Mediation Design by an Informed Party[☆]

Andrés Salamanca[1]

Department of Economics, Ca' Foscari University of Venice, Italy

Abstract

This paper investigates mediation design by an *informed* party—the expert—who selects a mediation mechanism at the *interim* stage, i.e., after observing her private information. We consider a basic strategic environment in which the expert's information is a binary state of the world and an uninformed decision-maker chooses an action on the real line. Preferences are quadratic, with state-contingent bliss points that differ across parties. Our framework is closely related to the informed-principal problem, and our analysis builds on its foundational results. To delimit the reasonable predictions of the interim mediation design game, we sequentially refine the set of perfect Bayesian equilibrium outcomes by applying several solution concepts: the strong solution, core mechanisms, neologism-proof equilibria, and the neo-optimum.

Keywords: Mediation design, informed principal, sender-receiver, core mechanism, strong solution, neologism proof, neo-optimum.

JEL Classification: D82, D83.

1. Introduction

Unlike the canonical principal–agent setting, where *only* the principal designs the agent's incentive scheme, the mediation framework allows either disputant to control the communication channels and thus, in principle, to hold full institutional authority over the choice of the mediation mechanism. Moreover, the mediator—an autonomous third party capable of shaping the parties' incentives—may also be delegated the design of the mechanism, although the mediator's objectives need not align with those of the disputants.

Accordingly, any analysis of mediation design must first specify which party (or parties) holds design authority and what objectives that designer pursues. In this paper, we study mediation design by an *informed* disputant. This scenario contrasts with the standard principal-agent setup, wherein the authority to design the coordination mechanism rests with the uninformed party—the principal.

[☆]This study received funding from the European Union through the Next-Generation EU/Italian National Recovery and Resilience Plan (NRRP)—Mission 4, Component 2, Investment 1.1: Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)—Project number 2022WS8AJY–CUP H53D23002600006. This manuscript reflects only the author's views and opinions, neither the European Union nor the European Commission can be considered responsible for them. I wish to express my sincere gratitude to Françoise Forges and Jin Yeub Kim for their insightful comments. I am particularly grateful to Frédéric Koessler for his stimulating questions and constructive discussions. This manuscript has been copy-edited with the assistance of AI tools, which were employed to enhance sentence structure and lexical variation.

Email address: andres.salamanca@unive.it (Andrés Salamanca)

[1]This version: January 12, 2026.

We study a basic strategic model with an informed party—the expert—who possesses private information about a binary state of the world. This information is relevant to an uninformed decision-maker who must choose an action on the real line, affecting the welfare of both parties. Preferences are quadratic, and the parties differ in their bliss points in each state, creating a conflict of interest. This conflict gives rise to two distinct commitment problems: *adverse selection*, stemming from the expert's inability to credibly commit to truthfully disclose her information; and *moral hazard*, arising from the decision-maker's inherent authority to choose his own action, which limits his ability to credibly commit to a particular decision.

In our setting, a mediator engages in behavior that is designed to elicit truthful information from the expert and exercise influence on the decision-maker by carefully communicating partially informative messages. The mediator first conducts a private meeting—a caucus—with the expert, aiming to incentivize honest revelation of her private information. Subsequently, the mediator distorts this information and transmits it to the decision-maker as a recommendation. We refer to any mediation mechanism structured in this manner as a *mediation plan*.

Under mediation, the expert's report is unverifiable to both the mediator and the receiver, allowing the expert to strategically misrepresent her private information. Moreover, the mediator's recommendation is non-binding; thus, the decision-maker remains free to select any action other than the mediator's suggested one. Consequently, for a mediation plan to be implementable, it must satisfy a set of constraints ensuring that the expert has no incentive to misrepresent the true state—known as *truth-telling incentive constraints*—and that the decision-maker has no incentive to deviate from the mediator's recommendation—known as *obedience incentive constraints*. A mediation plan satisfying these constraints is termed *incentive compatible*.

The model examined here was first introduced by Mitusch and Strausz (2005). In their framework, the decision-maker was assumed to hold full bargaining power and could therefore choose the mediation plan (or, equivalently, appoint the mediator) that best served his payoff-maximization objectives. Their analysis corresponds to a generalized principal–agent problem as formulated by Myerson (1982).[2] As is well known, this problem reduces to a well-defined optimization problem: selecting an incentive-compatible mediation plan that maximizes the decision-maker's *ex-ante* expected payoff.

In a more recent contribution, Salamanca (2024) examines the same model under an alternative scenario in which the mediator designs the mediation plan to favor the experts's interests.[3] This formulation likewise reduces to a straightforward optimization problem: selecting an incentive-compatible mediation plan that maximizes the experts's *ex-ante* expected payoff.

The term "ex-ante" here refers to the assumption that the mediation designer evaluates the parties' welfare behind a veil of ignorance—i.e., *before* acquiring any private information. This assumption reflects the fact that, in the settings studied by Mitusch and Strausz (2005) and Salamanca (2024), neither the decision-maker nor the mediator knows the actual state when selecting the mediation plan. In our framework, by contrast, the mediation plan is chosen at the *interim* stage—i.e., *after* the expert has privately observed the actual state. This shift introduces additional conceptual challenges, since the very act of selecting a mediation plan may itself reveal information about the state.

[2]This likely explains why Mitusch and Strausz (2005) refer to the parties as principal (decision-maker) and agent (expert). In the present paper, however, we adopt the terminology "expert" and "decision-maker" to avoid potential confusion arising from the fact that, in our framework, it is the agent who designs the mediation mechanism—contrary to what the principal-agent terminology suggests.

[3]When the mediator's objectives coincide with those of the decision-maker, the problem reduces to the original setting analyzed by Mitusch and Strausz (2005).

Choosing a mediation plan after acquiring private information leads the expert into an *inscrutability dilemma*: to conceal her private information, she must select a mediation plan independently of the actual state, yet her knowledge of the state influences her preferences over plans. If the expert designs the mediation plan to maximize her utility conditional on the actual state (i.e., her interim utility), the choice itself will inherently convey information about that state. With this information, the decision-maker may find new opportunities to profit by disregarding the mediator's recommendation. Consequently, even an incentive-compatible mediation plan may fail to be implementable. The expert's choice must therefore embody an inscrutable compromise between her payoff-maximizing objectives across different states.

Formally, the mediation design problem faced by the expert is a particular instance of the *informed-principal problem* studied by Myerson (1983). This setting presents a more complex strategic environment, embedding the mediation process within a signaling game. As a result, the expert's mediation design problem cannot be reduced to a simple optimization exercise. Myerson's (1983) seminal work develops a general theoretical framework for characterizing appropriate "inscrutable compromises" in such environments, introducing a variety of solution concepts grounded in both cooperative and non-cooperative game theory. In our setting, we apply three of these concepts: perfect Bayesian equilibrium (PBE), core mechanisms, and strong solutions.[4] Myerson also introduced an additional concept, the *neutral optimum*. However, this notion has only been defined in environments with finitely many types and actions, and it remains unclear how its analytic characterization might be appropriately extended to our model with a continuum of actions.

To present our results, it is useful to first introduce several notions. A mediation plan is said to be *fully revealing* if it discloses the actual state with certainty. Such a plan can always be constructed to satisfy the obedience constraints, though it may fail to induce the expert to report truthfully. Whenever the fully-revealing plan satisfies the truth-telling incentive constraints, we say that there is no *ex-ante misrepresentation problem*.[5] Another mediation plan central to our analysis is the *non-revealing* plan, which corresponds to the expert's option to remain silent throughout the mediation process and is, therefore, always incentive-compatible. Finally, our results will also hinge on a key statistic: the *relative degree of conflict* between the disputing parties, which captures the extent of the distance between their bliss points across states.

The informed-party problem studied here can be represented as an extensive-form game, where the expert first selects the mediation plan, followed by both parties participating in the induced mediation game. The informed-principal literature commonly uses variations of this multi-stage mechanism-selection game (Myerson, 1983; Maskin and Tirole, 1992; Severinov, 2008; Mylovanov and Tröger, 2012, 2014, 2025; Balkenborg and Makris, 2015). To establish that a mediation plan is a PBE outcome of the informed-party problem, one must verify that for every potential deviation to an alternative plan there exist off-path beliefs and a continuation equilibrium of the ensuing mediation game that render the deviation unprofitable. Identifying such belief–equilibrium pairs is generally nontrivial, complicating the task of characterizing all PBE outcomes. Nevertheless, the following key property facilitates broad characterizations: there exists an action that is optimal for the decision-maker when he attributes any deviation to a particular expert type and that

[4] Myerson (1983) uses the term *expectational equilibrium* to refer to a PBE. This notion differs from other definitions in two main respects: first, beliefs are required to be Bayes-consistent only immediately following the selection of the mediation plan, rather than at every information set where updating is feasible; and second, continuation equilibria are not constrained to satisfy sequential rationality. In contrast, in our framework we impose Bayes-consistency of beliefs whenever updating is possible and require actions to be sequentially rational given these beliefs.

[5] When there is no ex-ante misrepresentation problem, the truth-telling constraints may be omitted from the *ex-ante* formulation, reducing the problem to a standard Bayesian persuasion framework. However, this implication does not extend to the interim setting.

uniformly punishes both types, irrespective of the deviating plan.

Building on this property, we show that, in the absence of ex-ante misrepresentation, any incentive-compatible mediation plan that guarantees each expert type an interim utility no lower than under full revelation can be sustained as a PBE outcome. This is only a *partial* characterization—additional equilibria may still exist—but the expert can always secure her fully-revealing payoffs, regardless of whatever inference the decision-maker draws from the selection of the fully-revealing plan. Hence, she can exclude any outcome that assigns any of her types less than the full-disclosure utility level.

By contrast, when ex-ante misrepresentation is present, the same property yields a *complete* characterization of PBE outcomes: *all* incentive-compatible plans are PBE outcomes.

The PBE concept allows substantial flexibility in specifying posterior beliefs after an off-path deviation. As a result, a wide range of outcomes can be sustained in equilibrium. With the aim of narrowing the set of PBE outcomes—and ideally identifying a single mediation plan that constitutes the appropriate intertype compromise—we turn to alternative solution concepts.

As the name suggests, *core mechanisms* can be understood in terms of deviations by "coalitions" of types. An incentive-compatible plan is a core mechanism if no coalition of the expert's types can strictly benefit from deviating to an alternative plan that remains incentive-compatible whenever the decision-maker believes the expert's actual type lies within that coalition or any larger one. Applying the notion of core mechanisms to the informed-party problem refines the set of PBE outcomes characterized above by retaining only those equilibria that are *undominated*—that is, efficient for the different types of the expert among all incentive-compatible mediation plans. However, the set of undominated PBE outcomes often remains large, and some of these equilibria may still be sustained by implausible off-path beliefs. Therefore, to generate meaningful predictions, stronger equilibrium-selection criteria are required.

In our framework, a mediation plan is *safe* if it is fully revealing and incentive compatible. A *strong solution* is a safe plan that is also undominated. According to Myerson (1983), the strong solution is the most compelling outcome: it can be implemented regardless of what the decision-maker infers from its selection, and any alternative plan strictly preferred by some types would violate obedience when the decision-maker believes the expert's type is among those beneficiaries. However, a strong solution often fails to exist—either because no safe plan is possible due to an ex-ante misrepresentation problem, or because the safe plan is not efficient for the expert.

We found that, under the necessary condition of no ex-ante misrepresentation, a strong solution exists when the relative degree of conflict is sufficiently small; otherwise, the fully-revealing plan is dominated.

As a further attempt to identify a unique solution to the interim mediation design problem, we turn to Farrell's (1993) neologism-proofness criterion. A neologism is a deviation by the expert accompanied by the claim "my type lies in some set T." Such a claim is credible if precisely the types in T strictly prefer it to be believed over the outcome of the putative equilibrium. A *neologism-proof equilibrium* is thus a PBE outcome against which no credible neologism exists.

Our analysis shows that, in the absence of ex-ante misrepresentation, neologism-proof equilibria coincide with core mechanisms, implying that this credibility criterion offers no refinement beyond efficiency. In contrast, in the presence of ex-ante misrepresentation, neologism-proofness can narrow the set of core mechanisms: the set of neologism-proof equilibrium outcomes consists of all undominated plans that guarantee the *jeopardized type*—i.e., the type suffering from ex-ante misrepresentation—at least its fully-revealing payoff. For some parameter configurations, however, no such plan exists. To address this limitation, Mylovanov and Tröger (2025) introduced

a related refinement, the *neo-optimum*, which relaxes Farrell's (1993) credibility condition. A *neo-optimum* is defined as an incentive-compatible mediation plan yielding payoffs at least as high as the limits of neologism-proof payoffs. Every neologism-proof PBE outcome is a neo-optimum, and every neo-optimum is a core mechanism.

In our setting, whenever a neologism-proof PBE outcome does not exist, there is a unique neo-optimum, which coincides with the best incentive-compatible plan of the jeopardized type. Finally, when the payoff-maximization goals of both types align, the expert's best plan is well-defined and corresponds to the non-revealing plan.

Despite the simplicity of our binary-state setting, our results already highlight the limitations of the existing theory in providing a clear solution to the informed-party problem. While we were able to identify a unique "best" inscrutable compromise in many parameter configurations, there remains a non-degenerate set of cases where current solution concepts fail to deliver uniqueness. Unlike ex-ante mediation design, where a (generically) unique solution follows directly from a straightforward optimization problem, identifying an appropriate inscrutable compromise in the interim mediation design problem proves to be a far more subtle challenge—one that calls for further investigation.

Contribution and related literature. This paper contributes to the literatures on optimal mediation design, informed-principal problems, and communication in games. We integrate insights from these strands to study how an informed party can strategically design mediation mechanisms. In doing so, we connect the theoretical tools developed for informed-principal environments with the strategic communication issues central to mediation.

In our paper, we address the problem of *interim mediation design*—an open question largely overlooked in the optimal mediation literature on sender–receiver games (e.g., Mitusch and Strausz, 2005; Blume et al., 2007; Goltsman et al., 2009; Ivanov, 2010, 2014; Salamanca, 2021, 2024; Ganguly and Ray, 2023). Existing work has exclusively focused on mediation mechanisms designed by an *uninformed* party—either the decision-maker or the mediator—within an *ex-ante* framework, which reduces the analysis to a straightforward optimization problem. In contrast, we study optimal mediation design at the *interim* stage, where the mechanism choice itself can reveal information, embedding the mediation process within a signaling game.

Our framework is closely related to the informed-principal problem, and we draw on the theoretical foundations developed in that literature (e.g., Myerson, 1983; Maskin and Tirole, 1992; Severinov, 2008; Mylovanov and Tröger, 2012, 2014, 2025; Balkenborg and Makris, 2015). However, unlike most of these contributions, which are set in the principal–agent framework, we extend the analysis to mediation, where communication occurs through a neutral third-party making non-binding recommendations. By contrast, the principal–agent setting features an *arbitrator* (i.e., the mechanism) that renders a binding final decision. To our knowledge, this is the first systematic study applying the solution concepts developed for informed-principal environments to the broader problem of interim mediation design.

Within the informed-principal literature, the paper by Koessler and Skreta (2023) deserves special mention for its close connection to our contribution. The authors analyze an interim information design problem that differs from our interim mediation design problem primarily in that the expert does not face an adverse selection problem, and hence truth-telling constraints are not required. Put differently, their setting can be interpreted as an interim mediation design problem with an "omniscient" mediator, who issues recommendations to the decision-maker directly as a function of the *true* state.

The work most closely related to ours is Koessler and Skreta (2025), who analyze interim mediation design in a more general environment with multiple potentially informed decision-makers,

arbitrary utility functions, and finitely many types and actions. Their analysis is confined to characterizing the PBE outcomes of the interim mediation design problem. By contrast, we go further and investigate stronger solution concepts that refine the set of PBE outcomes. As we demonstrate in this paper, the set of PBE outcomes can be very large—an observation not unique to our model but common across other informed-principal settings. Following Myerson (1991, p. 107), the set of PBE outcomes can be interpreted as an *upper solution* to the interim mediation design problem: a solution concept that encompasses all reasonable predictions but may also admit some that are unreasonable. In this sense, PBE outcomes provide only a necessary condition for a satisfactory prediction. To obtain *exact solutions* capable of identifying appropriate inscrutable compromises, stronger solution concepts are required—concepts that rule out predictions that could never be deemed reasonable.

The remainder of the paper is organized as follows. Section 2 introduces the basic setup and defines the mediation process. Section 3 presents the informed-party problem and characterizes PBE outcomes. Section 4 studies core mechanisms, while Section 5 analyzes the existence of strong solutions. Section 6 considers neologism-proof equilibria and the neo-optimum. Section 7 concludes with a discussion and comparison of ex-ante and interim solutions.

2. The Model

We study a conflict scenario between an informed individual, referred to as the *expert*, and an uninformed decision-maker.[6] The payoffs of both parties depend on the underlying state of nature $\theta \in \Theta = \{1, 2\}$ and the action $y \in Y = \mathbb{R}$. The expert privately observes the realized state θ, whereas the decision-maker holds a prior belief $\pi \in (0, 1)$ that the state is $\theta = 2$. On the other hand, only the decision-maker has the authority to implement an action $y \in Y$.

Given the true state θ, the decision-maker's preferences are represented by the quadratic loss function $V_\theta(y) = -(y - y_\theta^d)^2$, where $y_\theta^d \in Y$ denotes his most preferred action in state θ—i.e., his *bliss point*. Similarly, the expert's utility is given by $U_\theta(y) = -(y - y_\theta^e)^2$, where $y_\theta^e \in Y$ is her bliss point in state θ.

A conflict of interest arises from the misalignment between the parties' preferences: in state θ, the expert favors the outcome y_θ^e, whereas the decision-maker prefers y_θ^d. Since only the expert is privately informed about the realized state, she has an incentive to misrepresent this information in order to steer the decision-maker's choice closer to her own preferred action.

The model is characterized by five parameters: the bliss points $\{y_\theta^i\}$ and the prior belief π. However, sometimes our results will be more succinctly expressed in terms of the following summary statistics: $\Delta^i := y_2^i - y_1^i$. When $\Delta^e \Delta^d < 0$, the preferences of the expert and the decision-maker are diametrically opposed, and communication yields no benefit to either party. Accordingly, we impose the *monotonicity assumption* $\Delta^e \Delta^d > 0$, which ensures that both players' bliss points move in the same direction across states. Without loss of generality, we adopt the convention that $\Delta^e > 0$ and $\Delta^d > 0$. Provided that the decision-maker benefits from learning the state (i.e., $\Delta^d \neq 0$), the monotonicity assumption rules out only the special case $\Delta^e = 0$, often referred to in the cheap-talk literature as *transparent motives*.

The decision-maker's choice depends solely on his belief about the state. Given an arbitrary belief $\rho \in [0, 1]$ that the true state is $\theta = 2$, he selects an action $y \in Y$ to maximize his expected utility:

$$\max_{y \in Y} (1 - \rho) V_1(y) + \rho V_2(y).$$

[6] We refer to the expert as female and the decision-maker as male.

Given the quadratic-loss preferences, the optimal action is then given by

$$\gamma(\rho) \coloneqq (1-\rho)y_1^d + \rho y_2^d. \tag{2.1}$$

This implies that the decision-maker's optimal action always lies in the interval between his two bliss points, i.e., within the *issue space* $[y_1^d, y_2^d]$. Moreover, under the assumption $\Delta^d > 0$, the function $\gamma(\cdot)$ is strictly increasing in ρ.

2.1. Mediation

We consider mediation mechanisms in which a trustworthy mediator privately collects non-verifiable reports from the expert and issues non-binding recommendations to the decision-maker. Formally, a *mediation plan* is a transition probability $\delta : \Theta \to \Delta(Y)$, where for each θ, $\delta(\cdot \mid \theta)$ is a probability distribution over the set of possible recommendations Y. The interpretation is that, upon receiving a report θ from the expert, the mediator draws a recommendation according to $\delta(\cdot \mid \theta)$.

The support of a mediation plan δ is defined as

$$supp\,(\delta) \coloneqq \bigcup_{\theta} supp\,(\delta(\cdot \mid \theta)),$$

where $supp\,(\delta(\cdot \mid \theta))$ denotes the support of the distribution $\delta(\cdot \mid \theta)$.

A mediation plan δ induces a *mediation game* between the expert and the decision-maker, denoted by $\Gamma_\delta(\pi)$. In this game, the expert privately communicates a report $\theta' \in \Theta$ to the mediator. Based on this report, the mediator draws a recommendation $y \in Y$ according to the distribution $\delta(\cdot \mid \theta')$. Finally, upon observing the recommendation, the decision-maker updates his prior belief π about the state and subsequently selects an action.

Any Nash equilibrium[7] of $\Gamma_\delta(\pi)$ induces an *equilibrium outcome* $\mu : \Theta \to \Delta(Y)$, where $\mu(y \mid \theta)$ denotes the probability that the action y is finally chosen by the decision-maker when the true state is θ. The expert's equilibrium payoffs are fully determined by the induced outcome and given by

$$U_\theta(\mu) \coloneqq \mathbb{E}_\mu[U_\theta(y) \mid \theta].$$

In the mediation game $\Gamma_\delta(\pi)$, the actual state is unverifiable to both the mediator and the decision-maker, which allows the expert to potentially misrepresent her private information. At the same time, the mediator's recommendation is not binding; the decision-maker remains free to ignore the recommendation and select any other action. Our analysis focuses on a particular class of equilibrium in which the expert finds it optimal to report the state truthfully, and the decision-maker finds it optimal to follow the mediator's recommendation.

Let $\pi_y(\delta)$ denote the posterior belief about state $\theta = 2$ that the decision-maker computes upon receiving the recommendation to choose y in the mediation game $\Gamma_\delta(\pi)$. A mediation plan δ is called *incentive-compatible* if and only if it satisfies

$$\mathbb{E}_\delta[U_\theta(y) \mid \theta] \geq \mathbb{E}_\delta[U_\theta(y) \mid \theta'], \quad \text{for all } \theta, \theta' \in \Theta, \tag{2.2}$$

$$y = \gamma\left(\pi_y(\delta)\right), \quad \text{for all } y \in supp\,(\delta). \tag{2.3}$$

The inequalities in (2.2) ensure that the expert has no incentive to misreport the state and are therefore referred to as *truth-telling incentive constraints*. Likewise, the equality in (2.3) guarantees that

[7] A Nash equilibrium of the game $\Gamma_\delta(\pi)$ is also called *communication equilibrium*.

the decision-maker has no incentive to disobey the mediator's recommendation and is thus known as the *obedience incentive constraint*.

Note that for any mediation plan δ, if the expert reports truthfully and the decision-maker follows the mediator's recommendation, the induced outcome coincides with the plan itself, i.e., $\mu = \delta$.

In general, the mediation game $\Gamma_\delta(\pi)$ may admit multiple Nash equilibria, even when the mediation plan δ is incentive compatible. However, for any such equilibrium, there exists a payoff-equivalent incentive-compatible mediation plan.[8] In this sense, it entails no loss of generality to restrict attention to mediation plans in which the expert truthfully reports the state to the mediator, and the decision-maker follows the mediator's recommendation. This insight is known as the *revelation principle* (see Myerson, 1982; Forges, 1986).

Two specific types of mediation plans will play a central role in our analysis. A mediation plan is said to be *non-revealing* if it does not induce any belief updating by the decision-maker—that is, if it leaves his prior belief unchanged. For such a plan to satisfy the obedience constraints, it must always recommend the action $\gamma(\pi)$ with probability one. This yields interim utilities $U_1(\gamma(\pi))$ and $U_2(\gamma(\pi))$, which depend solely on the prior π through the induced action $\gamma(\pi)$. Consequently, a non-revealing plan always satisfies the truth-telling constraints.

At the opposite end of the communication spectrum is the *fully-revealing* mediation plan, which takes the following form:

$\bar{\delta}(y \mid \theta)$	y_1^d	y_2^d
$\theta = 1$	1	0
$\theta = 2$	0	1

(2.4)

While the plan $\bar{\delta}$ satisfies the obedience constraints, it may fail to satisfy the truth-telling incentive constraints. We say that there is *no ex-ante misrepresentation problem* if the fully-revealing plan $\bar{\delta}$ is incentive compatible, namely,

$$U_1(y_1^d) \geq U_1(y_2^d), \qquad \text{and} \qquad U_2(y_2^d) \geq U_2(y_1^d). \tag{2.5}$$

This terminology reflects the fact that, if the expert were to select the mediation plan at the *ex-ante* stage—prior to observing the realized state—then, absent an ex-ante misrepresentation problem, the truth-telling constraints could be dispensed with (see Salamanca, 2024).[9] However, this implication does not carry over to the interim setting, where the mediation plan is chosen after the expert has observed the true state. As will become clear below, interim solutions typically correspond to mediation plans for which at least one truth-telling constraint binds.

Whenever type θ has an incentive to misrepresent type θ' under the fully-revealing plan, we say that type θ *jeopardizes* type θ', meaning that $U_\theta(y_\theta^d) < U_\theta(y_{\theta'}^d)$.

3. Interim Mediation Design

The preceding setup has described the mediation process under a fixed mediation plan δ. We now turn to the problem in which the expert selects the plan δ. Formally, we consider the following multistage game, which we refer to as the *informed-party problem*:

Stage 1. (choice of mediation plan) After learning the actual state θ, the expert publicly commits to a mediation plan δ.

[8] It suffices to treat the induced equilibrium outcome as a mediation plan.

[9] In that case, the ex-ante mediation-design problem collapses to a standard Bayesian persuasion problem.

Stage 2. (beliefs updating) The decision-maker updates his prior beliefs to some posterior beliefs ρ about the state $\theta = 2$ based on any information inferred from the experts's choice of the mediation plan.

Stage 3. (mediation) Given the updated belief ρ, the mediation game $\Gamma_\delta(\rho)$ is played.

The informed-party problem constitutes a specific case of the *informed-principal problem* as studied by Myerson (1983). Unlike the general framework in Myerson (1983), we do not permit richer *communication devices* that allow for arbitrary report and message spaces. In our setting, we restrict attention to so-called "direct" mediation plans. While the revelation principle allows us to focus on incentive-compatible (direct) mediation plans *on the equilibrium path*, expanding the set of communication devices *off the equilibrium path* may enable a broader set of profitable deviations for the expert, potentially reducing the set of sustainable equilibrium outcomes. Nonetheless, as shown by Forges et al. (2024), this restriction to direct mediation plans is without loss of generality in our present framework.

3.1. Perfect Bayesian Equilibria

To define a perfect Bayesian equilibrium (PBE) of the informed-party problem, we first observe that the expert need not convey information to the decision-maker through her choice of mediation plan. Any information she wishes to transmit can be embedded directly within the plan itself. As a result, any *separating* equilibrium strategy—where the expert selects different mediation plans in different states—can be replaced by an outcome-equivalent *pooling* strategy in which the same mediation plan is prescribed regardless of the realized state. Myerson (1983) refers to this claim as the *inscrutability principle*.

Combining the revelation and inscrutability principles allows for a streamlined characterization of PBE in the informed-party problem. On the equilibrium path, both types of the expert select the same incentive-compatible mediation plan δ^*, inducing the decision-maker to retain his prior belief π in stage 2. If the expert deviates by selecting an alternative mediation plan δ—not necessarily incentive-compatible—the decision-maker updates his belief to some posterior ρ, after which the parties interact in the mediation game $\Gamma_\delta(\rho)$. In a PBE, any such a deviation is unprofitable given the equilibrium played in $\Gamma_\delta(\rho)$.

Definition 1 (**Perfect Bayesian Equilibrium**)**.**
A mediation plan δ^* is a perfect Bayesian equilibrium outcome of the informed-party problem if

- *δ^* is incentive compatible (given the prior beliefs π).*
- *For every mediation plan δ there exists a belief $\rho \in [0, 1]$ and an equilibrium outcome μ of the continuation game $\Gamma_\delta(\rho)$, such that for every $\theta \in \Theta$,*

$$U_\theta(\delta^*) \geq U_\theta(\mu).$$

3.1.1. Equilibria in the Absence of Ex-ante Misrepresentation

Suppose that there is no ex-ante misrepresentation problem. Let δ be an incentive-compatible mediation plan that guarantees each type of the expert an interim utility at least as high as that obtained under the fully-revealing plan—that is, $U_2(\delta) \geq U_2(y_2^d)$ and $U_1(\delta) \geq U_1(y_1^d)$. Then δ can be supported as a PBE outcome of the informed-party problem.

To see this, suppose that if the expert deviates by choosing an alternative mediation plan, the decision-maker updates his belief to $\rho = 0$, thereby inferring that the deviator is type 1, and therefore optimally chooses the action $\gamma(\rho) = y_1^d$. Under this belief, the expert's continuation

payoffs following the deviation are $U_1(y_1^d)$ and $U_2(y_1^d)$, regardless of the deviation. Since there is no ex-ante misrepresentation problem, it follows that $U_2(\delta) \geq U_2(y_2^d) \geq U_2(y_1^d)$, implying that the deviation is unprofitable for both types, and the plan δ is sustained in equilibrium. In this equilibrium characterization, we have the decision-maker attributing any deviation entirely to type 1; however, the same conclusion obtains if the deviation were instead attributed to type 2.

Proposition 1 (**PBE in the absence of ex-ante misrepresentation**).
Suppose that there is no ex-ante misrepresentation problem. Then any incentive-compatible mediation plan δ satisfying $U_\theta(\delta) \geq U_\theta(y_\theta^d)$ for all $\theta \in \Theta$ is a PBE outcome of the informed-party problem. In particular, the fully-revealing plan—being incentive-compatible in the absence of ex-ante misrepresentation—constitutes a PBE outcome.

The preceding characterization is only partial and does not encompass the possibility of additional equilibrium outcomes arising whenever at least one type's utility falls below its fully-revealing payoff. To illustrate, consider parameters satisfying $y_1^d \leq y_1^e < y_2^e \leq y_2^d$. Let δ be an incentive-compatible mediation plan such that $U_1(\delta) < U_1(y_1^d)$ and $U_2(\delta) \geq U_2(y_2^d)$. To sustain δ as a PBE outcome, specify off-path beliefs that attribute any deviation entirely to type 2. The expert's continuation payoffs after a deviation are therefore $U_1(y_2^d)$ and $U_2(y_2^d)$. Under the standing parameter constellation, the absence of ex-ante misrepresentation implies $U_1(y_2^d) \leq U_1(y)$ for every $y \in [y_1^d, y_2^d]$.[10] That is, y_2^d is the least-preferred outcome for type 1 in the issue space. It follows that $U_1(y_2^d) \leq U_1(\delta)$, so no deviation is profitable for either type. An analogous argument implies that any incentive-compatible mediation plan under which only type 2 receives a payoff below its full-disclosure level can likewise be sustained as a PBE.

Notwithstanding the presence of additional equilibria, there are compelling reasons to regard such equilibria as implausible in the informed-party problem. Specifically, the fully-revealing payoffs constitute an *interim individual-rationality* level that the expert can secure, *irrespective of* any inference the decision-maker may draw about the state.

Formally, we call a mediation plan *safe* if it satisfies the truth-telling constraints and the decision-maker, upon learning the actual state, is willing to participate obediently. According to this definition, the only plan that can be safe is the fully-revealing plan, provided the truth-telling constraints hold. Hence, the absence of an ex-ante misrepresentation problem is both necessary and sufficient for the existence of a safe plan.

By definition, the safe plan is implementable by the expert *regardless of* the beliefs the decision-maker forms in response to the its selection. Consequently, by choosing the fully-revealing plan, the expert can exclude any outcome that yields *any* of her types less than the fully-revealing payoff. Moreover, as we will show in the subsequent analysis, when misrepresentation is absent, equilibrium outcomes that give any type less than the fully-revealing payoff do not survive stronger solution concepts. In light of these observations, we do not pursue a full characterization of all PBE outcomes. Readers interested in a comprehensive treatment are referred to Theorem 1 in Koessler and Skreta (2025).[11]

[10] A formal argument is provided in footnote 15.

[11] For any belief the decision-maker forms upon observing the expert's selection of a mediation plan, the expert can adopt a "babbling" strategy in the ensuing mediation game—equivalently, remain silent—which secures the non-revealing payoffs. Hence, the non-revealing payoffs constitute a *belief-contingent* interim individual-rationality level. This is the sense in which Koessler and Skreta (2025) employ the term "interim individual rationality" in their characterization of PBE outcomes. By contrast, our usage is stronger: the safe plan guarantees the expert the fully-revealing payoff *uniformly* across all beliefs the decision-maker may hold.

3.1.2. Equilibria When There is a Jeopardized Type

We now turn to the analysis of the PBE outcomes of the informed-party problem in the presence of ex-ante misrepresentation. To facilitate this analysis, we begin by stating the following result.

Lemma 1.
There is at most one jeopardized type.

Proof. Note that

$$U_\theta(y_\theta^d) - U_\theta(y_{\theta'}^d) = \left(y_{\theta'}^d - y_\theta^d\right)\left[y_{\theta'}^d + y_\theta^d - 2y_\theta^e\right].$$

Suppose both types jeopardize each other. Then

$$y_2^d + y_1^d - 2y_1^e < 0, \qquad \text{and} \qquad y_1^d + y_2^d - 2y_2^e > 0.$$

Adding these two inequalities we obtain $y_2^e - y_1^e < 0$, which contradicts the monotonicity assumption. □

According to Lemma 1, the ex-ante misrepresentation problem can arise from only one of the two types, but not both. Without loss of generality, we will assume that if a type is jeopardized, it is type 1.[12] Since only the jeopardizing type creates an incentive problem under the fully-revealing plan, the prior probability π captures the *ex-ante likelihood of misrepresentation*.

Let $\bar{y}^d := \frac{y_1^d + y_2^d}{2}$ denote the midpoint between y_1^d and y_2^d. Observe that type 2 jeopardizes type 1 if and only if $y_2^e < \bar{y}^d$. In this case, y_2^d is the action most distant from y_2^e in the issue space and therefore represents the worst outcome for type 2. This is a fortiori true for type 1, since $y_2^e > y_1^e$.

An immediate implication is that any mediation plan yields each type an interim utility no lower than that from y_2^d. Hence, a PBE outcome can be sustained for *any* incentive-compatible mediation plan by letting the decision-maker attribute any deviation entirely to type 2. In this way, y_2^d is the decision-maker's unique optimal response to any deviation, uniformly punishing both expert types and thereby deterring deviations.

Proposition 2 (PBE under ex-ante misrepresentation).
Suppose that type 2 jeopardizes type 1. Then every incentive-compatible mediation plan is a PBE outcome of the informed-party problem.

The greater the distance between $\bar{y}^d$ and y_2^e, the larger the benefit for type 2 from misrepresenting type 1. From the decision-maker's ex-ante perspective, this makes it increasingly likely that the expert manipulates her private information. As a result, the mediator faces greater difficulty in building trust between the parties, and any incentive-compatible mediation plan must be non-revealing. This idea is formally captured in Lemma 5 of Mitusch and Strausz (2005). To formalize the argument, we introduce the following critical cut-off beliefs:

$$\widehat{\pi}_\theta := \max_{\rho \in [0,1]} \left\{ \rho \;\middle|\; U_\theta(\gamma(\rho)) = U_\theta(y_1^d) \right\}.$$

We also define the corresponding induced action $\widehat{y}_\theta := \gamma(\widehat{\pi}_\theta)$. This cut-off point is illustrated in Figure 1. Note that $\widehat{\pi}_\theta > 0$ if and only if $y_1^d < y_\theta^e \leq \bar{y}^d$. Moreover, since $\Delta^e > 0$, then $\widehat{\pi}_1 \leq \widehat{\pi}_2$.

Corollary 1 (PBE under ex-ante misrepresentation and too high prior belief).
If type 2 jeopardizes type 1 and $\pi \geq \widehat{\pi}_2$, the unique PBE outcome of the informed-party problem is the non-revealing mediation plan.

[12] If instead type 1 jeopardizes type 2, all results can be recovered symmetrically by redefining actions as $y' = -y$ and swapping the roles of the two types.

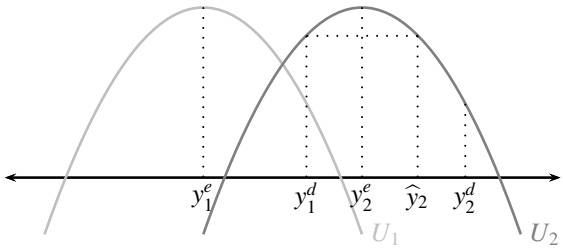


Figure 1: Experts's utility functions when $y_1^e < y_1^d < y_2^e < y_2^d$.

Consequently, the only case that remains relevant for identifying a solution to the informed-party problem under ex-ante misrepresentation is when $\pi < \widehat{\pi}_2$. This condition on the prior, together with the requirement that type 2 jeopardizes type 1, can be satisfied if and only if $y_1^d < y_2^e < \bar{y}^d$, yielding a parameter configuration as illustrated in Figure 1.[13]

4. Core Mechanisms

Regardless of whether an ex-ante misrepresentation problem is present, the set of PBE outcomes remains generally large—except when the prior belief is sufficiently low—and includes outcomes that are inefficient for the expert. To refine this set, we will consider how the expert might eliminate certain equilibria by arguing with the decision-maker on the unreasonableness of his off-equilibrium beliefs.

We say that a mediation plan δ is *dominated by* (or *interim Pareto inferior to*) another mediation plan μ if for all $\theta \in \Theta$

$$U_\theta(\mu) \geq U_\theta(\delta),$$

with at least one strict inequality. In case all inequalities are strict, then we say that δ is *strictly* dominated.

A mediation plan δ is (resp. *weakly*) *undominated* if it is incentive compatible and is not (resp. strictly) dominated by any other incentive compatible mediation plan.

For any given $\lambda \in [0,1]$ and a mediation plan δ, we let $U(\delta;\lambda) := (1-\lambda)U_1(\delta) + \lambda U_2(\delta)$. A mediation plan δ is undominated if and only if there exists $0 < \lambda < 1$ such that δ is a solution of the following optimization problem:

$$\begin{aligned} &\max_{\delta} U(\delta;\lambda) \\ &\text{s.t. } (2.2) - (2.3), \end{aligned} \tag{4.1}$$

We refer to this optimization problem as the *primal problem* for λ.

Whenever $\lambda = \pi$, the primal problem maximizes the expert's *ex-ante* expected utility among the incentive-compatible mediation plans. This formulation corresponds to the setting extensively studied by Salamanca (2024). Letting λ to vary in $[0,1]$, the optimal solutions to (4.1) cover the entire set of *weakly* undominated mediation plans.

Weakly undominated plans are important because the expert should be expected to propose only such plans. To see why, suppose the decision-maker anticipates that the expert will select a mediation plan δ that yields strictly lower utility for every type than some alternative incentive-compatible plan δ^*. In that case, following Myerson (1983), the expert could reason with the decision-maker as follows:

[13] The only other possible parameter constellation satisfying these assumptions is $y_1^d \leq y_1^e < y_2^e < y_2^d$ with $y_2^e < \bar{y}^d$.

> "I am going to choose δ^*. This mediation plan is *strictly* better for me in *both* states. Therefore, you should not draw any inference about the true state from my selection of δ^*. With no new information, you should maintain your prior beliefs and, consequently, participate obediently in the mediation plan δ^*."

The decision-maker has no rational basis to reject this argument. By offering such a justification, the expert can credibly exclude all inefficient plans from the set of possible outcomes.

A strengthening of this type of "objection" by the expert is the concept of a *core mechanism* introduced by Myerson (1983). The underlying idea is that the expert can influence the decision-maker's beliefs by credibly revealing that her type belongs to a subset T of types, with the restriction that she cannot claim to be a type that does not benefit from the deviation.

A mediation plan is said to be incentive-compatible *given a non-empty set of types* $T \subseteq \Theta$ if it satisfies the truth-telling incentive constraints and ensures that the decision-maker has no incentive to deviate from the mediator's recommendations, *given that he knows the expert's type lies in* T. In particular, when $T = \Theta$, this definition reduces to the constraints in (2.2)–(2.3). A mediation plan is safe if and only if it is incentive-compatible given $\{\theta\}$ for all $\theta \in \Theta$.

Now consider two mediation plans δ and μ, and let $T \subseteq \Theta$ denote the set of types strictly better off under μ than under δ. We say that μ *is an objection against* δ if μ is incentive-compatible given T.[14]

Definition 2 (**Core mechanisms**)**.**
A mediation plan δ is a core mechanism if the expert has no objection against δ.

Equivalently, if δ is not a core mechanism, then there exists some alternative plan μ that certain types strictly prefer, such that μ would be incentive-compatible given any information revealed by its selection, provided that all types preferring μ over δ are expected to select μ.

According to this definition, if both types strictly benefit from deviating to some plan μ, then μ cannot be incentive-compatible; hence, a core mechanism must be weakly undominated. On the other hand, if only type θ strictly benefits from deviating to μ, then either μ violates the truth-telling constraints or fails to satisfy the obedience constraints when the decision-maker knows that the actual state is θ.

Proposition 3 (**Core mechanisms**)**.**
Suppose that type 2 jeopardizes type 1. Then δ is a core mechanism if and only if it is weakly undominated. In case there is no ex-ante misrepresentation problem, δ is a core mechanism if and only if it is weakly undominated and $U_\theta(\delta) \geq U_\theta(y_\theta^d)$ for all $\theta \in \Theta$.

Proof. Assume type 2 jeopardizes type 1. Let μ be a mediation plan that is incentive compatible given $\{\theta\}$. Then $\mu(y_\theta^d \mid \theta) = 1$. Define $q := \mu(y_\theta^d \mid \theta')$, with $\theta' \neq \theta$. Note that the decision-maker's posterior beliefs are

$$\rho_1 = \frac{\pi\,(\theta - 1 + q(2-\theta))}{\pi\,(\theta - 1 + q(2-\theta)) + (1-\pi)\,(2 - \theta + q(\theta - 1))}$$

conditional on the recommendation y_θ^d, and $\rho_2 = 2 - \theta$ conditional on any recommendation $y \neq y_\theta^d$. To satisfy the obedience incentive constraints, it must hold that $y(\rho_1) = y_\theta^d$, or equivalently, $\rho_1 = \theta - 1$. Consequently, $q = 0$, which means that μ must be fully revealing. Thus, again the obedience

[14] In defining the core mechanisms, Myerson (1983) requires the objection to be incentive-compatible given *every* superset of T. In our present two-type framework, this requirement is equivalent to our definition.

constraints imply $y(\rho_2) = y^d_{\theta'}$, implying that $\mu(y^d_{\theta'} \mid \theta') = 1$. Yet, since type 2 jeopardizes type 1, μ cannot fulfill the truth-telling incentive constraints. We conclude that, for any $\theta \in \Theta$, a mediation plan μ cannot be incentive compatible given $\{\theta\}$. Hence, for δ to be a core mechanisms, it suffices that there exists no incentive-compatible mediation plan μ such that $U_\theta(\mu) > U_\theta(\delta)$ for all $\theta \in \Theta$. In other words, it is enough for δ to be weakly undominated.

In case there is no ex-ante misrepresentation problem, the only mediation plan that is incentive compatible given any $\theta \in \Theta$ is the fully-revealing plan. Thus, δ is a core mechanism iff δ is weakly undominated and $U_\theta(\delta) \geq U_\theta(y^d_\theta)$ for all $\theta \in \Theta$. □

5. Strong Solutions

As we have seen, the concept of a core mechanism substantially refines the set of PBE outcomes by ruling out equilibria that are interim Pareto inferior for the expert. However, the set of weakly undominated plans often remains large, and some of the resulting equilibrium outcomes may still rely on implausible beliefs. To identify a unique "best" inscrutable mediation plan for the expert, it is therefore necessary to appeal to stronger solution concepts.

Definition 3 (**Strong solution**).
A mediation plan is called a strong solution if it is both safe and undominated.

Myerson (1983) argues that the strong solution is the most compelling outcome of the informed-party problem. Regardless of what the decision-maker infers about the actual state, if the expert selects the strong solution, he will still find it optimal to participate obediently because the plan is safe. Furthermore, if the decision-maker infers that the expert's type belongs to the set of types that strictly prefer some alternative plan to the strong solution, that alternative must violate the obedience constraints given those posterior beliefs. Consequently, any incentive-compatible mediation plan other than the strong solution becomes non-implementable as soon as the decision-maker concludes that the "deviator" is one of the expert's types that would be strictly better off. In addition, the strong solution is always a PBE of the informed-party problem.

Despite its appeal as a solution concept, a strong solution often does not exist—either because no safe plan exists or because the safe plan is dominated. In particular, the absence of an ex-ante misrepresentation problem is a necessary (but not sufficient) condition for the existence of a strong solution, as it ensures that a safe plan exists.

The following result directly follows from Lemma 1 in Salamanca (2024).

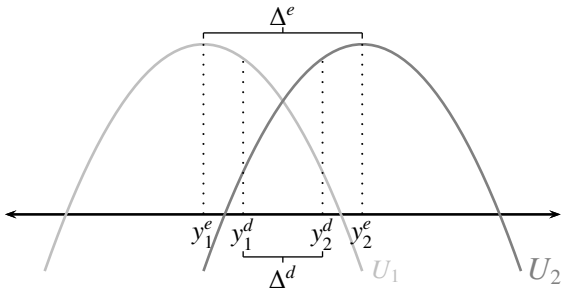
Lemma 2.
Assume there is no ex-ante misrepresentation problem. The fully-revealing mediation plan is undominated if $\Delta^d \leq 2\Delta^e$.

The inequality $\Delta^d \leq 2\Delta^e$ indicates that the decision-maker's preferences across states cannot diverge too greatly from the expert's preferences across states. It therefore captures the *relative* degree of conflict of interest between the parties. If Δ^d is large relative to Δ^e, the decision-maker will choose an action that differs substantially from those preferred by the expert. In such a case, the expert gains no *ex-ante* benefit from revealing her private information.

Lemma 2 provides a sufficient condition for the existence of a strong solution. To explore whether this condition is also necessary, it is helpful to split the analysis into several cases according to the various possible parameter configurations.

Case 1: $y_1^e \leq y_1^d < y_2^d \leq y_2^e$.

In this case, there is no ex-ante misrepresentation problem. Moreover, $\Delta^d \leq \Delta^e < 2\Delta^e$. Therefore, this parameter constellation always satisfies the sufficient conditions in Lemma 2.



Case 1: $y_1^e \leq y_1^d < y_2^d \leq y_a^2$

Before considering the following parameter configuration, we first define the cut-off beliefs

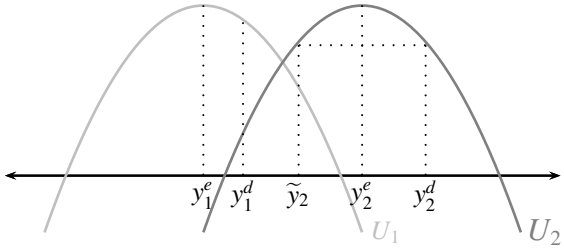
$$\widetilde{\pi}_\theta \coloneqq \min_{\rho \in [0,1]} \left\{ \rho \mid U_\theta(\gamma(\rho)) = U_\theta(y_2^d) \right\}.$$

We likewise define the corresponding optimal action as $\widetilde{y}_\theta \coloneqq \gamma(\widetilde{\pi}_\theta)$.

Case 2: $y_1^e \leq y_1^d \leq y_2^e \leq y_2^d$.

In this case, the absence of ex-ante misrepresentation requires that $y_1^d \leq \widetilde{y}_2$. Hence, the following chain of inequalities hold:

$$\begin{aligned}
\Delta^d &= (y_2^d - y_2^e) + (y_2^e - y_1^d), \\
&= (y_2^e - \widetilde{y}_2) + (y_2^e - y_1^d), \\
&\leq (y_2^e - y_1^d) + (y_2^e - y_1^d), \quad \text{(since } \widetilde{y}_2 \geq y_1^d\text{)} \\
&\leq 2\Delta^e. \quad \text{(since } y_1^d \geq y_1^e\text{)}
\end{aligned}$$



Case 2: $y_1^e \leq y_1^d \leq y_2^e \leq y_2^d$

The analysis of the configuration $y_1^d \leq y_1^e \leq y_2^d \leq y_a^2$ is symmetric, leading to the same inequality $\Delta^d \leq 2\Delta^e$. This can be seen by redefining the actions as $y' = -y$ and interchanging the roles of the two types.

We conclude that the absence of a misrepresentation problem is both a necessary and sufficient condition for the existence of the strong solution in Cases 1 and 2.

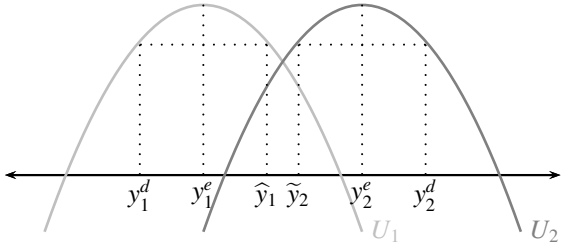
Proposition 4 (Strong solution—part 1).
Suppose that $y_\theta^d \in [y_1^e, y_2^e]$ for some $\theta \in \Theta$. The strong solution exists if and only if there is no ex-ante misrepresentation problem.

The previous result still leaves open the question of cases in which *both* of the decision-maker's bliss points lie outside the interval $[y_1^e, y_2^e]$. Any parameter configuration satisfying either $y_2^d \leq y_1^e$ or $y_2^e \leq y_1^d$ results in one type being jeopardized. Consequently, no strong solution can exist in these cases. The only remaining case is therefore the following:

Case 3: $y_1^d \leq y_1^e < y_2^e \leq y_2^d$.

To rule out ex-ante misrepresentation, it must be that $y_1^e \leq \bar{y}^d \leq y_2^e$. We can then write Δ^e as:

$$\begin{aligned}
\Delta^e &= (y_2^e - \widetilde{y}_2) + (\widetilde{y}_2 - \widehat{y}_1) + (\widehat{y}_1 - y_1^e), \\
&= (y_2^d - y_2^e) + (\widetilde{y}_2 - \widehat{y}_1) + (y_1^e - y_1^d), \\
&= \Delta^d - \Delta^e + (\widetilde{y}_2 - \widehat{y}_1),
\end{aligned}$$



Case 3: $y_1^d \leq y_1^e < y_2^e \leq y_2^d$

From this equality, it follows that $\Delta^d \leq 2\Delta^e$ if and only if $\widetilde{y}_2 \geq \widehat{y}_1$—or equivalently, $\widetilde{\pi}_2 \geq \widehat{\pi}_1$.[15] Hence, in what follows, we examine Case 3 under the restrictions $\widetilde{\pi}_2 < \widehat{\pi}_1$ and $y_1^e \leq \bar{y}^d \leq y_2^e$, as depicted in Figure 2. We will refer to this parameter configuration as *Case 3a*.

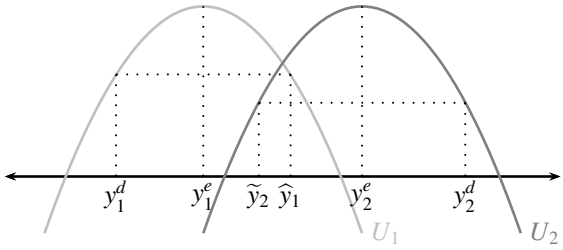


Figure 2: $y_1^d \leq y_1^e < y_2^e \leq y_2^d$

Suppose that the prior belief, π, satisfies $\widetilde{\pi}_2 \leq \pi \leq \widehat{\pi}_1$. Then the non-revealing plan dominates the fully-revealing plan. To see this, note that since $\gamma(\pi) \leq \widehat{y}_1$, it follows that $U_1(\gamma(\pi)) \geq U_1(\widehat{y}_1) = U_1(y_1^d)$. Similarly, because $\gamma(\pi) \geq \widetilde{y}_2$, we have $U_2(\gamma(\pi)) \geq U_2(\widetilde{y}_2) = U_2(y_2^d)$. Given that $\widetilde{y}_2 < \widehat{y}_1$, at least one of these inequalities must be strict. Hence, the strong solution cannot exist when $\widetilde{\pi}_2 \leq \pi \leq \widehat{\pi}_1$.

Now consider a prior belief $\pi < \widetilde{\pi}_2$. Define the partially-revealing mediation plan $\widehat{\delta}_1$ as follows:

$\widehat{\delta}_1(y \mid s)$	y_1^d	$\widehat{y}_1$
$s = 1$	$\widehat{p}_1$	$1 - \widehat{p}_1$
$s = 2$	0	1

, $\qquad \widehat{p}_1 := \dfrac{\widehat{\pi}_1 - \pi}{\widehat{\pi}_1(1 - \pi)}$.

Since $\pi < \widetilde{\pi}_2 < \widehat{\pi}_1$, the mediation plan $\widehat{\delta}_1$ is well defined. It is easily verified that this mediation plan is incentive-compatible. Moreover, it yields:

$$U_1(\widehat{\delta}_1) = U_1(y_1^d), \qquad U_2(\widehat{\delta}_1) = U_2(\widehat{y}_1) > U_2(y_2^d),$$

where the strict inequality holds because $\widetilde{y}_2 < \widehat{y}_1$, so that $U_2(\widehat{y}_1) > U_2(\widetilde{y}_2) = U_2(y_2^d)$. Thus, the plan $\widehat{\delta}_1$ dominates the fully-revealing mediation plan in this case as well.

Finally, consider a prior belief $\pi > \widehat{\pi}_1$. As in the previous case, it is straightforward to verify that the following mediation plan is incentive-compatible and dominates the fully-revealing mediation plan:

$\widetilde{\delta}_2(y \mid s)$	$\widetilde{y}_2$	y_2^d
$s = 1$	1	0
$s = 2$	$1 - \widetilde{q}_2$	$\widetilde{q}_2$

, $\qquad \widetilde{q}_2 := \dfrac{\pi - \widetilde{\pi}_2}{\pi(1 - \widetilde{\pi}_2)}$.

In summary, in Case 3 the condition $\Delta^d \leq 2\Delta^e$ is not only sufficient but also necessary for the existence of the strong solution.

Proposition 5 (Strong solution—part 2).
Suppose that $y_\theta^d \notin [y_1^e, y_2^e]$ for each $\theta \in \Theta$. The strong solution exists if and only if there is no ex-ante misrepresentation problem and $\Delta^d \leq 2\Delta^e$.

[15] Under the Case 3 parameter constellation, we have that $y_1^e \leq \bar{y}^d \leq y_2^e$ if and only if $\widehat{y}_1 \leq y_2^d$ and $y_1^d \leq \widetilde{y}_2$. Thus, y_2^d (resp. y_1^d) is the least-preferred outcome in the issue space for type 1 (resp. type 2).

6. Neologism-proofness and the Neo-optimum

To identify a unique "best" inscrutable mediation plan, we sequentially applied stronger solution concepts that, in our setting, refined the set of PBE outcomes.[16] However, a unique outcome could not be identified in all parameter configurations since the set of core mechanisms often remained large and a strong solution failed to exist. In this section, we adopt a different approach by applying a refinement criterion that requires any observed deviation from an equilibrium plan to be accompanied by a "credible" belief.

The expert's speech in Section 4, used to justify why she should never be expected to implement a strictly dominated plan, already contains the essence of *neologism-proofness*, a refinement introduced by Farrell (1993).[17] In our context, a neologism is a pair (T, δ) representing the statement: "I am going to implement the mediation plan δ because my type is in the set $T \subseteq \Theta$." Such a statement is deemed *credible* by the decision-maker if only the types in T could benefit from inducing the belief that the actual type lies in T. A set $T \subseteq \Theta$ with this property is called *self-signaling*. According to Farrell (1993), a claim that the expert's type belongs to a self-signaling set should be believed.

Formally, if believed, a neologism (T, δ) causes the decision-maker to adopt beliefs $\pi|_T$ that are Bayesian consistent, namely, put zero probability on types not in T and retain likelihoods across all types in T. The neologism (T, δ) is *credible relative to an incentive-compatible mediation plan* δ^* if there exists an equilibrium outcome μ of the mediation game $\Gamma_\delta(\pi|_T)$ such that

- $U_\theta(\delta^*) < U_\theta(\mu)$, for all $\theta \in T$.
- $U_\theta(\delta^*) \geq U_\theta(\mu)$, for all $\theta \notin T$.

Definition 4 (**Neologism-proof**).
A PBE outcome is called neologism-proof if no neologism is credible relative to it.

An immediate observation is that any neologism-proof PBE outcome δ^* must be weakly undominated. Otherwise, there would exist an incentive-compatible mediation plan δ that strictly dominates δ^*. In that case, the neologism (Θ, δ) would be credible, leading to a contradiction. Conversely, no credible neologism (Θ, δ) can exist relative to a weakly undominated plan. If such a neologism did exist, there would be an equilibrium outcome μ of $\Gamma_\delta(\pi)$ that strictly improves upon δ^* for both types. By the revelation principle, this would imply the existence of an incentive-compatible mediation plan that strictly dominates δ^*, again yielding a contradiction.

In light of this observation, the search for neologism-proof equilibria can be restricted to PBE outcomes that are weakly undominated and for which no credible neologism of the form $(\{\theta\}, \delta)$, with $\theta \in \{1, 2\}$, exists.

6.1. Neologism-Proof Equilibria in the Absence of Ex-ante Misrepresentation

As we have already shown, in the absence of ex-ante misrepresentation the set of core mechanisms consists of all weakly undominated plans that guarantee each type of the expert an interim utility at least as high as under full revelation. Within this setting, we were able to identify a unique solution (i.e., the strong solution) in all cases except Case 3a, characterized by the parameter restrictions $y_1^d \leq y_1^e < y_2^e \leq y_2^d$, $y_1^e \leq \bar{y}^d \leq y_2^e$, and $\Delta^d > 2\Delta^e$ (i.e., $\widetilde{\pi}_2 < \widehat{\pi}_1$), as illustrated in Figure 2. In this

[16] Unlike the strong solution, core mechanisms may fail to be PBE outcomes.

[17] Similar ideas were also employed by Grossman and Perry (1986) in motivating their concept of perfect sequential equilibrium.

case, however, the set of core mechanisms typically contains a continuum of mediation plans. For example, consider the numerical parameter specification

$$y_1^d = \tfrac{1}{2}, \quad y_1^e = 1, \quad y_2^e = 2, \quad y_2^d = 3.$$

Here, $\bar{y}^d = \frac{7}{4} \in [1, 2]$ and $\widetilde{\pi}_2 = \frac{1}{5} < \frac{2}{5} = \widehat{\pi}_1$.

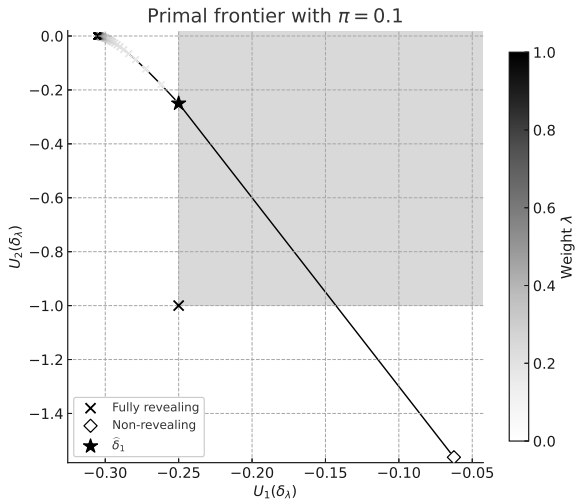


Figure 3: Interim primal-Pareto frontier for $y_1^d = \frac{1}{2}$, $y_1^e = 1$, $y_2^e = 2$, $y_2^d = 3$, and $\pi = \frac{1}{10}$.

Figure 3 shows the expert's interim Pareto frontier obtained by solving the primal problem for each $\lambda \in [0, 1]$ when $\pi = \frac{1}{10} < \widetilde{\pi}_2$. The shaded region depicts all interim utility vectors that give each type at least its fully-revealing payoff. As it can be deduced, the set of core mechanisms corresponds to all convex combinations of the non-revealing plan and the plan $\widehat{\delta}_1$ that yield type 2 at least $U_2(y_2^d)$.

A symmetric situation arises when $\pi = \frac{1}{2} > \widehat{\pi}_1$, as illustrated in Figure 4. In this case, the set of core mechanisms corresponds to all convex combinations of the non-revealing plan and the plan $\widetilde{\delta}_2$ that guarantee type 1 at least $U_1(y_1^d)$.

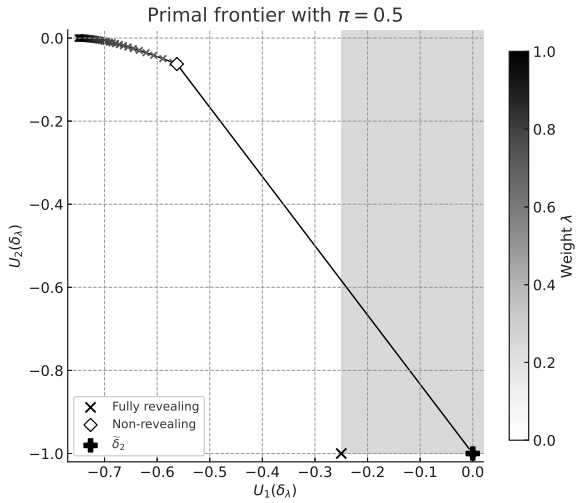


Figure 4: Interim primal-Pareto frontier for $y_1^d = \frac{1}{2}$, $y_1^e = 1$, $y_2^e = 2$, $y_2^d = 3$, and $\pi = \frac{1}{2}$.

When the prior is $\pi = \frac{3}{10}$, so that $\widetilde{\pi}_2 < \pi < \widehat{\pi}_1$, the Pareto frontier exhibits a "kink" at the non-revealing plan, as shown in Figure 5. In this case, the set of core mechanisms consists of all

convex combinations of the non-revealing plan with $\widehat{\delta}_1$, together with all convex combinations of the non-revealing plan with $\widetilde{\delta}_2$.

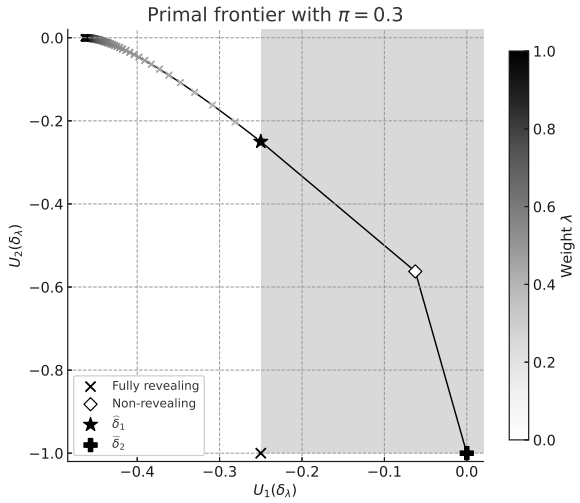


Figure 5: Interim primal-Pareto frontier for $y_1^d = \frac{1}{2}$, $y_1^e = 1$, $y_2^e = 2$, $y_2^d = 3$, and $\pi = \frac{3}{10}$.

The qualitative features of the sets of weakly undominated plans shown in Figures 3–5 are not specific to this example but arise generically for any parameter constellation satisfying the assumptions of Case 3a.

Unfortunately, in the absence of ex-ante misrepresentation, the neologism-proof property does not provide any further refinement beyond the set of core mechanisms. Any neologism of the form $(\{\theta\}, \delta)$ with $\theta \in \{1, 2\}$ is not credible. If the decision-maker were to believe such a claim, his optimal response would be to choose y_θ^d, regardless of the plan δ, yielding utility $U_\theta(y_\theta^d)$ to type θ. However, for any core mechanism δ^* we have $U_\theta(y_\theta^d) \leq U_\theta(\delta^*)$. Thus, type θ cannot benefit from inducing such a belief. We conclude that the set of core mechanisms coincides with the set of neologism-proof equilibria.

We summarize the previous finding in the following proposition:

Proposition 6 (Neologism-proof equilibria in the absence of ex-ante misrepresentation). *Suppose that there is no ex-ante misrepresentation problem. A PBE outcome of the informed-party problem is neologism-proof if and only if it is a core mechanism.*

In view of this result, all neologism-proof equilibria can be supported by reasonable off-path beliefs. To see this, let δ^* be any neologism-proof equilibrium. Since δ^* is undominated, for any alternative mediation plan δ and for all equilibria of the continuation game $\Gamma_\delta(\pi)$, there can be at most one type of the expert that strictly benefits relative to δ^*; that is, $U_\theta(\delta^*) > U_\theta(\delta)$ and $U_{\theta'}(\delta^*) \leq U_{\theta'}(\delta)$. Thus, it is reasonable that after observing a deviation to δ, the decision-maker infers that such a deviation could only have been made by the unique type θ that benefits from it. The decision-maker would then optimally select y_θ^d in the ensuing mediation game, yielding continuation payoffs $U_\theta(y_\theta^d)$ and $U_{\theta'}(y_\theta^d)$. Yet, $U_\theta(y_\theta^d) \leq U_\theta(\delta^*)$ and $U_{\theta'}(y_\theta^d) \leq U_{\theta'}(y_{\theta'}^d) \leq U_{\theta'}(\delta^*)$, since δ^* is a core mechanism and there is no ex-ante misrepresentation problem. Hence, any such deviation is rendered unprofitable once the decision-maker infers that the deviator is the unique type that could have gained from it.

If instead the decision-maker observes a deviation to a mediation plan that strictly reduces the payoff of *both* types, then he should draw no inference. With no new information, the rational

response is to maintain his prior beliefs, which ensures that the deviating plan yields strictly worse payoffs for both types in the continuation game.

6.2. Neologism-proof Equilibria when Type 2 Jeopardizes Type 1

The only other case in which a unique solution could not be identified arises when type 2 jeopardizes type 1 and $\pi < \widehat{\pi}_2$. In this situation, the concept of core mechanisms refines the set of PBE outcomes by retaining only the weakly undominated mediation plans. However, as we will show, unless additional restrictions are imposed on the model's parameters, either no neologism-proof equilibrium exists to further narrow the set of weakly undominated plans, or all weakly undominated plans turn out to be neologism-proof.

First, we observe that type 2 cannot be self-signaling. Inducing the belief that the actual type is $\theta = 2$ would lead the decision-maker to choose y_2^d, which under the current assumptions is the worst possible outcome in the issue space. Hence, type 2 cannot benefit from such a statement. In view of this, the search for neologism-proof equilibria can be restricted to weakly undominated plans for which type 1 is not self-signaling.

To identify all neologism-proof equilibrium outcomes, it is useful to first characterize the best incentive-compatible plan for each expert's type. These correspond to the solutions of the primal problem for $\lambda \in \{0, 1\}$. To this end, we define the critical threshold

$$\bar{\pi} := \frac{1}{\Delta^d} \max\left\{0, \Delta^d - 2\Delta^e\right\}.$$

Whenever $y_1^d < y_1^e$, it follows that $\Delta^d - 2\Delta^e \in (0, \Delta^d)$. Otherwise, if $y_1^d \geq y_1^e$ then $\Delta^d - 2\Delta^e \in (-\infty, \Delta^d)$. Then $0 \leq \bar{\pi} < 1$.

Next, let $\widehat{\delta}_2$ denote the mediation plan

$\widehat{\delta}_2(y \mid s)$	y_1^d	$\widehat{y}_2$
$s = 1$	$\widehat{p}_2$	$1 - \widehat{p}_2$
$s = 2$	0	1

, $\qquad \widehat{p}_2 := \dfrac{\widehat{\pi}_2 - \pi}{\widehat{\pi}_2(1 - \pi)}$.

By construction, this mediation plan makes the truth-telling constraint of type 2 binding. On the other hand, $\widehat{\delta}_2$ satisfies the truth-telling constraint for type 1 if and only if $U_1(y_1^d) \geq U_1(\widehat{y}_2)$. Because $\widehat{\pi}_1 < \widehat{\pi}_2$, it follows that $U_1(y_1^d) = U_1(\widehat{y}_1) > U_1(\widehat{y}_2)$. We therefore conclude that $\widehat{\delta}_2$ is incentive compatible.

Proposition 7 (**Type 1's optimal mediation plan**)**.**
Suppose type 2 jeopardizes type 1. Then:

i) If $\pi \leq \bar{\pi}$ or $\pi \geq \widehat{\pi}_2$, the unique optimal solution to the primal problem for $\lambda = 0$ is the non-revealing mediation plan.

ii) If $\bar{\pi} < \pi < \widehat{\pi}_2$, the unique optimal solution to the primal problem for $\lambda = 0$ is the mediation plan $\widehat{\delta}_2$.

In order to characterize the best mediation plan for type 2, we need to define the following critical threshold:

$$\pi_\theta^* := \frac{y_\theta^e - y_1^d}{\Delta_p}.$$

Note that $\widehat{\pi}_\theta = 2\pi_\theta^*$ and $\gamma(\pi_\theta^*) = y_\theta^e$.

Let δ_2^* denote the mediation plan

$\delta_2^*(y \mid s)$	y_1^d	y_2^e
$s = 1$	p_2^*	$1 - p_2^*$
$s = 2$	0	1

, $\qquad p_2^* := \dfrac{\pi_2^* - \pi}{\pi_2^*(1-\pi)}$.

Proposition 8 (**Type 2's optimal mediation plan**).
Suppose type 2 jeopardizes type 1. Then:

i) If $\pi_2^* \leq \pi$, the unique optimal solution to the primal problem for $\lambda = 1$ is the non-revealing mediation plan.

ii) If $\pi_2^* > \pi$ and $\pi_2^* \geq \widehat{\pi}_1$, the unique optimal solution to the primal problem for $\lambda = 1$ is the mediation plan δ_2^*.

iii) If $\widehat{\pi}_1 > \pi_2^* > \pi$ and $\widehat{\pi}_2 - \widehat{\pi}_1 \leq \pi$, the unique optimal solution to the primal problem for $\lambda = 1$ is the non-revealing mediation plan.

iv) If $\widehat{\pi}_1 > \pi_2^* > \pi$ and $\widehat{\pi}_2 - \widehat{\pi}_1 > \pi$, the unique optimal solution to the primal problem for $\lambda = 1$ is the mediation plan $\widehat{\delta}_1$.

The proofs of Propositions 7–8 are largely technical and are therefore deferred to the appendix.

As we will see, combining these two results provides a "lower" bound on the Pareto frontier in the space of the expert's interim payoffs, allowing us to identify which efficient payoff vectors are immune to neologisms. We now proceed by analyzing several cases that together exhaustively cover all parameter configurations in Propositions 7–8 that are mutually compatible. The detailed results of this analysis are established through a sequence of claims and synthesized in Proposition 9.

Case A: $\pi < \min\left\{\bar{\pi}, \pi_2^*\right\}$.

In this case, the non-revealing plan is optimal for type 1. Moreover, $U_1(\gamma(\pi)) \geq U_1(y_1^d)$ if and only if $\pi \leq \widehat{\pi}_1$. Also, $U_2(\gamma) > U_2(y_1^d)$, since $\pi < \pi_2^* < \widehat{\pi}_2$. Thus, the non-revealing plan is neologism-proof if and only if $\pi \leq \widehat{\pi}_1$.

To proceed, we divide the analysis into subcases, each identifying the optimal plan for type 2.

Subcase A1: $\min\left\{\bar{\pi}, \pi_2^*\right\} = \bar{\pi}$. If $\widehat{\pi}_1 > 0$, then $\bar{\pi} = \widehat{\pi}_1 + (1 - \widehat{\pi}_2)$. Since type 2 jeopardizes type 1, we have $\widehat{\pi}_2 < 1$, which implies $\bar{\pi} > \widehat{\pi}_1$. Otherwise, $\bar{\pi} > \pi > 0 = \widehat{\pi}_1$. Hence, $\pi_2^* > \widehat{\pi}_1$. We conclude that δ_2^* is optimal for type 2. The expert's interim payoffs from δ_2^* are

$$U_2^* := U_2(\delta_2^*) = U_2(y_2^e) > U_2(y_1^d), \quad \text{and} \quad U_1^* := U_1(\delta_2^*) = p_2^* U_1(y_1^d) + (1 - p_2^*) U_1(y_2^e).$$

Note that

$$U_1^* - U_1(y_1^d) = (1 - p_2^*)\left[U_1(y_2^e) - U_1(y_1^d)\right].$$

Since $\pi_2^* > \widehat{\pi}_1$ and δ_2^* is incentive compatible, it follows that $U_1(y_2^e) < U_1(y_1^d)$. We therefore conclude that $U_1^* < U_1(y_1^d)$, and thus type 1 is self-signaling relative to δ_2^*. Hence δ_2^* is not neologism-proof.

Figure 6 summarizes the results of the previous analysis.

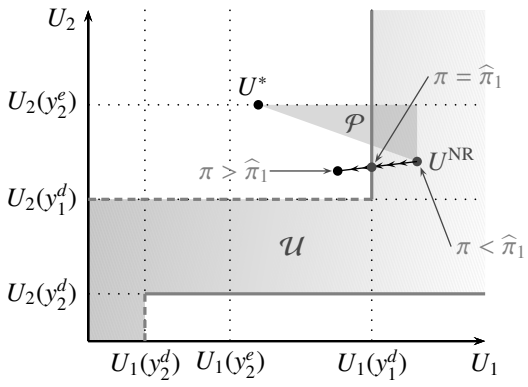


Figure 6: Illustration of Claim 1

The shaded area $\mathcal{U}$ represents all interim payoff vectors (feasible or not) for which no credible neologism of the form (θ, δ) exists. Equivalently, $U \in \mathcal{U}$ if no type is self-signaling relative to U. As the prior π increases from 0 to $\bar{\pi}$, the value of $U_1(\gamma(\pi))$ decreases until it reaches $U_1(y_1^d)$ and then continues to fall further, while $U_2(\gamma(\pi))$ remains bounded below by $U_2(y_1^d)$. Hence, the vector $U^{\mathrm{NR}} = (U_1(\gamma(\pi)), U_2(\gamma(\pi)))$ moves from inside $\mathcal{U}$ to outside it as π increases from 0 to $\bar{\pi}$. The exact point at which U^{NR} reaches the boundary of $\mathcal{U}$ occurs when $\pi = \widehat{\pi}_1$.

Moreover, any weakly undominated plan generates payoffs satisfying $U \geq \theta U^* + (1-\theta)U^{\mathrm{NR}}$ for any $\theta \in [0,1]$, as well as $U \leq U^*$ and $U \leq U^{\mathrm{NR}}$. Consequently, the interim Pareto frontier lies within the convex hull $\mathcal{P}$ of the points U^*, U^{NR}, and $(U_1^{\mathrm{NR}}, U_2^*)$. It follows that the interim Pareto frontier intersects $\mathcal{U}$ if and only if $\pi \leq \widehat{\pi}_1$.[18]

Subcase A2: $\min\left\{\bar{\pi}, \pi_2^*\right\} = \pi_2^* > \widehat{\pi}_1$. The same analysis applies as in Subcase A1, leading to the same conclusion.

Claim 1. *Suppose type 2 jeopardizes type 1 and assume that* $\min\left\{\bar{\pi}, \pi_2^*\right\} > \widehat{\pi}_1$.

a. *If $\pi \leq \widehat{\pi}_1$, then an equilibrium outcome δ is neologism-proof if and only if it is weakly undominated and satisfies $U_1(\delta) \geq U_1(y_1^d)$. In particular, the non-revealing plan is neologism-proof.*

b. *If $\widehat{\pi}_1 < \pi < \min\left\{\bar{\pi}, \pi_2^*\right\}$, then no equilirium outcome is neologism-proof.*

At this point, it is useful to compare our application of the neologism-proofness criterion to a related solution concept recently proposed by Mylovanov and Tröger (2025), namely the *neo-optimum*. As Claim 1 illustrates, neologism-proofness can be overly demanding: a credible neologism should destroy its putative equilibrium, and in some cases every equilibrium is overturned by some credible neologism.[19] To address this limitation, they relax the neologism-proofness requirement and define *neo-optima* as the set of incentive-compatible mediation plans whose payoff vectors for the expert are at least as good as the limits of neologism-proof payoff vectors.

[18] We conjecture, though do not prove, that whenever type 2 jeopardizes type 1 and the best incentive-compatible plans of the two types differ, the interim Pareto frontier coincides with the line segment connecting the payoff vectors generated by the respective optimal plans of each type. In that case, the set of weakly undominated plans consists of all convex combinations of these two optimal plans.

[19] The non-existence of neologism-proof equilibria is not specific of our model. This same phenomenon is illustrated by Mylovanov and Tröger (2025) in their application to the job-market signaling model of Spence (1973). See also Farrell (1993) for another example of non-existence in a basic signaling game.

Definition 5 (Neo-optimum).
A mediation plan δ is a neo-optimum if δ is incentive compatible (given π) and there exists an interim payoff vector $V \leq U(\delta)$ (not necessarily feasible) such that V is a limit of neologism-proof payoff vectors.

It follows immediately that every neologism-proof PBE outcome is a neo-optimum. Moreover, every neo-optimum is also a core mechanism (see Proposition 4 in Mylovanov and Tröger, 2025). In particular, neo-optima are weakly undominated.[20]

To compute the neo-optima under the assumptions of Claim 1b, we first need to determine the full set of neologism-proof payoff vectors. This is obtained by intersecting the set $\mathcal{U}$ with all payoff vectors for which no credible neologism of the form (Θ, δ) exists. This latter set consists of all payoff vectors V such that $V \geq U$ for any weakly undominated payoff vector U, together with all payoff vectors V satisfying $V_1 \geq U_1^{\text{NR}}$, and all payoff vectors V satisfying $V_2 \geq U_2^*$. The resulting set of neologism-proof payoff vectors, denoted $\mathcal{U}^{\text{neo}}$, is illustrated in Figure 7.

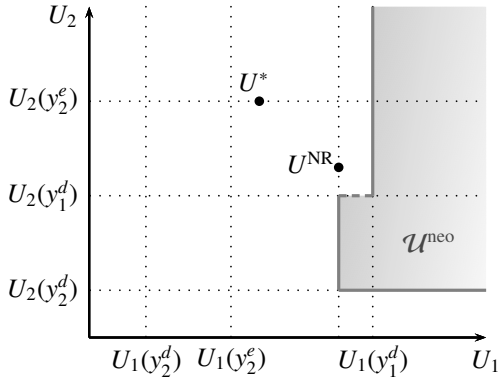


Figure 7: Illustration of Claim 2

Since the non-revealing plan is the unique optimal solution for type 1, no weakly undominated plan can have an associated payoff vector U with $U_2 \neq U_2^{\text{NR}}$. As illustrated in Figure 7, U^{NR} is therefore the only weakly undominated payoff vector lying above a point in the topological closure of $\mathcal{U}^{\text{neo}}$. It follows that the non-revealing plan is the unique neo-optimum.

Claim 2. *Suppose type 2 jeopardizes type 1. Assume that $\widehat{\pi}_1 < \pi < \min\left\{\bar{\pi}, \pi_2^*\right\}$. Then the only neo-optimum is the non-revealing mediation plan.*

Subcase A3: $\min\left\{\bar{\pi}, \pi_2^*\right\} = \pi_2^* = \widehat{\pi}_1$. In this case, the plan δ_2^* remains optimal for type 2. However, $U_1(\delta_2^*) = U_1(y_1^d)$, so that $U^* \in \mathcal{U}$. Moreover, since $\widehat{\pi}_1 > \pi$, then $U^{\text{NR}} \in \text{int}(\mathcal{U})$. We therefore conclude that all weakly undominated mediation plans are neologism-proof, as illustrated in Figure 8.

[20] When there is no ex-ante misrepresentation, the set of neo-optima coincides with the set of core mechanisms

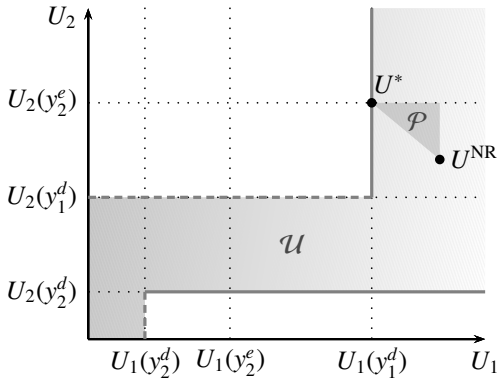


Figure 8: Illustration of Claim 3

Claim 3. *Suppose type 2 jeopardizes type 1. Assume that $\pi < \widehat{\pi}_1 = \pi_2^* \leq \bar{\pi}$. A PBE outcome is neologism-proof if and only if it is weakly undominated.*

Subcase A4: $\min\{\bar{\pi}, \pi_2^*\} = \pi_2^* < \widehat{\pi}_1$. If $\widehat{\pi}_2 - \widehat{\pi}_1 \leq \pi$, the non-revealing plan is also optimal for type 2. It follows that the non-revealing plan is the unique undominated plan and therefore the unique core mechanism.[21] Moreover, since $\pi < \widehat{\pi}_1$, the non-revealing plan is also the only neologism-proof equilibrium outcome. In this case, the payoff-maximization goals of both types coincide, so the expert faces no trade-off between the objectives of her two types.

If $\widehat{\pi}_2 - \widehat{\pi}_1 > \pi$, then $\widehat{\delta}_1$ becomes optimal for type 2. Moreover, $U_1(\widehat{\delta}_1) = U_1(y_1^d)$, implying that $\widehat{\delta}_1$ is neologism-proof. Consequently, we obtain a situation analogous to the one described in Subcase A3, as illustrated in Figure 8.

Claim 4. *Suppose type 2 jeopardizes type 1 and assume that $\pi < \pi_2^* < \widehat{\pi}_1 < \bar{\pi}$.*

a. If $\widehat{\pi}_2 - \widehat{\pi}_1 \leq \pi$, then the non-revealing plan is the only neologism-proof PBE outcome.

b. If $\widehat{\pi}_2 - \widehat{\pi}_1 > \pi$, then a PBE outcome is neologism-proof if and only if it is weakly undominated.

Case B: $\pi_2^* \leq \pi \leq \bar{\pi}$.

In this case, the non-revealing plan is optimal for both types of the expert. It is therefore the unique undominated plan and, a fortiori, the unique core mechanism. However, unlike in Claim 4a, the non-revealing plan is neologism-proof if and only if $\pi \leq \widehat{\pi}_1$. Regardless of whether this condition holds, the non-revealing plan remains the unique compelling solution to the informed-party problem.

Claim 5. *Suppose type 2 jeopardizes type 1. Assume that $\pi_2^* \leq \pi \leq \bar{\pi}$. The non-reveling plan is the unique core mechanism.*

Case C: $\bar{\pi} \leq \pi \leq \pi_2^*$.

Consider first the situation where $\bar{\pi} < \pi$. Because $\pi_2^* < \widehat{\pi}_2$, the mediation plan $\widehat{\delta}_2$ is optimal for type 1. The corresponding expert's interim payoffs are

$$\widehat{U}_2 := U_2(\widehat{\delta}_2) = U_2(\widehat{y}_2) = U_2(y_1^d), \quad \text{and} \quad \widehat{U}_1 := U_1(\widehat{\delta}_2) = \widehat{p}_2 U_1(y_1^d) + (1 - \widehat{p}_2) U_1(\widehat{y}_2).$$

[21] Unlike the case $\pi \geq \widehat{\pi}_2$, where the non-revealing plan is the only incentive-compatible mediation plan, here infinitely many partially revealing plans remain incentive compatible—for instance, the plan $\widehat{\delta}_2$, as well as any convex combination of $\widehat{\delta}_2$ and the non-revealing plan.

Observe that

$$\widehat{U}_1 - U_1(y_1^d) = (1 - \widehat{p}_2)\left[U_1(\widehat{y}_2) - U_1(y_1^d)\right].$$

Since $\widehat{\pi}_2 > \widehat{\pi}_1$, it follows that $U_1(\widehat{y}_2) < U_1(y_1^d)$. Hence, $\widehat{U}_1 < U_1(y_1^d)$, which implies that type 1 is self-signaling relative to $\widehat{\delta}_2$. Therefore, $\widehat{\delta}_2$ is not neologism-proof.

On the other hand, $\widehat{\pi}_1 \leq \bar{\pi}$, which implies $\widehat{\pi}_1 < \pi_2^*$. Consequently, δ_2^* is optimal for type 2. As shown earlier in Subcase A1, however, δ_2^* is not neologism-proof. The situation is thus analogous to that described in Claim 1b, where no PBE outcome is neologism-proof.

Whenever $\bar{\pi} = \pi$, the situation mirrors the one just described, except that the optimal plan for type 1 is now the non-revealing plan.

Extending the same reasoning as in Claim 2 leads to the following result:

Claim 6. *Suppose type 2 jeopardizes type 1 and $\bar{\pi} \leq \pi \leq \pi_2^*$. Then the unique neo-optimum is the best incentive-compatible mediation plan for type 1. Specifically:*

a. *If $\bar{\pi} < \pi$, the unique neo-optimum is the mediation plan $\widehat{\delta}_2$.*

b. *If $\bar{\pi} = \pi$, the unique neo-optimum is the non-revealing plan.*

Case D: $\max\left\{\bar{\pi}, \pi_2^*\right\} < \pi$.

Clearly, the mediation plan $\widehat{\delta}_2$ is optimal for type 1, though we have already shown that it is not neologism-proof. On the other hand, the non-revealing plan is optimal for type 2. Since $\widehat{\pi}_1 \leq \bar{\pi}$, it follows that $\pi > \widehat{\pi}_1$, and hence the non-revealing plan is also not neologism-proof. As in Case C, no neologism-proof equilibrium outcome exists, and the unique neo-optimum is $\widehat{\delta}_2$.

Claim 7. *Suppose type 2 jeopardizes type 1. Assume that* $\max\left\{\bar{\pi}, \pi_2^*\right\} < \pi$. *The unique neo-optimum is the mediation plan $\widehat{\delta}_2$.*

This completes the analysis of all possible parameter configurations arising from the conditions in Propositions 7 and 8.

Proposition 9.
The following statements summarize the general findings established in Claims 1–7:

a. *The set of neologism-proof equilibrium outcomes consists of all weakly undominated plans δ that guarantee type 1 an interim utility at least as high as under full revelation, i.e., $U_1(\delta) \geq U_1(y_1^d)$.*

b. *If no such plan exists, there is a unique neo-optimum, which coincides with the best incentive-compatible mediation plan for type 1.*

c. *Whenever the payoff-maximization goals of both types align, the expert's "best" incentive-compatible plan is the non-revealing plan.*

The foregoing analysis highlights the pivotal role of the optimal plan for type 1 as a solution to the informed-party problem. First, the existence of a neologism-proof equilibrium depends on whether type 1's optimal plan is itself neologism-proof; in fact, whenever such an equilibrium exists, type 1's optimal plan must be neologism-proof. Second, whenever a neologism-proof equilibrium does not exist, type 1's optimal plan emerges as the unique neo-optimum.

There are also compelling reasons to expect that the expert would implement type 1's optimal mediation plan. This mediation plan is the best incentive-compatible plan for type 1, so that only

type 2 would have an incentive to deviate and propose an alternative plan—consistent with the off-path beliefs underlying the PBE characterization in Proposition 2. Moreover, it would seem counterintuitive for the decision-maker to infer that the expert is type 2 when type 1's optimal plan is unexpectedly announced—yet such an inference was required to sustain any other mediation plan as a PBE under Proposition 2.

Overall, the optimal plan for type 1 acquires a special preeminence that sets it apart from other potential solutions to the informed-party problem. It is our conjecture that this plan is the unique *neutral optimum* in the sense of Myerson (1983). Unfortunately, this solution concept has only been formally defined in environments with finitely many types and actions, and it remains unclear how its analytical characterization might be extended to models such as ours with a continuum of actions.

7. Ex-ante vs. Interim

The *ex-ante* solutions studied in Salamanca (2024) correspond to solving the primal problem with $\lambda = \pi$. These solutions are appealing when the expert is placed under a veil of ignorance at the moment of choosing the mediation plan—i.e., *before* learning the true state—so that her choice does not make any explicit use of private information. By contrast, in the *interim* setting, the expert is already informed when selecting the plan and could, in principle, exploit this information to her advantage. The dilemma is that her choice must remain inscrutable, preventing the decision-maker from inferring her type. This does not imply that she should entirely ignore her private information, but rather that the chosen plan must look as though either type could have selected it. Put differently, the plan must appear to be a "fair compromise" between the objectives of both types, rather than simply reflect the preference of the actual type.

Such compromises can be interpreted through the utility weights λ for which a given weakly undominated plan solves the primal problem. The parameter λ captures the relative importance attached to type 2's payoff-maximization goals. In some cases, the fair compromise is reflected in $\lambda = \pi$, meaning the best inscrutable strategy is to ignore private information altogether. In other cases, however, inscrutability requires systematically mimicking one of the types. For instance, when type 2 jeopardizes type 1, and a unique neo-optimum exists, the prescribed interim solution solves the primal problem for $\lambda = 0$, thus resolving the inscrutability dilemma in favor of type 1.

Ex-ante solutions can sometimes fail to satisfy even minimal requirements for being considered valid solutions to the interim mediation design problem—specifically, when they do not correspond to PBE outcomes. Koessler and Skreta (2025) provide such an example, highlighting the difficulty of justifying why a rational player would ever select the ex-ante solution under those circumstances.

That said, in most environments—including ours—ex-ante solutions are PBE outcome and therefore cannot be ruled out *a priori* as candidate solutions to the informed-party problem. However, as our analysis shows, not all PBE outcomes can be regarded as reasonable. Consequently, in some situations, ex-ante solutions may fail to withstand reasonableness tests. Table 7.1 compares the ex-ante and interim solutions when there is no ex-ante misrepresentation.[22]

[22]The reader is referred to Salamanca (2024) for specific results regarding the ex-ante solutions.

	$\Delta^d \leq 2\Delta^e$	$\Delta^d > 2\Delta^e$
Ex-ante Interim	FR	NR Core

Table 7.1: Ex-ante vs. interim solutions in the absence of ex-ante misrepresentation

In the absence of ex-ante misrepresentation, when $\Delta^d \leq 2\Delta^e$, a strong solution exists, rendering the fully-revealing (FR) plan the most compelling outcome in both the ex-ante and interim mediation design problems. In contrast, when $\Delta^d > 2\Delta^e$, the non-revealing (NR) plan qualifies as a core mechanism only if $\widetilde{\pi}_2 \leq \pi \leq \widehat{\pi}_1$. For prior beliefs outside this interval, the non-revealing plan fails to guarantee at least one type the utility it would receive under full disclosure (see Figures 3–4). Hence, by Proposition 3, the ex-ante solution cannot be regarded as a reasonable solution to the informed-party problem.

A similar tension between the ex-ante and interim solutions also arises when type 2 jeopardizes type 1. Consider the case where $0 < \bar{\pi} < \pi \leq \pi_2^*$. Since $\bar{\pi} > 0$, it follows that $\Delta^d > 2\Delta^e$. Lemma 1 in Salamanca (2024) then implies that the ex-ante solution corresponds to the non-revealing plan. Thus, the ex-ante solution prescribes withholding private information *both* at the stage of selecting the mediation plan and during its implementation. By contrast, our Claim 6a shows that the only reasonable interim solution is the plan $\widehat{\delta}_2$. Under this plan, the expert systematically pretends to be type 1 when selecting the mediation plan. Furthermore, to prevent the decision-maker from perfectly inferring state 2 during the mediation process, $\widehat{\delta}_2$ provides cover to type 2 by randomizing over recommendations when the true state is 1. Hence, unlike the ex-ante solution, the interim solution allows for partial disclosure of private information in the course of mediation.

8. Appendix

This appendix provides detailed proofs of Propositions 7 and 8. For this purpose, we adapt the general methodology developed by Salamanca (2021) to the present model.

Let $\alpha(\theta' \mid \theta) \geq 0$ denote the dual variable (or Lagrange multiplier) for the truth-telling incentive constraint (2.2) asserting that type θ should not gain by reporting θ' in the primal problem (4.1). Multiplying the truth-telling incentive constraints by their corresponding dual variables and adding them into the objective function, we obtain the following Lagrangian relaxation of (4.1):

$$\mathcal{L}(\delta; \pi, \lambda, \alpha) := U(\delta; \lambda) + \sum_{\theta=1}^{2} \alpha(3-\theta \mid \theta) \Big[\mathbb{E}_\delta[U_\theta(y) \mid \theta] - \mathbb{E}_\delta[U_\theta(y) \mid 3-\theta] \Big],$$

where δ is a mediation plan satisfying only the obedience constraints (2.3).

For any given action y, we define

$$\begin{aligned} W_1(y; \pi, \lambda, \alpha) &:= \frac{1}{1-\pi} \Big[\left((1-\lambda) + \alpha(2 \mid 1)\right) U_1(y) - \alpha(1 \mid 2) U_2(y) \Big], \\ W_2(y; \pi, \lambda, \alpha) &:= \frac{1}{\pi} \Big[\left(\lambda + \alpha(1 \mid 2)\right) U_2(y) - \alpha(2 \mid 1) U_1(y) \Big]. \end{aligned}$$

Myerson (1991) refers to $W_\theta(y; \pi, \lambda, \alpha)$ as the expert's *virtual utility* from action y, when the state is θ, w.r.t. the prior π, the utility weights λ, and the dual variables α. With this definition, the above Lagrangian can be written as

$$\mathcal{L}(\delta; \pi, \lambda, \alpha) = (1-\pi)\mathbb{E}_\delta \left[W_1(y; \pi, \lambda, \alpha) \mid 1 \right] + \pi \mathbb{E}_\delta \left[W_2(y; \pi, \lambda, \alpha) \mid 2 \right]. \tag{8.1}$$

That is, the Lagrangian in (8.1) is simply the expert's (ex-ante) expected *virtual* utility from a mediation plan δ.

Now we consider the problem of maximizing (8.1) over all mediation plans satisfying only the obedience incentive constraints:

$$\begin{aligned} &\max_{\delta} \mathcal{L}(\delta; \pi, \lambda, \alpha) \\ &\text{s.t. (2.3)}. \end{aligned} \tag{8.2}$$

This optimization problem corresponds to a Bayesian persuasion problem, except that the expert's preferences are measured in the virtual utility scales. Given π, λ, and α, the *indirect* virtual utility function is defined as follows:

$$\widehat{W}(\rho; \pi, \lambda, \alpha) := (1-\rho) W_1(\gamma(\rho); \pi, \lambda, \alpha) + \rho W_2(\gamma(\rho); \pi, \lambda, \alpha), \quad \rho \in [0, 1]. \tag{8.3}$$

Hence, the optimal value of (8.1) is $cav\, \widehat{W}(\pi\,; \pi, \lambda, \alpha)$, where $cav\, \widehat{W}(\cdot\,; \pi, \lambda, \alpha)$ denotes the concavification of $\widehat{W}(\cdot\,; \pi, \lambda, \alpha)$. The following result relates the value of (8.1) to the value of (4.1). In doing so, it provides sufficient conditions under which a candidate mediation plan is optimal in (4.1).

Proposition 10 (**Weak Duality**).
Let π and λ be fixed. Let δ^* be an incentive-compatible mediation plan such that

$$U(\delta^*; \lambda) = cav\, \widehat{W}(\pi; \pi, \lambda, \alpha^*),$$

for some $\alpha^* \geq 0$. Then, δ^* is an optimal solution of (4.1) for λ.

Proof. Let δ be an incentive-compatible mediation plan. Then the following chain of inequalities hold:

$$U(\delta;\lambda) \le \mathcal{L}(\delta;\pi,\lambda,\alpha^*) \le cav\,\widehat{W}(\pi;\pi,\lambda,\alpha^*) = U(\delta^*;\lambda),$$

where the first inequality holds because δ is incentive-compatible and $\alpha^* \ge 0$; the second inequality comes from the fact that $cav\,\widehat{W}(\pi;\pi,\lambda,\alpha^*)$ is the optimal value of (8.2); and finally, the last equality holds by hypothesis. □

With this result in hand, we can now proceed to prove the statements in Propositions 7 and 8. Specifically, we show that for each candidate plan in the propositions there exists a vector α^* satisfying weak duality. Uniqueness follows directly from the strict concavity of the utility functions and the convexity of the set of incentive-compatible mediation plans.

Because $\pi < \widehat{\pi}_2$ and type 2 jeopardizes type 1, it must hold that $y_1^d < y_2^e < \widehat{y}_2 < y_2^d$ and $y_2^e < \bar{y}^d$. These inequalities will be often used in the following analysis.

8.1. Proof of Proposition 7

We set the utility weight to $\lambda = 0$ and take the dual variables as $\alpha(1 \mid 2) = a \ge 0$ and $\alpha(2 \mid 1) = 0$. Define $\Delta_s := y_s^e - y_1^d$. Then $\widehat{\pi}_2 = \frac{2\Delta_2}{\Delta^d} > 0$.

The indirect virtual utility simplifies to

$$\widehat{W}(\rho) := \widehat{W}(\rho;\pi,\lambda,\alpha) = -\frac{1-\rho}{1-\pi}\big(\rho\Delta_p - \Delta_1\big)^2 - \frac{a(\rho-\pi)}{\pi(1-\pi)}\big(\rho\Delta_p - \Delta_2\big)^2.$$

The derivatives of $\widehat{W}(\cdot)$ are:

$$\widehat{W}'(\rho) = \frac{A_2\rho^2 + A_1\rho + A_0}{1-\pi}, \qquad \widehat{W}''(\rho) = \frac{2A_2\rho + A_1}{1-\pi}.$$

$$A_2 := 3(\Delta^d)^2\left(1 - \tfrac{a}{\pi}\right),$$
$$A_1 := 2(\Delta^d)^2\left[(a-1) + \frac{2}{\Delta^d}\left(\tfrac{a\Delta_2}{\pi} - \Delta_1\right)\right],$$
$$A_0 := -2a\Delta^d\Delta_2 + 2\Delta^d\Delta_1 - \tfrac{a(\Delta_2)^2}{\pi} + (\Delta_1)^2,$$

Case 1: $0 < \pi \le \bar{\pi}$. Then $\bar{\pi} = \frac{\Delta^d - 2\Delta^e}{\Delta^d} \in (0,1)$. Let $a = \pi$. Then $A_2 = 0$ and $A_1 \le 0$. Hence, the indirect virtual utility is concave and, therefore, $cav\,\widehat{W}(\rho) = \widehat{W}(\rho)$ for all $\rho \in [0,1]$. In particular,

$$\widehat{W}(\pi) = U_1(\gamma(\pi)).$$

Thus, by weak duality, the non-revealing plan is an optimal solution of the primal problem for any $\pi \le \bar{\pi}$.

Case 2: $\bar{\pi} < \pi < \widehat{\pi}_2$. Define the dual variable a by

$$a = \frac{\pi}{\Delta^d}\frac{\Delta^d\widehat{\pi}_2 - (\Delta^d - 2\Delta^e)}{\widehat{\pi}_2 - \pi}.$$

Under the current restrictions on prior beliefs, a is well defined and $a > \pi$.

The second derivative of $\widehat{W}(\cdot)$ is affine in ρ, thus it gives the unique inflection point

$$\rho_{\text{inf}} = \frac{2}{3}\widehat{\pi}_2.$$

Moreover,

- $A_2 < 0$ because $\Delta^d - 2\Delta^e < \Delta^d\pi$ and $\widehat{\pi}_2 > \pi$.
- $A_1 > 0$ because $\Delta^d - 2\Delta^e < \Delta^d\pi$ and $\widehat{\pi}_2 > 0$.

Therefore, the curvature of $\widehat{W}(\rho)$ switches from convex to concave at ρ_{inf}.

On the other hand, the value of a has been chosen so that the slope of the secant line starting at $\big(0, \widehat{W}(0)\big)$ and going through $\big(\widehat{\pi}_2, \widehat{W}(\widehat{\pi}_2)\big)$ equals $\widehat{W}'(\widehat{\pi}_2)$. Thus, $\widehat{W}(\cdot)$ and its concavification $cav\,\widehat{W}(\cdot)$ look like in Figure 9.

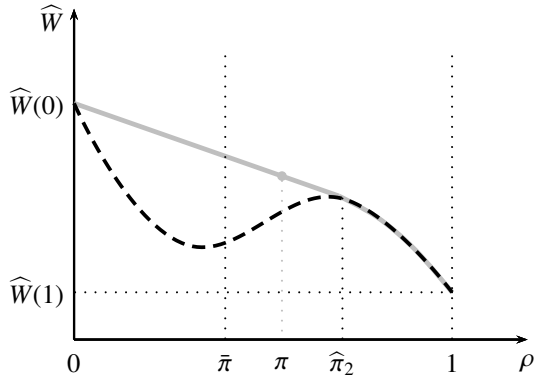


Figure 9: Functions $\widehat{W}(\cdot)$ (black dashed line) and $cav\,\widehat{W}(\cdot)$ (gray solid line)

As a result, for any $\pi \in (\bar{\pi}, \widehat{\pi}_2)$, we have that

$$\begin{aligned}
cav\,\widehat{W}(\pi) &= \left(1 - \frac{\pi}{\widehat{\pi}_2}\right)\widehat{W}(0) + \frac{\pi}{\widehat{\pi}_2}\widehat{W}(\widehat{\pi}_2),\\
&= -\frac{2\Delta_2 - \pi\Delta^d}{2\Delta_2(1-\pi)}(\Delta_1)^2 - \frac{\pi(\Delta^d - 2\Delta_2)}{2\Delta_2(1-\pi)}(2\Delta_2 - \Delta_1)^2,\\
&= \frac{2\Delta_2 - \pi\Delta^d}{2\Delta_2(1-\pi)}U_1(y_1^d) + \frac{\pi(\Delta^d - 2\Delta_2)}{2\Delta_2(1-\pi)}U_1(\widehat{y}_2),\\
&= \frac{\widehat{\pi}_2 - \pi}{\widehat{\pi}_2(1-\pi)}U_1(y_1^d) + \left(1 - \frac{\widehat{\pi}_2 - \pi}{\widehat{\pi}_2(1-\pi)}\right)U_1(\widehat{y}_2),\\
&= U_1(\widehat{\delta}_2).
\end{aligned}$$

Hence, by weak duality, $\widehat{\delta}_2$ is an optimal solution of the primal problem for any $\pi \in (\bar{\pi}, \widehat{\pi}_2)$.

8.2. Proof of Proposition 8

Let $\lambda = 1$. We set $\alpha = (\alpha(1 \mid 2), \alpha(2 \mid 1)) = (0, a)$, with $a \geq 0$ to be determined. The indirect virtual utility function is

$$\widehat{W}(\rho) := \widehat{W}(\rho; \pi, \lambda, \alpha) = \frac{a(\pi - \rho)}{\pi(1-\pi)}U_1(\gamma(\rho)) + \frac{\rho}{\pi}U_2(\gamma(\rho)).$$

Differentiating this function yields:

$$\begin{aligned}
\widehat{W}'(\rho) =& \frac{\Delta^d}{\pi(1-\pi)}\left[a(\pi - \rho)U_1'(\gamma(\rho)) + \rho(1-\pi)U_2'(\gamma(\rho))\right]\\
&+ \frac{1}{\pi(1-\pi)}\left[(1-\pi)U_2(\gamma(\rho)) - aU_1(\gamma(\rho))\right],
\end{aligned}$$

and

$$\widehat{W}''(\rho) = \frac{2(\Delta^d)^2}{\pi(1-\pi)}\left[(1-\pi)\widehat{\pi}_2 + 3\rho(a - 1 + \pi) - a(\widehat{\pi}_1 + \pi)\right].$$

Case 1: $\pi \geq \pi_2^*$. Set $a = 0$.

$$\widehat{W}'(\rho) = \frac{U_2(\gamma(\rho))}{\pi} + \frac{\rho}{\pi} U_2'(\gamma(\rho))\Delta^d,$$

$$\widehat{W}''(\rho) = \frac{2\Delta^d}{\pi}\left[U_2'(\gamma(\rho)) - \Delta^d \rho\right]$$

Note that $U_2'(\gamma(\rho)) \leq 0$ if and only if $\rho \geq \pi_2^*$. Hence, $\widehat{W}(\cdot)$ is locally concave for all $\rho \geq \pi_2^*$. Moreover, $U_2(y(\pi_2^*)) = U_2'(\gamma(\pi^*)) = 0 = \widehat{W}(0)$. Thus, $\rho = 0$ and $\rho = \pi_1^*$ are global maxima. Moreover, there exists a unique inflection at $\rho_{\text{inf}} = \frac{\widehat{\pi}_2}{3} < \pi_2^*$. Hence, it is concluded that $\widehat{W}(\cdot)$ is initially convex and then it turns concave at ρ_{inf}, as illustrated in Figure 10.

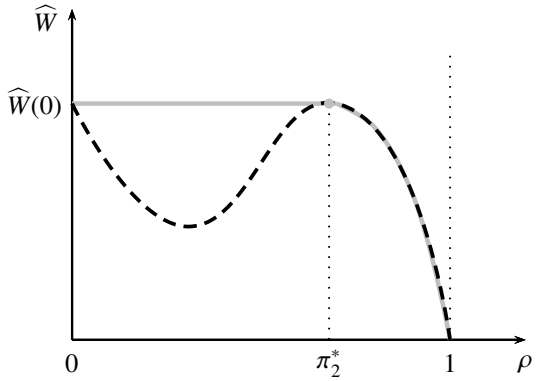


Figure 10: Functions $\widehat{W}(\cdot)$ (black dashed line) and $cav\,\widehat{W}(\cdot)$ (gray solid line)

Therefore, for any $\pi \geq \pi_2^*$, we have that

$$cav\,\widehat{W}(\pi) = \widehat{W}(\pi) = U_2(y(\pi)).$$

Thus, by weak duality, we conclude that the non-revealing plan is a solution to the primal problem for $\lambda = 1$ when $\pi \geq \pi_2^*$.

Case 2: $\pi_2^* > \pi$ *and* $\pi_2^* \geq \widehat{\pi}_1$. Then the mediation plan δ_2^* is well-defined. Moreover, since $U_2(y_2^e) > U_2(y_1^d)$ then δ_2^* satisfies the truth-telling incentive constraints for type 2. On the other hand, δ_2^* satisfies the truth-telling incentive constraints for type 1 if and only if

$$U_1(y_2^e) \leq U_1(y_1^d) \qquad \Leftrightarrow \qquad (y_2^e - y_1^d)(y_2^e + y_1^d - 2y_1^e) \geq 0.$$

The assumption $\pi_2^* \geq \widehat{\pi}_1$ implies that $y_2^e + y_1^d \geq 2y_1^e$. Since $\pi_2^* > \pi > 0$, we have that $y_2^e - y_1^d > 0$. Hence, δ_2^* is incentive compatible.

Because $U_2(\delta_2^*) = U_2(y_a^2)$ and y_a^2 is the expert's bliss point in state $\theta = 2$, then δ_2^* is an optimal solution to the primal problem for $\lambda = 1$.

Case 3: $\widehat{\pi}_1 > \pi_2^* > \pi$ *and* $\widehat{\pi}_2 - \widehat{\pi}_1 \leq \pi$. Let $a = 1 - \pi$. Then the second derivative of the indirect virtual utility function simplifies to

$$\widehat{W}''(\rho) = \frac{2(\Delta^d)^2}{\pi}\left[\widehat{\pi}_2 - \widehat{\pi}_1 - \pi\right]$$

By assumption, $\widehat{\pi}_2 - \widehat{\pi}_1 - \pi \leq 0$. Thus, the indirect utility function is concave. Consequently, $cav\,\widehat{W}(\pi) = \widehat{W}(\pi) = U_2(y(\pi))$. Thus, by weak duality, we conclude that the non-revealing plan is a solution to the primal problem for $\lambda = 1$ in this case.

Case 4: $\widehat{\pi}_1 > \pi_2^* > \pi$ *and* $\widehat{\pi}_2 - \widehat{\pi}_1 > \pi$. Set the dual variable a to

$$a = \frac{U_2'(\widehat{y}_1)}{\widehat{p}_1 U_1'(\widehat{y}_1)}.$$

Because $\widehat{\pi}_2 - \widehat{\pi}_1 > \pi$, we have that

$$\widehat{W}''(\rho) \leq 0 \quad \Leftrightarrow \quad \rho \geq \rho_{\text{inf}} := \frac{2}{3}\widehat{\pi}_1.$$

Therefore, the curvature of $\widehat{W}(\rho)$ switches from convex to concave at $\rho_{\text{inf}} < \widehat{\pi}_1$.

On the other hand, the value of a has been chosen so that the slope of the secant line starting at $\left(0, \widehat{W}(0)\right)$ and going through $\left(\widehat{\pi}_1, \widehat{W}(\widehat{\pi}_1)\right)$ equals $\widehat{W}'(\widehat{\pi}_1)$. Thus, $\widehat{W}(\cdot)$ and its concavification $cav\,\widehat{W}(\cdot)$ look like in Figure 11.

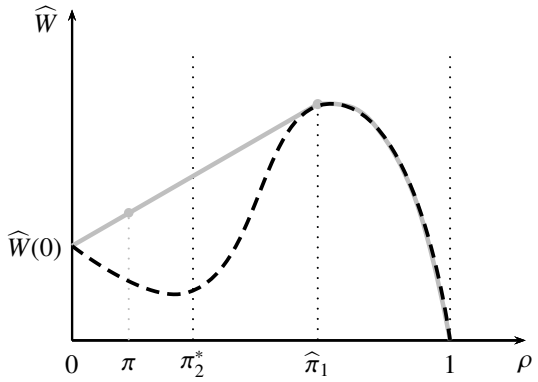


Figure 11: Functions $\widehat{W}(\cdot)$ (black dashed line) and $cav\,\widehat{W}(\cdot)$ (gray solid line)

As a result, we have that

$$\begin{aligned}
cav\,\widehat{W}(\pi) &= \left(1 - \frac{\pi}{\widehat{\pi}_1}\right)\widehat{W}(0) + \frac{\pi}{\widehat{\pi}_1}\widehat{W}(\widehat{\pi}_1), \\
&= \left(\frac{\widehat{\pi}_1 - \pi}{\widehat{\pi}_1}\right)\left[-\frac{(\Delta^d)^2}{4}\frac{\widehat{\pi}_1^2(2\widehat{\pi}_1 - \widehat{\pi}_2)}{\widehat{\pi}_1 - \pi}\right] + \frac{\pi}{\widehat{\pi}_1}\left[\frac{(\Delta^d)^2}{4\pi}\widehat{\pi}_1(\widehat{\pi}_2 - \widehat{\pi}_1)(2\widehat{\pi}_1 - \widehat{\pi}_2)\right], \\
&= -\frac{(\Delta^d)^2}{4}\left[\widehat{\pi}_1(2\widehat{\pi}_1 - \widehat{\pi}_2)\right] + \frac{(\Delta^d)^2}{4}(\widehat{\pi}_2 - \widehat{\pi}_1)(2\widehat{\pi}_1 - \widehat{\pi}_2), \\
&= -\frac{(\Delta^d)^2}{4}(2\widehat{\pi}_1 - \widehat{\pi}_2)^2, \\
&= U_2(\widehat{\delta}_1).
\end{aligned}$$

Hence, by weak duality, $\widehat{\delta}_1$ is an optimal solution of the primal problem for $\lambda = 1$.

References

Balkenborg, Dieter and Miltiadis Makris, "An undominated mechanism for a class of informed principal problems with common values," *Journal of Economic Theory*, 2015, *157*, 918–958.

Blume, Andreas, Oliver J. Board, and Kohei Kawamura, "Noisy talk," *Theoretical Economics*, 2007, *2*, 395–440.

Farrell, Joseph, "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior*, 1993, *5* (4), 514–531.

Forges, Françoise, "An approach to communication equilibria," *Econometrica*, 1986, *54* (6), 1375–85.

_ , Frédéric Koessler, and Andrés Salamanca, "Interacting mechanisms: A perspective on generalized principal–agent problems," *Journal of Mathematical Economics*, 2024, *114*, 103023.

Ganguly, Chirantan and Indrajit Ray, "Simple mediation in a cheap-talk game," *Games*, 2023, *14* (3), 47.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani, "Mediation, arbitration and negotiation," *Journal of Economic Theory*, 2009, *144* (4), 1397–1420.

Grossman, Sanford J and Motty Perry, "Perfect sequential equilibrium," *Journal of Economic Theory*, 1986, *39* (1), 97–119.

Ivanov, Maxim, "Communication via a Strategic Mediator," *Journal of Economic Theory*, 2010, (145), 869–84.

_ , "Beneficial mediated communication in cheap talk," *Journal of Mathematical Economics*, 2014, *55* (C), 129–135.

Koessler, Frédéric and Vasiliki Skreta, "Informed Information Design," *Journal of Political Economy*, 2023, *131* (11), 3186–3232.

_ **and** _ , "Informed Communication Equilibrium," Working paper 2025.

Maskin, Eric and Jean Tirole, "The Principal-Agent Relationship with an Informed Principal, II: Common Values," *Econometrica*, 1992, *60* (1), 1–42.

Mitusch, Kay and Roland Strausz, "Mediation in situations of conflict and limited commitment," *Journal of Law, Economics and Organization*, 2005, *21* (2), 467–500.

Myerson, Roger B., "Optimal coordination mechanisms in generalized principal-agent problems," *Journal of Mathemtical Economics*, 1982, *10* (1), 67–81.

_ , "Mechanism design by an informed principal," *Econometrica*, 1983, *51* (6), 1767–1797.

_ , *Game Theory: Analysis of Conflict*, Harvard University Press, 1991.

Mylovanov, Tymofiy and Thomas Tröger, "Informed-principal problems in environments with generalized private values," *Theoretical Economics*, 2012, *7* (3), 465 – 488.

_ **and** _ , "Mechanism Design by an Informed Principal: Private Values with Transferable Utility," *The Review of Economic Studies*, 2014, *81* (4), 1668–1707.

_ **and** _ , "Neo-optimum: A Unifying Solution to the Informed-Principal Problem," Discussion paper 643, Collaborative Research Center Transregio 224 2025.

Salamanca, Andrés, "The value of mediated communication," *Journal of Economics Theory*, 2021, *192*, 105191.

_ , "Biased Mediators in Conflict Resolution," *American Law and Economics Review*, 2024.

Severinov, Sergei, "An efficient solution to the informed principal problem," *Journal of Economic Theory*, 2008, *141* (1), 114–133.

Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, *87* (3), 355–374.