employ signal-contingent mechanisms to induce separate decisions by different buyers. Because $p_v = p_a$ is assumed, the seller in this application will not receive an infinite benefit from the information value. We suppose the seller chooses the minimum θ when she is indifferent, which ensures a compact Θ.

The condition in Proposition 6 is satisfied because there are only two types of buyers. According to Proposition 6, the seller can always establish the necessary participation constraint and incentive compatibility to exclude only buyers with p_b regardless of the persuasion strategy employed. She then selects the persuasion strategy $\tau_{p_a}^{CM}$ with support $\big(q_a(s_1), q_a(s_2)\big) = (q_1, q_2)$ to maximize the sum of persuasion value and information value of buyers with p_a. In an optimal persuasion strategy, $q_a(s_2) \geq x$ must be satisfied; otherwise, buyers with p_a are associated with zero persuasion value and information value. Given the conditions of $q_1 < p_a$ and $q_2 \geq x$, the total value a seller can obtain from buyers with p_a is $f(p_a)\frac{x(p_a - q_1)}{q_2 - q_1}$. This value is maximized when $\tau_{p_a}^{CM} = \tau_{p_v}^{*}$, which induces $(q_1, q_2) = (0, x)$.

To ensure the buyers with q_b are excluded when $(q_1, q_2) = (0, x)$ is adopted at the second stage, the seller needs

$$\begin{aligned} U_{p_a}(\tau_{p_a}^{*}) &= \bar{g}\frac{x - p_a}{x} + \underline{g}\frac{p_a}{x} \geq 0 \\ U_{p_b}(\tau_{p_a}^{*}) &= \left[q_b(s_2) - x + \underline{g}\right]\frac{p_b}{q_b(s_2)} + \bar{g}\frac{q_b(s_2) - p_b}{q_b(s_2)} - (p_b - x) \leq -\epsilon, \end{aligned} \tag{15}$$

where $\bar{g} > 0$ and $\underline{g} < 0$ represents reward and punishment in the mechanism, and $q_b(s_2) = \frac{x\frac{p_b}{p_a}}{x\frac{p_b}{p_a} + (1-x)\frac{1-p_b}{1-p_a}}$.

Rearranging the conditions in (15) and set $\epsilon = 0$ gives

$$\begin{aligned} \bar{g}(x - p_a) &\geq -\underline{g}p_a \\ \bar{g} - \underline{g} &\geq \frac{x(1 - p_b)}{p_b - p_a}, \end{aligned} \tag{16}$$

where the first condition ensures the participation constraint and the second guarantees the incentive compatibility.

Given the assumption that the seller always sets $\bar{g} - \underline{g}$ to the minimum level when she is indifferent, the optimal signal-contingent mechanism is $\bar{g}^{*} = \frac{p_a(1-p_b)}{p_b - p_a}$ and $\underline{g}^{*} = -\frac{(1-p_b)(x-p_a)}{p_b - p_a}$.

This mechanism allows for $q_a(s_1)^{MC} = q_b(s_1)^{MC} = 0$, $q_a(s_1)^{MC} = x$, and $q_b(s_2)^{MC} = \frac{x\frac{p_b}{p_a}}{x\frac{p_b}{p_a}+(1-x)\frac{1-p_b}{1-p_a}}$. As shown in panel (iii) of Figure 7, the signal-contingent mechanism further enlarges the shaded region to the maximum extent, even allowing $\tau^*_{p_a}(\mathcal{P}^*)$ that generates the highest possible persuasion value payoff for the seller.

Signal-contingent mechanism becomes more important when the distribution $f(p)$ does not provide ideal support for the signal-independent mechanism. For example, consider the values $p_a = 0.2$, $p_b = 0.85$, and $x = 0.75$, which lie outside the region depicted in panel (iii) of Figure 8. $U_{p_a} > U_{p_a}$ requires $q_1 > \frac{0.0195-0.0815q_2}{0.0095+0.0325q_2}$. Since $q_2 > 0.75$ is necessary, it follows that q_1 cannot be smaller than $\frac{x\frac{p_b}{p_a}}{x\frac{p_b}{p_a}+(1-x)\frac{1-p_b}{1-p_a}} = 0.117$. This implies that with a signal-independent mechanism, the shaded region in panel (i) of Figure 7 cannot be expanded even to the extent depicted in panel (ii). Nevertheless, the seller can still find a pair of $(\bar{g}, \underline{g})$ satisfying (16) for this set of parameters in order to make the signal-contingent mechanism feasible, thereby preserving the persuasiveness enhancement.

8 Concluding Remarks

When attempting to persuade a group of Receivers with heterogeneous prior beliefs, an over-persuasion issue is likely to arise and may result in a loss of persuasiveness. This issue may offer an explanation for the backfire effect at a collective level. As the underlying cause of the over-persuasion issue, Receivers' heterogeneous prior beliefs also provide the foundation for mechanism design as a potential solution. Specifically, an identical information structure can yield varying information values for the Receivers who hold different prior beliefs. In addition, receivers with heterogeneous prior beliefs have differing estimations of the probability that a given signal will be realized. These two characteristics can be employed to establish the participation constraint and incentive compatibility for signal-independent and signal-contingent mechanisms, respectively.

This study has generated several potential topics that cannot be comprehensively addressed within the scope of this discussion. Our discussion of mechanism design is based

on predetermined $v(q)$ and $u^*(q)$. In the real world, they are more often given as baselines. With certain constraints, the Sender may be able to modify them to optimally facilitate the mechanism design. Second, to emphasize the impact of the mechanism on the persuasion outcome, we simplify the assumption of β and minimize its value whenever the topic permits. However, it may have more complicated properties that reflect the transaction cost associated with the market structure and can therefore be regulated. Its elaboration may prove useful when examining the regulation of persuasion contracts, particularly those that are contingent upon realized signals.

Persuasion can be an effective technique for influencing individuals' behavior when employed with benevolent intentions. However, it also gives rise to concerns regarding the potential belief manipulation when wielded maliciously. Over-persuasion issue may be one of the primary limitations on its power, particularly in terms of the scope to which it can be applied. The design of a mechanism, especially one that is contingent upon signals, can significantly unleash such a power. As a manifestation of the condition that establishes the feasibility of signal-contingent mechanisms, when a particular religion or ideology is exclusively prevalent in society, persuasions through the stories associated with these beliefs can be extremely effective. A healthy society, however, must embrace a variety of voices and beliefs to avoid being vulnerable to belief manipulation. We advocate persuasion and pursue its effectiveness, but we must also be alert to maximized persuasiveness, which is not always optimal.

References

Alonso, Ricardo and Odilon Câmara, "Bayesian persuasion with heterogeneous priors," *Journal of Economic Theory*, 2016, *165*, 672–706.

Au, Pak Hung, "Dynamic information disclosure," *The RAND Journal of Economics*, 2015, *46* (4), 791–823.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Bickart, Barbara A, "Carryover and backfire effects in marketing research," *Journal of Marketing Research*, 1993, *30* (1), 52–62.

Boyaci, Tamer, Soudipta Chakraborty, and Huseyin Gurkan, "Persuading Skeptics and Fans: Information Design for New Experience Goods," *Available at SSRN*, 2022.

Dworczak, Piotr and Alessandro Pavan, "Preparing for the worst but hoping for the best: Robust (Bayesian) persuasion," *Econometrica*, 2022, *90* (5), 2017–2051.

Gitmez, A Arda and Pooya Molavi, "Polarization and media bias," *arXiv preprint arXiv:2203.12698*, 2022.

Gu, Mofan, Bruce Taylor, Harold A Pollack, John A Schneider, and Nickolas Zaller, "A pilot study on COVID-19 vaccine hesitancy among healthcare workers in the US," *PLoS One*, 2022, *17* (6), e0269320.

Guo, Yingni and Eran Shmaya, "The interval structure of optimal disclosure," *Econometrica*, 2019, *87* (2), 653–675.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Khubchandani, Jagdish, Elizabeth Bustos, Sabrina Chowdhury, Nirbachita Biswas, and Teresa Keller, "COVID-19 vaccine refusal among nurses worldwide: review of trends and predictors," *Vaccines*, 2022, *10* (2), 230.

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li, "Persuasion of a privately informed receiver," *Econometrica*, 2017, *85* (6), 1949–1964.

Laclau, Marie, Ludovic Renou et al., "Public persuasion," *Working Paper*, 2017.

Lewandowsky, Stephan, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological science in the public interest*, 2012, *13* (3), 106–131.

Nyhan, Brendan and Jason Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 2010, *32* (2), 303–330.

_ and _ , "Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information," *Vaccine*, 2015, *33* (3), 459–464.

Pham, Hien, "Screening with Information Design and Heterogeneous Priors," *Available at SSRN 4304837*, 2023.