Online Appendix:
Toward an Understanding of Optimal Mediation Choice

Jin Yeub Kim[†] Wooyoung Lim[‡]

Appendix C provides discussions of our modeling choices, along with the relevant theoretical characterizations and results. Appendix D presents additional analyses of data from Experiments I and II. Appendix E reports results from an auxiliary experiment that follows the same design as Experiment II, except that it does not include the bonus-payment scheme.

Appendix C: Supplements to Sections 2 and 3

C.1 Discussion of the Environment and the Mediation Protocol

Our baseline environment can be formulated as a special case of a Bayesian incentive problem à la Myerson (1983), in which outcomes are combinations of enforceable and private actions. Let D_0 denote the set of all possible *enforceable actions*, and for each player i, let D_i denote the set of all possible *private actions* that are controlled by player i. Let $D = D_0 \times D_1 \times D_2$ denote the set of all possible combinations of enforceable and private actions, with d denoting a typical *outcome* in D. For each player i, T_i is the set of possible types; let $T = T_1 \times T_2$ denote the set of type profiles $t = (t_1, t_2)$. In this problem, a mediator chooses an outcome $d = (d_0, d_1, d_2) \in D$ as a function of the reported types $t \in T$. Then the enforceable action d_0 is carried out, and each player i is confidentially recommended the private action d_i. The incentive constraints are characterized to ensure that the players report their types truthfully and carry our their recommended private actions obediently.

In our environment, $D_0 = \{d^a, d^*\}$, where d^a represents agreement (the equal split) and d^* represents disagreement (war), and each player i's set of private actions is simply $D_i = \{\text{accept, reject}\}$. If any player chooses "reject," disagreement occurs and war payoffs are realized; and disagreement itself is an enforceable action that also gives war payoffs to all players. We can therefore, without loss of generality, restrict attention to mechanisms in which no player is ever recommended "reject," because the enforceable disagreement action d^* can be used instead. Hence, the outcome space D effectively reduces to D_0: given type reports m, the mediator recommends d^a (the equal split) with probability $p(m)$ and d^* (war)

[†]Associate Professor, School of Economics, Yonsei University, 50 Yonsei-ro Seodaemun-gu, Seoul 03722, Republic of Korea, E-mail: jinyeub@yonsei.ac.kr

[‡]Professor, Department of Economics, The Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong, E-mail: wooyoung@ust.hk

with probability $1-p(m)$. Restricting attention to symmetric mechanisms, we use notations $p_H \equiv p(h,h)$, $p_M \equiv p(h,l) = p(l,h)$, and $p_L \equiv p(l,l)$.

Given this formulation, the incentive constraints that ensure truthful reporting and obedience in the mechanism reduce to the following two sets of constraints. First, the mechanism (p_H, p_M, p_L) is (Bayesian) *incentive compatible* iff it satisfies the following informational incentive constraints for the H and L types, respectively:

$$\begin{aligned} &q(p_H(1/2) + (1-p_H)\theta/2) + (1-q)(p_M(1/2) + (1-p_M)\delta\theta) \\ &\geq q(p_M(1/2) + (1-p_M)\theta/2) + (1-q)(p_L(1/2) + (1-p_L)\delta\theta); \\ &q(p_M(1/2) + (1-p_M)(1-\delta)\theta) + (1-q)(p_L(1/2) + (1-p_L)\theta/2) \\ &\geq q(p_H(1/2) + (1-p_H)(1-\delta)\theta) + (1-q)(p_M(1/2) + (1-p_M)\theta/2). \end{aligned} \tag{C.1}$$

That is, no player has any incentive to lie in the mechanism. The revelation principle (Myerson, 1982) applies: there is no loss of generality in considering only incentive compatible mechanisms that satisfy (C.1). Second, the mechanism (p_H, p_M, p_L) is *individually rational* iff it satisfies the following participation constraints for the H and L types, respectively:

$$\begin{aligned} &q(p_H(1/2) + (1-p_H)\theta/2) + (1-q)(p_M(1/2) + (1-p_M)\delta\theta) \\ &\geq q\theta/2 + (1-q)\delta\theta; \\ &q(p_M(1/2) + (1-p_M)(1-\delta)\theta) + (1-q)(p_L(1/2) + (1-p_L)\theta/2) \\ &\geq q(1-\delta)\theta + (1-q)\theta/2. \end{aligned} \tag{C.2}$$

That is, no player has any incentive to reject in the mechanism. With no loss of generality, we can assume that the players will agree to participate in the mechanism that satisfies (C.2) (see Myerson, 1991, p.267). We say the mechanism (p_H, p_M, p_L) is (incentive) *feasible* if it satisfies (C.1) and (C.2), which ensure truthful and obedient participation in mediation.

In the mediation game of Hörner, Morelli and Squintani (2015) (HMS), given type reports, the mediator publicly recommends a split $(x, 1-x)$ according to a cumulative distribution function, with some recommendations leading to war. Each player then decides independently whether to accept or reject the recommended split; unless both players accept, war occurs. Implementing this mediation protocol in the lab, Casella, Friedman and Perez Archila (2025) (CFP) set $\delta = 1$ and constrain the mediator's recommendations to lie in a restricted set containing only those that appear in the optimal mediation mechanism: $\{(1-\theta, \theta), (1/2, 1/2), (\theta, 1-\theta), w\}$, where w denotes "walking out," equivalent to a recommendation of war. We constrain them further to $\{(1/2, 1/2), w\}$. This restriction serves two important purposes in our paper.

First, it simplifies the representation of mediators, facilitating subjects' comparison across options. If a mediator could recommend different splits $(x, 1-x)$ for $x \in [0,1]$, the description of mediator would involve both the recommended splits and the probabilities of recommending those splits, as functions of type reports. Instead, when the only possible split recommendation is $(1/2, 1/2)$, each mediator can be represented solely by the probability of recommending agreement given type reports.

Second, the restriction allows us to stipulate the mediation protocol without a separate stage where the players decide whether to accept or reject the mediator's recommendations, as reflected in the formulation of our environment above. In our mediator selection game implemented in the lab, the subordinate has the option to decline the mediator after the principal's chosen mediator is announced. This concerns participation in mediation rather than disobedience or rejection of the mediator's recommendation.

Importantly, the absence of an acceptance/rejection stage does not imply that our mediator has enforcement power (what HMS term the arbitrator). When the mediator's agreement recommendation is restricted to an equal split, the optimal recommendation strategy is the same regardless of enforcement power. For mediators without enforcement power, the relevant incentive constraints are incentive compatibility constraints *involving double deviations* and *ex post* participation constraints. With only one possible agreement outcome, however, Proposition 1 of Kim (2017) proves that the set of feasible arbitration and mediation mechanisms coincide. The intuition behind the proof is as follows. The participation constraints (C.2) ensure that the feasible probabilities (p_H, p_M, p_L) induce posterior beliefs under which players are willing to accept the equal split when recommended, in addition to being willing to participate in mediation. Together with the incentive compatibility constraints (C.1), this ensures that players have incentives to honestly report their types and to "voluntarily" obey the equal-split recommendation. That is, the optimal recommendation of the equal split by the "arbitrator" is self-enforcing rather than reliant on enforcement power. Accordingly, unlike in HMS, whether the mediator has enforcement power is immaterial in our environment; our objective is not to compare mediation with arbitration.

C.2 Discussion of the Mediator Selection Game

In Stage 2 of the mediator selection game described in Section 2.2, the subordinate simultaneously decides whether to go to war or to participate in mediation and, if participating, what report to send to the mediator. By contrast, the principal only decides what report to send. This asymmetry in participation strategies reflects the requirement that a feasible mediator must satisfy the participation constraints (C.2) *under the prior beliefs*.

After the announcement of her choice among feasible mediators, the principal acquires no additional information about the subordinate's type. Hence, her participation constraints remain satisfied in Stage 2 given her prior beliefs about the subordinate. When the mediator is to be played, the principal therefore only needs to decide what message to send. By contrast, the subordinate may update his beliefs about the principal's type based on the announcement, potentially violating his participation constraints for the chosen mediator. We therefore give the subordinate the additional option of going to war, allowing his participation decision to reflect his updated beliefs.

One could alternatively separate the subordinate's decisions in Stage 2 as follows. The subordinate first chooses whether to go to war (declining the mediator) or to participate in the mediation. Conditional on participation, the two players then proceed to a subsequent stage in which they send reports to the mediator. This structure resembles the three-stage mechanism selection game studied by Maskin and Tirole (1992). In their setting, the subordinate has no private information, so only the subordinate updates beliefs about the principal's type following the principal's Stage 1 choice. In our setting, however, the subordinate has private information; so the principal could infer information about the subordinate's type from his participation decision, in addition to the subordinate inferring information about the principal's type from her mediator choice. Introducing a three-stage structure would therefore add an additional layer of information leakage, complicating the informed principal's mediator selection problem both theoretically and experimentally without yielding commensurate insights. The two-stage game described in Section 2.2 allows us to focus squarely on the informed principal's dilemma arising from her private information.

C.3 Mediator Characterization

To describe our theoretical characterizations, we use the following three statistics, the first two of which were used in HMS: $\lambda \equiv \frac{q}{1-q}$, $\gamma_H \equiv \frac{\delta\theta - 1/2}{1/2 - \theta/2}$, and $\gamma_L \equiv \frac{1/2-\theta/2}{1/2-(1-\delta)\theta}$. The parameter $\lambda > 0$ is the H/L type odds ratio; $\gamma_H > 0$ measures the H type's net benefit of war against an L type relative to its net cost of war against an H type, and $\gamma_L > 0$ measures the L type's net benefit of agreement with an L type relative to that with an H type. With this simplification, the assumption of $q\theta/2 + (1-q)\delta\theta > 1/2$ can be rewritten as $\lambda < \gamma_H$.

C.3.1 Interim Incentive Efficient Mechanisms

We apply the concept of efficiency to delineate mediators that the players could reasonably consider choosing. The proper concept of efficiency is *interim incentive efficiency* for games in which the players know their private information when the game begins; and is *ex ante*

incentive efficiency for games in which the players learn their private information during the game (Holmström and Myerson, 1983).

The following proposition characterizes the set of all interim incentive efficient mechanisms, adapted from Proposition 2 in Kim (2017) stated here without proof.

Proposition C.1 (Kim 2017)**.** *For the environment with $\gamma_H > \gamma_L$, any mediation mechanism (p_H, p_M, p_L) that satisfies the following characteristics is interim incentive efficient (IIE):*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M \in [0, \lambda/\gamma_H]$, $p_H = 1$.*
2. *For $\gamma_L \leq \lambda < \gamma_H$, $p_L = 1$, $p_M \in \left[0, \frac{\gamma_L}{\gamma_L+\gamma_H-\lambda}\right]$, $p_H = p_M + (1-p_M)\gamma_L/\lambda$.*

For the environment with $\gamma_H \leq \gamma_L$, the characterization of IIE mechanisms is subsumed by Case 1 in Proposition C.1, in which case the upper bound on p_M is λ/γ_H if $\lambda < \gamma_H$ and 1 if $\gamma_H \leq \lambda < \gamma_L$. Note that p_L is always one, and p_H is determined given p_M and is increasing in p_M. Therefore, we can use p_M as the sole parameter that represents each IIE mediator.

The following proposition characterizes the ex ante incentive efficient mechanism.

Proposition C.2. *For the environment with $\gamma_H > \gamma_L$, there is a unique ex ante incentive efficient mechanism such that:*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M = \lambda/\gamma_H$, $p_H = 1$.*
2. *For $\gamma_L \leq \lambda < \gamma_H$, $p_L = 1$, $p_M = \frac{\gamma_L}{\gamma_L+\gamma_H-\lambda}$, $p_H = p_M + (1-p_M)\gamma_L/\lambda$.*

Proof. Because ex ante incentive efficiency implies interim incentive efficiency, we solve for the ex ante incentive efficient mechanism among all IIE mechanisms. A player's ex ante expected utility in mechanism $p \equiv (p_H, p_M, p_L)$, denoted by $U(p)$, is:

$$\begin{aligned} U(p) \equiv\ & q^2(p_H(1/2) + (1-p_H)(\theta/2)) + q(1-q)(p_M(1/2) + (1-p_M)(\delta\theta)) \\ & + (1-q)q(p_M(1/2) + (1-p_M)(1-\delta)\theta) + (1-q)^2(p_L(1/2) + (1-p_L)(\theta/2)) \\ =\ & (1/2 - \theta/2)Q(p) + \theta/2, \end{aligned}$$

where $Q(p) \equiv q^2 p_H + 2q(1-q)p_M + (1-q)^2 p_L$, which is the ex ante probability of agreement in mechanism $p \equiv (p_H, p_M, p_L)$ given q. Maximizing $U(p)$ is therefore equivalent to maximizing $Q(p)$, because the two differ only by a positive linear transformation. For any given IIE mechanism, $p_L = 1$ and p_H increases with p_M. The mechanism that maximizes $Q(p)$ is thus the one with the highest feasible value of p_M among all IIE mechanisms. This mechanism is uniquely ex ante incentive efficient. □

Proposition 3 in Kim (2017) can be used to characterize neutral optima for our environment, stated here without proof.

Proposition C.3 (Kim 2017). *For the environment with $\gamma_H > \gamma_L$, there is a unique neutral optimum such that:*

1. *For $\lambda < \gamma_L$, $p_L = 1$, $p_M = 0$, $p_H = 1$.*
2. *For $\gamma_L \leq \lambda < \gamma_H$, $p_L = 1$, $p_M = 0$, $p_H = \gamma_L/\lambda$.*

Theoretical Benchmarks in Section 2.3 Given the experimental parameter values $\theta = 0.75$, $\delta = 0.8$, and $q = 1/4$ or $2/5$, the IIE mediators for our experimental environment can be characterized as:

1. For $q = 1/4$, $p_L = 1$, $p_M \in [0, 5/12]$, and $p_H = 1$.
2. For $q = 2/5$, $p_L = 1$, $p_M \in [0, 75/103]$, and $p_H = 15/28 + (13/28)p_M$.

We focus on two extreme mediators in terms of the value of p_M as theoretical benchmarks and provide subjects with close approximations to these mediators in the lab. These two extremes have clear theoretical interpretations. The ex ante incentive efficient mediator is the IIE mediator with the highest feasible value of p_M, and thus with the highest ex ante probability of agreement among all IIE mediators. We label this mediator the *P-Max Mediator*. It corresponds to the optimal mediation program studied in HMS and implemented in CFP. At the other extreme, the neutral optimum is the IIE mediator with the lowest feasible value of p_M among all IIE mediators. We label this mediator the *Neutral Mediator*. Corollary 1 in Kim (2017) further implies that, among all IIE mediators, the P-Max Mediator is the best feasible mechanism for an L type player, whereas the Neutral Mediator is the best feasible mechanism for an H type player.

C.3.2 Alternative Mediator Choice

We treat interim efficiency as a requirement for the set of mediators that players may consider. Maskin and Tirole (1992) (henceforth, MT92) study the problem of mechanism selection by an informed principal using a different approach than Myerson (1983). In their analysis, the *Rothschild-Stiglitz-Wilson (RSW) allocation* plays a central role (see MT92 for the formal definition), which corresponds to the *best safe mechanism* (Myerson, 1983). MT92 characterize the equilibrium set of the mechanism selection game, which consists of the allocations that weakly Pareto dominate the RSW allocation. Thus, the RSW allocation can be interpreted as the worst equilibrium outcome for every type of principal. MT92 further show that, when only the principal has private information, the RSW allocation is the unique allocation that passes the intuitive criterion.[1] However, as MT92 note, the RSW

[1] Nishimura (2022) extends this result to a trading environment with bilateral asymmetric information.

allocation need not be interim efficient relative to the prior (or any strictly positive belief), and there are many equilibria allocations that are not even weakly interim efficient.

We now characterize the RSW allocation, which we call the *RSW Mediator*, in our setting.

Proposition C.4. *For the environment with $\gamma_H > \gamma_L$, the RSW Mediator is characterized by $p_L = 1$, $p_M = p_H = 0$.*

Proof. The RSW allocation is defined as an allocation in which each type of the principal maximizes her own utility within the set of allocations that are incentive compatible (for the principal) and, regardless of the principal's type, yields the subordinate at least his reservation utility. In our setting, the constraints translate into the principal's interim feasibility constraints and the subordinate's *ex post* feasibility constraints, which require that the subordinate be willing to participate and report truthfully when he knows the principal's true type. Because the incentive structure is the same for both principal and subordinate, the subordinate's ex post feasibility constraints imply the principal's interim feasibility constraints. Therefore, the RSW Mediator is the solution to the following program:

$$
\begin{aligned}
& \text{For all } t \in \{H, L\}, \max_{p=(p_H,p_M,p_L)} U_1(p|t) \\
\text{subject to: } & p_H(1/2) + (1-p_H)\theta/2 \geq p_M(1/2) + (1-p_M)\theta/2, && \text{(HH-EPIC)} \\
& p_M(1/2) + (1-p_M)(1-\delta)\theta \geq p_H(1/2) + (1-p_H)(1-\delta)\theta, && \text{(LH-EPIC)} \\
& p_M(1/2) + (1-p_M)\delta\theta \geq p_L(1/2) + (1-p_L)\delta\theta, && \text{(HL-EPIC)} \\
& p_L(1/2) + (1-p_L)\theta/2 \geq p_M(1/2) + (1-p_M)\theta/2, && \text{(LL-EPIC)} \\
& p_H(1/2) + (1-p_H)\theta/2 \geq \theta/2, && \text{(HH-EPIR)} \\
& p_M(1/2) + (1-p_M)(1-\delta)\theta \geq (1-\delta)\theta, && \text{(LH-EPIR)} \\
& p_M(1/2) + (1-p_M)\delta\theta \geq \delta\theta, && \text{(HL-EPIR)} \\
& p_L(1/2) + (1-p_L)\theta/2 \geq \theta/2, && \text{(LL-EPIR)}
\end{aligned}
$$

where $U_1(p|t)$ is the t-type principal's expected utility in mechanism $p = (p_H, p_M, p_L)$. That is, $U_1(p|H) = q(p_H(1/2) + (1-p_H)\theta/2) + (1-q)(p_M(1/2) + (1-p_M)\delta\theta)$ and $U_1(p|L) = q(p_M(1/2) + (1-p_M)(1-\delta)\theta) + (1-q)(p_L(1/2) + (1-p_L)\theta/2)$. The HH-EPIC and LH-EPIC constraints together imply $p_H = p_M$. The HL-EPIC and LL-EPIC constraints imply $p_L \geq p_M$. The four EPIR constraints imply $p_H \geq 0$, $p_M \geq 0$, $p_M = 0$, and $p_L \geq 0$. Hence, $p_H = p_M = 0$. Substituting into $U_1(p|L)$ shows that it is maximized at $p_L = 1$. □

The RSW Mediator prescribes agreement only when the reported types are both L. The associated ex ante probability of peace is $1/4 = 0.25$. The RSW Mediator is not only interim

incentive *inefficient* but also inferior to the best separating equilibrium of the unmediated communication game (see Proposition D.1 in online Appendix D.1). Thus, the equilibrium set of mediators that weakly Pareto dominate the RSW Mediator yields too large a set of predictions for our informed mediator selection problem.

In our experiments, subjects' choice set is restricted to the two IIE mediators. Any IIE mediator can be supported as a sequential equilibrium of the informed principal's mechanism selection game. Hence, adding a third IIE mediator, or allowing subjects to choose from the entire set of IIE mediators, would unnecessarily expand the choice set without substantively affecting the results. Alternatively, one could include the RSW Mediator as a third option but because it is strictly worse than the two IIE mediators for both types of the principal, we conjecture that subjects would hardly ever entertain such an option.

C.4 No Separating Equilibrium

In Section 3.2, we provided an intuitive explanation for why no separating equilibrium with honest participation exists in the informed mediator selection game. We state and prove this result formally.

Proposition C.5. *In the informed mediator selection game, given the set of IIE mediators as the available mediators, there exists no separating sequential equilibrium in which the two types of the principal choose distinct IIE mediators, followed by participation and truthful reporting by both players in the mediation stage.*

Proof. Suppose that there are two IIE mediators, denoted by μ_H and μ_L, such that the H type principal is expected to choose μ_H and the L type principal is expected to choose μ_L. (Our argument here can be extended to allow for randomized mediator selection.) The principal would choose mediators in this way only if they satisfy $U_1(\mu_t|t) \geq U_1(\mu_{t'}|t)$ for each $t = H, L$ and $t' \neq t$, and are incentive feasible for the principal, where $U_1(\cdot|t)$ denotes the principal's interim expected payoff given type t. By Corollary 1 in Kim (2017), the two types have conflicting incentives, reflected in their opposite preference orderings over all IIE mediators. Thus, it suffices to consider separating equilibria, if any exist, in which the L type principal chooses an IIE mediator μ_L with $p_M \neq 0$, while the H type principal chooses an IIE mediator μ_H with a strictly lower value of p_M.

If the subordinate expects the principal to choose mediators in this manner, the chosen mediator can be implemented on the equilibrium path only if it is incentive feasible given the information revealed about the principal's type. By Bayes' consistency, when μ_t is selected, the subordinate would rationally infer that the principal's type is t. Hence, each μ_t must be incentive feasible when the subordinate knows that the principal's type is t.

When μ_H is chosen and announced, the subordinate would infer that the principal is of H type. An L type subordinate would then deviate from truthful reporting because $p_M(1/2) + (1-p_M)(1-\delta)\theta < p_H(1/2) + (1-p_H)(1-\delta)\theta \longleftrightarrow (p_M - p_H)(1/2) < (p_M - p_H)(1-\delta)\theta$, where $p_M < p_H$ for any IIE mediator characterized in Proposition C.1, and $1/2 > (1-\delta)\theta$. This violates the L type's incentive compatibility constraint under the updated belief implied by Bayes' consistency. Similarly, when μ_L is chosen and announced, the subordinate would infer that the principal is of L type. An H type subordinate would then refuse to participate and choose war, because $p_M(1/2) + (1-p_M)\delta\theta < \delta\theta$ for $p_M \neq 0$, violating the H type's individual rationality constraint under the updated belief. That is, whichever mediator, μ_H or μ_L, is selected becomes infeasible once chosen. This completes the proof. □

Appendix D: Additional Analyses

D.1 Experiment I: Unmediated Communication and Mediation

D.1.1 Experimental Procedures for Parts 1–3

In the environment described in Section 2.1, the players can communicate without mediation. We consider the following simple unmediated communication game. After privately learning their own types, both players simultaneously send unverifiable messages $m_i \in \{h, l\}$ to each other. After messages are sent and received, the two players simultaneously choose either "in" or "out." If both choose "in," the equal split (agreement) is implemented; if either player chooses "out," war occurs and the shrunk surplus is divided according to the players' types. To keep the comparison to the mediation game consistent, we have restricted players' demands to either an equal split or walking out.

In Experiment I, we implement this game in Part 1 of each session for all treatments. The experimental procedures of Part 1, together with those for Parts 2–3, are described below.

Part 1 (Unmediated Communication). In each round, subjects were randomly and anonymously matched in pairs and independently assigned types by the computer according to q. After learning their own types, subjects each sent a message chosen from $\{H, L\}$ to the other participant in the pair. After messages were exchanged, each subject simultaneously chose either "In" or "Out." If both subjects chose "In," agreement was reached and the surplus was split equally. If either subject chose "Out," disagreement followed, giving payoffs according to subjects' true types. At the end of each round, subjects received feedback on both subjects' types, messages exchanged, the final outcome, and their own payoff.

Parts 2–3 (Mediation). At the beginning of each part, subjects were introduced to a computer mediator with a mediation plan that applied to all rounds in that part. In eight sessions under the Simple treatment, either the 40-Mediator or the 0-Mediator was used; in eight sessions under the Complex treatment, either the 70-Mediator or the 0-Mediator was used. In each round, subjects were randomly and anonymously matched in pairs and independently assigned types by the computer according to q. After learning their own types, subjects each sent a confidential message from $\{H, L\}$ to the mediator. Based on the reported types, the mediator either prescribed agreement or walked out of the mediation, according to its mediation plan. The mediator's plan remained visible on subjects' screens throughout the message-sending stage. If the mediator prescribed agreement, the equal split was implemented; otherwise, disagreement occurred, shrinking the surplus and allocating payoffs according to subjects' true types. At the end of each round, subjects received feedback on both subjects' types, messages sent, the mediation outcome, and their own payoff.

D.1.2 Theoretical Predictions

As in HMS, the unmediated communication game can be set up as a public correlation device. With probability $p(m)$, the device coordinates the players on both choosing "In," leading to the equal split. With probability $1 - p(m)$, the coordination fails, resulting in war. We restrict attention to pure-strategy separating equilibria in which players report their types truthfully, and to equilibria in which probabilities $p(m)$ are symmetric across players. Let $\tilde{p}_H \equiv p(h,h)$, $\tilde{p}_M \equiv p(h,l) = p(l,h)$, and $\tilde{p}_L \equiv p(l,l)$.

We characterize the optimal equilibrium of unmediated communication that maximizes the ex ante probability of agreement, subject to the constraints that the players communicate their types truthfully and agree to the equal split whenever they are coordinated on by the public correlation device. As introduced in online Appendix C.3, define $\lambda \equiv \frac{q}{1-q}$, $\gamma_H \equiv \frac{\delta\theta - 1/2}{1/2 - \theta/2}$, and $\gamma_L \equiv \frac{1/2 - \theta/2}{1/2 - (1-\delta)\theta}$.

Proposition D.1. *For the environment with $\gamma_H > \gamma_L$, the optimal equilibrium of the unmediated communication game is characterized by $\tilde{p}_L = 1$, $\tilde{p}_M = 0$, and $\tilde{p}_H = 1$ if $\lambda < \gamma_L$, and by $\tilde{p}_L = 1$, $\tilde{p}_M = 0$, and $\tilde{p}_H = \gamma_L/\lambda$ if $\gamma_L \leq \lambda < \gamma_H$.*

Proof. The optimal separating equilibrium solves the following problem:

$$\min_{\tilde{p}_H, \tilde{p}_M, \tilde{p}_L} q^2(1 - \tilde{p}_H) + 2q(1-q)(1 - \tilde{p}_M) + (1-q)^2(1 - \tilde{p}_L)$$

subject to the incentive compatibility (IC) constraints with double deviations and the ex-post individual rationality (IR) constraints for both types. Because messages reveal types in

a separating equilibrium, players must find it optimal to accept the equal split when offered. However, because $\delta\theta > 1/2$, an H type player facing a self-reported L type opponent strictly prefers war to the equal split. This implies that in any optimal separating equilibrium, $\tilde{p}_M = 0$, in order to satisfy the ex-post IR constraint for the H type. Because $1/2 \geq (1-\delta)\theta$, the ex-post IR constraint for L type is always satisfied. Given $\tilde{p}_M = 0$, the IC constraints with double deviations for types H and L can be written as follows:

$$\begin{aligned} & q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)\theta/2) + (1-q)\delta\theta \\ \geq\ & q\theta/2 + (1-q)(\tilde{p}_L \max\{1/2, \delta\theta\} + (1-\tilde{p}_L)\delta\theta); \end{aligned} \tag{H-IC*}$$

$$\begin{aligned} & q(1-\delta)\theta + (1-q)(\tilde{p}_L(1/2) + (1-\tilde{p}_L)\theta/2) \\ \geq\ & q(\tilde{p}_H \max\{1/2, (1-\delta)\theta\} + (1-\tilde{p}_H)(1-\delta)\theta) + (1-q)\theta/2. \end{aligned} \tag{L-IC*}$$

The max operators capture the possibility of double deviations, in which a player both misreports her type and deviates from the recommendation of the equal split, opting instead for war. Because $\max\{1/2, \delta\theta\} = \delta\theta$, constraint (H-IC*) simplifies to $q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)\theta/2) \geq q\theta/2$, which holds for all $\tilde{p}_H \in [0,1]$. In (L-IC*), the double deviation in which an L type misreports and then wages war against an H type is never profitable, because $\max\{1/2, (1-\delta)\theta\} = 1/2$. Setting $\tilde{p}_L = 1$ minimizes the objective function only to relax the constraint (L-IC*) without violating any other constraints. Rewriting (L-IC*), we obtain $q(1-\delta)\theta + (1-q)(1/2) \geq q(\tilde{p}_H(1/2) + (1-\tilde{p}_H)(1-\delta)\theta) + (1-q)\theta/2$, which implies $\tilde{p}_H \leq \frac{1-q}{q}\frac{1/2-\theta/2}{1/2-(1-\delta)\theta} = \gamma_L/\lambda$. Because the objective function is decreasing in $\tilde{p}_H$, the optimal solution sets $\tilde{p}_H = 1$ when $\gamma_L > \lambda$, and $\tilde{p}_H = \gamma_L/\lambda$ when $\gamma_L \leq \lambda$. □

Propositions C.3 and D.1 show that the optimal equilibrium of unmediated communication coincides with the Neutral Mediator, which immediately implies the following corollary.

Corollary D.1. *The Neutral Mediator achieves the same ex ante probability of agreement as the optimal equilibrium of the unmediated communication game, while all other IIE mediators (including the P-Max Mediator) achieve a strictly higher probability of agreement.*

In theory, optimal mediation always yields a weakly higher probability of agreement than unmediated communication. HMS compare the optimal mediation mechanism with the optimal equilibrium of unmediated communication game and show that, for a subset of their parameter space, optimal mediation yields a strictly higher probability of agreement. CFP experimentally test this theoretical prediction. Likewise, we can compare the performance of unmediated communication (UC), mediation under the P-Max Mediator (P-Max Mediation), and mediation under the Neutral Mediator (Neutral Mediation). The preceding theoretical analysis leads to the following testable hypotheses.

Hypothesis D.1 (UC vs. P-Max Mediation vs. Neutral Mediation)**.**

(a) P-Max Mediation yields a higher rate of agreement than UC.

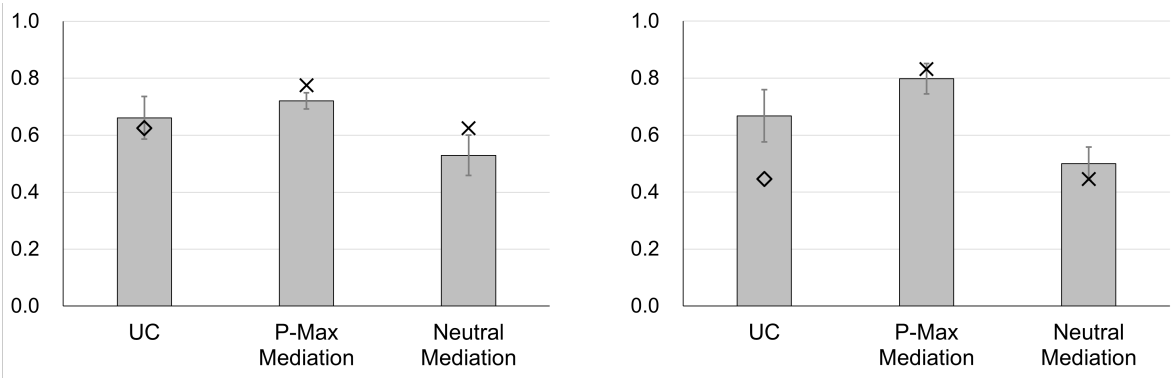
(b) P-Max Mediation yields a higher rate of agreement than Neutral Mediation.

(c) There is no difference in the rates of agreement between Neutral Mediation and UC.

D.1.3 Experimental Results

The Informed and Uninformed treatments in Experiment I differ only in Part 4, which concerns mediator selection. So for Parts 1–3, we pool data across the Informed and Uninformed treatment sessions within each Simple or Complex environment, aggregating all rounds from the eight relevant sessions (four Informed and four Uninformed) for each part.

Figure D.1 reports the average rate of agreement across the three parts—UC, P-Max Mediation, and Neutral Mediation—for the Simple and the Complex treatments. We show 95% confidence intervals calculated from standard errors clustered at the session level. Table D.1 on the next page reports the non-parametric test results for agreement rates.



(a) Simple (b) Complex

Figure D.1: Rate of Agreement

Note: The ⋄ markers indicate the ex ante probability of agreement in the optimal equilibrium of the unmediated communication game, and the × markers indicate that under each mediator.

The average agreement rates, ordered as {UC, P-Max Mediation, Neutral Mediation}, are {66%, 72%, 53%} in the Simple treatment and {67%, 80%, 50%} in the Complex treatment.

In both treatments, *P-Max Mediation yields higher agreement rates than both UC and Neutral Mediation*, consistent with Hypothesis D.1(a)-(b). The effect of P-Max Mediation on agreement relative to UC is small in the Simple treatment (one-sided p =0.091, Wilcoxon signed-rank test). Importantly, the direction of the effect aligns with the theoretical prediction that P-Max Mediation yields a higher probability of agreement than UC. This finding

contrasts with CFP's experimental result, which shows that unmediated communication results in slightly more frequent peace than (peace-maximizing) mediation on average, with no statistically significant difference between the two.

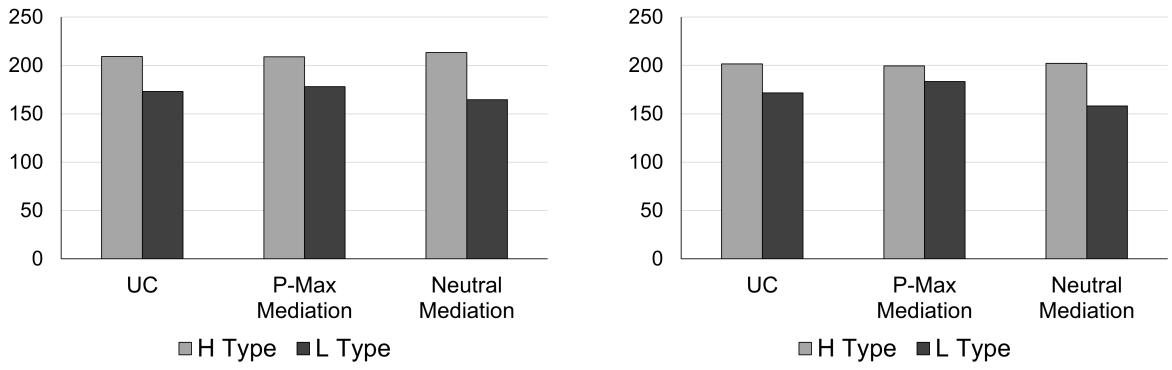
By contrast, *Neutral Mediation yields significantly lower agreement rates than UC* in both treatments, contrary to Hypothesis D.1(c). The theoretical predictions for agreement rates are indicated by the diamond and cross markers in the figure. The average agreement rate under UC exceeds the optimal equilibrium prediction of the unmediated communication game, particularly in the Complex treatment. This is surprising, given that in the lab a public correlation device is absent; hence, unmediated communication as played in the lab would be expected to achieve a weakly lower probability of agreement than the theoretical optimal equilibrium. In comparison, P-Max Mediation slightly underperforms relative to its theoretical prediction, and the performance of the Neutral Mediator relative to its theoretical prediction is mixed across the two treatments.

Table D.1: **Non-parametric Tests for Experiment I (Parts 1–3)**

Null Hypothesis	p-values	
	Simple	Complex
The agreement rate is the same across P-Max Mediation and UC.	0.1829	0.0781
The agreement rate is the same across P-Max and Neutral Mediation.	0.0078	0.0078
The agreement rate is the same across Neutral Mediation and UC.	0.0781	0.0422

■ We use Wilcoxon signed-rank tests. All null hypotheses are two-sided.

For reference, Figure D.2 reports the average payoffs of each subject type across the three parts—UC, P-Max Mediation, and Neutral Mediation—for the Simple and the Complex treatments.



(a) Simple (b) Complex

Figure D.2: Average Payoff of Each Type (Parts 1–3, Experiment I)

D.2 Experiment II: Fixed Mediator-Selection Rule

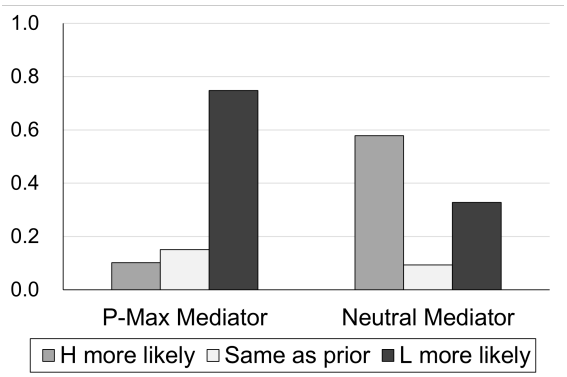
D.2.1 Experimental Procedures for Parts 1–3

In Experiment II, the experimental procedures across parts 1–3 are identical and described below, except that in each part subjects were given one of the three mediator-selection rules to play all rounds in that part: the Uncompromising Rule for Part 1, the P-Max Compromising Rule for Part 2, and the Neutral Compromising Rule for Part 3.

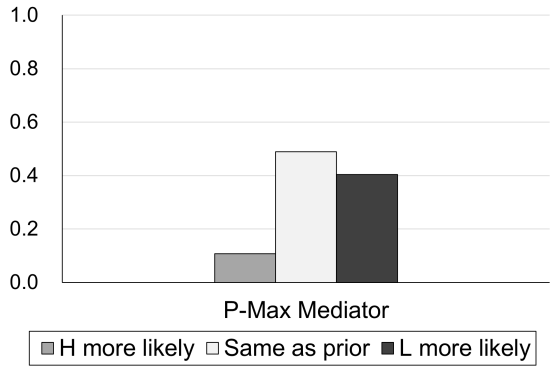
In each round, subjects were randomly and anonymously matched in pairs. Within each pair, one subject was randomly chosen, with equal probability, and announced to be the selector for that round. Subjects were then independently assigned private types by the computer according to $q = 1/4$. Subjects were reminded of the given mediator-selection rule. The selector's mediator choice was then automatically assigned based on the given mediator-selection rule and the selector's type (which the selector knows privately). After the chosen mediator was announced, the non-selector was asked to report her inference about the selector's type based on the chosen mediator. Next, both subjects sent a confidential message from $\{H, L\}$ to the chosen mediator, except that the non-selector had an additional option to decline the mediator. If the non-selector declined, disagreement occurred immediately and payoffs were determined by the subjects' true types. Otherwise, given the reported types, the mediator prescribed agreement or walked out, according to its mediation plan; and payoffs were realized. Throughout all rounds, the mediator's mediation plan, as well as the mediator-selection rule that was given for the part, remained visible on subjects' screens. At the end of each round, subjects received feedback on the given mediator-selection rule, both subjects' types, the selector's mediator choice, whether the non-selector declined, messages sent (if any), the final outcome, and their own payoffs.

D.2.2 Experimental Results

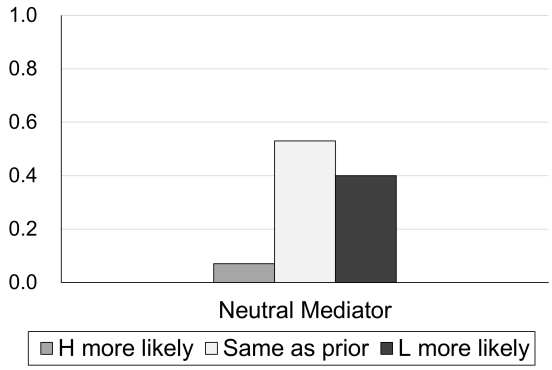
Figure D.3 reports the average frequencies of three inferences by subordinates, conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal according to the given mediator-selection rule for each part. Given the Uncompromising rule (Part 1), most subordinates infer that the principal is more likely to be of L type than under the prior after observing the principal's choice of the P-Max Mediator (75%), and more likely to be of H type after observing the choice of the Neutral Mediator (58%). Inference patterns are similar when subjects are given the P-Max Compromising rule (Part 2) and when they are given the Neutral Compromising rule (Part 3): Given the P-Max Compromising rule, the inferences "same as prior" and "L more likely" occur at rates of 49% and 40%, respectively; given the



(a) Uncompromising Rule (Part 1)



(b) P-Max Compromising Rule (Part 2)

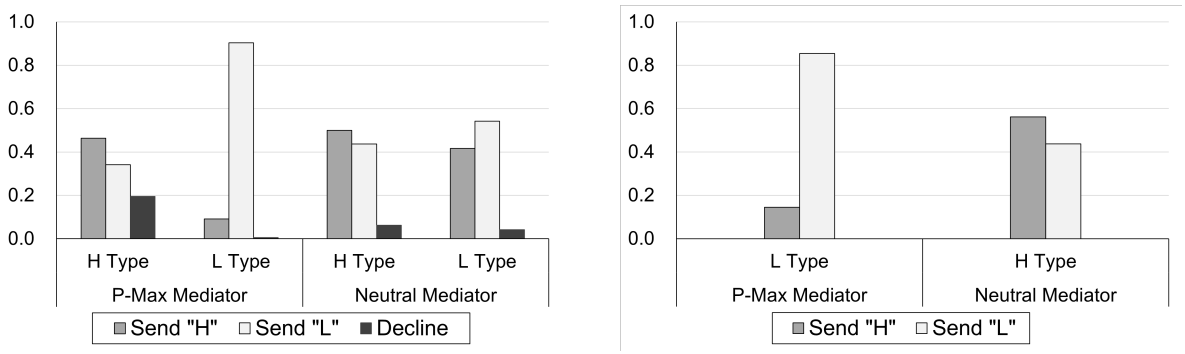


(c) Neutral Compromising Rule (Part 3)

Figure D.3: Subordinate's Inference Conditional on Mediator Choice given Selection Rule

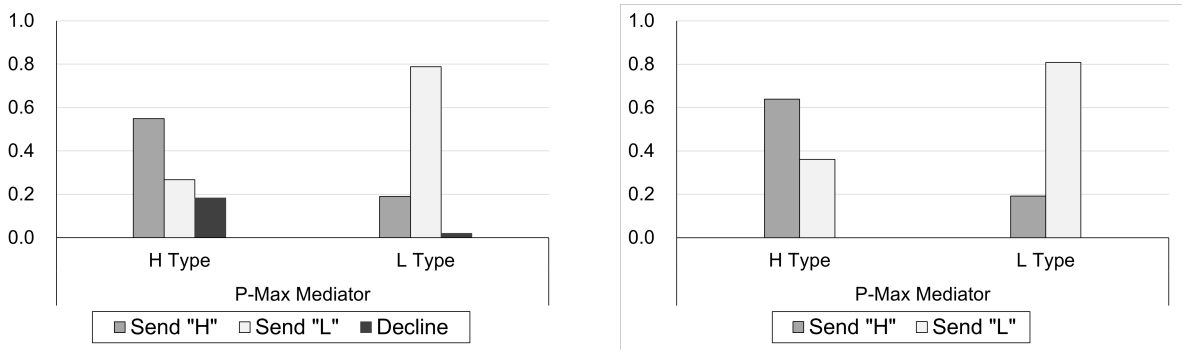
Neutral Compromising rule, the corresponding frequencies are 53% and 40%. We conjecture that some subjects interpreted the "L more likely" inference as effectively equivalent to "same as prior," given that the prior probability of the L type was already 3/4. With this caveat in mind, the inference patterns suggest that subjects correctly understood that mediator choices are informative about the principal's type under the Uncompromising rule, but uninformative under the two Compromising rules.

Figures D.4–D.6 report the average frequencies of subject strategies by type, conditional on the choice of the P-Max Mediator or the Neutral Mediator according to the given mediator-selection rule for each of the three parts. When the Uncompromising rule is given, the P-Max Mediator (resp. Neutral Mediator) is chosen when the principal's realized type is L (resp. H). Accordingly, in panel (b) of Figure D.4, only the strategy of L type (resp. H type) principals is shown when the P-Max Mediator (resp. Neutral Mediator) is chosen. When the P-Max Compromising and the Neutral Compromising rules are given, the mediator choice is fixed at the P-Max Mediator and the Neutral Mediator, respectively.



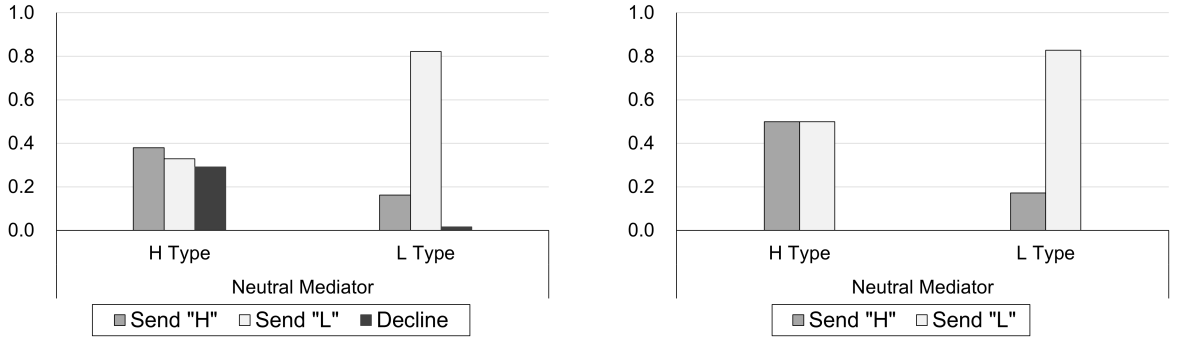
(a) Subordinate (b) Principal

Figure D.4: Subject's Strategy in Chosen Mediator by Type given Uncompromising Rule (Part 1)



(a) Subordinate (b) Principal

Figure D.5: Subject's Strategy in Chosen Mediator by Type given P-Max Compromising Rule (Part 2)



(a) Subordinate (b) Principal

Figure D.6: Subject's Strategy in Chosen Mediator by Type given Neutral Compromising Rule (Part 3)

Appendix E: Experiment II without Bonus Payments

In Experiment II, informed subjects were prompted to take the perspective of their unrealized type when making their mediator choice. This design feature was introduced because the mediator-choice data from Experiment I do not reveal whether subjects engage in forming an intertype compromise that underlies the theory of inscrutable selection. Such a compromise is an internal deliberative process that takes place prior to mediator selection, which Experiment II is designed to elicit. However, by making the unrealized type's mediator choice payoff-relevant through the bonus-payment scheme, the design exogenously induces subjects to undertake a cognitive step that is not an explicit stage of the theoretical model.

To address this concern, we conducted an auxiliary experiment that mirrors Experiment II but removes the bonus-payment scheme. Although subjects were still prompted to take the perspective of their unrealized type, the mediator choice specified for the unrealized type in each mediator-selection rule was payoff-irrelevant in the absence of bonus payments. Thus, engaging in the intertype compromise process was no longer materially required. This auxiliary experiment enables us to assess whether the main behavioral patterns observed in Experiment II are driven by the incentivization of counterfactual reasoning or whether they persist even when such reasoning is elicited in a weaker, non-incentivized manner.

E.1 Experimental Design and Procedure

As in Experiment II, the experiment focused on the Simple-Informed environment and was conducted in English using oTree in real-time online mode via Zoom (with videos on) at HKUST. A total of 90 subjects were recruited. We conducted six sessions. Each subject participated in one session. Session sizes ranged from 6 to 22 participants. Payment procedures were identical to those in Experiment II. On average, subjects earned HKD 225.7 ($\approx$ USD 29), including a HKD 40 show-up fee, for participating in a session that lasted approximately 1.5 hours.

Each session consisted of four parts, the procedures of which were identical to those in Experiment II, with three design differences. First, Parts 1–3 comprised 8 rounds each and Part 4 comprised 16 rounds, rather than 4 rounds each and 25 rounds, respectively, in Experiment II. Second, when subordinate-subjects were asked to report their inferences about the principal's type, they were given five response options instead of three: (i) Surely H, (ii) More likely to be H, (iii) Same as the prior, (iv) More likely to be L, and (v) Surely L. Because the shorter Part 4 affects only the length of exposure and the inference reports were not incentivized, these first two modifications are unlikely to affect the qualitative behavioral

patterns. Third, and most importantly, the bonus-payment scheme in Part 4 (explained at the end of Section 4.2) was removed.

E.2 Experimental Results

We report the experimental results from this auxiliary experiment and compare them with those from Experiment II reported in Sections 5 and 6. Table E.1 at the end of this appendix reports the non-parametric test results for reference.

E.2.1 Inscrutable Intertype Compromise

Figure E.1 reports the proportions of the three mediator-selection rules chosen by all principal-subjects (cf. Figure 4). The Wilcoxon signed-rank tests confirm that, on average, the Uncompromising rule is chosen significantly less often (23%) than the two Compromising rules combined and the P-Max Compromising rule is chosen significantly more often (57%) than the Neutral Compromising rule (20%). Finding 4 continues to hold.

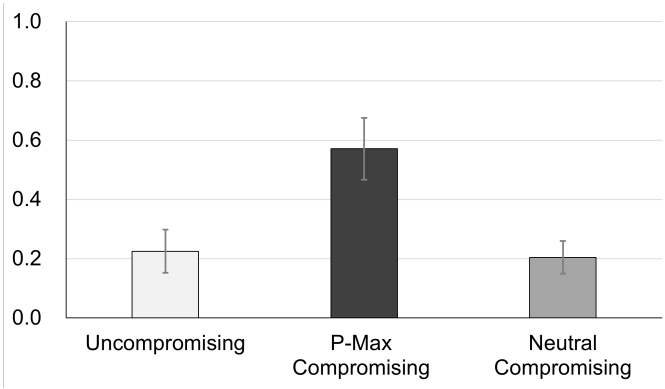
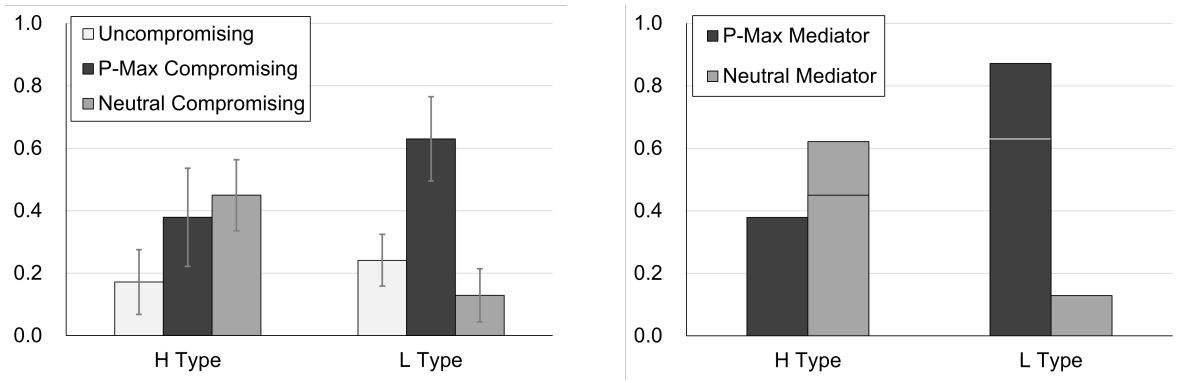


Figure E.1: Proportion of Mediator-Selection Rule Chosen (All Principals)

Panel (a) of Figure E.2 reports the proportions of mediator-selection rule choices by type, and panel (b) shows the proportions of the resulting mediator choices (cf. Figure 5). Compared to Experiment II, the average behavior of L type principals is virtually identical, whereas that of H type principals differs. In particular, L type principal-subjects choose the P-Max Compromising rule significantly more often (63%) than the Neutral Compromising (13%) (recall that the corresponding proportions in Experiment II were 63% and 9%, respectively). By contrast, H type principal-subjects choose the Neutral Compromising more often (45%) than the P-Max Compromising (38%), although this difference is statistically insignificant (cf. the corresponding proportions in Experiment II were 23% and 53%, respectively, and the difference was statistically significant).



(a) Mediator-Selection Rules Chosen (b) Resulting Mediator Choices

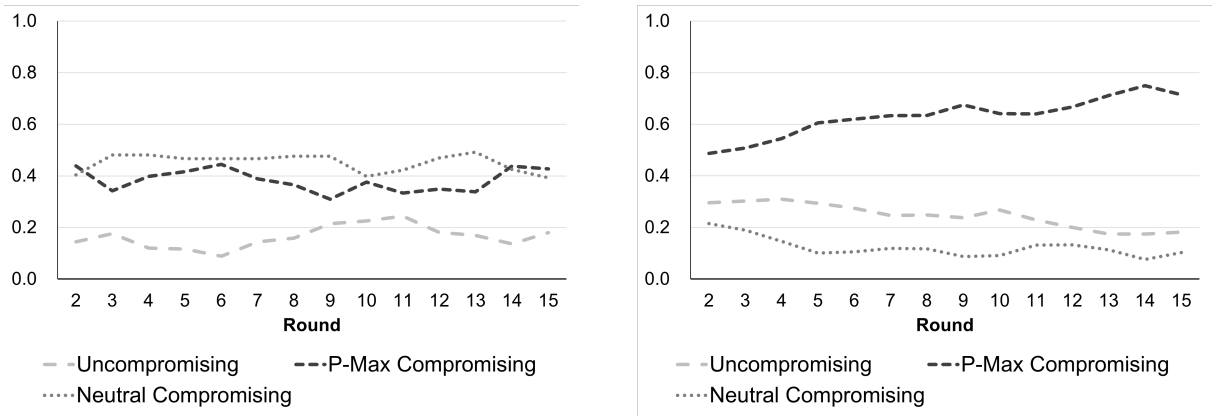
Figure E.2: Proportion of Mediator-Selection Rule/Mediator Chosen By Type (All Principals)

Thus, among the two Compromising rules, a majority of L types resolve the intertype compromise toward the P-Max Mediator, whereas H types are more evenly split between the P-Max and the Neutral Mediators; Finding 5 does not hold. Nevertheless, the resulting pattern of mediator choices is closer to that observed in Experiment I (Figure 3(a)) than the pattern observed in Experiment II (Figure 5(b)) is to that in Experiment I.

These observations suggest that the asymmetric type-dependent mediator choices arise primarily from disagreement among H type principals over how the intertype compromise should be resolved: 45% of H types favor pooling on the Neutral Mediator, while 38% favor pooling on the P-Max Mediator. This raises the question of why H types exhibit greater disagreement in this auxiliary experiment than in Experiment II. One plausible explanation is that, although prompting subjects to take the perspective of their unrealized type encouraged the formation of an intertype compromise, they were not incentivized to maintain consistency between the choices associated with their realized and unrealized types. H type subjects, in particular, may have been more responsive to the absence of such incentives and therefore more inclined to select compromises that better served the interests of their realized type.

Figure E.3 compares principal behavior over time across the two types. The figure presents three-round moving averages of the frequencies of three mediator-selection rules chosen, conditional on each principal type.[2] H type principals consistently choose the Neutral Compromising rule at frequencies between 39% and 49%, but a persistent and comparable fraction of H types choose the P-Max Compromising rule, with frequencies between 31% and 44%. Thus, even over time, H types do not converge toward a single compromise. By contrast, L type principals exhibit a clear tendency toward the P-Max Compromising rule

[2]The moving average for round n is calculated as the average frequency across rounds $n-1$, n, and $n+1$. Accordingly, the plotted data begin at Round 2 and end at Round 15.



(a) H type (b) L type

Figure E.3: Trends of Frequencies of Mediator-Selection Rules Chosen Conditional on Types (3-Round Moving Averages)

from the outset. The frequency of the P-Max Compromising rule chosen by L types starts at about 49% and gradually rises up to 71%–75% in later rounds. These patterns further reinforce the conclusion that the neutral optimum fails to emerge in the lab.

The observed tendency for each type to favor a Compromising rule that benefits its own type can be explained by *projection bias* (Loewenstein, O'Donoghue and Rabin, 2003). When forming an intertype compromise, principals evaluate the other type's interim payoff through the lens of their own realized type, effectively overweighting their own preferences. Consequently, H types tilt the compromise toward the H-type preferred mediator, while L types tilt in the opposite direction. The k-means clustering analysis below provides individual-level evidence qualitatively consistent with this interpretation. Myerson's (1983) theory conceptualizes intertype compromise as a virtual bargaining process across types. Our findings in the auxiliary experiment, however, suggest that such compromises are behaviorally distorted by projection bias in the absence of incentives for the unrealized type's choice, leading to systematic deviations from the neutral optimum.

E.2.2 Classifications of Individual Behavior

We examine individual behavior in the auxiliary experiment in the same manner as in Section 5.3 for Experiment II. Each subject's decisions as a Principal over 16 rounds can be illustrated, as in Figure 6; 90 such figures are available upon request. Using the two measures, inscrutability and type-polarization, we consolidate all 90 individual observations and classify them into five clusters using k-mean clustering. Figure E.4 presents the clustering result. We compare this to that of Experiment II (see Figure 7).

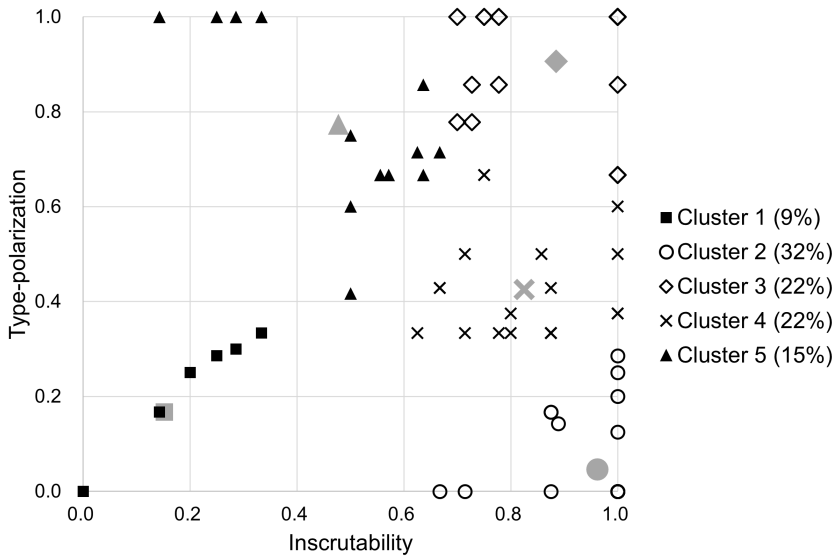


Figure E.4: k-means Clustering with Five Clusters

Note: The horizontal axis measures the frequencies of inscrutable selections, represented by choices of the P-Max and Neutral Compromising rules. The vertical axis measures the distributional divergence in choice frequencies across the three mediator-selection rules between the two types. Due to multiple subjects sharing the same pair of measures, certain markers represent more than one subject, notably at $(0, 0)$, $(1, 0)$, and $(1, 1)$. Cluster centroids are indicated by the corresponding marker shapes shaded with gray.

The proportion of individuals in Cluster 2 is smaller (32%) than in Experiment II (43%). However, this does not imply lower overall inscrutability. Both experiments exhibit a similar overall mass at "Inscrutability=1" (42% vs. 47% in Experiment II), consistent with our conclusion that Finding 4 continues to hold. The difference lies in how this mass is internally structured.

In Experiment II, the mass at "Inscrutability=1" is less differentiated, with Cluster 2 capturing much of the right-hand mass. In contrast, in the auxiliary experiment, Cluster 4 expands (22% compared to 15% in Experiment II) and captures a more distinct moderate type-polarization segment within the high-inscrutability population ("Inscrutability>0.6"). Cluster 5 nearly doubles from 8% to 15% but is more visibly separated from Cluster 4.[3] As a result, high-inscrutability individuals are more evenly distributed across type-polarization levels (Clusters 2, 3, and 4), yielding clearer vertical stratification between low, moderate, and high type-polarization clusters; whereas in Experiment II these regimes were more blended (Clusters 2 and 3). Overall, the auxiliary experiment exhibits less dominance of Cluster 2 but greater internal heterogeneity in type-polarization conditional on high inscrutability.

[3]Individuals with "Inscrutability<0.5" either mostly select the Uncompromising rule regardless of type (belonging to Cluster 1) or fully polarize, with only one type choosing the Uncompromising rule (belonging to Cluster 5).

Despite these differences, the clustering results remain broadly consistent with those of Experiment II. A large majority of principals (Clusters 2, 3, and 4; 69 individuals, 76%) engage in inscrutable intertype compromise. In addition, 38 individuals (42% of the sample) exhibit "Inscrutability=1." Within this group, 23 individuals (belonging to Cluster 2) converge on a stable, type-independent intertype compromise, while 15 individuals (belonging to Clusters 3 and 4) display type-contingent intertype compromise. These correspond to 61% and 39% of the "Inscrutability=1" group, compared to 75% and 25%, respectively, in Experiment II. Thus, at the individual level, the auxiliary experiment shows a larger share of high-inscrutability subjects exhibiting disagreement over how the intertype should be resolved, failing to converge across types. This pattern is qualitatively consistent with the aggregate-level observation that, among those making largely inscrutable selections, divergence among H type principals is more pronounced in the auxiliary experiment than in Experiment II.

E.2.3 Subordinate's Inference and Strategy

Figure E.5 reports the frequencies of five inference categories chosen by subordinates, conditional on either the P-Max Mediator or the Neutral Mediator chosen by the principal according to her chosen mediator-selection rule (cf. Figure 9).

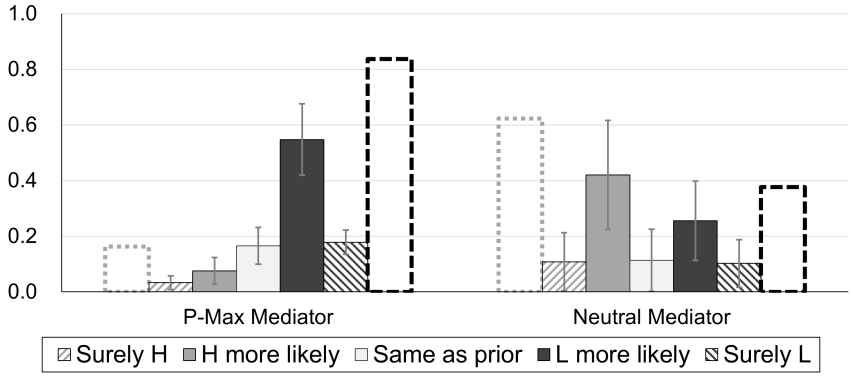


Figure E.5: Subordinate's Inference Conditional on Mediator Choice

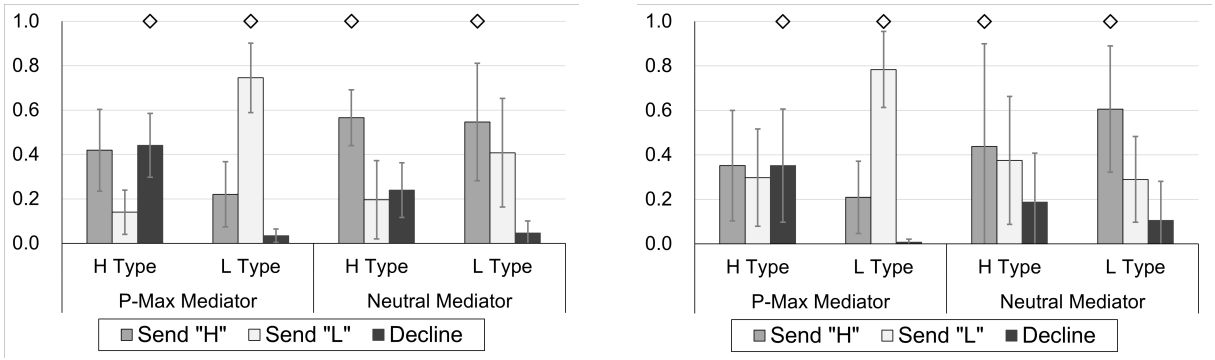
Note: The bars with gray dotted borders and black dashed borders show the empirical posterior probabilities that the principal is H type and L type, respectively, conditional on each mediator choice.

Subordinates infer that the principal is more likely to be the L type than the prior after observing the principal's choice of the P-Max Mediator with a significantly higher frequency (55%) than any other inference category in all pairwise comparisons (one-sided $p < 0.02$, Wilcoxon signed-rank tests). They infer that the principal is more likely to be the H type than the prior after observing the Neutral Mediator chosen with a significantly higher

frequency (42%) than any other category in all pairwise comparisons (one-sided $p < 0.05$, Wilcoxon signed-rank tests), except for the 'L more likely' inference for which the difference is statistically insignificant.

Again, as in Experiment I, belief reporting was not incentivized and should therefore be interpreted as supplementary evidence only. Nonetheless, the observed patterns suggest that subjects perceive the informed principal's mediator choice as revealing information. Reported beliefs were roughly in the direction of the empirically updated posteriors of the principal's type following each mediator choice. Providing more response options was intended to allow subjects to fine-tune their inferences, but the extreme categories "Surely H" and "Surely L" were rarely selected. This pattern in fact supports our argument in Section 6 that any change in beliefs was small.

Figure E.6 reports the frequencies of subordinate strategies by type, conditional on whether the principal selected the P-Max Mediator or the Neutral Mediator, aggregated over all rounds in panel (a) and over the last five rounds in panel (b).



(a) All Rounds (b) Last 5 Rounds

Figure E.6: Subordinate's Strategy in Chosen Mediator by Type

Note: The diamond markers indicate the subordinate's best responses for each type given updated posterior beliefs about the principal's type, specifically, $q' > q$ following the choice of the P-Max Mediator and $q' < q$ following the choice of the Neutral Mediator, which are consistent with the predominant beliefs reported on average. These best responses are computed against the principal sending truthful messages, and are formally characterized given any arbitrary posterior beliefs in Appendix B.4.

As shown in panel (a), L type subordinates are, on average, significantly more sincere when the P-Max Mediator is chosen than when the Neutral Mediator is chosen (75% vs. 41%, respectively; one-sided $p < 0.02$, Wilcoxon signed-rank test); H type subordinates decline both the P-Max Mediator and the Neutral Mediator at rates 44% and 24%, respectively, whereas L types rarely decline mediation. These patterns are consistent with those observed in panel (a) of Figure 11 for Experiment II and closely resemble those in panel (a) of Figure 10 for the Simple-Informed treatment in Experiment I.

A notable difference emerges when comparing the last five rounds of this auxiliary experiment with those of Experiment II (cf. panel (b) of Figure E.6 and panel (b) of Figure 11). Under the P-Max Mediator, behavior of both types of subordinates is similar across the two experiments and remains stable over time in both experiments. However, under the Neutral Mediator, H type subordinates in this auxiliary experiment become relatively less sincere (57% in all rounds and 44% in the last five rounds) and continue to decline mediation at a rate of 19% in the last five rounds, albeit with large heterogeneity. By contrast, in Experiment II, H type subordinates become substantially more sincere in later rounds and no longer decline the Neutral Mediator. One plausible explanation is that H type subjects were particularly sensitive to the absence of incentives for maintaining consistency between the choices associated with their realized and unrealized types; and were therefore more inclined to seek divergent compromises and less inclined to fine-tune their subsequent strategies.

Remark. While the average by-type behavior in inscrutable intertype compromise appears to be influenced by the explicit incentivization associated with the unrealized-type choice, the core behavioral result of Experiment II remains intact. In particular, Finding 4 is robust to the removal of the bonus-payment scheme: subjects recognize the need for inscrutable mediator selection, engage in an intertype compromise between their realized and unrealized types, and systematically view the P-Max Mediator as the appropriate intertype compromise, counter to the theory of neutral optimum.

Table E.1: **Non-parametric Tests for the Auxiliary Experiment (Part 4)**

Reference	Test	Null Hypothesis	p-values
Fig. E.1	WSR	1.1 The rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	1.2 The rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	0.03125
	WSR	1.3 The rate of choosing Compromising (both P-Max and Neutral) = the rate of Uncompromising.	0.03125
Fig. E.2	WSR	2.1 For H type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.21875
	WSR	2.2 For H type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	2.3 For H type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp..	0.28071
	WSR	2.4 For L type, the rate of choosing P-Max Comp. = the rate of choosing Uncompromising.	0.03125
	WSR	2.5 For L type, the rate of choosing Neutral Comp. = the rate of choosing Uncompromising.	0.0625
	WSR	2.6 For L type, the rate of choosing P-Max Comp. = the rate of choosing Neutral Comp.	0.03125
Fig. E.5	WSR	3.1 Given P-Max announced, the rate of 'More L' inference = the rate of 'Surely H' inference.	0.03125
	WSR	3.2 Given P-Max announced, the rate of 'More L' inference = the rate of 'More H' inference.	0.03125
	WSR	3.3 Given P-Max announced,the rate of 'More L' inference = the rate of 'Same' inference.	0.03125
	WSR	3.4 Given P-Max announced, the rate of 'More L' inference = the rate of 'Surely L' inference.	0.03125
	WSR	3.5 Given Neutral announced, the rate of 'More H' inference = the rate of 'Surely H' inference.	0.03125
	WSR	3.6 Given Neutral announced, the rate of 'More H' inference = the rate of 'Same' inference.	0.0625
	WSR	3.7 Given Neutral announced, the rate of 'More H' inference = the rate of 'More L' inference.	0.21875
	WSR	3.8 Given Neutral announced, the rate of 'More H' inference = the rate of 'Surely L' inference.	0.03125
Fig. E.6	WSR	4.1 (All rounds) The rate of rejecting P-Max = the rate of rejecting Neutral, by H type.	0.03552
	WSR	4.2 (All rounds) The rate of rejecting P-Max= the rate of rejecting Neutral, by L type.	0.18145
	WSR	5.1 (All rounds) The rate of "H" message in P-Max = that in Neutral, by H type	0.03125
	WSR	5.2 (Last five rounds) The rate of "H" message in P-Max = that in Neutral, by H type	0.3125
	WSR	5.3 (All rounds) The rate of "L" message in P-Max = that in Neutral, by L type.	0.03125
	WSR	5.4 (Last five rounds) The rate of "L" message in P-Max = that in Neutral, by L type.	0.03125

■ WSR refers to the Wilcoxon signed-rank test. All null hypotheses are two-sided.
■ With six independent sessions, the minimum attainable p-value is 0.03125 (two-sided).
■ P-Max Comp., Neutral Comp., and Uncompromising refer to the P-Max Compromising rule, the Neutral Compromising rule, and the Uncompromising rule, respectively. P-Max and Neutral refer to the P-Max and Neutral Mediators, respectively.
■ In 3.1–3.8, inference labels are abbreviated ('More L'='More likely to be L'; 'Same'='Same as prior'; 'More H'='More likely to be H').
■ For 1.1–2.6, we test for principal-subjects; For 3.1–5.4, we test for subordinate-subjects.

References

Casella, Alessandra, Evan Friedman and Manuel Perez Archila. 2025. "On the Fragility of Mediation: Theory and Experimental Evidence." *Journal of the European Economic Association* jvaf049.

Holmström, Bengt and Roger B Myerson. 1983. "Efficient and Durable Decision Rules with Incomplete Information." *Econometrica* 51(6):1799–1819.

Hörner, Johannes, Massimo Morelli and Francesco Squintani. 2015. "Mediation and Peace." *Review of Economic Studies* 82(4):1483–1501.

Kim, Jin Yeub. 2017. "Interim Third-Party Selection in Bargaining." *Games and Economic Behavior* 102:645–665.

Loewenstein, George, Ted O'Donoghue and Matthew Rabin. 2003. "Projection bias in predicting future utility." *the Quarterly Journal of economics* pp. 1209–1248.

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

Myerson, Roger B. 1982. "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems." *Journal of Mathematical Economics* 10:67–81.

Myerson, Roger B. 1983. "Mechanism Design by an Informed Principal." *Econometrica* 51(6):1767–1797.

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, M.A.: Harvard University Press.

Nishimura, Takeshi. 2022. "Informed Principal Problems in Bilateral Trading." *Journal of Economic Theory* 204:105498.