Biased Mediators in Conflict Resolution[☆]

Andrés Salamanca[1]

Ca'Foscari University of Venice, Italy

Abstract

One of the most important and disputed questions within the fields of international relations and conflict studies concerns the issue of mediator impartiality. Should mediators be biased—supportive of one but not both of the main disputants—or should mediators always be impartial? This paper contributes to this debate by studying the effectiveness of mediation with regard to the role of mediator bias in a game-theoretic model of cheap talk. This study shows that the institutional design of a mediation process is affected by two factors: the relative degree of conflict and the incentives to misrepresent private information. We find that a necessary (but not always sufficient) condition for the success of any mediation process is a sufficiently low likelihood of a misrepresentation problem. If in addition, the relative degree of conflict is low enough, mediation is effective and the institutional design of a mediation process is the same *regardless of the bias*. Otherwise, the design will be quite different depending on the direction of the bias.

Keywords: Biased mediation, conflict resolution, information provision, partial commitment.

JEL Classification: D63, D74, D82.

1. Introduction

Interstate disputes, internal conflicts, and civil cases are frequently subject to mediation—negotiation facilitated by a third-party. In some instances, mediation has proven to be a successful device for reaching negotiated agreements, such as the U.S. President Jimmy Carter mediation effort to end the Egyptian-Israeli conflict that culminated in the 1979 Israel-Egypt Peace Treaty (Cohen and Azar, 1981). In other cases, mediation has failed to produce successful outcomes, as the U.S. Secretary of State Alexander Haig unfruitful attempt to mediate Britain and Argentina during the Falklands/Malvinas dispute (Jones, 2013). Why do some mediation processes help to overcome the barriers to conflict resolution and others fail to do so? Although scholars have succeeded in identifying various mechanisms whereby mediators are effective to reach dispute settlements, the literature

[☆]I am deeply grateful to Frédéric Koessler for his helpful comments that helped me to improve the paper. I acknowledge valuable feedback and suggestions from audiences at the 11th UECE Lisbon Meetings in Game Theory and Applications, the 7th Game Theory, Economics and Mathematics Workshop (University of Southern Denmark), the NEAT Seminar (Ca'Foscari University of Venice and University of Padua), the Paris Game Theory Seminar, the Theory, Organization, and Markets Seminar (Paris School of Economics), the ECOBAS Seminar (University of Vigo), and the Economics Seminar series of the University of Salamanca.

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

[1]This version: July 18, 2022.

has not yet come to a general consensus about what makes mediation successful. One prominent debate has concerned the role of *mediator bias*. The present article contributes to this debate by comparing the effectiveness of mediation, with regard to the role of mediator bias, in a model of cheap talk.

Traditionally, it has often been said that one major factor determining the success of mediation is the mediator's perceived impartiality. Scholars who advocate this idea argue that mediator impartiality is crucial for disputants' confidence in the mediator, which is necessary for acceptability of the mediation process (Young, 1967, p. 81, Assefa, 1987, p. 22; Miall, 1992, p. 62; Hume, 1994). Some other analysts, however, have questioned the importance of mediator impartiality and have made a case in favour of biased mediation. Touval (1975), Touval and Zartman (1989), and Bercovitch et al. (1991) claim that mediators are often biased and impartiality is neither indispensable nor necessary for the success of mediation. Kydd (2003) argues that mediators have strong incentives to strategically manipulate information, so only biased mediators will be believed. Svensson (2009, 2015) claims that impartial mediators will face difficulties to provide credible security guarantees for a durable agreement. In contrast, biased mediators will make sure that the parties will conform to the provisions that favor the interests of "their" side.

One major reason why the role of mediator bias is still an issue of debate is the inability of the existing research to provide a systematic deductive theory of mediation. On the one hand, qualitative studies allow for detailed and in-depth examination of specific conflicts, but they do not offer reproducible findings consistent with other similar cases. On the other hand, much of the quantitative work, although rigorous, fails to produce conclusive empirical results. In the absence of formal deductive theories of mediation, it is difficult to demonstrate the role that mediation bias can play in mediation success. This observation has given rise to a game-theoretic analysis of mediation bias (Kydd, 2003; Smith and Stam, 2003; and Rauchhaus, 2006). This article follows this strand of the literature and considers a model where an *agent* possesses private information about a binary state of the world that is relevant for a *principal* in order to make a decision that affects the welfare of both parties. Individuals disagree about what everyone considers should be the final decision. Consequently, a conflict of interests arises. This, in turn, leads to a mistrust problem, because the agent may strategically manipulate her (unverifiable) private information, which gives the principal reasons to be skeptical about the veracity of the information disclosed by the agent. A mediator engages in behavior that is designed to elicit information from the agent and exercise influence on the principal by judiciously communicating garbled information. When doing so, the mediator helps the parties build trust and reach a compromise. The model considered in this paper was originally introduced by Mitusch and Strausz (2005) for the study of the efficiency of mediation when compared to unfacilitated negotiation.

According to Savun (2008), another important source of disagreement about the role of mediator bias is related to the wide variety of third-party activities that scholars consider to be mediation. Mediators in practice can assist negotiators by: incentivizing parties to share confidential information, probing positions, establishing an agenda, suggesting strategic and tactical approaches, helping the parties clarify their values, deflating unreasonable claims, mitigating commitment problems, providing reality checks, articulating a rational for agreement, etc.[2] Given the wide range of services that a mediator can

[2] Doyle and Haydock (1991, 88–92), Raiffa (1985, 108–109), and Singer (1990, p. 20) provide extensive lists of the tactics, techniques, and services provided by a mediator.

provide, it is possible that mediator bias affects the effectiveness of some *but not all* mediation activities. Therefore, it is necessary to restrain the analysis to a specific kind of mediation activity. We chose to focus on information provision, as it is the most commonly used and least costly mediation strategy (Bercovitch and Houston, 2000). In the model here studied, the mediator controls the flow of information between the disputants, which alleviates the conflict of interests by mitigating the commitment problems they face. Commitment problems are recognized to be one of the major strategic obstacles to reach negotiated settlements in conflicts (Fearon, 1995; Svensson, 2007). Two different commitment problems arise in the model: *Adverse selection*, caused by the agent's hidden information that limits her ability to commit to tell the truth and, therefore, generates a misrepresentation problem—one of the agent types would be tempted to impersonate the other.[3] The second commitment problem is *moral hazard*, caused by the principal's unalienable right to control his own actions, which makes it difficult for him to credibly commit to uphold a mutually beneficial deal.

A mediator in the model first holds a "caucus" with the agent—confidential, private meeting in which the mediator tries to incentivize the agent to disclose her private information honestly. Then the mediator selectively distorts and communicates part of this information to the principal in the form of a recommendation. Such a communication strategy is called *mediation plan*. It describes the institutional design of the mediation process. As shown by Mitusch and Strausz (2005) and Goltsman et al. (2009), adding noise may enhance communication between the parties. By committing to translate imprecisely the agent's information, the mediator may induce the agent to make truthful statements about her information. This is so because the imprecise translation reduces the principal's ability to use such an information to the agent's detriment. Similarly, given the agent's greatest honesty and the mediator's commitment, the principal has less incentives to renege on a settlement and will therefore more readily follow the mediator's recommendations.[4]

There exist infinitely many different mediation plans that a mediator can adopt. The outcome of any such plan may be more or less favorable to each disputant. We say that the mediator is biased in favor of a disputant if he is committed to a mediation plan that maximizes the *ex-ante* welfare of that disputant. The term "ex-ante" refers to the fact that the mediator evaluates both parties' welfare behind a veil of ignorance—*before* any private information is acquired. Mitusch and Strausz (2005) assume that the principal has all the bargaining ability—the principal has effective control over the communication channels. Hence, they only consider principal-biased mediation. In this article, we characterize mediation outcomes for agent-biased mediation. We show that the optimal mediation plan implemented by an agent-biased mediator depends on two factors: (*i*) the relative degree of conflict, and (*ii*) the likelihood of a misrepresentation problem. The relative degree of conflict measures the extent to which the preferences of disputants are misaligned. On the other hand, the likelihood of a misrepresentation problem refers to the probability that the agent finds it favorable to manipulate the settlement using her private information. The mediation success depends on both the parameter configuration and the direction of mediator bias as summarized in Table 1.1.

When comparing mediation outcomes with regard to the role of mediator bias, our re-

[3] The agent's information is represented by her *type*: We say that the agent's type is s when she knows that the actual state is s.

[4] Brown and Ayres (1994) also recognize the role of controlling the flow of information in mediation to mitigate the commitment problems.

		Relative degree of conflict: Low	Relative degree of conflict: High
Likelihood of misrepresentation: Low		Effective mediation regardless of bias	Effective principal-biased mediation Ineffective agent-biased mediation
Likelihood of misrepresentation: High		Ineffective mediation regardless of bias	

Table 1.1: Mediation success with regard to mediator bias

sults show that, for most parameter configurations, mediator bias is inconsequential—mediation yields the same outcome regardless of the bias. A necessary (but not always sufficient) condition for the success of *any* mediation process is a sufficiently low likelihood of misrepresentation: When the principal has a strong belief that the agent will use her private information to manipulate the agreement, the mediator is unable to build trust between the parties. Provided that the likelihood of misrepresentation is low enough, mediation is effective *regardless of the bias* only when the relative degree of conflict is sufficiently low. Otherwise, only a principal-biased mediator will be effective. That is, the direction of the bias is only relevant for mediation success whenever the relative extent of the conflict is large but a misrepresentation problem remains unlikely (so that mediation can overcome mistrust problems). This occurs because a strong conflict of interests makes any information disclosure detrimental for the agent. Thus, seeking to protect its protégé, an agent-biased mediator will make sure not to convey any information to the principal. The consequence is that agent-biased mediation hinders communication, even if reaching a settlement would have been preferred by any of the agent's types.

Our conclusions can be understood in the context of intrastate armed conflicts. The government takes over the role of the principal and the rebels assume the role of the agent. The rebels have private information about their power and resolve to fight. On the other hand, the government needs to implement reforms to convince the revels to lay down their arms. In this context, Svensson (2007) provides statistical evidence that biased mediators are effective, but that it is important to distinguish between rebel-biased and government-biased mediators since he finds that, unlike government-biased mediators, rebel-biased mediators have no significant effect. Our findings thus provide the formal conditions under which Svensson's conclusions are theoretically sound. We will elaborate on the relationship between our results and Svensson's thesis in Section 7.1.

Naturally, the principal always prefers a mediator biased towards himself. However, the same is not true for the agent, who, depending on her knowledge of the actual state, may prefer a principal-biased mediator instead. When only one of the two agent types is better-off keeping quiet (and therefore prefers an agent-biased mediator), the involvement of a principal-biased mediator would be detrimental for that type. If the acceptability of a principal-biased mediator depends in any way on the agent's type, the participation decision will convey information about her type to the principal. In other words, the agent's acceptance of a principal-biased mediator is a signal that serves to separate her various types. With this new information, the principal may find new opportunities to gain by disobeying the mediator. As a result, to conceal her information, the agent's participation decision must not depend on her information; but her information may influence what she prefers. To address this dilemma, we extend the model to include an earlier stage of voluntary mediation in which the agent decides whether to accept a principal-biased mediator. The principal forms new *posterior* beliefs to reflect any information conveyed by the agent's participation decision. In case of rejection, the principal is left to make a

decision given his posterior beliefs.

We show that there is no (separating) Perfect Bayesian Equilibrium (PBE) in which either type agrees to participate and the other rejects. That is, to remain inscrutable, both agent types must either accept or reject mediation. A (pooling) PBE always exists in which both agent types undergo principal-biased mediation (acceptance equilibrium). This equilibrium is justified by the (off-path) posterior beliefs that an unexpected rejection is always attributed to the jeopardizing type (i.e., the one generating the misrepresentation problem). Because the jeopardizing type always prefers an agent-biased mediator, this equilibrium prescribes a sensible outcome only when the other type (i.e., the jeopardiz*ed* type) favors principal-biased mediation. In such a situation, the beneficial effect of mediation occurs at expenses of the jeopardizing type. On the other hand, when both types prefer an agent-biased mediator, the off-path beliefs do not appear to be reasonable. In this case, we show that the acceptance equilibrium can be "destroyed" when we require credibility constraints on beliefs analogous to those imposed by the notions of core in Myerson (1983), neologism proofness in Farrell (1993), and perfect sequential equilibrium in Grossman and Perry (1986). Indeed, the agent can address the principal with a message of the form "I'm rejecting because none of my types are better-off mediating." Such a statement is credible, since both types benefit from uttering it when the principal believes it. The use by the agent of this speech breaks the equilibrium and causes both parties to coordinate on a pooling PBE where both agent types reject and the principal holds his prior beliefs both on and off the equilibrium path (rejection equilibrium).

When the relative degree of conflict is too large, principal-biased mediation becomes harmful for the agent *regardless of her private information*. Therefore, our analysis of the voluntary mediation suggests that principal-biased mediation will be rejected.[5] Therefore, the agent prefers what Touval and Zartman (1985) call a "hurting stalemate," a situation in which parties find themselves locked in a conflict where unilateral success is unattainable. In contrast, when the relative conflict of interests is low enough, the actual preferences of the agent are in conflict with her need to be inscrutable. That is, both agent types contradict each other regarding their desire to mediate. The agent resolves this tension accepting an unfavorably aligned mediator. The acceptability of the mediator only depends on how large the relative degree of the conflict is. The agent does not reject a principal-biased mediator because of his partiality, but because mediation efforts cannot provide an acceptable outcome for either of the agent types. This result provides support to Touval and Zartman's (1985) hypothesis that a biased mediator may be acceptable if the disputants perceive that the mediator can provide contributions that each party wants.

The rest of the paper is organized as follows. In Section 2 we frame our contributions in light of the relevant literature. Section 3 introduces the basic setup and formally defines a mediation plan. Section 4 characterizes the optimal mediation plans of a biased mediator. In Section 5 we compare the effectiveness of mediation with regard to the role of mediator bias. In Section 6 we analyse the issue of mediator acceptability when the effectiveness of mediation depends on the mediator bias. We conclude the paper in Section 7 with some final discussions.

[5] Because the principal is uninformed, he will always (weakly) prefer mediation regardless of the bias.

2. Related Literature

This paper contributes to the literature on alternative dispute resolution in Law and Economics and on international relations and conflict studies in Political Sciences. It provides a formal deductive theory to understand the role mediator bias plays for the effectiveness of mediation. This is not the first study addressing this question from a formal theoretical perspective. In a seminal work, drawing on Fearon's (1995) argument about the role of private information, Kydd (2003) examines the effect of biased versus impartial mediation in a model of armed conflict with asymmetric information about resolve to fight. Kydd shows that, because it is in the interest of an impartial mediator to achieve a peaceful settlement, it has a strong incentive to lie about the primary parties' resolve. Accordingly, only biased mediators will be effective. Kydd's results are thus not entirely similar to ours. On the one hand, our findings imply that, whenever the mediator bias is inconsequential, *any* (optimal) mediator is as effective as a biased mediator. On the other hand, when mediator bias matters, it is important to distinguish between principal-biased and agent-biased mediators since we found that, unlike principal-biased mediators, agent-biased mediators are ineffective.

This discrepancy of results is a consequence of differing formal approximations of the mediation process. In Kydd's model, a mediator is strategic in the sense that it has preferences over the outcome of the issue in dispute. In particular, it always prefers an issue solution rather than the permanence of the conflict. Additionally, the mediator cannot commit to translate information according to some mediation plan (i.e., communication strategy). That is, the mediator has no intrinsic commitment to honesty and, therefore, it may strategically manipulate its private information. For this reason, the mediator may have incentives to make statements that will induce one of the parties to make a compromise, even when those statements are not truthful. Under such circumstances, a mediator can only be credible when its preferences are fully aligned with either party. In contrast, in our study, a mediator is always indifferent over all outcomes of mediation *even* when biased toward either party.[6] Moreover, a mediator commits to transmit information according to a (commonly known) mediation plan. These two features of our model imply that a mediator is trustworthy.

Kydd's formulation of the mediator preferences conflates two closely related but distinct issues. The first issue is the mediator's ability to create value by controlling the flow of information. That is, the capacity of the mediator to enhance communication. The second issue relates to the mediator credibility, which constitutes a subsidiary adverse selection problem between the primary parties and the mediator. Kydd (2003) avoids the double adverse selection problem by treating the mediator's information as exogenously determined. That is, the mediator is (reasonably well) informed about the parties resolve without the need to communicate with them. A priori, there is no reason for a third party to be better informed about the disputants than what the disputants themselves know of each other, unless the third party has already communicate with them. For instance, we cannot think that the US Secretary of State Haig was already convinced that Britain had a high resolve and would fight *before* caucusing with the British commanders. Therefore, we believe treating the mediator's information as exogenous does not seem to be a good formulation. Instead, what it does appear to be reasonable is that Haig was convinced by

[6] Ivanov (2010) refers to this property of the mediator's preferences as mediator *neutrality*. To avoid any confusion, we use the terms *impartiality* or *unbiasedness* to designate the tendency of the mediator not to side in a conflict. For a discussion about impartiality the reader is referred to Section 7.2.

Thatcher's administration that they had a high resolve. In this way, the mediator's information is endogenously acquired in meetings with individual disputants. In our model, the rationale for mediation centers on caucusing because it is here that the mediator most clearly controls the flow of information between the disputants.

Kydd's (2003) model features—an informed strategic mediator—are also present in two related papers by Smith and Stam (2003) and Rauchhaus (2006). It follows from a formal model employed by Rauchhaus (2006) that both biased and impartial mediators are effective but impartial mediators outperform biased ones. Although Rauchhaus and Kydd's models are very similar, their conclusions are opposite. Kydd (2006) explains that this discrepancy is due to differing definitions of bias. While Kydd (2003) defines a mediator as biased when it shares one side's preference ordering over the issue space, Rauchhaus sees a mediator as biased if it has an ideal point on the issue space outside the range between the bliss points of the primary parties. On the other hand, Smith and Stam (2003) detach from Fearon's (1995) argument about private information and misrepresentation problems, and propose a stochastic model of conflict consisting of a series of battles over a set of territorial units. The authors assume that each disputant has different beliefs about the probability of victory in the next battle. However, each disputant thinks the other's beliefs are wrong, hence making beliefs not consistent with a common prior. In contrast to the previous cited studies, Smith and Stam (2003) found that biased mediators are not effective.

From a more methodological point of view, our paper also relates to the literature on optimal mediation in sender-receiver games (e.g., Mitusch and Strausz, 2005; Blume et al., 2007; Goltsman et al., 2009; Ganguly and Ray, 2009; Ivanov, 2010, 2014; and Salamanca, 2021). This literature studies mediation as a mechanism for enhancing communication. Therefore, it focuses on mediation that is beneficial to the uninformed party (i.e., the receiver/principal).[7] Very often this literature builds on Crawford and Sobel's (1982) model of information transmission, and more particularly on its uniform-quadratic case (i.e., quadratic preferences and uniform type distribution). A well known result in this model states that a mediation plan maximizes the sender's (ex-ante) welfare if and only if it also maximizes the receiver's welfare. Therefore, the restriction to receiver-biased mediation is usually inconsequential. One important assumption behind this finding requires that the degree of the conflict be independent of the sender's private information. In contrast, in our framework, the conflict of interest varies across states. However, when the *relative* conflict of interests in our model is close to 1, the *absolute* conflict is nearly constant across states. In this case, mediation is effective regardless of the mediator bias (provided that the likelihood of misrepresentation is low). A prominent contribution in this literature has been provided by Mitusch and Strausz (2005). As we have previously mentioned, we elaborate on their model and borrow some of their results. In a more recent contribution Salamanca (2021) proposed a methodology to characterize optimal mediation plans that maximize the (ex-ante) welfare of the sender (i.e., the agent). We exploit his approach to compute mediation outcomes for agent-biased mediation in Mitusch and Strausz's model.

Alternative settings have also been studied by the literature on mediation. Fey and Ramsay (2010), Hörner et al. (2015), and Meirowitz et al. (2019) analyze mediation in one-shot conflict games. In this setting, disputants contest the distribution of a "pie" (e.g., a territory). A mediator collects information privately from disputants

[7]The reader is referred to Section 7.3 for discussion about mutually beneficial mediation.

(e.g., their political resolve or military strength) and recommends a split of the pie. If the proposed division is rejected, parties go to war, which reduces the value of the pie. Jarque et al. (2003) and Čopič and Ponsatí (2008) investigate mediation in which intermediate concessions are not observable by the primary parties: disputants send private messages to a mediator, whose only active roles is to make an agreement public as soon as it is reached. Fanning (2021a,b) explore the effectiveness of mediation in a dynamic reputational bargaining models. Finally, Gottardi and Mezzetti (2022) study the value of shuttle diplomacy in a model where a mediator gradually provides "reality checks" about the parties' costs and benefits from a settlement.

3. The Model

We consider a situation of conflict between a privately informed *agent* and an uninformed decision-maker called the *principal*. The principal must implement an action $y \in \mathbb{R}$ affecting the welfare of both individuals.[8] There are two possible states of nature $s = 1, 2$, one of which is randomly chosen according to a probability distribution; the probability of state $s = 2$ being denoted $\pi \in (0, 1)$. The probability π is commonly known by both individuals, but only the agent is informed about the chosen state (her *type*). We refer to π as the principal's prior belief. When the actual state is $s = 1, 2$, the principal's preferences are represented by the utility function $V_s(y) = -(y - y_p^s)^2$, where $y_p^s \in \mathbb{R}$ denotes his most preferred action in state s (*bliss point*). Similarly, we write the agent's utility function as $U_s(y) = -(y - y_a^s)^2$, with y_a^s her bliss point in state s.

There is a conflict of interests due to the incompatibility between the preferences of both parties—while the agent prefers the final decision to be y_a^s (in state s), the principal would like to choose y_p^s. Therefore, the difference $y_p^s - y_a^s$ measures the *absolute* extent of the conflict in state s. Provided that the principal benefits from learning the actual state (i.e., $y_p^1 \neq y_p^2$), the agent will try to misrepresent such information to influence the principal's decision. We adopt the following *monotonicity condition*: $\Delta_a := y_a^2 - y_a^1 > 0$, that is, the agent's bliss points are increasing in the state. This assumption is sometimes referred to as the *single-crossing property* and will serve to ensure that the agent wants to misrepresent her information only in one state but not in both. Otherwise, the misrepresentation problem may become so significant that no credible information transmission may occur. We also assume the same ordering for the principal's bliss points, that is, $\Delta_p := y_p^2 - y_p^1 > 0$. This will guarantee that the preferences of both individuals are *minimally aligned*, so that the conflict of interests is not too prominent for precluding any possibility of communication.[9]

The model possesses five parameters: $\{y_i^s\}_{s=1,2;i=a,b}$ and π. However, a more parsimonious description of our results can be given in terms of only two statistics: π and $\sigma := \frac{\Delta_p}{\Delta_a} > 0$. The coefficient σ is a two-fold measure of the *relative* intensity of the conflict. First, it shows to what extent the disputants' preferences are misaligned in state 2 relative to state 1. Note that $\sigma \geq 1$ iff $y_p^1 - y_a^1 \leq y_p^2 - y_a^2$. In other words, $\sigma \geq 1$ whenever the conflict of interests is more severe in state 2 as compared to state 1. Moreover, the degree of conflict (in state 2 relative to state 1) increases with the value of σ. The relative degree of conflict might be low even when the absolute degree in both states is very large. As we will show below, only the relative, and not the absolute, degree of conflict matters. Second, the

[8] We refer to the agent as female and the principal as male.

[9] Given $y_a^2 > y_a^1$, the principal's monotonicity assumption is a necessary condition for communication to occur. A formal reasoning for this statement is given in footnote 14.

coefficient σ also provides an ex-ante measure of the principal's ability to use the agent's information to her detriment. Notice that the principal chooses action y_p^s when he learns that the actual state is $s = 1, 2$. Therefore, from an ex-ante perspective (i.e., before any information is acquired), learning the state when $\Delta_p \gg \Delta_a$ leads to great dispersion in actions as compared to the agent's bliss points. Consequently, the agent's risk aversion implies that her (ex-ante) welfare will greatly deteriorate.

The principal's decision will only depend on his beliefs. For an arbitrary belief $\rho \in [0, 1]$ about state $s = 2$, the principal chooses an action y to maximize his expected utility:

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

Given the strict concavity of the utility function, the principal implements the optimal action

$$y(\rho) := (1 - \rho) y_p^1 + \rho y_p^2. \tag{3.1}$$

The principal's optimal action lies between y_p^1 and y_p^2. That is, the effective issue space is the interval $[y_p^1, y_p^2]$. Moreover, the condition $\Delta_p > 0$ implies that $y(\cdot)$ is monotonically increasing.

3.1. Mediation

We consider mediation protocols in which a trustworthy mediator gathers non-verifiable caucus reports from the agent and makes non-binding recommendations to the principal. Formally, a *mediation plan* specifies: a (countable) set $Y \subset \mathbb{R}$ of recommendations;[10] and, a family of probability distributions $\delta = (\delta^s)_{s=1,2}$, with $\delta^s \in \Delta(Y)$ for every $s = 1, 2$.[11] Generic elements of Y are denoted y_j, where $j \in \mathbb{N}$. We denote δ_j^s the conditional probability that the mediator recommends the action $y_j \in Y$ when the agent's report is $s = 1, 2$.

When the parties communicate through a mediator implementing the mediation plan δ, one obtains a mediated game, denoted $\Gamma_\delta(\pi)$, which is played as follows:

Stage 1 (mediation phase). The agent transmits a confidential report s' to the mediator, which may differ from the actual state s. Then the mediator recommends action $y_j \in Y$ to the principal with probability $\delta_j^{s'}$.

Stage 2 (action phase). The principal updates his beliefs and chooses an action, which may not coincide with the mediator's recommendation.

In the previous mediated game, neither the mediator nor the principal can verify the actual state, which allows the agent to strategically manipulate her private information. On the other hand, the mediator's recommendation is not binding; the principal is free to choose any action different from the recommended one. We shall focus on a particular equilibrium of the mediated game in which the agent has no incentives to lie about the state, and the principal has no incentives to disobey the mediator's recommendation.

[10] Because the number of states is finite, the restriction to a countable set of recommendations is without loss of generality. Indeed, as it will be clear later on, no more than two different recommendations are required in an optimal mediation plan.

[11] With a slight abuse of terminology, we simply refer to a mediation plan by its transition probabilities δ, with not explicit mention of the underlying recommendations set.

A mediation plan δ is called *incentive-compatible for the agent* if and only if the sincere reporting is a best-response for the agent when the principal is obedient, that is,

$$\sum_{y_j \in Y} \delta_j^s U_s(y_j) \geq \sum_{y_j \in Y} \delta_j^{s'} U_s(y_j) \quad \forall s, s' = 1, 2. \tag{3.2}$$

The inequalities in (3.2) are called *truth-telling incentive constraints*.

Suppose action y_j is recommended to the principal according to the mediation plan δ, provided that the agent is sincere. Then, using the Bayes rule, the principal computes the following posterior belief about state 2:

$$\pi_j(\delta) := \frac{\delta_j^2 \pi}{\delta_j^2 \pi + \delta_j^1 (1 - \pi)}.$$

The mediation plan δ is said to be *incentive-compatible for the principal* if and only if following the recommendation is a best-response for the principal whenever the agent is sincere. Formally,

$$y_j = y\left(\pi_j(\delta)\right), \quad \text{for all } j \text{ such that } \delta_j^s > 0 \text{ for some } s. \tag{3.3}$$

We refer to (3.3) as the *obedience incentive constraints*.

A mediation plan δ is called *incentive-compatible* if and only if it satisfies (3.2)-(3.3), namely, the sincere and obedient strategies form a Nash equilibrium of the mediated game $\Gamma_\delta(\pi)$.

In general, there can be many different Nash equilibria of a mediated game $\Gamma_\delta(\pi)$, even when δ is incentive compatible. However, given any such equilibrium, there always exists a payoff-equivalent incentive-compatible mediation plan. In this sense, there is no loss of generality in assuming that both individuals communicate through a mediator who induces the sender to report the state truthfully, and the principal to follow the prescribed recommendation. This result is known as the Revelation Principle for Bayesian games (see Myerson, 1982; Forges, 1986).

4. Biased Mediators

The mediator in the game $\Gamma_\delta(\pi)$ does not have preferences over the issue in dispute.[12] His participation is limited only to facilitate the communication between the agent and the principal. However, this does not mean that the mediator is impartial, as the mediation plan δ may have been chosen to specifically favor one of the two parties. In this section we shall characterize the mediation plans that an *uninformed* mediator would choose when serving the particular interests of either the principal or the agent. Specifically, we characterize optimal mediation plans that maximize the *ex-ante* welfare of each party. The case of principal-biased mediation plans has been fully analyzed by Mitusch and Strausz (2005). Therefore, we shall focus on the case of agent-biased mediation.

4.1. Agent-Biased Mediation

Let δ be an incentive-compatible mediation plan. If mediation is carried out according to δ, then the agent's expected utility is

$$U(\delta; \pi) := \sum_{y_j \in Y} \left[(1 - \pi) \delta_j^1 U_1(y_j) + \pi \delta_j^2 U_2(y_j) \right] \tag{4.1}$$

[12] Alternatively, one may also argue that the mediator is indifferent over all outcomes of mediation.

An agent-biased mediator chooses a mediation plan to maximize the agent's expected utility among all incentive-compatible mediation plans. Formally, it solves the optimization problem[13]

$$\max_{Y,\delta} U(\delta;\pi) \quad \text{s.t. (3.2) – (3.3)} \tag{4.2}$$

A mediation plan $\delta^* := \delta^*(\pi)$ solving the problem (4.2) is called an *optimal agent-biased mediation plan*. The corresponding expected utility $U(\delta^*;\pi)$ is the *value of agent-biased mediation at* π and will be denoted $U^*(\pi)$.

4.1.1. "Omniscient" Mediation

In order to characterize an optimal agent-biased mediation plan, it is instructive to first suppose that mediation plans are implemented by an "omniscient mediator" that directly observes the actual state of the nature. Such a mediator sends recommendations to the principal conditional on the *true* state, regardless of the agent's report. Consequently, omniscient mediation is only characterized by the obedience incentive constraints in (3.3). We thus consider the following relaxed optimization problem:

$$\max_{Y,\delta} U(\delta;\pi) \quad \text{s.t. (3.3)} \tag{4.3}$$

The previous omniscient mediation problem is known to be a *Bayesian persuasion problem* (see Kamenica and Gentzkow, 2011). The expected utility to the agent from a mediation plan solving (4.3), called *value of persuasion at* π, can be derived from the concave envelop of the agent's indirect utility function over beliefs. Formally, let $\widehat{U}(\rho)$ denote the utility that the agent expects when the principal makes a decision while holding the beliefs ρ:

$$\widehat{U}(\rho) = (1-\rho)U_1(y(\rho)) + \rho U_2(y(\rho)). \tag{4.4}$$

Hence, $\widehat{U}(\pi)$ is the value to the agent when she abstains from mediating. In this case, the agent simply remains silent and let the principal choose an action given his prior beliefs π. The value of persuasion at π is given by $cav\,\widehat{U}(\pi)$, where $cav\,\widehat{U}$ denotes the *concavification* of $\widehat{U}$, namely, the pointwise lowest concave function that is everywhere larger than or equal to $\widehat{U}$.

Because an omniscient mediator can always implement an incentive-compatible mediation plan, it must be clear that

$$\widehat{U}(\pi) \leq U^*(\pi) \leq cav\,\widehat{U}(\pi), \quad \forall \pi. \tag{4.5}$$

Lemma 1.
The indirect utility function $\widehat{U}$ is either concave or convex. Moreover, it is strictly convex if and only if the realtive degree of conflict satisfies $\sigma < 2$.

Proof. Twice-differentiate $\widehat{U}$ with respect to ρ to obtain:

$$\widehat{U}''(\rho) = 2\Delta_p\left[2\Delta_a - \Delta_p\right].$$

[13]The maximization in (4.2) is performed not only over all mediation plans for a given set Y, but also over all possible (countable) sets of recommendations $Y \subset \mathbb{R}$.

Notice that $\widehat{U}''$ is a constant (i.e., independent of ρ). Thus, $\widehat{U}$ is either concave or convex. Because $\Delta_p > 0$, $\widehat{U}$ is strictly convex if and only if $2\Delta_a - \Delta_p > 0$.[14] □

Whenever $\sigma \geq 2$, the indirect utility function, $\widehat{U}$, is concave, and therefore, we have that $\widehat{U} = cav\,\widehat{U}$. This means that releasing private information is detrimental for the agent when the relative conflict of interests is too prominent. Hence, the best the mediator can do for the agent is to hinder communication. Of course, in view of (4.5), this is *a fortiori* also true when the mediator is *not* omniscient. Therefore, $\sigma < 2$ is a necessary condition for mediation to help alleviate the conflict.

Because the omniscient mediator sends recommendations based on the actual state, the agent cannot strategically manipulate her private information to influence the outcome of the mediation plan. Therefore, the analysis of the omniscient mediation problem (4.3) reveals that the conflict of interests may be severe even when the agent cannot misrepresent her information.

4.1.2. The Misrepresentation Problem

Lemma 1 indicates that mediation can only be helpful provided that the conflict of interests does not aggravate too much in state 2 relative to state 1. Thus, we shall assume in the following that condition $\sigma < 2$ holds. Therefore, the indirect utility function, $\widehat{U}$, is strictly convex and its concavification, $cav\,\widehat{U}$, is a straight line as depicted in Figure 1.

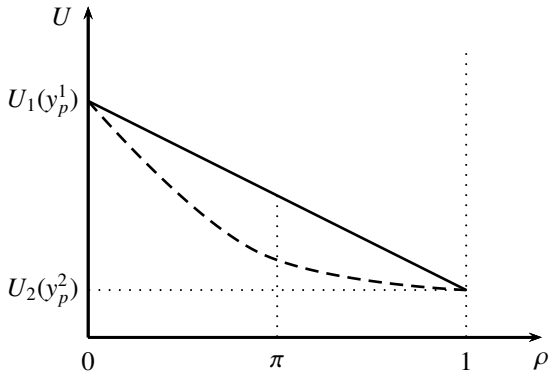

Figure 1: Indirect utility $\widehat{U}$ (black dashed line) and its concavification $cav\,\widehat{U}$ (black solid line)

From Figure 1 it is deduced that the value of persuasion at π, $cav\,\widehat{U}(\pi)$, can be achieved inducing posterior beliefs $\rho_1 = 0$ and $\rho_2 = 1$. The unique mediation plan that satisfies the obedience incentive constraints and generates this set of posterior beliefs is the *fully-revealing mediation plan*: In state s the mediator recommends action y_p^s, the principal correctly infers the actual state, and finds it optimal to follow the prescribed recommendation.

The value of persuasion can be seen as a first-best benchmark for an agent-biased mediator, since it constitutes the ideal outcome in the absence of the informational distortions that the agent may introduce by manipulating her private information. The question then arises as to whether the value of persuasion can be attained when the mediator is not omniscient. We say that there is no *misrepresentation problem* if the fully-revealing mediation

[14] Since $\Delta_a > 0$, we have that $\Delta_p \leq 0$ would imply that $2\Delta_a - \Delta_p > 0$, hence making $\widehat{U}$ concave. Therefore, $\Delta_p > 0$ is a necessary condition for $\widehat{U}$ to be convex.

plan satisfies the truth-telling incentive constraints in (3.2), namely,

$$U_s(y_p^s) \geq U_s(y_p^{s'}), \quad s, s' = 1, 2. \tag{4.6}$$

The following result readily follows:

Proposition 1.
Suppose $\sigma < 2$ holds. For every prior π, the fully-revealing mediation plan is agent-biased optimal if and only if there is no misrepresentation problem.

A type s for which $U_s(y_p^s) < U_s(y_p^{s'})$ $(s \neq s')$ imposes an efficiency loss that prevents the mediation process from achieving the value of persuasion. In this case, we expect the principal and the mediator to have difficulty preventing type s from claiming that the actual state is s'. Consequently, the informational advantage of the agent generates a misrepresentation problem that exacerbates the conflict of interests. This observation motivates the following definition: we say that type s *jeopardizes* type s' if $U_s(y_p^s) < U_s(y_p^{s'})$.[15]

When neither type causes a misrepresentation problem, both parties can dispense with the mediator and still achieve the value of persuasion through a non-facilitated negotiation. Indeed, both parties can communicate directly in the following way: The agent sends a message of the form "the state is s", the principal hears the message, updates his beliefs (using the Bayes rule), and optimally chooses an action. It is important to notice that the agent has no possibility to commit herself to tell the truth. In other words, the agent may argue that the state is s when it is actually s'. However, the condition (4.6) guarantees the existence of a truthful Nash equilibrium. Suppose the principal expects the agent to tell the truth. Then after the message "the state is s" is sent, the principal forms posterior beliefs $\pi_s = s - 1$ and chooses action y_p^s. The agent then obtains a utility $U_s(y_p^s)$. If the agent deviates in state s and announces that the state is $s'(\neq s)$, then she obtains a utility $U_s(y_p^{s'})$. Therefore, the deviation is not profitable and the agent has no incentives to misrepresent the actual state. Since mediation is no more helpful than face-to-face negotiation when there is no misrepresentation problem, we assume hereinafter that there exists a jeopardized type.

Lemma 2.
There is at most one jeopardized type.

Proof. Suppose, for instance, that type 1 jeopardizes type 2. Then, the monotonicity assumption implies that

$$0 < U_1(y_p^2) - U_1(y_p^1) = \int_{y_p^1}^{y_p^2} U_1'(y)dy < \int_{y_p^1}^{y_p^2} U_2'(y)dy = U_2(y_p^2) - U_2(y_p^1).$$

□

According to Lemma 2, the misrepresentation problem can only be caused by one of the types, but not both. As already anticipated, the key condition to obtain this result is

[15] The term "jeopardizing type" is used by Myerson (1991, p. 498) to indicate that such a type imposes a signaling cost that prevents the agent from achieving her first-best. In contrast, Mitusch and Strausz (2005) used the term "incompatible type" indicating that the interests of such a type are incompatible with those of the principal, which prevents the principal from achieving his first-best. We follow Myerson's terminology as it better reflects the misrepresentation problem when mediation is biased toward the agent.

the agent's monotonicity assumption. In the remainder of this article, we shall assume, without loss of generality, that type 2 jeopardizes type 1.[16] Because only the jeopardizing type leads to an incentive problem, the prior probability π measures the *likelihood of the misrepresentation problem*.

We shall now construct a candidate incentive-compatible mediation plan, $\hat{\delta}$, solving the mediator's problem in (4.2) when $\sigma < 2$ and type 2 jeopardizes type 1. Consider a configuration of bliss points that satisfies our assumptions, as illustrated in Figure 2. The fully-revealing mediation plan gives incentives to type 2 to misrepresent type 1. The following question arises: How to increase the incentives for type 2 to tell the truth while leaving type 1's incentives unchanged? This objective can be achieved if instead of y_p^2, the mediator recommends an action sufficiently close to y_a^2. In particular, action $\hat{y}$ as illustrated in Figure 2 leaves type 2 exactly indifferent.

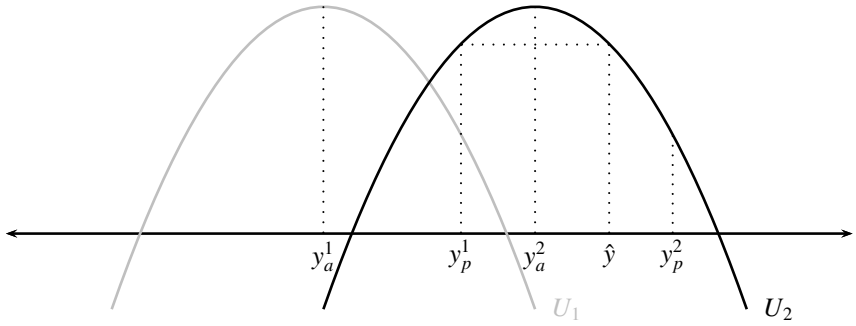

Figure 2: Agent's utility functions when type 2 jeopardizes type 1

With this idea in mind, we thus define the belief $\hat{\pi}$ as follows:

$$\hat{\pi} := \max_{\rho \in [0,1]} \left\{ \rho \mid U_2(y(\rho)) = U_2(y_p^1) \right\}.$$

Note that $\hat{\pi} > 0$ iff $y_p^1 < y_a^2$. By construction, $\hat{y} = y(\hat{\pi})$. Consider now a mediation plan inducing the posterior beliefs $\rho_1 = 0$ and $\rho_2 = \hat{\pi}$. The Bayes rule implies that such a mediation plan can only be constructed provided that $\pi < \hat{\pi}$, since the expected posterior belief must equal the prior belief (*Bayes plausibility*). The unique mediation plan that satisfies the obedience incentive constraints and induces the above distribution of posteriors is

$$\begin{aligned} \hat{\delta}^1_{y_p^1} = \hat{\theta} := \frac{\hat{\pi} - \pi}{\hat{\pi}(1-\pi)}, \quad \hat{\delta}^1_{\hat{y}} = 1 - \hat{\theta}, \\ \hat{\delta}^2_{y_p^1} = 0, \quad \hat{\delta}^2_{\hat{y}} = 1. \end{aligned} \tag{4.7}$$

After receiving the recommendation to choose y_p^1 the principal infers that the actual state is $s = 1$, thus it is optimal for him to follow the recommendation. Upon receiving the recommendation to choose $\hat{y}$, the principal deduces that the current state is $s = 2$ with probability $\hat{\pi}$. Thus, it is optimal for him to follow the recommendation. On the other hand, by construction, type 2 of the agent is indifferent between y_p^1 and $\hat{y}$, while type

[16] If instead it is type 1 that jeopardizes type 2, we can mirror our problem by redefining the actions $y' = -y$ and exchanging the roles of both types.

1 strictly prefers y_p^1. Hence, the truth-telling incentive constraint of type 2 (resp. 1) is binding (resp. slack) and the mediation plan $\hat{\delta}$ is incentive-compatible.

Theorem 1.
Suppose $\sigma < 2$ holds. Assume type 2 jeopardizes type 1.[17] *Then*

- *$\pi < \hat{\pi}$ if and only if $\hat{\delta}$ (as defined in (4.7)) is agent-biased optimal. Moreover, the value of agent-biased mediation is*

$$U^*(\pi) = \frac{\hat{\pi} - \pi}{\hat{\pi}} \widehat{U}(0) + \frac{\pi}{\hat{\pi}} \widehat{U}(\hat{\pi}).$$

- *If $\pi \geq \hat{\pi}$ mediation cannot facilitate communication between the parties.*

Theorem 1 reveals two important features of agent-biased mediation. First, whenever the likelihood of the misrepresentation problem is too high (i.e., $\pi \geq \hat{\pi}$), the mediator is unable to build trust between the parties and no credible information transmission is possible. Second, optimal mediation reveals type 1 with probability $\hat{\theta} < 1$. If both types were perfectly revealed by the mediation plan, type 2 would have incentives to misrepresent the current state. The principal would then anticipate this misleading behavior and the communication would cease to be informative. Therefore, although type 1 has no problem to reveal itself, it must provide cover for type 2.

Under non-facilitated negotiation, there also exists a Nash equilibrium exhibiting underrevelation of type 1 (see Proposition 3 in Mitusch and Strausz, 2005). Formally, let $\tilde{\pi} := \max_{\rho \in [0,1]} \left\{ \rho \mid U_1(y(\rho)) = U_1(y_p^1) \right\}$. Define $\tilde{\theta} := \frac{\tilde{\pi} - \pi}{\tilde{\pi}(1-\pi)}$. Consider the following signalling strategy for the agent. In state 1, the agent sends the message "the state is $s = 1$ (resp. $s = 2$)" with probability $\tilde{\theta}$ (resp. $1 - \tilde{\theta}$). In state 2, the agent truthfully communicates the actual state. Upon receiving the message that the state is $s = 1$, the principal correctly infers the state and chooses y_p^1. After receiving the message that the state is $s = 2$, the principal deduces that the current state is $s = 2$ with probability $\tilde{\pi}$ and optimally chooses $\tilde{y} := y(\tilde{\pi})$. Type 2 has no incentives to announce that the state is $s = 1$. Similarly, by construction of $\tilde{\pi}$, type 1 is indifferent between the two messages.

The unfacilitated equilibrium exhibits the same information structure as the agent-biased mediation plan. In both equilibria the jeopardized type is underrevealed. However, because $\tilde{\pi} \leq \hat{\pi}$, the degree of underrevelation is less when both parties use a mediator (i.e., $\tilde{\theta} \leq \hat{\theta}$). The mediator thus is able to provide cover more efficiently, which enhances communication.

Theorem 1 also shows how the value of agent-biased mediation depends on the likelihood of a misrepresentation problem. Figure 3 depicts and compares the value of agent-biased mediation, U^*, the value of persuasion, $cav\, \widehat{U}$, and the value in the absence mediation, $\widehat{U}$.

4.2. Principal-Biased Mediation

Now we turn our attention to mediation plans that serves the interests of the principal. For any given incentive-compatible mediation plan δ, the expected utility of the principal is:

$$V(\delta; \pi) := \sum_{y_j \in Y} \left[(1 - \pi) \delta_j^1 V_1(y_j) + \pi \delta_j^2 V_2(y_j) \right]$$

[17] The proof of Theorem 1 is relegated to the Appendix. It builds on the methodology developed by Salamanca (2021). This methodology combines two different strands of literature: On the one hand, the concavification approach from Bayesian persuasion (see Kamenica and Gentzkow 2011) and, on the other hand, Myerson's (1983) virtual utility approach to the informed principal problem.

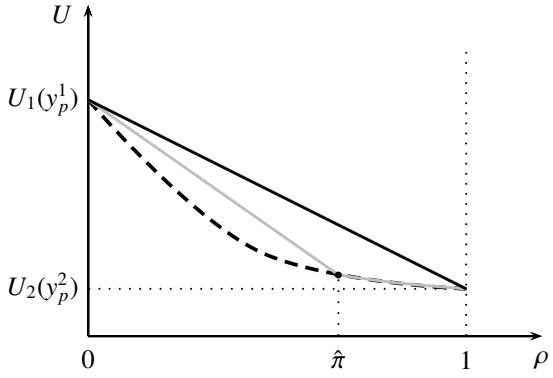

Figure 3: Functions U^* (gray solid line), $\widehat{U}$ (black dashed line), and $cav\,\widehat{U}$ (black solid line)

A principal-biased mediator chooses a mediation plan to maximize the principal's expected utility among all incentive-compatible mediation plans. Formally, it solves

$$\max_{Y,\delta} V(\delta;\pi)$$
$$\text{s.t. } (3.2) - (3.3)$$

This problem was extensively analyzed by Mitusch and Strausz (2005). We summarize their main findings in the following theorem.

Theorem 2 (Mitusch and Strausz, 2005)**.**
(i) If there is no misrepresentation problem the fully-revealing mediation plan is principal-biased optimal.

(ii) Assume type 2 jeopardizes type 1.

- *If $\pi < \hat{\pi}$, the mediation plan $\hat{\delta}$ (as defined in (4.7)) is principal-biased optimal.*
- *If $\pi \geq \hat{\pi}$ mediation cannot facilitate communication between the parties.*

5. Effectiveness of the Biased Mediators

The following table summarizes the results of Theorems 1 and 2. It compares the outcomes of mediation depending on the direction of the bias for the various possible parameter constellations.

	$\pi < \hat{\pi}$		$\pi \geq \hat{\pi}$
	$\sigma < 2$	$\sigma \geq 2$	
Agent	$\hat{\delta}$	Uninformative	Mediation cannot build trust
Principal	$\hat{\delta}$	$\hat{\delta}$	

Table 5.1: Biased mediators

When the principal strongly believes that the agent will use her private information to manipulate the outcome of the mediation plan (i.e., $\pi \geq \hat{\pi}$), the mediator is unable to build trust between the parties. In this case, no credible communication can occur. Therefore, a necessary (but not always sufficient) condition for mediation to be strictly beneficial is a sufficiently low likelihood of a misrepresentation problem (i.e., $\pi < \hat{\pi}$). Provided that the relative degree of conflict is also low enough (i.e., $\sigma < 2$), mediation is effective *regardless of the mediator bias*. Otherwise, only principal-biased mediation will be effective. When

the relative degree of conflict is too large (i.e., $\sigma \geq 2$), mediation hinders communication when the mediator attends the agent's interests. From the mediator's perspective, given the strong conflict of interests, any information disclosure is detrimental for the agent in *at least* one of the two states, and such a loss cannot be offset by a welfare improvement in the other state. Consequently, seeking to protect its protégé, an agent-biased mediator will hinder communication even if some information transmission would have been preferred by some agent type.

6. Deciding Whether to Mediate

The possibility that the direction of the bias affects the outcome of the mediation underlines the significance of *self-determination* in mediation. Self-determination is the act of coming to a voluntary, uncoerced decision in which each party makes free choices as to mediator selection, process design, participation in or withdrawal from the process. Note that the agent can always achieve the outcome of optimal agent-biased mediation without the assistance of a mediator: all she has to do is keep quiet. In this way, when invited to mediate in a process that is biased toward the principal, the agent may exercise self-determination either by refraining from making any statement during the process or by refusing to participate in it. In case the agent undergoes mediation, and if (as seems reasonable) she expects the principal to conform to the mediator's recommendation, then she does not have the incentives to hide information. This is so because an optimal principal-biased mediation plan is incentive compatible. Therefore, the agent does not do well in participating and remaining silent. Instead, the agent may decline mediation in order to let the principal choosing according to his prior beliefs. However, if the acceptability of the mediator depends in any way on the agent's type, refusing to participate can be interpreted by the principal as evidence about the actual state. In other words, the agent's participation decision is a signal that serves to separate her various types. With this new information, the principal may find new opportunities to gain by disobeying the mediator if mediation occurs, or may take advantage of it when mediation is refused. The agent then faces the following dilemma: to conceal her information, the agent's participation decision cannot depend on her type, but her type may determine whether she prefers to mediate. To understand this predicament, it is useful first to compare the equilibrium outcomes that would be obtained under mediation and in the absence of communication.[18]

In this section we will focus on the situation in which type 2 jeopardizes type 1, so that misrepresentation is problematic, but the likelihood of an incentive problem is low (i.e., $\pi < \hat{\pi}$). We assume in the following that a principal-biased mediator wants to implement the plan $\hat{\delta}$. We say that type s of the agent *a priori benefits from mediation* if $U_s(\hat{\delta}^s) \geq U_s(y(\pi))$, that is, type s prefers the outcome of mediation under the plan $\hat{\delta}$ rather than what it could get in the absence of mediation. Here the term *a priori* refers to the fact that mediation gains are assessed based on the principal's prior knowledge (i.e, under the prior beliefs π). Notice that the outcome of the mediated game $\Gamma_{\hat{\delta}}(\rho)$ depends on the prior beliefs ρ. In particular, $\hat{\delta}$ may cease to be incentive compatible under prior beliefs different from π.

[18] We focus on the agent's participation decision for various reasons. First, unlike the agent, the principal's participation decision is uninformative. Second, as it is well known, better information is characterized by higher (ex-ante) payoffs to a decision maker (Blackwell 1951, 1953). Consequently, the principal will always reject the optimal agent-biased mediation when confronted with some more informative mediation plan. Third, the principal will always accept an optimal principal-biased mediator.

We first remark that $y_p^1 < y(\pi) < \hat{y}$, since $0 < \pi < \hat{\pi}$. Therefore, $U_2(y(\pi)) > U_2(y_p^1) = U_2(\hat{\delta}^2)$. That is, type 2 a priori *never* benefits from mediation. On the other hand, the gains from mediation for type 1 depend on the parameter configuration. For instance, suppose in addition that $\sigma < 2$. Then $\hat{\delta}$ is also the unique optimal agent-biased mediation plan. Consequently, we must have that $U_1(y(\pi)) < U_1(\hat{\delta}^1)$ and, thus type 1 a priori benefits from mediation.[19]

In case type 1 a priori benefits from mediation, it is tempting to argue that type 1 would undergo mediation, while type 2 would refuse to participate in the mediation process. If the agent behaves this way, the principal would correctly infer the actual state from the agent's choice. Whenever the agent agrees to participate, the principal must believe that the true state is 1, thus rendering the mediation process ineffective: no matter what the mediator can recommend, the principal will always choose y_p^1. On the other hand, a rejection by the agent is taken as evidence that the actual state is 2, hence leading the principal to choose y_p^2, which is the worst action for type 2 in the effective issue space.

A more formal analysis of the agent's participation decision requires an explicit treatment of the principal's beliefs. To this end, we extend the mediated game $\Gamma_{\hat{\delta}}(\pi)$ to include an earlier stage of voluntary participation:

Stage 1 (participation phase). The mediation plan $\hat{\delta}$ is publicly proposed and the agent decides whether or not to participate in the mediation process.

Stage 2 (learning phase). The principal updates his prior belief about the state 2 based on any information revealed by the agent's participation decision. We let μ_A (resp. μ_R) denote the resulting posterior beliefs whenever the agent accepts (resp. refuses) to participate.

Stage 3 (mediation phase). If the agent agrees to participate, the mediated game $\Gamma_{\hat{\delta}}(\mu_A)$ is played. On the contrary, in case of refusal, the principal is left to choose an action based on his posterior beliefs μ_R.

The above informal comparison of equilibrium outcomes suggests that a separating strategy in which one type accepts mediation, while the other rejects it, releases too much information. This mitigates the agent's informational advantage and increases the incentives of type 2 to deviate from its participation strategy. This intuition is formalized in the following result.

Proposition 2 (No Separating Equilibrium).
There is no perfect Bayesian equilibrium of the voluntary mediation game in which either type agrees to participate and the other rejects.

Proof. Consider the separating strategy for the agent in which type 1 accepts and type 2 rejects. Then the Bayes rule implies that $\mu_A = 0$ and $\mu_R = 1$. After acceptance, the continuation game is $\Gamma_{\hat{\delta}}(0)$. Regardless of the mediator's recommendation in $\Gamma_{\hat{\delta}}(0)$, the principal always chooses $y(\mu_A) = y_p^1$. On the other hand, after rejection, the principal

[19] Mitusch and Strausz (2005, p. 490) compare the welfare of both agent's types under the equilibrium outcomes achieved with mediation and with direct communication. In their analysis, they argue that if $y_a^1 < y_p^1$, then type 1 prefers the outcome of $\hat{\delta}$ over the non-revealing equilibrium (which in this case is the *unique* cheap-talk equilibrium). As Example 1 below shows, this assertion is *not* correct. In particular, $\sigma < 2$ is a strictly stronger condition than $y_a^1 < y_p^1$, since the former implies the latter, but not viceversa.

chooses $y(\mu_R) = y_p^2$. However, because type 2 jeopardizes type 1, type 2 has incentives to deviate and accept.

A symmetrical analysis yields the same conclusion when we consider the separating strategy in which type 2 accepts and type 1 rejects. □

Proposition 2 tells us that the actual preferences of the agent are in conflict with her need to be inscrutable. In order to conceal her type, the agent must accept or reject mediation in a way that is independent of her type. Hence, for inscrutability, both agent's types must either accept or reject mediation.

Proposition 3 (Pooling Acceptance Equilibrium).
There exists a perfect Bayesian equilibrium of the voluntary mediation game in which both agent's types undergo mediation and $\hat{\delta}$ is truthfully and obediently implemented.

Proof. On the equilibrium path the Bayes rule implies that $\mu_A = \pi$. Thus, the continuation game is $\Gamma_{\hat{\delta}}(\pi)$. Since $\hat{\delta}$ is incentive-compatible given the beliefs π, then the sincere and obedient strategies form a Nash equilibrium of $\Gamma_{\hat{\delta}}(\pi)$. The corresponding utilities are $U_2(\hat{\delta}^2) = U_2(\hat{y})$ and $U_1(\hat{\delta}^1) = \hat{\theta}U_1(y_p^1) + (1 - \hat{\theta})U_1(\hat{y})$.

Consider now the situation after a rejection by some type. In this case, the (off-path) beliefs μ_R are not restricted by the Bayes rule. Let $\mu_R \geq \hat{\pi}$. Then $y(\mu_R) \geq \hat{y}$ and, therefore, $U_2(y(\mu_R)) \leq U_2(\hat{y})$. That is, type 2 does not have incentives to deviate. On the other hand, because $\Delta_a > 0$, then $U_1(y_p^1) > U_1(\hat{y}) \geq U_1(y(\mu_R))$. Therefore, $U_1(y(\mu_R)) < U_1(\hat{\delta}^1)$. Hence, type 1 has no incentive to deviate either. □

This acceptance equilibrium depends on out-of-equilibrium beliefs $\mu_R \geq \hat{\pi}$. That is, after observing an unexpected rejection, the principal revises his prior probability assessment that the agent is type 2 to $\hat{\pi}$ or more. No other beliefs can sustain such an equilibrium, since acceptance by type 2 is sequentially rational only when $\mu_R \geq \hat{\pi}$. These "posterior" beliefs are not computable using the Bayes rule, since there is zero prior probability of observing a refusal (in equilibrium). However, in case type 1 a priori benefits from mediation, one may argue that the principal's off-path beliefs are sensible only when $\mu_R = 1$. As the following example shows, there are also situations in which *both* types a priori strictly loose from mediation.

Example 1. Consider the following parameter configuration: $\pi = \frac{1}{10}$, $y_a^1 = 0$, $y_a^2 = 2$, $y_p^1 = 1$, and $y_p^2 = 11$. Straightforward computations yield $\sigma = 5$, $U_2(y_p^2) = -81 < -1 = U_2(y_p^1)$ (type 2 jeopardizes type 1), and

$$\hat{\pi} = \max\left\{0, \frac{2(y_a^2 - y_p^1)}{\Delta_p}\right\} = \frac{1}{5} > \pi.$$

Moreover, $U_1(\hat{\delta}^1) = -\frac{41}{9} < -4 = U_1(y(\pi))$ and $U_2(\hat{\delta}^2) = -1 < 0 = U_2(y(\pi))$. Namely, both types are a priori (strictly) worse-off under mediation. In spite of this, under the acceptance equilibrium, the agent cannot refuse to mediate, because in doing so she runs the risk that such a behavior will be interpreted by the principal as strong evidence in favor of state 2. Such a belief will cause the principal to choose an action that is much more detrimental for both types than the outcome of mediation. Nevertheless, the agent may address the principal as follows:

> *Speech A:* "I am rejecting the proposition to mediate. Notice that all of my types prefer the outcome without mediation. Thus, you shouldn't infer anything about my type from the fact that I have chosen not to continue with

mediation. With no new information about my type, you should find it optimal to hold your prior beliefs π and choose $y(\pi)$."

If both agent types are a priori better-off under mediation, the statement in speech A is *credible* in the sense of Farrell (1993): whenever the principal expects both types to send this unexpected message, he cannot infer anything from the speech, so he maintains his prior beliefs π, and all agents types benefit from sending the message. Then, if (as seems reasonable) the principal understands such an argument and validates it, the acceptance equilibrium is destroyed. Both parties then move to a situation where mediation does not occur and the principal holds passive beliefs whatever the agent's participation decision. This new specification constitutes also an equilibrium.

Proposition 4 (Pooling Rejection Equilibrium (part 1)).
Assume that no agent's type a priori benefits from mediation. Then there exists a perfect Bayesian equilibrium of the voluntary mediation game in which both agent's types refuse to mediate and the principal holds his prior beliefs in and out the equilibrium path.

Proof. On the equilibrium path the Bayes rule implies that $\mu_R = \pi$. Thus, after rejection the principal chooses $y(\pi)$. Let $\mu_A = \pi$ be the principal's off-path beliefs after observing an unexpected acceptance. Since, $\hat{\delta}$ is incentive compatible (given π), then the sincere and obedient strategies form a Nash equilibrium of the continuation $\Gamma_{\hat{\delta}}(\mu_A)$. Because $U_s(\hat{\delta}^s) < U_s(y(\pi))$ for $s = 1, 2$, then no type has incentives to deviate and accept mediation. □

Because the principal's prior beliefs remain unchanged after the agent's participation decision, the rejection equilibrium in Proposition 4 can only be sustained when both types a priori loose from mediation. However, an alternative rejection equilibrium exists.

Proposition 5 (Pooling Rejection Equilibrium (part 2)).
There exists a perfect Bayesian equilibrium of the voluntary mediation game in which both agent's types refuse to mediate and the principal attributes an unexpected acceptance entirely to type 2.

Proof. On the equilibrium path the principal maintains his prior beliefs (i.e., $\mu_R = \pi$) and optimally chooses $y(\pi)$. Let $\mu_A = 1$ be the off-path beliefs of the principal after observing acceptance. Then in the continuation game $\Gamma_{\hat{\delta}}(\mu_A)$ mediation is ineffective: no matter what the mediator recommends, the principal will always choose y_p^2. Under our current assumptions that type 2 jeopardizes type 1 and $\pi < \hat{\pi}$, we have that $U_s(y_p^2) < U_s(y(\pi))$ for $s = 1, 2$. Therefore, no type has incentives to deviate and accept mediation. □

Consider the situation in which only type 1 a priori benefits from mediation. Then, under the rejection equilibrium in Proposition 5, in principle, there is no reason to think that type 2 is responsible for an unexpected deviation. Hence, the off-path belief $\mu_A = 1$ seems unjustified.

We can now articulate our results to prescribe the outcome of the voluntary mediation game. As we have seen, according to Proposition 2, the principal should not expect one type to accept mediation, while the other rejects it. If that were the case, type 2 would have incentives to mimic type 1's behavior. Consequently, the principal should anticipate that both types will choose the same participation decision. Is it reasonable to think that the principal envisages both types to reject mediation? If this were true, then the principal would simply refrain from mediating in the first place. Instead of wasting time initiating a mediation process that is going to be rejected anyway, he would simply choose an action from the outset. More formally, because mediation is unanimous, if the

voluntary mediation game also contemplates the principal's participation decision, then it is a (weak) dominant strategy for the principal to accept mediation. Therefore, he either accepts mediation with the expectation that both types will do the same, or he is indifferent whenever both types reject, in which case he can simply reject and make a decision on the basis of his prior beliefs. On the grounds of the previous reasoning, we shall argue that the principal engages in mediation in the hope that both agent's types will accept. Drawing on this postulate, we claim that both parties concert their expectations on the acceptance equilibrium. That is, we consider the acceptance equilibrium to be a focal point. This does not mean, however, that parties in conflict will always coordinate effectively in such an equilibrium. Consider, for instance, the situation in which neither type a priori benefits form mediation (for instance because the relative degree of conflict is very large, i.e., $\sigma \gg 2$). In this case, the agent can approach the principal using speech A, which reshapes the principal's initial perception and diverts both parties to the rejection equilibrium in Proposition 4. In this way, the rejection equilibrium acquires a particular preeminence. On the other hand, if only type 1 a priori benefits from mediation (for instance if $\sigma < 2$), then only type 2 would (a priori) reject mediation. Therefore, coordinating on the acceptance equilibrium with off-path beliefs $\mu_R = 1$ predicts a sensible outcome. The beneficial effect of mediation here occurs at the expenses of type 2. Namely, the conflict of interests between the agent's types is resolved in favor of the jeopardized type.

7. Concluding Discussion

7.1. How rebels commit to peace

The results in this study could be interpreted in the context of a more specific discussion on the role of biased mediators for the achievement of peaceful settlements in civil wars (Svensson, 2015). Consider a situation of internal armed conflict with asymmetric information about the rebels' resolve to continue fighting—rebels (i.e., the agent) may have either high or low resolve, which is not directly observable by the government (i.e., the principal). Belligerents anticipating their post-agreement disarmament vulnerabilities, will pretend to be more bellicose than they actually are to get a better deal in negotiations (misrepresentation problem). The government wants an end to the confrontation, but it could not afford to stop fighting without being reasonably sure that the rebels would also do so (moral hazard problem). Hence, a mediator can only convince the government to stop fighting and cede some of its political power to the rebels by providing information about the rebels' resolve to fight.

Suppose that the likelihood of a misrepresentation problem is low, so that rebels are more likely to have a high resolve. In this case, there is room for a mediator to play a trust-building role by providing information. When the relative intensity of the conflict is too high, both types of rebels would be greatly harmed should the government learn their resolve. The reason is that, since it would be in the government's interest to renege on a peace deal, to ensure that the government comply with the stipulations in the agreement, a mediator needs the rebels to accept very costly concessions. This puts the rebels in a position of vulnerability that is detrimental even to the more violent type of rebel. Therefore, seeking to protect its protégé, a rebel-biased mediator has no option but to hamper communication, which makes mediation ineffective.

A government-biased mediator, on the other hand, serving the interests of its side, engages in a more efficient information provision strategy. Yet the misrepresentation problem restricts the mediator in inducing full information disclosure. Consequently, the high-resolve type does not reveal itself completely, but provides cover for the problematic

low-resolve type. In this way, the rebels may provide reassurance for the government in their commitment to peace, while at the same time not exposing themselves entirely.

The above contextualization of our results leads to the theoretical prediction that, *when the relative intensity of the conflict is high and the misrepresentation problem is low*, "government-biased mediators, rather than rebel-biased mediators, should have a positive effect on the likelihood of reaching a negotiated settlement in internal conflicts" (Svensson, 2007, p. 183). Our results thus provide the formal theoretical conditions for the internal consistency of Svensson's (2007) thesis and quantitative analysis. However, whether these conditions actually have an explanatory power remains an open empirical question for future work.

7.2. Impartial Mediators

There is a conceptual confusion about the definition of *impartiality*. Impartiality may refer to the mediator's interest in the issue of conflict. For instance, Kydd (2003) defines a mediator as impartial when it does not get any benefit from any particular solution. This definition is in accordance with the principle that an impartial decision is one in which a certain sort of considerations (e.g. about the issue in dispute) have no influence. Ivanov (2010) refers to this form of impartiality as *neutrality*. In line with this precept, a biased mediator in our setting is impartial, since, despite having strong ties to one of the parties, it does not have preferences over the issue in dispute.

Impartiality could also relate to the mediator's tendency to *side* in a conflict. Rauchhaus (2006) sees a mediator as impartial if its ideal point is located in the middle of the issue space, so that its preferences do not display any favoritism for some side. Rauchhaus' definition is based on an idea of impartiality that ensures "moderate" concessions.

Analysts agree, though, that impartiality is essentially a matter of perceptions of the parties in conflict (Touval, 1975). The mediator's impartiality can be assessed from the mediation process itself or from the outcomes that it produces. In the first case, we refer to perceived fairness from a *procedural justice* point of view. The second case pertains to the domain of *distributive justice* (as in Rauchhaus, 2006).

According to the participation model of procedural justice, a fair procedure is one that affords those who are involved the same chance to participate in the making of decisions. Such a procedure would be comparable to randomly choosing the mediator bias uniformly. In this manner, each disputant has equal probability of enjoying a mediation process that is most favorable to him/her. In case the outcome of mediation is affected by the direction of the bias (i.e., $\pi < \hat{\pi}$ and $\sigma \geq 2$), this procedure amounts to implement the mediation plan $\hat{\delta}$ with probability $^1/_2$ and hinder communication with probability $^1/_2$. Because the set of incentive-compatible mediation plans is convex, the outcome of such a "random dictatorship" procedure is a convex combination of the (ex-ante) expected utilities obtained from both optimal biased mediation plans. Moreover, this outcome is (ex-ante) Pareto efficient. Its proof is that the Pareto frontier is a straight line from $\left(U(\hat{\delta};\pi), V(\hat{\delta};\pi)\right)$ to $\left(\widehat{U}(\pi), \widehat{V}(\pi)\right)$.[20]

Alternatively, one may also advocate the egalitarian principles of distributive justice, according to which all individuals are fundamentally equal and therefore must be treated equally. This idea suggests that an impartial mediator should give equal consideration to the interests of each party. This amounts to maximize a social welfare function that puts

[20] We have used the notation $\widehat{V}(\pi) := (1-\pi)V_1(y(\pi)) + \pi V_2(y(\pi))$.

equal weight on the preferences of each disputant. Formally, the mediator should choose an incentive-compatible mediation plan, δ, to maximize

$$\frac{1}{2}U(\delta;\pi) + \frac{1}{2}V(\delta;\pi)$$

Because the Pareto frontier is linear, it can be described by a supporting hyperplane of the form $\lambda U + (1-\lambda)V = W$, for some utility weight $\lambda \in (0,1)$ that depends on the parameters of the model. If $\lambda < {}^1/_2$ (resp. $\lambda > {}^1/_2$), the outcome of such an impartial mediation process coincides with an optimal principal-biased mediation plan (resp. agent-biased mediation plan). Therefore, even a "moderate" mediator who weighs both disputants equally will adopt behavior that favors one side over the other.

7.3. Mutually Beneficial Mediation

The vast majority of the literature on optimal mediation in sender-receiver games centers its analysis on the conditions under which mediation facilitates communication between the parties. Therefore, the beneficial effect of mediation is measured as the mediator's ability to improve upon the informativeness of non-facilitated negotiation. Blackwell's (1951; 1953) theorems on the ranking of statistical experiments imply that beneficial mediation can be equivalently quantified by the capacity of a mediator to increase the welfare of the decision maker (i.e., the receiver/principal). Because a mediator can always mimic the outcome of a non-facilitated negotiation, the focus of this literature has been on mediation plans that maximize the receiver's expected utility. Such an approach implicitly assumes that the receiver has all the bargaining ability—he possesses effective control over all communication channels. It is for this reason that Mitusch and Strausz (2005) refer to the decision maker as the "principal".

An exception to the previous approach was provided by Salamanca (2021), who, as we do in this work, focuses his analysis on the informed party (i.e., the sender/agent). Of course, this means that the previous measure of beneficial mediation is no longer adequate. In fact, as shown in this article, there are situations in which, in order to maximize the agent's welfare, the mediator must minimize the amount of disclosed information. For this reason, in this paper we adopt a notion of *effectiveness*, which is defined as the ability of the mediator to improve upon the uninformed outcome. This is a very weak notion of mediation success. However, all our results also hold when rephrased in terms of the stronger definition of beneficial mediation described above.

8. Appendix

This appendix contains a detailed proof of the statements in Theorem 2.

In a recent contribution, Salamanca (2021) developed a general approach to study optimal mediation. We shall exploit this methodology to characterize an agent-biased optimal mediation plan when misrepresentation is problematic.

Let $\alpha(s' \mid s) \geq 0$ denote the shadow price (or Lagrange multiplier) for the truth-telling incentive constraint (3.2) asserting that type s should not gain by reporting s' in the problem (4.2). The shadow prices measure the expected marginal cost of strengthening the truth-telling incentive constraints. Thus, they quantify the efficiency loss incurred due to the incentive problem. Multiplying the truth-telling incentive constraints by their corresponding shadow prices and adding them into the objective function, we obtain the following

Lagrangian relaxation of (4.2):

$$\mathcal{L}(\delta;\pi,\alpha) \coloneqq U(\delta;\pi) + \sum_{s}\sum_{s'}\alpha(s' \mid s)\sum_{y_j\in Y} U_s(y_j)\left[\delta_j^s - \delta_j^{s'}\right]. \tag{8.1}$$

For any given action y, we define

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

Myerson (1991) refers to $W_s(y;\pi,\alpha)$ as the agent's *virtual utility* from action y, when the state is s, w.r.t. the prior π and the shadow prices α. The virtual utility of a type s equals its actual utility plus the information rents it obtains from misreporting the actual state, minus the signaling costs incurred because of the misrepresentation by type s'. It is worth noticing that what is a cost for one type becomes a rent for the other type. Thus, we can interpret the shadow prices as signaling costs (rather than information rents).

With this definition, the above Lagrangian can be written as

$$\mathcal{L}(\delta;\pi,\alpha) = \sum_{y_j\in Y}\left[(1-\pi)\delta_j^1 W_1(y_j;\pi,\alpha) + \pi\delta_j^2 W_2(y_j;\pi,\alpha)\right].$$

That is, the Lagrangian in (8.1) is simply the agent'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 in (3.3):

$$\begin{aligned} &\max_{Y,\delta}\ \mathcal{L}(\delta;\pi,\alpha) \\ &\text{s.t. (3.3)}. \end{aligned} \tag{8.2}$$

This optimization problem corresponds to an omniscient mediation problem (as in (4.3)), except that now the agent's preferences are measured in the virtual utility scales. Given π and α, the *indirect virtual utility function* is defined as follows:

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

Hence, the optimal value of (8.2) is $cav\,\widehat{W}(\pi\,;\pi,\alpha)$, where $cav\,\widehat{W}(\cdot\,;\pi,\alpha)$ denotes the concavification of $\widehat{W}(\cdot\,;\pi,\alpha)$. The following result relates the value of (8.2) to the value of mediation for the agent. In doing so, it provides sufficient conditions under which a candidate incentive-compatible mediation plan is optimal in (4.2).

Proposition 6 (Weak Duality, Salamanca, 2021).
Fix a prior π and let (Y^*,δ^*) be an incentive-compatible mediation plan such that $U(\delta^*;\pi) = cav\,\widehat{W}(\pi;\pi,\alpha^*)$ for some $\alpha^* \geq 0$. Then, (Y^*,δ^*) is an optimal solution of (4.2) and, therefore, the value of mediation to the agent equals $U^*(\pi) = cav\,\widehat{W}(\pi;\pi,\alpha^*)$.

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

$$U(\delta;\pi) \leq \mathcal{L}(\delta;p,\alpha^*) \leq cav\,\widehat{W}(\pi;\pi,\alpha^*) = U(\delta^*;p), \tag{8.4}$$

where the first inequality holds because (Y, δ) is incentive-compatible and $\alpha^* \geq 0$; the second inequality comes from the fact that $cav\,\widehat{W}(\pi; \pi, \alpha^*)$ is the optimal value of (8.2); and finally, the last equality holds by hypothesis. We then conclude that $U(\delta; \pi) \leq U(\delta^*; \pi)$ for all (Y, δ) incentive-compatible mediation plans. □

Let $\hat{\delta}$ be defined as in (4.7). We shall now construct a vector of shadow prices, $\hat{\alpha}$, verifying the conditions in Proposition 6 for $\hat{\delta}$. Because only the truth-telling incentive constraint of type 2 is binding under $\hat{\delta}$, we expect that $\alpha(2 \mid 1) = 0$ and $a := \alpha(1 \mid 2) > 0$. For a given vector $\alpha = (0, a)$, and under the current assumptions, the indirect virtual utility, $\widehat{W}(\cdot\,; \pi, a)$, looks as in Figure 4. That is, there exists a belief $\varpi := \varpi(a)$ such that, for any $0 \leq \rho \leq \varpi$, $cav\,\widehat{W}(\rho; \pi, a)$ coincides with the linear segment which is tangent to $\widehat{W}(\cdot; \pi, a)$ at ϖ, and whose intercept with the ordinate is $\widehat{W}(0; \pi, a)$. Hence, for any prior $\pi < \varpi$, $cav\,\widehat{W}(\pi; \pi, a)$ is achieved by splitting the prior belief into posterior beliefs $\rho_1 = 0$ and $\rho_2 = \varpi(a)$.

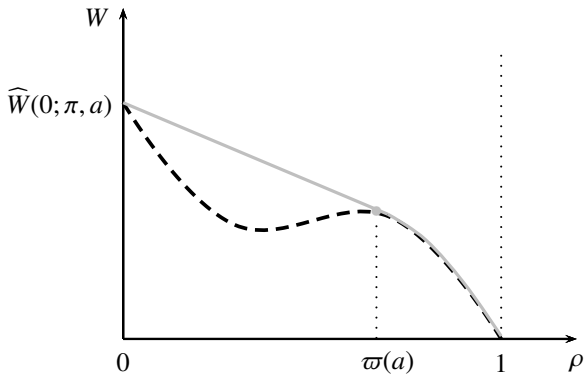

Figure 4: Functions $\widehat{W}(\cdot\,; \pi, a)$ (black dashed line) and $cav\,\widehat{W}(\cdot\,; \pi, a)$ (gray solid line)

According to weak duality, in order to construct the appropriate shadow prices $\hat{\alpha} = (0, \hat{a})$, we require the agent's expected payoff from $\hat{\delta}$ to equal the concavification of the indirect virtual utility at π. Hence, we need the posteriors generated by $\hat{\delta}$ to achieve $cav\,\widehat{W}(\pi; \pi, \hat{a})$. This suggest defining implicitly $\hat{a}$ as follows:

$$\varpi(\hat{a}) = \hat{\pi}. \tag{8.5}$$

Lemma 3.
Suppose $\sigma < 2$ holds. Then equation (8.5) has a unique non-negative solution, $\hat{a}$, provided that $\pi < \hat{\pi}$.

Proof. By definition, $\hat{a}$ is a solution of (8.5) if and only if $\widehat{W}(\cdot\,; \pi, \hat{a})$ coincides with $cav\,\widehat{W}(\cdot\,; \pi, \hat{a})$ at $\hat{\pi}$. Formally,

$$\widehat{W}(\hat{\pi}; \pi, \hat{a}) = \widehat{W}(0; \pi, \hat{a}) + \frac{\partial \widehat{W}(\hat{\pi}; \pi, \hat{a})}{\partial \rho}\hat{\pi} \tag{8.6}$$

Computing each term in (8.6) yields:

$$\frac{\partial \widehat{W}(\hat{\pi}; \pi, \hat{a})}{\partial \rho} = \widehat{U}'(\hat{\pi}) + \frac{\hat{a}}{\pi(1-\pi)}\left[U_2(y_p^1) + (\hat{\pi} - \pi)U_2'(y(\hat{\pi}))\Delta_p\right] \tag{8.7}$$

$$\widehat{W}(0; \pi, \hat{a}) = U_1(y_p^1) - \frac{\hat{a}}{1-\pi}U_2(y_p^1) \tag{8.8}$$

$$\widehat{W}(\hat{\pi}; \pi, \hat{a}) = \widehat{U}(\hat{\pi}) + \hat{a}\frac{\hat{\pi} - \pi}{\pi(1-\pi)}U_2(y_p^1) \tag{8.9}$$

Plugging (8.7)-(8.9) into (8.6) and solving for $\hat{a}$ we obtain:

$$\hat{a} = \frac{\pi(1-\pi)}{\hat{\pi}} \frac{\widehat{U}(\hat{\pi}) - U_1(y_p^1) - \hat{\pi}\widehat{U}'(\hat{\pi})}{(\hat{\pi}-\pi)U_2'(y(\hat{\pi}))\Delta_p} \tag{8.10}$$

Because $\sigma < 2$, $\widehat{U}$ is strictly convex, which implies that

$$0 > \widehat{U}(\hat{\pi}) - \widehat{U}(0) + \widehat{U}'(\hat{\pi})[0 - \hat{\pi}] = \widehat{U}(\hat{\pi}) - U_1(y_p^1) - \hat{\pi}\widehat{U}'(\hat{\pi}),$$

Thus, the numerator of (8.10) is strictly negative. On the other hand, notice that $U_2'(y(\hat{\pi})) < 0$ if and only if $y_p^1 < y_a^2$ if and only if $\hat{\pi} > 0$. Therefore, $\hat{\pi}(\hat{\pi} - \pi)U_2'(y(\hat{\pi})) < 0$ only if $0 < \pi < \hat{\pi}$. Hence, $\hat{a} > 0$ provided $\pi < \hat{\pi}$. □

We set the vector of shadow prices to be $\hat{\alpha} = (\hat{\alpha}(2 \mid 1), \hat{\alpha}(1 \mid 2)) = (0, \hat{a})$, where $\hat{a}$ is the solution to (8.5).

Suppose $\sigma < 2$ holds. Assume $\pi < \hat{\pi}$. Then

$$\begin{aligned}
cav\,\widehat{W}(\pi\,;\pi,\hat{a}) &= \widehat{W}(0\,;\pi,\hat{a}) + \frac{\partial \widehat{W}(\hat{\pi}\,;\pi,\hat{a})}{\partial \rho}\pi \\
&= U_1(y_p^1) - \frac{\hat{a}}{1-\pi}U_2(y_p^1) + \pi\widehat{U}'(\hat{\pi}) \\
&\quad + \frac{\hat{a}}{1-\pi}\left[U_2(y_p^1) + (\hat{\pi}-\pi)U_2'(y(\hat{\pi}))\Delta_p\right], \\
&= U_1(y_p^1) + \pi\widehat{U}'(\hat{\pi}) + \frac{\hat{a}}{1-\pi}(\hat{\pi}-\pi)U_2'(y(\hat{\pi}))\Delta_p, \\
&= U_1(y_p^1) + \pi\widehat{U}'(\hat{\pi}) + \frac{\pi}{\hat{\pi}}\left[\widehat{U}(\hat{\pi}) - U_1(y_p^1) - \hat{\pi}\widehat{U}'(\hat{\pi})\right], \\
&= \widehat{U}(0)\frac{\hat{\pi}-\pi}{\hat{\pi}} + \widehat{U}(\hat{\pi})\frac{\pi}{\hat{\pi}}, \\
&= U(\hat{\delta};\pi),
\end{aligned}$$

where the equality in the fourth line is obtained from the definition of $\hat{a}$ in (8.10). Hence, by Proposition 6, $\hat{\delta}$ is optimal in (4.2).

Assume now that $\pi \geq \hat{\pi}$. Then all incentive-compatible mediation plans are non-revealing. For a proof of this statement, the reader is referred to Lemma 5 in Mitusch and Strausz (2005). This completes the proof.

References

Assefa, Hizkias, *Mediation of Civil Wars: Approaches and Strategies—the Sudan Conflict*, Taylor and Francis, 1987.

Bercovitch, Jacob and Allison Houston, "Why Do They Do It like This? An Analysis of the Factors Influencing Mediation Behavior in International Conflicts," *The Journal of Conflict Resolution*, 2000, *44* (2), 170–202.

_ , J. Theodore Anagnoson, and Donnette L. Wille, "Some Conceptual Issues and Empirical Trends in the Study of Successful Mediation in International Relations," *Journal of Peace Research*, 1991, *28* (1), 7–17.

Blackwell, David, *Comparison of experiments*, University of California Press,

_ , "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 1953, *24* (2), 265–272.

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

Brown, Jennifer Gerarda and Ian Ayres, "Economic Rationales for Mediation," *Virginia Law Review*, 1994, *80* (2), 323–402.

Cohen, Stephen P. and Edward E. Azar, "From War to Peace: The Transition between Egypt and Israel," *The Journal of Conflict Resolution*, 1981, *25* (1), 87–114.

Crawford, Vincent and Joel Sobel, "Strategic information transmission," *Econometrica*, 1982, *50* (6), 1431–51.

Doyle, Stephen P. and Roger S. Haydock, *Without the Punches: Resolving Disputes without Litigation*, Minneapolis: Equilaw, 1991.

Fanning, Jack, "Mediation in Reputational Bargaining," *American Economic Review*, 2021, *111* (8), 2444–72.

_ , "Optimal Danymic Mediation," Mimeo 2021.

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

Fearon, James D., "Rationalist Explanations for War," *International Organization*, 1995, *49* (3), 379–414.

Fey, Mark and Kristopher W. Ramsay, "When is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation," *World Politics*, 2010, *62* (4), 529–560.

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

Ganguly, Chirantan and Indrajit Ray, "Simple mediation in a cheap-talk game," Discussion Papers 05-08, Department of Economics, University of Birmingham 2009.

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

Gottardi, Piero and Claudio Mezzetti, "Shuttle Diplomacy," CEPR Discussion Paper No. DP16934 2022.

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

Hörner, Johannes, Massimo Morelli, and Francesco Squintani, "Mediation and Peace," *The Review of Economic Studies*, 05 2015, *82* (4), 1483–1501.

Hume, Cameron R., *Ending Mozambique's War: The Role of Mediation and Good Offices*, United States Institute of Peace Press, 1994.

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.

Jarque, Xavier, Clara Ponsatí, and József Sákovics, "Mediation: incomplete information bargaining with filtered communication," *Journal of Mathematical Economics*, 2003, *39* (7), 803–830.

Jones, Frank Leith, "Haig's Waterloo: Lessons from a Failure in International Mediation," *International Journal on World Peace*, 2013, *30* (3), 7–29.

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

Kydd, Andrew H., "Which Side Are You On? Bias, Credibility, and Mediation," *American Journal of Political Science*, 2003, *47* (4), 597–611.

_ , "When Can Mediators Build Trust?," *The American Political Science Review*, 2006, *100* (3), 449–462.

Meirowitz, Adam, Massimo Morelli, Kristopher W. Ramsay, and Francesco Squintani, "Dispute Resolution Institutions and Strategic Militarization," *Journal of Political Economy*, 2019, *127* (1), 378–418.

Miall, Hugh, *The Peacmakers: Peaceful Settlements of Disputes since 1945*, Palgrave Macmillan, 1992.

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.

Raiffa, Howard, *The Art and Science of Negotiation*, Cambridge, Massachusetts: Harvard University Press, 1985.

Rauchhaus, Robert W., "Asymmetric Information, Mediation, and Conflict Management," *World Politics*, 2006, *58* (2), 207–241.

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

Savun, Burcu, "Information, Bias, and Mediation Success," *International Studies Quarterly*, 2008, *52* (1), 25–47.

Singer, Linda, *Settling Disputes: Conflict Resolution in Business, Families, and the Legal System*, Taylor and Francis, 1990.

Smith, Alastair and Allan Stam, "Mediation and Peacekeeping in a Random Walk Model of Civil and Interstate War," *International Studies Review*, 2003, *5* (4), 115–135.

Svensson, Isak, "Bargaining, Bias and Peace Brokers: How Rebels Commit to Peace," *Journal of Peace Research*, 2007, *44* (2), 177–194.

_ , "Who Brings Which Peace? Neutral versus Biased Mediation and Institutional Peace Arrangements in Civil Wars," *The Journal of Conflict Resolution*, 2009, *53* (3), 446–469.

_ , *International Mediation Bias and Peacemaking: Taking Sides in Civil Wars*, Taylor and Francis, 2015.

Touval, Saadia, "Biased Intermediaries: Theoretical and Historical Considerations," *Jerusalem Journal of International Relations*, 1975, *1* (51), 51–69.

_ **and I. William Zartman**, *International Mediation in Theory and Practice*, Washington: Johns Hopkins Foreign Policy Institute: Westview Press,

_ **and** _ , *Mediation Research: The Process and Effectivencess of Third Party Intervention* Jossey-Bass Social and Behavioral Sience Series, 1st ed., San Francisco: Jossey-Bass,
ΩČopič and Ponsatí

Čopič, Jernej and Clara Ponsatí, "Robust Bilateral Trade and Mediated Bargaining," *Journal of the European Economic Association*, 2008, *6* (2/3), 570–580.

Young, Oran, *The Intermediaries: Third Parties in International Crises*, N.J: Published for the Center of International Studies, Princeton University: Princeton University Pres, 1967.