The Honest Broker:

Mediation and Mistrust

Andrew H. Kydd*

May 18, 2005

Abstract

Mediation is one of the most widespread techniques for preventing conflict and promoting cooperation. Unfortunately, the literature on mediation has not yet reached consensus on what makes mediation work. For instance, some have argued that mediators should be unbiased, while others argue that biased mediators are effective. This paper examines the conditions under which mediators can facilitate cooperation by building trust between the two parties. Mediators can be credible trust builders in one round interactions only if they prefer mutual non-cooperation to either side exploiting the other. A biased mediator or one who is solely interested in promoting cooperation will be ineffective. If the mediator is involved in an ongoing relationship with the parties, biased mediators can function as trustbuilders, provided that the degree of bias is not too great.

*Associate Professor of Government, Harvard University. Andrew Kydd, WCFIA, 1033 Massachusetts Ave., Cambridge, MA 02138. akydd@wcfia.harvard.edu, 617-495-5422.

Mediation is one of the most important tools for promoting cooperation and conflict resolution. It is applied in fields as diverse as international and civil war, economic exchange, labor negotiations and divorce proceedings. Mediation can make the difference between a war that continues to take thousands of lives or a successfully negotiated peace deal, between a strike that costs thousands of hours of lost production and a contract that puts people back to work. Given the importance of mediation, it is not surprising that scholars have devoted considerable attention to studying what makes mediation work.[1]

Any understanding of how mediation can facilitate cooperation must be built on a diagnosis of what causes conflict in the first place and what mediators can do about it. One widely discussed cause of conflict that seems amenable to mediation is uncertainty or asymmetric information. Blainey famously argued that wars begin when states disagree about their relative power and end when they agree again (Blainey 1988, 122). More generally, in models of bargaining, disputants that share complete information can typically reach agreement without delay or conflict because the cost of conflict provides an incentive to agree sooner rather than later (Rubinstein 1982). However, when each side has private information about its own resolve or relative power, it may have an incentive to pretend to be stronger than it is in order to get a better deal, which can prevent a deal being struck (Fearon 1995). This has led to a large literature in economics and international relations on the relationship between bargaining, private information, and conflict.[2] Studies of mediation have begun to be grounded in this theory as well, analyzing how mediators can facilitate cooperation by providing or withholding information about the parties' resolve or power (Jarque, Ponsati,

[1] For a recent literature review, see (Wall, Stark, and Standifer 2001).

[2] For a review see (Powell 2002).

and Sakovics 2003; Kydd 2003; Smith and Stam 2003; Rauchhaus 2005).[3]

Another form of uncertainty that can cause conflict is mistrust. Mistrust arises when actors are uncertain about the other side's preferences or intentions, specifically, about whether the other side will reciprocate cooperation or exploit it (Gambetta 1988; Hardin 2002). If two sides mistrust each other they can fall into conflict because they believe that the other side will take advantage of them or betray them in some way. Mistrust has been held responsible for negative outcomes in a variety of settings, from general economic performance to the Cold War (Fukuyama 1995; Larson 1997). Trust is also argued to be a key component of good governance (Braithwaite and Levi 1998).

Mediators may also be able to help with mistrust problems. For instance, if two disputants are engaged in an ongoing civil or international conflict but wish to conclude a peace treaty, a mediator can reassure each side that the other is genuinely interested in peace, and not attempting to deceive and exploit them. Kelman argues in the context of the Oslo negotiations that the Israelis and Palestinians "had to be persuaded that there was a genuine readiness on the other side to make the necessary concessions" and that unofficial mediation efforts "contributed to the gradual development within the two political communities of . . . a degree of working trust – i.e., trust that the other side is genuinely committed, largely out of its own interest, to finding an accommodation" (Kelman 1997; Kriesberg 2001). The fact that they were able to develop this degree of trust enabled the parties to cooperate, at least for a while.

Similarly, in an economic context, where two parties are contemplating a mutually beneficial exchange, a mediator can vouch for the trustworthiness of each side of a risky transaction.

[3] Foundational articles on mediated communication in economics include (Myerson 1986; Forges 1986).

For instance, Milgrom, North and Weingast argue that the revival of trade in the middle ages was facilitated by the evolution of a system of private law and judges, the *Lex Mercatoria* or merchant law (Milgrom, North, and Weingast 1990). A private judge would keep a record of any merchants accused of wrongdoing. At a fair, any merchant could consult the judge about a prospective trading partner to learn if they had honored their contracts in the past. This system facilitated exchange between actors who knew little about each other and might have been too mistrustful to cooperate without some reassurance that the other side was willing to fulfill its promises.

Mistrust, then, is a form of uncertainty that can cause conflict but is susceptible to mediation. However, unlike in the case of uncertainty about resolve or relative power, the theory of how mediation can overcome mistrust problems is not well developed. Basic questions remain to be addressed. What characteristics should a mediator have to build trust? In particular, what is the impact of the mediator's preferences over the issue in dispute or the possible outcomes from a successful or unsuccessful trade? Should a mediator simply want to promote cooperation, without caring what form that cooperation takes? Must mediators be unbiased, and what is the role of the shadow of the future and reputation? Without answers to questions such as these, theories of mediation cannot provide well grounded hypotheses for empirical evaluation nor effective prescriptive advice for practitioners of conflict resolution.

To begin to answer these questions, I develop a model of mediation in situations of mistrust. Two sides must decide whether or not to cooperate with each other, each fearing that the other side may exploit it. The mediator has some information about the trustworthiness of the two sides, and can share this information with the parties in an effort to reassure them. The key question is, when will the mediator honestly communicate its information, particularly if that information indicates that the parties are untrustworthy and need to be warned against each other. The central result of the model is that in order to be credible in one round situations a mediator must be *exploitation averse* or prefer mutual non-cooperation to either side exploiting the other. In the conflict resolution context, this is a product of preferring a moderate solution to the dispute, for instance, a deal which splits the difference and gives something to both sides. In a one round game, *bias*, or a preference for exploitation over mutual non-cooperation, is fatal to a mediator's credibility. So too is indifference over the different issue resolutions or caring too much about avoiding conflict. That is, a mediator who is solely motivated by a desire to prevent conflict will not be credible, and hence will fail to have any impact on the likelihood of conflict. However, if the game is repeated, the mediator can acquire a reputational incentive to be honest which is sufficient to overcome a limited amount of bias. Too much bias, however, undermines honesty even in the repeated game. Overall, unbiased mediators make the best trustbuilders.

1 Mistrust and Mediation

Mistrust causes conflict in a variety of settings. Two contexts have received particular attention, international and civil war, and the problem of exchange under uncertainty. In each, mediation can provide reassurance that can lead to cooperation.

There is a long standing tradition of explaining international and civil conflict as a function of vulnerability combined with distrust. Hobbes explained civil strife by arguing that all men are vulnerable and that given the inherent uncertainty about the intentions and ambitions of others it makes sense even for defensively motivated individuals to act pre-emptively to destroy the power of others who might threaten them (Hobbes 1651, 184). Herz developed this argument in international relations and argued that states face a security dilemma, in which anarchy plus uncertainty about the intentions of others leads states to arm themselves, which harms the security of others leading them to respond in kind in a vicious circle (Herz 1950). This argument has become a cornerstone of realist thought in international relations (Jervis 1976; Jervis 1978; Glaser 1995) and has also been used to explain ethnic conflict (Posen 1993). Walter argues that many civil wars are difficult to end because each side fears being exploited in the post conflict phase, particularly a rebel group that must disarm as part of the peace process (Walter 2002). All of these approaches imply that conflict could be prevented or resolved through strategies of reassurance or trustbuilding, and that this could be a role for mediators in the conflict resolution or prevention process. Walter finds that the presence of a mediator increases the likelihood of successful negotiations in civil wars by 39%.

Another social context in which trust is important is the problem of exchange under uncertainty. Many economic and other transactions depend on a certain degree of trust. If one side must perform their end of the bargain first and then wait for the other side to fulfill their promise, the first mover must trust that the second one will actually do their part (Coleman 1990, 91). Similarly, if the quality of the products is not immediately apparent, then each side must trust that the other side has not unloaded shoddy goods on them if they are to be willing to pay full price. Akerlof argued that in the used car market, “lemons” may drive out good cars because buyers know less about a car’s quality than sellers. If buyers will only pay what an average car in the used car market is worth, then sellers of top quality cars will not get enough for their cars and will withdraw from the market, driving down the average quality level still further, until the market consists only of very low quality cars (Akerlof 1970). This has led to a large literature on how this kind of information problem can be overcome through signaling devices such as brand names which invoke a reputation.[4] While often this problem is asymmetrical, such that one side bears all the risk, other times the risk is mutual. In the used car market, for instance, sellers use credit rating agencies to vouch for the ability and willingness of buyers to make their payments, while buyers rely on branding, or word of mouth to evaluate dealers. In general, economic intermediaries such as retailers and rating agencies can be seen as a way of overcoming mistrust problems that could prevent exchange under uncertainty (Spulber 1996). Biglaiser analyses such a model and shows how intermediaries have an incentive to invest in the ability to discern quality and preserve a reputation for selling high quality goods (Biglaiser 1993). Lizzeri advances the analysis further by analyzing the strategic incentives for the intermediary to reveal or conceal information, a key focus of the model presented below (Lizzeri 1999).

The literature on mediation has not ignored the issues of trust and reassurance. A research tradition begun by Burton and Walton argues that conflict is driven or exacerbated by stereotypes and mistrust and that a form of unofficial mediation involving scholars of conflict resolution can help to overcome these problems in special workshop style sessions (Burton 1969; Walton 1969). Fisher develops the idea and summarizes empirical applications (Fisher 1972; Fisher 1983; Fisher and Keashly 1991), while Kelman has led one of the most sustained efforts along these lines focused on the Israeli-Palestinian dispute (Kelman 2000). Others have borrowed the concept of confidence building measures from international security and applied it in the mediation of family and divorce proceedings (Landau and Landau 1997).

[4] For a review see (Riley 2001).

Ross and Weiland found that mediators in an experimental setting facing situations of low trust resorted to trustbuilding strategies including the use of humor (Ross and Wieland 1996). Wehr and Lederach argue that in Central America cultural factors favor mediators who are trusted members of the community, even if they are partial to one side (Wehr and Lederach 1991).

However, the mediation literature has not yet come to consensus on what makes for successful mediation, in general or in the specific case of trustbuilding. One prominent debate has concerned the role of mediator bias. Some scholars, such as Young, have included impartiality in the very definition of mediation (Young 1967). Others, such as Saadia Touval, have argued that mediators are often biased and can perform their tasks as well if not better for it (Touval 1975; Touval and Zartman 1989). Thomas Princen has argued that weak mediators do better when neutral but strong mediators from great powers are biased but effective (Princen 1991; Princen 1992).

This debate has arisen anew in the formal literature on mediation and asymmetric information about resolve and power. Drawing on the cheap talk literature in economics (Farrell and Rabin 1996), Kydd argued that mediators must be biased to be effective (Kydd 2003). In his model, a mediator is trying to convince one side to make a concession by providing information about the resolve of the other side. Only a mediator who is biased towards one party can credibly tell them that the adversary will not make peace without the concession, because such a mediator would not urge a concession on the party unless the mediator thought it was truly necessary. However, Rauchhaus, based on a similar model, finds that "neutral" mediators can be credible about resolve (Rauchhaus 2005). Meawhile, Smith and Stam, adapting their random walk model of war to the mediation question, find that biased mediators are not credible on the subject of the disputants' power (Smith and Stam 2003).

The discrepency is a result of differing definitions of bias. For Kydd, a mediator is biased if it shares one side's preference ordering over the issue space and unbiased if it is indifferent over the various issue resolutions. All mediators are assumed to find war costly. In Rachhaus's case the mediator is neutral if its ideal point on the issue space is located between the two possible negotiated solutions, and biased if it is outside that range. Rauchhaus's neutral mediator is therefore not the same as Kydd's unbiased mediator, so the results do not necessarily conflict. For Smith and Stam, a mediator is biased towards peace if it prefers peace to war but has no preferences over the issue in dispute, and biased towards a player if it prefers one side's favorite issue resolution but does not care about the cost of war. A mediator that is biased towards peace according to Smith and Stam is unbiased according to Kydd, and both find such a mediator ineffective (as does Rauchhaus). However, Smith and Stam do not consider a mediator who is both biased towards one side and finds war to be costly, and so do not evaluate if a mediator who is biased according to Kydd could be credible in their model.[5]

The dispute about mediator bias highlights the need for careful and transparent definitions of terms and rigorous analysis. Indeed the appropriate definition of bias, and its impact on mediation, may depend on the context.[6] In the model below, I develop a definition of

[5] Other recent game theoretic approaches to mediation include (Mitusch and Strausz 2000; Jarque, Ponsati, and Sakovics 2003; O'Neill 2003; O'Neill 2004; Crescenzi, Kadera, Mitchell, and Thyne 2005; Favretto 2005). Related models analyze the problem of third party intervention in conflict (Carment and Rowlands 1998).

[6] Holger Schmidt analyses when bias is good or bad via a model of third party monitoring of peace agreements and an analysis of postwar cases of peace implementation (Schmidt 2004).

Table 1: The Mediation Game

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	$1, 1, \rho$	$-a_1, b_2, \beta_2$
	Defect	$b_1, -a_2, \beta_1$	$0, 0, 0$

bias appropriate to the trustbuilding context and show that it is harmful to the mediator's credibility. Instead the mediator should be averse to either side being exploited.

2 The Model

There are three players, player 1, player 2 and the mediator. Player 1 and 2 face a mistrust problem. They may cooperate or defect as illustrated in Table 1. Player 1's payoffs are listed first, player 2's second, and the mediator's third. I normalize the payoff for successful cooperation to 1 for each player and that for mutual defection to zero. If either side defects unilaterally, it receives b_i while its opponent receives $-a_j$. There are two varieties of player, trustworthy types for whom $b_i \leq 1$ and untrustworthy types for whom $b_i > 1$. In terms of the familiar two by two games, trustworthy types have Assurance payoffs and untrustworthy types have Prisoner's Dilemma payoffs.

If the players cooperate, the mediator receives a reward $\rho > 0$. This could reflect a desire for peace on the part of a mediator in a conflict situation, or a bonus for a successful exchange in a trading scenario. If both sides defect the mediator receives zero. Thus ρ is a measure of how much the mediator prefers mutual cooperation or successful trade to mutual defection or no deal.

The mediator's attitude towards exploitation of one side by the other is defined with respect to the payoff the mediator gets if one side defects while the other cooperates. If player 1 unilaterally defects, the mediator receives β_1 and if player 2 unilaterally defects, the mediator receives β_2. If the mediator prefers mutual defection to either side being exploited, $\beta_1 < 0$ and $\beta_2 < 0$, I will call the mediator *exploitation averse* because it prefers no deal to exploitation by either side. If $\beta_1 > 0$ and $\beta_2 > 0$ I term the mediator *exploitation loving* because it prefers to see either side exploit the other rather than have mutual defection[7] If $\beta_1 > 0 \geq \beta_2$ the mediator is *biased* towards player 1, if $\beta_2 > 0 \geq \beta_1$ the mediator is biased towards player 2.[8] Finally, if $\beta_1 = \beta_2 = 0$, the mediator is said to be *exploitation indifferent*, since it does not care what happens if the players do not cooperate. Such a mediator would be solely concerned with increasing the likelihood of cooperation.

Nature starts the game by determining the player's types. I assume the payoff for exploiting the other side, b_i, is private information to each player. I also assume that this payoff is partly determined by the player's "character" or aspects unique to the player and partly determined by factors specific to the "issue" in question. Formally, assume that the payoff for exploitation has two components, $b_i = u_i + v_i$, where u_i is fixed and corresponds to the player's type, and v_i is variable and corresponds to some characteristic of the current issue.

[7] This preference ordering would seem to be rare empirically, and I show below that if the mediator's preferences are single peaked on some issue space the mediator cannot be exploitation loving.

[8] An alternative definition of bias would be to say that the mediator is biased towards player 1 if $\beta_1 > \beta_2$ and biased towards player 2 if the reverse holds. However, in this context, if both β_1 and β_2 are positive or negative, the difference between them is not strategically important. That is, the relationship between β_i and zero is more important than the relationship between β_1 and β_2.

Let u_i be distributed according to the probability density function (PDF) f_i (with cumulative density function (CDF) F_i) defined on the real line, while v_i is distributed according to g_i (CDF G_i), where g_i is single peaked with mean zero. Then b_i will be distributed according to the PDF

$$h_i(b_i) = \int_{-\infty}^{+\infty} f_i(u_i) g_i(b_i - u_i) du_i$$

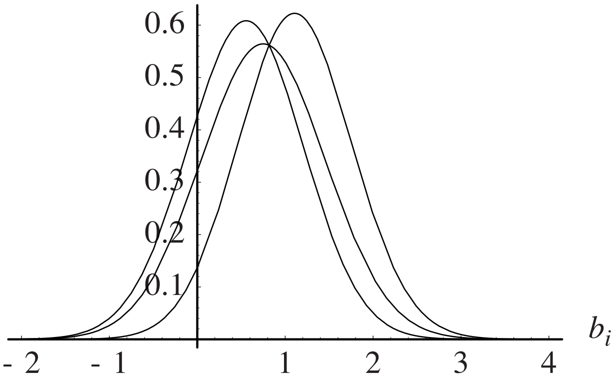
and the CDF

$$H_i(b_i^0) = \int_{-\infty}^{b_i^0} \int_{-\infty}^{+\infty} f_i(u_i) g_i(b_i - u_i) du_i db_i.$$

The central assumption behind the model is that the mediator will have some information to share with the parties about the other side's type. One way to conceive of this is to posit that the mediator observes the behavior of the parties on a previous issue, or hears a report about such behavior from some other party. The other context is related to the present one, but not perfectly predictive, so the mediator has a better idea of the parties' types, but not perfect information. Formally, assume that the mediator receives a signal from nature about the behavior of the parties in some previous context in which players with payoffs below a certain threshold, call it b_i^0, do one thing and types with payoffs above that threshold do another. If player i behaved such that $b_i < b_i^0$, the mediator receives a signal T_i indicating player i was trustworthy (had a lower payoff for exploiting the other side) while if player i behaved as the types with $b_i > b_i^0$ did then the mediator gets the U_i signal indicating player i was untrustworthy. Let $S_i = \{T_i, U_i\}$ be an ordered set where the order corresponds to the exploitation payoff, so $T_i \prec U_i$. I assume that the players know the mediator's information about themselves, but do not know the mediator's information about the other side.

What are the mediator's posterior beliefs? After receiving the T_i message the posterior

Figure 1: The Prior and Posterior PDFs



PDF over u_i is, from Bayes' rule,

$$f_i(u_i|T_i) = \frac{G_i(b_i^0 - u_i)}{H_i(b_i^0)} f_i(u_i)$$

and after getting the U_i message it is

$$f_i(u_i|U_i) = \frac{1 - G_i(b_i^0 - u_i)}{1 - H_i(b_i^0)} f_i(u_i).$$

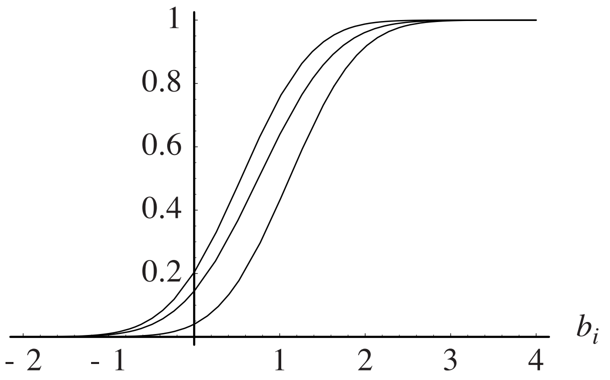
These can be substituted into the expressions for h_i and H_i to derive posterior beliefs about b_i, noting that g_i is not updated because it is issue specific and hence new each time.

These posterior beliefs are illustrated in Figure 1.[9] The central PDF is the prior belief over b_i. The curve to the left is the posterior belief after getting the T_i message so greater weight is given to lower payoff values. The curve to the right is the posterior belief after receiving the U_i message. Note the posterior beliefs are more concentrated about their means, reflecting greater certainty on the mediator's part after getting new information.

An important fact about the posterior CDFs, $H_i(b_i|S_i)$ is the following.

[9] In the illustration, f_i is normal with $\mu = 0.75$ and $\sigma = 0.5$, g_i is normal with $\mu = 0$ and $\sigma = 0.5$ and $b_i^0 = 1$.

Figure 2: The Prior and Posterior CDFs



Lemma 1 *The posterior CDF, $H_i(b_i^1|S_i)$, is decreasing in S_i, that is, $H_i(b_i^1|T_i) \geq H_i(b_i^1|U_i)$, provided that the variance of g_i is sufficiently small.*

Proof: See Appendix. ■

The lemma says that if the signal the mediator received is sufficiently informative, then if the mediator sees an indicator that the player has costs less than a certain threshold, the mediator will think it more likely that in a future instance the player will have low costs as well. The prior and posterior CDFs are illustrated in Figure 2. The central CDF is the prior beliefs about b_i. The curve to the left is the belief after receiving the T_i signal, the curve to the right is the posterior belief after getting the U_i signal.

The mediator then communicates with each player privately about the trustworthiness of the other player. The mediator can say that the other side is likely to be trustworthy, t_i, corresponding to the T_i signal, or that the party is likely to be untrustworthy, u_i, reporting the U_i signal. Let the mediator's communication be denoted $s_i \in \{t_i, u_i\}$. After the mediator communicates with each player, if the mediator is believed to be telling the truth, the parties' beliefs will shift to mirror the mediator's, so $h_i(b_i|s_i) = h_i(b_i|S_i)$, and $H_i(b_i|s_i) = H_i(b_i|S_i)$.

The mediator will of course take this into account when deciding whether or not to be honest.

After the mediator's announcement, the two players simultaneously choose to cooperate or defect. The notation in the game is summarized in the Appendix.

3 Equilibria in the Game

I solve for perfect Bayesian equilibria. In cheap talk games of this kind there are two broad categories of equilibria, "babbling" and "truthtelling." In babbling equilibria, the mediator's signals are uncorrelated with its information, and the players retain and act upon their prior beliefs. Babbling equilibria are possible regardless of the initial conditions.[10] In truthtelling equilibria, the mediator faithfully passes on its information, the players update their beliefs and act accordingly. For such an equilibrium to hold, the mediator's payoff for sending a signal must be conditional on the mediator's beliefs, such that if the mediator received the T_i signal, it would prefer to send the t_i signal rather than the u_i signal, but if it received the U_i signal, it would prefer to send the u_i signal rather than the t_i signal. The main focus of analysis is to determine the conditions under which truthtelling is sustainable in equilibrium. I first discuss the player's behavior and then the mediator's.

[10] The game can only go off the equilibrium path in a babbling equilibrium in which the mediator is expected to send one signal regardless of the signal received, but actually sends the other one. Therefore assumptions about off equilibrium path beliefs do not affect the conditions under which truthtelling is possible. For completeness, I assume that signals that have probability zero are perceived to be uncorrelated with the player's types. This keeps beliefs unchanged, supporting the babbling equilibrium.

3.1 The Player's Behavior

All types with $b_i > 1$ have a dominant strategy to defect. Types with $b_i \leq 1$ could cooperate given certain beliefs and expectations about behavior from the other side. Equilibrium behavior of the players is described in the following theorem.

Theorem 1 *In any equilibrium there will be cutoff points,* $b_i^* \in (-\infty, 1)$ *below which types will cooperate and above which types will not. One possible equilibrium in which no type cooperates is* $b_1^* = b_2^* = -\infty$*. If an equilibrium exists in which some types cooperate, the cutoff points must satisfy the following relation.*

$$b_i^* = 1 + a_i - \frac{a_i}{H_j(b_j^*|s_j)} \tag{1}$$

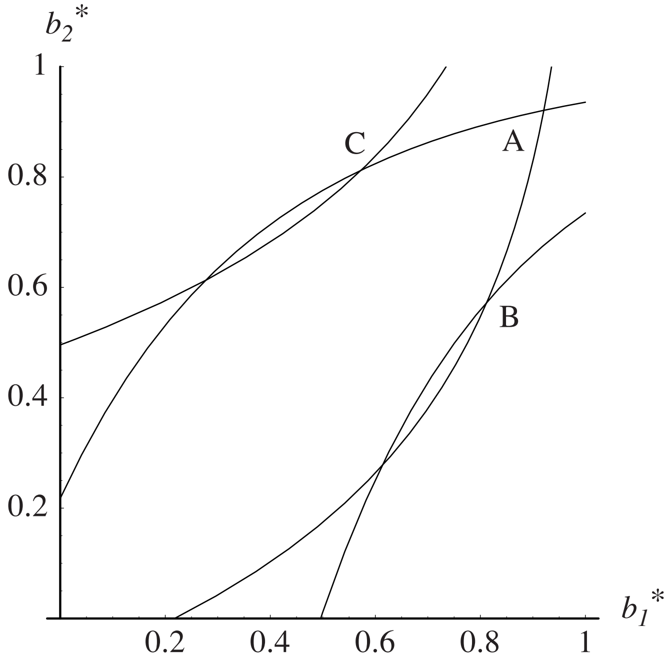
Proof: Let the likelihood that player j cooperates in equilibrium be θ_j. Player i's payoff for cooperation is $\theta_j 1 + (1-\theta_j)(-a_i)$ while the payoff for defection is $\theta_j b_i + (1-\theta_j)0$. Cooperation beats defection if player i's payoff for exploitation is below a threshold:

$$b_i \leq \frac{\theta_j + (1-\theta_j)(-a_i)}{\theta_j}$$

In a truthtelling equilibrium the other side's likelihood of cooperation is $H_j(b_j^*|s_j) = \theta_j$. In a babbling equilibrium the prior beliefs $H_j(b_j^*)$ can be substituted. ■

A "non-cooperative" equilibrium with $b_i^* = -\infty$ is possible regardless of beliefs. If the other side is expected to defect regardless of its payoffs, each side has an incentive to defect as well, to avoid the $-a_i$ payoff. If the h_i are especially weighted towards values greater than 1, this may be the only equilibrium. More cooperative equilibria may be possible in which there is some chance that each player cooperates. Since each equilibrium cutoff point is an increasing function of the other one, if there are multiple equilibria they can be ordered

Figure 3: Player Reaction Functions



by the likelihood of mutual cooperation. I will focus on the "most cooperative" one in which the likelihood of mutual cooperation is highest, given that its Pareto superiority has obvious focal appeal, and it has convenient and quite general comparative static properties (Milgrom and Roberts 1994; Milgrom and Shannon 1994). I restrict attention to the case where the posterior distributions, after the mediator's information is revealed, could support an equilibrium with $b_i^* > -\infty$ and hence some chance of cooperation. If this were not the case, mediation would be pointless.

The reaction functions are illustrated in Figure 3.[11] If both sides receive positive signals, the two players become more trusting and the reaction curves intersect at A, indicating a high likelihood of mutual cooperation. If player 2 gets the u_1 signal about player 1, player 2 becomes less trusting and the curves intersect at B, indicating that player 2 is less likely to cooperate than player 1. If instead player 1 received the u_2 signal, then the intersection would be at C, with player 2 more likely to cooperate than player 1. Finally, if both sides receive bad news about each other, then the reaction functions do not intersect and the players will defect regardless of type.

The central comparative static of interest with respect to the player's behavior is how the thresholds for cooperation respond to the mediator's information, assuming it to be honestly relayed. This is given in the following theorem.

Theorem 2 *In a truthtelling equilibrium, each player is more likely to cooperate if it is reassured about the other side, $b_i^*(s_j)$, is declining in s_j.*

Proof: From Lemma 1, we know that $H_j(b_j^*|s_j)$ is decreasing in s_j in a truthtelling equilibrium. Since b_i^* is increasing in $H_j(b_j^*|s_j)$ from Equation 1, the result is obtained. ■

What this means is that if the mediator tells a player that the other side is trustworthy, that player will be more likely to cooperate than if the mediator told it that the other side was untrustworthy. This is a natural result of the player becoming more trusting of the other side, and hence more willing to take a risk and cooperate.

[11] These are based on the same assumptions about information that underly Figure 1 and in addition $a_i = 0.2$.

3.2 The Mediator's Behavior

The mediator's payoff, denoted η, as a function of the signals it receives, S_i, and those it sends, s_i, is

$$\begin{aligned}\eta(s_1, s_2|S_1, S_2) \quad = \quad & \rho \times H_1(b_1^*(s_2)|S_1)H_2(b_2^*(s_1)|S_2) + \\ & \beta_1 \times (1 - H_1(b_1^*(s_2)|S_1))H_2(b_2^*(s_1)|S_2) + \\ & \beta_2 \times H_1(b_1^*(s_2)|S_1)(1 - H_2(b_2^*(s_1)|S_2)).\end{aligned} \tag{2}$$

In a babbling equilibrium, the H_i will not be affected by the s_j because the players will disregard the mediator and act on their prior beliefs. The mediator's payoff is therefore unaffected by its communication and it will be willing to play its role in the equilibrium. In a truthtelling equilibrium, the mediator's payoff will depend on its communication, and we can determine the conditions under which the mediator will tell the truth. These are reflected in the following theorem.

Theorem 3 *In the single round mediation game, for a truthtelling equilibrium to be possible the mediator must be exploitation averse.*

Proof: If the mediator is not exploitation averse, for at least one player i, $\beta_i \geq 0$. In that case, the mediator's payoff is increasing in $H_j(b_j^*|s_i)$, and from Theorem 2, sending the t_i signal will increase the payoff regardless of the signal received, eliminating truthtelling equilibria. ■

In the one round game, therefore, unless the mediator is exploitation averse, there cannot be a truthtelling equilibrium. That is, the mediator must prefer that the two sides both defect rather than that either side exploit the other, if the mediator is to be honest to both players. If the mediator is exploitation averse, there may be truthtelling equilibria. For instance, consider the symmetrical case where $\beta_1 = \beta_2 = \beta < 0$. In this case, given that $\rho - 2\beta > 0 > \beta$, the response of the payoff to changes in s_1 and s_2 will be conditional on how much each term in equation 2 increases, which will be conditional on the signal received. If we consider the simple case in which both parties need to be vouched for for any possibility of cooperation to exist, the problem reduces to verifying that $\eta(t_1, t_2 | T_1, T_2) > 0$ and that the reverse holds for any other set of signals from Nature.

An important corollary of this result concerns mediators that are solely concerned with achieving cooperation.

Corollary 1 *In the single round mediation game, if the mediator is exploitation indifferent, a truthtelling equilibrium will not be possible.*

If the mediator just wants to prevent conflict and achieve cooperation, so that $\rho > 0$ and $\beta_1 = \beta_2 = 0$, its payoff from Equation 2 will reduce to a simple function of the likelihood of mutual cooperation. It will therefore have an incentive to say whatever will maximize the chance that both sides cooperate. If the players will believe the mediator, it will have an incentive to vouch for both of them, because this makes them more likely to cooperate. This incentive will hold regardless of the mediator's information. Therefore the mediator who just wants to promote cooperation will vouch for the players regardless of its beliefs about their trustworthiness, and so it will not be capable of sustaining a truthtelling equilibrium. Hence it will have no ability to reassure the players, and no impact on the likelihood of cooperation. Note, this result in the trustbuilding context is the same as that found in case of communication about power and resolve. In the models by Kydd and Smith and Stam, a mediator who just wants to avoid war (biased in favor of peace according to Smith and Stam, unbiased according to Kydd) will not be effective.

A final corollary concerns biased mediators.

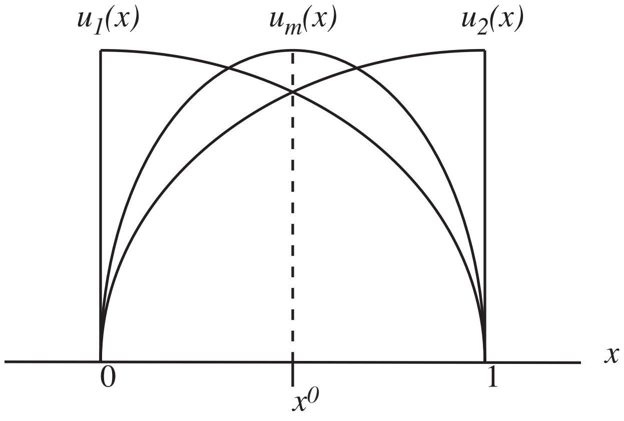
Corollary 2 *In the single round mediation game, if the mediator is biased, a truthtelling equilibrium will not be possible.*

Bias towards a player makes the mediator wish to encourage the other side to cooperate, regardless of the truth. Biased mediators, in the one round game, are therefore incapable of providing mutual reassurance. Note, if the mediator was biased towards player 1, such that $\beta_1 > 0 > \beta_2$, the mediator could be honest to player 1 about player 2. Since the mediator does not want player 1 to be exploited, it could be trusted to tell player 1 the truth about player 2. However, it could not be honest to player 2 about player 1, since the mediator would prefer that player 2 be exploited, and so would always wish to encourage player 2 to cooperate. Only exploitation aversion enables the mediator to be honest to both parties in a trustbuilding role.

4 Sources and Substitutes for Exploitation Aversion

While exploitation aversion may seem like a desirable quality because it sustains honesty and facilitates trustbuilding, it may also seem to be a rare commodity, given that it involves some degree of sympathy for both parties. This raises the question, what are the sources of exploitation aversion? Why might some mediators be exploitation averse and others not? Also, are there functional substitutes for exploitation aversion, factors which could encourage exploitation indifferent or biased mediators to behave as if they were exploitation averse? I

Figure 4: The Bargaining Space



will focus on one source and one substitute: moderate ideal points in bargaining contexts and repeated play respectively.

4.1 Moderate Preferences, Low Costs of Conflict

Consider the conflict resolution context. Assume that two states or a government and a rebel group have fought a conflict to a stalemate and are contemplating a peace accord. The issue space over which they have been fighting is denoted x conceived of as the unit interval, and player 1's ideal point is 0 whereas player 2's is 1. Each player has a utility function over the interval, $u_1(x)$, $u_2(x)$, and $u_m(x)$, which is non-increasing in either direction from their ideal points. Let the utilities for the ideal points be normalized to one and the utilities of the bargainers for the other side's ideal point be normalized to zero. Let the negotiated deal on the table be $x^0 \in (0, 1)$. The bargaining space is illustrated in Figure 4.

Model conflict as having three possible outcomes, victory, defeat, and stalemate. Assume that the stalemate outcome is identical with the deal under consideration, so that player i's utility for stalemate is $u_i(x^0)$. Let player i's chance of winning the conflict if both sides reject the deal be $\pi_i \in (0, 1)$, where $\pi_1 + \pi_2 < 1$ because of the possibility of stalemate. However, assume that if player j accepts the deal and implements it, while player i reneges and resumes the conflict, player i's chance of winning goes up to $\pi_i + \phi_i$, while the chance of stalemate declines to $1 - \pi_i - \pi_j - \alpha\phi_i$ and player j's chance of victory decreases to $\pi_j - (1 - \alpha)\phi_i$, where the three probabilities are bounded by zero and 1. The α term just represents the fraction of the increase in likelihood of winning that comes at the expense of the likelihood of stalemate, while $1 - \alpha$ is the fraction that comes from the likelihood of defeat, $\alpha \in (0, 1)$. Each player suffers a cost of conflict, c_1, c_2, and c_m. The payoffs in this game are shown in Table 2 for player 1 and the mediator for each of the four possible outcomes: mutual cooperation, mutual defection, player 1 cooperating and player 2 defecting, and vice versa.

These payoffs can easily take on the same ordering as the payoffs in the game just analyzed. First consider the players. If we assume that conflict is sufficiently costly and the deal on the table reflects the current balance of power, then both sides will prefer mutual cooperation to mutual defection, $u_i(x^0) > \pi_i + (1 - \pi_i - \pi_j)u_i(x^0) - c_i$, as assumed above. Given that $\phi_i > 0$ and $u_i(x^0) \leq 1$, each side will prefer to renege unilaterally rather than return to conflict simultaneously, and cooperating while the other side reneges will be the worst outcome, as before. Finally, a player could prefer to renege rather than cooperate, if the costs of war are low enough,

$$c_i < \pi_i + \phi_i - (\pi_i + \pi_j + \alpha\phi_i)u_i(x^0).$$

If we then assume that the costs for conflict are private information, distributed according

Table 2: The Payoffs in the Bargaining Context

Player 1	
CC	$u_1(x^0)$
DD	$\pi_1 + (1 - \pi_1 - \pi_2)u_1(x^0) - c_1$
DC	$\pi_1 + \phi_1 + (1 - \pi_1 - \pi_2 - \alpha\phi_1)u_1(x^0) - c_1$
CD	$\pi_1 - (1 - \alpha)\phi_2 + (1 - \pi_1 - \pi_2 - \alpha\phi_2)u_1(x^0) - c_1$
Mediator	
CC	$u_m(x^0)$
DD	$\pi_1 u_m(0) + (1 - \pi_1 - \pi_2)u_m(x^0) + \pi_2 u_m(1) - c_m$
DC	$(\pi_1 + \phi_1)u_m(0) + (1 - \pi_1 - \pi_2 - \alpha\phi_1)u_m(x^0) + (\pi_2 - (1 - \alpha)\phi_1)u_m(1) - c_m$
CD	$(\pi_1 - (1 - \alpha)\phi_2)u_m(0) + (1 - \pi_1 - \pi_2 - \alpha\phi_2)u_m(x^0) + (\pi_2 + \phi_2)u_m(1) - c_m$

to $k_i(c_i)$ (CDF $K_i(c_i)$) over $[0, +\infty]$, the game closely resembles the previous one, with the exception that the value for mutual defection is private as well. Types with low enough costs will prefer unilateral defection, types with higher costs will prefer mutual cooperation, but fear exploitation. If types with costs greater than c_j^* cooperate, cooperation beats defection if $(1-K_j(c_j^*))u_i(x^0)+K_j(c_j^*)(\pi_i-(1-\alpha)\phi_j+(1-\pi_i-\pi_j-\alpha\phi_j)u_i(x^0)-c_i) > (1-K_j(c_j^*))(\pi_i+\phi_i+(1-\pi_i-\pi_j-\alpha\phi_i)u_i(x^0)-c_i)+K_j(c_j^*)(\pi_i+(1-\pi_i-\pi_j)u_i(x^0)-c_i))$ so that the cutoff point for player i will be

$$c_i^* = \pi_i + \phi_i - (\pi_i + \pi_j + \alpha\phi_i)u_i(x^0) + \frac{K_j(c_j^*)}{1-K_j(c_j^*)}(1-\alpha+\alpha u_i(x^0))\phi_j \tag{3}$$

These critical values are positive functions of each other as before, the only difference is that here the Pareto superior equilibrium is the one in which the players with the lowest possible costs cooperate. Otherwise the analysis is identical, so the parties' behavior will respond to information from the mediator in the same way, namely, that vouching for a player will increase the likelihood that the other player cooperates.

Now consider the mediator's payoffs. If the mediator does not care about the issue dimension at all, so that $u_m(x^0) = u_m(0) = u_m(1) = 1$, but does care about preventing a return to conflict, so that $c_m > 0$, its payoff will be 1 for mutual cooperation and $1 - c_m$ for any of the conflict outcomes. Strategically, this means that the mediator will be exploitation indifferent. Therefore, the mediator will have an incentive to vouch for the parties even if it gets bad signals about them. If the mediator has no preferences over the issue in dispute, therefore, the mediator cannot be trusted to tell the truth. This also implies that there will be an upper bound on the costs of war in a truthelling equilibrium. If the costs of war, c_m, are too great, the mediator's preferences will approximate the case in which the mediator is indifferent among the issue resolutions but just wants to avoid war. In this case the temptation to say whatever maximizes the chance of peace, namely that the two parties are trustworthy, will be irresistible.

If the mediator's ideal point is far enough to the left, then the mediator's preference order over the issue space will be $u_m(0) > u_m(x^0) > u_m(1)$, and its preferences over the outcomes in the game will be that player 1 exploiting player 2 will be preferred to mutual defection, which beats player 2 reneging on player 1. As defined earlier, the mediator will be biased towards player 1. This is because reneging increases the chance of one side winning, so if the mediator prefers player 1's ideal point to the stalemate, then it will prefer that player 1 renege rather than that both defect. The same logic applies if the mediator's ideal point is far to the right, the mediator will prefer to have player 2 exploit player 1 rather than have both defect. Therefore, if the mediator's ideal point is too extreme, then the mediator will be biased and will be incapable of supporting a truthtelling equilibrium.[12] I call such preferences *extreme*, otherwise the mediator's preferences are *moderate*. The concept of moderate preferences is similar to Rauchhaus's definition of neutrality, having an ideal point between the ideal points of the players, or between two relevant offers on the table (Rauchhaus 2005, 14).

Finally, consider a mediator with moderate preferences that is not indifferent over the outcomes. Such a mediator would have preferences such that $u_m(x^0) > \max\{u_m(0), u_m(1)\}$. In this case, the mediator could prefer mutual cooperation first, mutual defection second, and exploitation by either side third and fourth. This is because unilateral exploitation makes

[12] For instance, if the mediator's ideal point lies outside the interval between the two parties' ideal points, then it will have extreme preferences. However, the ideal point need not fall outside this interval to generate extreme preferences.

extreme outcomes, 1 or 0, more likely in expectation than mutual defection, because the chance of a stalemate is reduced by some fraction α of the advantage of reneging unilaterally. For instance, if $u_m(x^0) = 1$ and $u_m(0) = u_m(1) = 0$, as illustrated in Figure 4, then the mutual defection payoff will be $1-\pi_1-\pi_2-c_m$ which will beat the payoff for player 1 exploiting player 2, $1-\pi_1-\pi_2-\alpha\phi_1-c_m$, and player 2 exploiting player 1, $1-\pi_1-\pi_2-\alpha\phi_2-c_m$. Reneging reduces the likelihood of moderate outcomes, so if the mediator has moderate preferences it will prefer mutual defection to exploitation by either side, producing exploitation aversion.[13]

These results are summed up in the following theorem.

Theorem 4 *In the conflict resolution context, a truthtelling equilibrium is possible only if a mediator has sufficiently low costs of war, c_m, and moderate preferences, such that $u_m(x^0) > \max\{u_m(0), u_m(1)\}$.*

In the conflict resolution setting, therefore, if a mediator is to be credible in a trust building role, it must not prefer one side's ideal point to the deal on the table. Instead, it must have preferences for moderate outcomes, so that it has an incentive to prevent exploitation by either side, even while it also seeks to promote cooperation. Finally, it cannot be too sensitive to the costs of conflict. In particular, it cannot be so averse to conflict that it will say anything to reduce its likelihood. Once again, we see the seeming paradox that those who are most concerned to avoid war will be least capable of preventing it.

Bias is fatal to trustbuilding in the one round setting. This result contrasts with Kydd's finding that bias is necessary for honesty in conveying information about resolve (Kydd

[13] Note, given that the mediator's preferences are single peaked and $x^0 \in (0, 1)$, it cannot be that $u_m(x^0) < \min\{u_m(0), u_m(1)\}$, which means the mediator cannot be exploitation loving.

2003). This discrepancy is a result of certain differences in the models. In Kydd's model, the mediator was assumed to share one side's preference ranking over the possible issue resolutions, or be indifferent to the issue. The possibility of an interior ideal point was not considered, and this turns out to be the key to truthtelling in the trust building context. With an interior ideal point, the mediator can be in a sense biased towards both sides, in that the mediator shares with each a preference that they not be exploited.

A moderate ideal point and a low cost of conflict can therefore produce exploitation aversion in a mediator, supporting honesty in a truthtelling equilibrium. But what if the mediator is not exploitation averse?

4.2 Repetition and the Shadow of the Future

If a mediator is not exploitation averse, a possible substitute for exploitation aversion is the desire to preserve a reputation for honesty, so as to ensure future employment. Consider a repeated version of the first game. The players are chosen anew each round, so the information structure is the same. The mediator is the same each round, with a discount factor δ. The mediator is "hired" for the round provided it has never said that a player is trustworthy who subsequently defects. If the mediator does make this mistake, it is never hired again.[14] For all rounds in which the mediator is not hired, it receives a payoff of zero. I consider perfect Bayesian equilibria, so that the mediator's choice at each round must be optimal given its information. As in the one round game, untrustworthy types will defect in any equilibrium while trustworthy types can cooperate if they are trusting enough.

[14] We can imagine a pool of mediators available such that the parties can select a different mediator if the previous mediator has been fired.

The payoff to the mediator for the equilibrium strategy is a function of the per round payoff and the likelihood of passing to the next round. Equation 2 gives the mediator's payoff conditional on the signals received and sent. To derive the ex-ante per period payoff, this must be expanded assuming that signals sent match signals received. The expansion is shown in the Appendix. Denote the ex-ante per-round payoff for telling the truth η_t. The likelihood that the mediator gets to the next round is the likelihood that the player(s) the mediator vouches for cooperates. The ex-ante likelihood of getting to the next round by telling the truth is denoted γ_t and the formula is also in the Appendix. The equilibrium payoff in the game can then be written as

$$\eta_t + \delta\gamma_t\eta_t + (\delta\gamma_t)^2\eta_t + \cdots = \frac{\eta_t}{1-\delta\gamma_t}.$$

This is also the continuation payoff from any point forward, assuming honest behavior from then on.

To determine when the equilibrium is sustainable, we must consider the payoffs for deviations ex-post, after the mediator receives its information. Denote the payoff from telling the truth in any round given the signals received as $\eta_t(S_1, S_2)$ and the corresponding payoff from a contemplated deviation as $\eta_d(S_1, S_2)$. The likelihood of getting to the next round will also be a function of the signals and the contemplated deviation, denote these as $\gamma_t(S_1, S_2)$ and $\gamma_d(S_1, S_2)$. The payoff for sticking to the truth can then be written as

$$\eta_t(S_1, S_2) + \delta\gamma_t(S_1, S_2)\frac{\eta_t}{1-\delta\gamma_t}$$

while the payoff for a one round deviation[15] is

$$\eta_d(S_1, S_2) + \delta\gamma_d(S_1, S_2)\frac{\eta_t}{1-\delta\gamma_t}$$

[15] If no one round deviation is profitable, no longer one will be. If a one round deviation is not profitable,

Following the equilibrium payoff will beat the deviation if

$$\eta_d(S_1, S_2) - \eta_t(S_1, S_2) \leq (\gamma_t(S_1, S_2) - \gamma_d(S_1, S_2))\delta \frac{\eta_t}{1 - \delta\gamma_t} \tag{4}$$

Note that for the equilibrium to hold, if a deviation improves the payoff in the current round it must reduce the likelihood of getting to the next round, or if it improves the likelihood of getting to the next round, it must reduce the payoff in the current round. If a deviation improved both the current round payoff and the likelihood of getting to the next round, the equilibrium would be unsustainable. Also, as usual in such games, the discount factor, δ, must not be too small or the mediator will have an incentive to do whatever maximizes the current round payoff.

In order to determine when this condition holds for various possible deviations, we need to consider how the payoffs and likelihood of getting to the next round vary with the signals sent and received. We can write out the likelihood of getting to the next round more fully as a function of the signals received and sent as follows.

$$\begin{aligned} \gamma(t_1, t_2|S_1, S_2) &= H_1(b_1^*(t_2)|S_1)H_2(b_2^*(t_1)|S_2) \\ \gamma(u_1, t_2|S_1, S_2) &= H_2(b_2^*(u_1)|S_2) \\ \gamma(t_1, u_2|S_1, S_2) &= H_1(b_1^*(u_2)|S_1) \\ \gamma(u_1, u_2|S_1, S_2) &= 1 \end{aligned}$$

an n round deviation will be beat by an $n-1$ round deviation where the mediator returns to honesty one round earlier. This will unravel to the first round, so any finite deviation will be beat by the equilibrium. The payoff for an infinite deviation will be the limit of the payoffs of finite deviations of increasing lengths, which are decreasing, all of which are beat by the equilibrium, so the infinite deviation will be beat by the equilibrium as well.

The likelihood of getting to the next round is increasing in the signal sent, s_i, regardless of the signal received. The signal received determines how much it increases, however. The key result is that the likelihood of getting to the next round exhibits increasing differences in the signals sent and received, that is, $\gamma(u_i|T_i) - \gamma(t_i|T_i) < \gamma(u_i|U_i) - \gamma(t_i|U_i)$. This means that the increase in the likelihood of getting to the next round from warning a player that the other side is untrustworthy is bigger if the mediator has received the signal that the other side actually is untrustworthy.

Now return to Equation 2, and consider the case where the mediator is biased towards player 1, $\beta_1 > 0 \geq \beta_2$. The mediator's payoff is decreasing in s_1, because sending the u_1 signal causes player 2 to be less likely to cooperate, reducing the likelihood of mutual cooperation and player 1 exploiting player 2. However, if $\rho > \beta_1$, the payoff exhibits decreasing differences, that is, $\eta(t_1, s_2|U_1, S_2) - \eta(u_1, s_2|U_1, S_2) < \eta(t_1, s_2|T_1, S_2) - \eta(u_1, s_2|T_1, S_2)$. In words, the increase in the payoff of vouching for player 1 is stronger when the signal received indicates that player 1 is trustworthy, than when the signal indicates otherwise.

Putting these two facts together means that after receiving the U_1 signal, the increase in the payoff for deviating from the u_1 signal to the t_1 signal will be small and the decrease in the likelihood of getting to the next round will be large. This provides a disincentive to lie and means that the condition in Equation 4 may hold. Conversely, if the T_1 signal is received, the payoff loss from deviating from t_1 to u_1 will be large and the increase in likelihood of getting to the next round will be small, once again punishing dishonesty and making Equation 4 more likely to be satisfied. These considerations have the potential to make the overall payoff conditional on the signals received, making truthtelling equilibria possible.

Intuitively, the fact that the game is repeated can give the mediator an incentive to be honest. The mediator knows that if it is caught in a lie, saying a player is trustworthy who then defects when a trustworthy player would have cooperated, it will not be hired again. Being hired again is valuable because of the opportunity to get more rewards ρ for successful cooperation. So the mediator may not wish to jeopardize its chances of getting to the next round by vouching for the players when it has received information that they are not that trustworthy. Saying that they are untrustworthy is a sure way to get to the next round because the mediator has not vouched for players who subsequently defect.

However, too much bias on the part of the mediator can still make the truthtelling equilibrium break down, rendering the mediator incapable of fostering cooperation. For instance, if $\beta_i > \rho$, the increase in payoff for vouching for player i will actually be higher when the player is believed to be untrustworthy, after the U_i signal. If the effect is small it could still be overwhelmed by the decrease in likelihood of getting to the next round, but as β_i increases this effect will eventually dominate, making a truthtelling equilibrium impossible to sustain. Excess bias will lead to the mediator not being honest about the side they are biased towards, rendering the mediator useless. These results are summarized in the following theorem.

Theorem 5 *In the repeated mediation game, there can be a truthelling equilibrium with mediators who are biased or exploitation indifferent.*[16] *However, there will be an upper limit on the payoffs for exploitation, β_1 and β_2, beyond which there will be no truthtelling equilibrium.*

[16] Or exploitation loving.

5 Discussion

Mediators in a one round game must be exploitation averse to be honest brokers when it comes to trust building. That is, they must prefer that no deal be struck rather than that one side exploit the other, no matter which side does the exploiting. The more biased a mediator is, the greater danger that it will lie about the side towards whom it is biased, in order to encourage cooperation by the other in spite of the risk that its favored side will exploit the other side. For this reason, biased mediators themselves will be less trusted by the party against whom they are biased because they have an incentive to deceive them about the trustworthiness of the other party. In situations where the mediator is trying to build trust, therefore, the mediator should be exploitation averse, not biased.

The model thus supports the intuition that mediators should be unbiased to build trust, however, it clarifies the issue by providing a precise definition of bias, and showing that it is not actually bias but exploitation aversion that is the key property. Countless scholars and practitioners have noted the advantage of neutrality in mediation without being very rigorous about what they mean by it and why it works. Young argues that "a meaningful role for a third party will depend on the party's being perceived as an impartial participant, (in the sense of having nothing to gain from aiding either protagonist)" (Young 1967, 81). In fact, however, the mediator must have "something to gain" because in the conflict resolution context it must have clear preferences for moderate outcomes and so wish to achieve some favored issue resolution. Senator George Mitchell recalled that when he began mediating the conflict in Northern Ireland, "I again pledged to act in a fair and impartial manner and assured them that my only interest was to be helpful to them and to the people of Northern

Ireland." He also recounts how the previous negotiator had been viewed as "too close to the British government" and how two Unionist parties walked out initially, viewing Mitchell as "the equivalent of appointing an American Serb to preside over talks on the future of Croatia" (Mitchell 1999, 47,53). Mitchell would certainly be incapable of trustbuilding if he were perceived as too close to one side or the other. However, Mitchell's statement could be read as saying he had no preference over the issues in dispute and only wanted the conflict to be resolved, which the model shows would also have made him non-credible as a trust builder. Indeed, of all the findings of the model, the fact that mediators who simply want to avoid war will be incapable of lessening its likelihood seems the most counterintuitive and absent from the writings of practitioners.

Repetition can provide an acceptable substitute for exploitation aversion. Mediators that are biased, but not too biased, can still be honest brokers if they could lose their position by being caught vouching for players who subsequently betray the other side. This provides a theoretical rationale for Wehr and Lederach's observation that the parties to conflict in Central America prefer mediators who have ties to the community even if they are somewhat biased, over outsiders who are neutral. Wehr and Lederach write,

> The insider-partial is the "mediator from within the conflict" whose acceptability to the conflictants is rooted not in distance from the conflict or objectivity regarding the issues, but rather in connectedness and trusted relationships with the conflict parties. The trust comes partly from the fact that the mediators do not leave the postnegotiation situation. . . They must continue to relate to conflictants who have trusted their commitment to a just and durable settlement.

(Wehr and Lederach 1991, 87).

The model clarifies the strategic logic of this case. A somewhat biased mediator in a repeated setting may be honest, whereas in a one round game, a mediator who is indifferent over the issue outcomes and just wants to resolve the conflict will not be credible. Therefore the parties may be right to prefer a moderately biased intermediary who has lasting ties to the community.

Bercovitch and Houston also find, in a quantitative study of 364 post-1945 mediation efforts, that "another strong effect in our data concerns the importance of a continuing relationship, especially one that may extend into the future. Mediators from the same bloc tend to be more successful and to use a different pattern of strategies than other mediators. Mediation works best when the parties and the mediator share some bonds and are part of a recognizable network of interdependence" (Bercovitch and Houston 1993, 317). The different strategies mentioned are "communication facilitation" strategies, which correspond more closely to trustbuilding than the other two categories, "procedural," and "directive".

Repetition is perhaps more likely to be a feature in the exchange under uncertainty scenario than in conflict resolution. In the exchange setting, potential mediators have many interactions over time and no incentive to privilege one over the others. The merchant judges of the Champaign fairs had long term investments in their honesty and little incentive to vouch for an untrustworthy actor. In international conflict resolution, mediators may have a much greater payoff in case of success, in terms of political rewards, and much less concern with the future (high ρ, low δ). When President Clinton mediated between Arafat and Barak at Camp David in 2000, the reward for a successful settlement would have been tremendous, while Clinton's concern about future mediation opportunities would have been minimized by uncertainty that there would ever be a next time, at least of such magnitude. However, long term mediators associated with international organizations such as the UN could perhaps have the same incentive to preserve a reputation for honesty. States, such as the Scandinavian countries, that specialize in international conflict resolution could also establish incentives for their diplomats to preserve the national reputation for exploitation aversion.

6 Conclusion

In attempting to promote cooperation, mediators must work on the causes of conflict that are within their power to affect. Uncertainty is one such cause of conflict; information can be provided by actors that are otherwise powerless. One form of uncertainty is mistrust, a belief that the other side may prefer to exploit one's cooperation rather than reciprocate it. When mediators attempt to overcome mistrust, they must be exploitation averse, or prefer that neither side cooperate rather than one side exploit the other. In a conflict resolution context, they must actively prefer moderate issue resolutions and not be overly concerned with avoiding war at all costs, such a mediator would say anything to avoid conflict, and hence could not be trusted. Repetition may create a shadow of the future which is sufficient to overcome a limited amount of bias, so that exploitation indifferent or biased mediators will act as if they are exploitation averse.

Appendix

Notation in the game is summarized in Table 3.

Proof of Lemma 1

The posterior CDF, $H_i(b_i^1|S_i)$, is decreasing in S_i, that is, $H_i(b_i^1|T_i) \geq H_i(b_i^1|U_i)$, provided that the variance of g_i is sufficiently small.

Proof: We wish to show that

$$\int_{-\infty}^{b_i^1}\int_{-\infty}^{+\infty}\frac{G_i(b_i^0-u_i)}{H_i(b_i^0)}f_i(u_i)g_i(b_i-u_i)du_idb_i >$$
$$\int_{-\infty}^{b_i^1}\int_{-\infty}^{+\infty}\frac{1-G_i(b_i^0-u_i)}{1-H_i(b_i^0)}f_i(u_i)g_i(b_i-u_i)du_idb_i$$

Simplifying, we get

$$\int_{-\infty}^{b_i^1}\int_{-\infty}^{+\infty}\left(\frac{G_i(b_i^0-u_i)}{H_i(b_i^0)}-\frac{1-G_i(b_i^0-u_i)}{1-H_i(b_i^0)}\right)f_i(u_i)g_i(b_i-u_i)du_idb_i > 0$$

$$\int_{-\infty}^{b_i^1}\int_{-\infty}^{+\infty}\frac{G_i(b_i^0-u_i)-H(b_i^0)}{H_i(b_i^0)(1-H_i(b_i^0))}f_i(u_i)g_i(b_i-u_i)du_idb_i > 0$$

$$\int_{-\infty}^{b_i^1}\int_{-\infty}^{+\infty}(G_i(b_i^0-u_i)-H(b_i^0))f_i(u_i)g_i(b_i-u_i)du_idb_i > 0$$

$$\int_{-\infty}^{+\infty}(G_i(b_i^0-u_i)-H(b_i^0))G_i(b_i^1-u_i)f_i(u_i)du_i > 0$$

$$\int_{-\infty}^{+\infty}G_i(b_i^0-u_i)G_i(b_i^1-u_i)f_i(u_i)du_i > H(b_i^0)H(b_i^1)$$

$$\int_{-\infty}^{+\infty}G_i(b_i^0-u_i)G_i(b_i^1-u_i)f_i(u_i)du_i >$$
$$\int_{-\infty}^{+\infty}G_i(b_i^0-u_i)f_i(u_i)du_i\int_{-\infty}^{+\infty}G_i(b_i^1-u_i)f_i(u_i)du_i$$

The smaller the variance of g_i, the more closely the G_i functions approach step functions that go down from 1 to zero at b_i^0 and b_i^1. The left hand side integral then approximates

Table 3: Notation in the Game

$-a_i$	Payoff for unilateral cooperation
b_i, $h_i(b_i)$, $H_i(b_i)$	Payoff for unilateral defection, its PDF and CDF
ρ	Mediator's reward for successful agreement
β_i	Mediator's payoff if player i exploits player j
u_i, $f_i(u_i)$, $F_i(u_i)$	Fixed component of b_i, its PDF and CDF
v_i, $g_i(v_i)$, $G_i(v_i)$	Issue specific component of b_i, its PDF and CDF
$S_i = \{T_i, U_i\}$, $s_i = \{t_i, u_i\}$	Signals from nature and the mediator about player i's type
x^0	The deal on the table and issue resolution from stalemate
c_i, k_i, K_i	Player i's costs of conflict, its PDF and CDF
π_i	Player i's chance of winning a conflict
ϕ_i	Player i's first strike advantage
α	Fraction of the first strike advantage deducted from likelihood of stalemate
δ	Mediator's discount factor
η	Mediator's per round payoff in repeated game
γ	Mediator's likelihood of passing to the next round

$F(b_i^0)$ if $b_i^0 < b_i^1$, and $F(b_i^1)$ if the reverse holds. The right hand side approaches $F(b_i^0)F(b_i^1)$, which is smaller than either.

■

The Values of η_t and γ_t

The mediator's ex-ante per round payoff assuming truthtelling is

$$\begin{aligned}
\eta_t \quad = \quad & \rho \times [H_1(b_1^*(t_2)|T_1)H_2(b_2^*(t_1)|T_2)H_1(b^0)H_2(b^0) + \\
& H_1(b_1^*(t_2)|U_1)H_2(b_2^*(u_1)|T_2)(1 - H_1(b^0))H_2(b^0) + \\
& H_1(b_1^*(u_2)|T_1)H_2(b_2^*(t_1)|U_2)H_1(b^0)(1 - H_2(b^0)) + \\
& H_1(b_1^*(u_2)|U_1)H_2(b_2^*(u_1)|U_2)(1 - H_1(b^0))(1 - H_2(b^0))] + \\
& \beta_1 \times [(1 - H_1(b_1^*(t_2)|T_1))H_2(b_2^*(t_1)|T_2)H_1(b^0)H_2(b^0) + \\
& (1 - H_1(b_1^*(t_2)|U_1))H_2(b_2^*(u_1)|T_2)(1 - H_1(b^0))H_2(b^0) + \\
& (1 - H_1(b_1^*(u_2)|T_1))H_2(b_2^*(t_1)|U_2)H_1(b^0)(1 - H_2(b^0)) + \\
& (1 - H_1(b_1^*(u_2)|U_1))H_2(b_2^*(u_1)|U_2)(1 - H_1(b^0))(1 - H_2(b^0))] + \\
& \beta_2 \times [H_1(b_1^*(t_2)|T_1)(1 - H_2(b_2^*(t_1)|T_2))H_1(b^0)H_2(b^0) + \\
& H_1(b_1^*(t_2)|U_1)(1 - H_2(b_2^*(u_1)|T_2))(1 - H_1(b^0))H_2(b^0) + \\
& H_1(b_1^*(u_2)|T_1)(1 - H_2(b_2^*(t_1)|U_2))H_1(b^0)(1 - H_2(b^0)) + \\
& H_1(b_1^*(u_2)|U_1)(1 - H_2(b_2^*(u_1)|U_2))(1 - H_1(b^0))(1 - H_2(b^0))]
\end{aligned}$$

The ex-ante likelihood of getting into the next round is

$$\begin{aligned}
\gamma_t \quad = \quad & 1 - \{H_1(b^0)[(1 - H_1(b_1^*(t_2)|T_1))H_2(b^0) + (1 - H_1(b_1^*(u_2)|T_1))(1 - H_2(b^0))] + \\
& H_2(b^0)[(1 - H_2(b_2^*(t_1)|T_2))H_1(b^0) + (1 - H_2(b_2^*(u_1)|T_2))(1 - H_1(b^0)))]\}.
\end{aligned}$$

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality, Uncertainty and the Market Mechanism. *Quarterly Journal of Economics 84*(3), 488–500.

Bercovitch, J. and A. Houston (1993). Influence of Mediator Characteristics and Behavior on the Success of Mediation in International Relations. *The International Journal of Conflict Management 4*(4), 297–321.

Biglaiser, G. (1993). Middlemen as Experts. *RAND Journal of Economics' 24*(2), 212–223.

Blainey, G. (1988). *The Causes of War* (Third ed.). New York: The Free Press.

Braithwaite, V. and M. Levi (Eds.) (1998). *Trust and Governance*. New York: Russel Sage Foundation.

Burton, J. W. (1969). *Conflict and Communication: the Use of Controlled Communication in International Relations*. New York: Free Press.

Carment, D. and D. Rowlands (1998). Three's Company: Evaluating Third-Party Intervention in Intrastate Conflict. *Journal of Conflict Resolution 42*(5), 572–599.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: Belknap Press.

Crescenzi, M. J. C., K. M. Kadera, S. M. Mitchell, and C. L. Thyne (2005). A Supply Side Theory of Third Party Conflict Management. *Manuscript*.

Farrell, J. and M. Rabin (1996). Cheap Talk. *Journal of Economic Perspectives 10*(3), 103–118.

Favretto, K. (2005). Does Love Make War Fair? Mediator Bias and Coercion in International Mediation. *manuscript UCLA*.

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

Fisher, R. J. (1972). Third Party Consultation: A Method for the Study and Resolution of Conflict. *Journal of Conflict Resolution 16*(1), 67–94.

Fisher, R. J. (1983). Third Party Consultation as a Method of Intergroup Conflict Resolution: A Review of Studies. *Journal of Conflict Resolution 27*(2), 301–334.

Fisher, R. J. and L. Keashly (1991). The Potential Complementarity of Mediation and Consultation within a Contingency Model of Third Party Intervention. *Journal of Peace Research 28*(1), 29–42.

Forges, F. (1986). An Approach to Communication Equilibria. *Econometrica 54*(6), 1375–1385.

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity.* New York: The Free Press.

Gambetta, D. (Ed.) (1988). *Trust: Making and Breaking Cooperative Relations.* New York: Basil Blackwell.

Glaser, C. L. (1995). Realists as Optimists: Cooperation as Self Help. *International Security 19*(3), 50–90.

Hardin, R. (2002). *Trust and Trustworthiness.* New York: Russell Sage Foundation.

Herz, J. H. (1950). Idealist Internationalism and the Security Dilemma. *World Poli-*

tics 2(2), 157–180.

Hobbes, T. (1968 (1651)). *Leviathan*. New York: Penguin.

Jarque, X., C. Ponsati, and J. Sakovics (2003). Mediation: Incomplete Information Bargainign with Filtered Communication. *Journal of Mathematical Economics 39*(7), 803–830.

Jervis, R. (1976). *Perception and Misperception in International Politics*. Princeton: Princeton University Press.

Jervis, R. (1978). Cooperation under the Security Dilemma. *World Politics 30*(2), 167–214.

Kelman, H. C. (1997). Some Determinants of the Oslo Breakthrough. *International Negotiation 2*, 183–194.

Kelman, H. C. (2000). The Role of the Scholar-Practitioner in International Conflict Resolution. *International Studies Perspectives 1*, 273–288.

Kriesberg, L. (2001). Mediation and the Transformation of the Israeli-Palestinian Conflict. *Journal of Peace Research 38*(3), 373–392.

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

Landau, D. and S. Landau (1997). Confidence Building Measures in Mediation. *Mediation Quarterly 15*, 97–103.

Larson, D. W. (1997). *Anatomy of Mistrust: U.S.-Soviet Relations During the Cold War*. Cornell Studies in Security Affairs. Ithaca, NY: Cornell University Press.

Lizzeri, A. (1999). Information revelation and certification intermediaries. *RAND Journal of Economics' 30*(2), 214–231.

Milgrom, P. and J. Roberts (1994). Comparing Equilibria. *American Economic Review 84*(3), 441–459.

Milgrom, P. and C. Shannon (1994). Monotone Comparative Statics. *Econometrica 62*(1), 157–180.

Milgrom, P. R., D. C. North, and B. R. Weingast (1990). The Role of Institutions in the Revival of Trade: The Medieval Law Merchant, Private Judges and Champaign Fairs. *Economics and Politics 2*, 1–23.

Mitchell, G. J. (1999). *Making Peace*. Berkeley: University of California Press.

Mitusch, K. and R. Strausz (2000). Mediation in Situations of Conflict. *Manuscript Berlin*.

Myerson, R. (1986). Multistage Games with Communication. *Econometrica 54*(2), 323–358.

O'Neill, B. (2003). Mediating National Honor: Lessons from the Era of Duelling. *Journal of Institutional And Theoretical Economics 159*(1), 229–247.

O'Neill, B. (2004). What Can a Disinterested Powerless Mediator Do for Strategic Negotiators? *Manuscript*.

Posen, B. R. (1993). The Security Dilemma and Ethnic Conflict. *Survival 35*(1), 27–47.

Powell, R. (2002). Bargaining Theory and International Conflict. *Annual Review of Political Science 5*, 1–30.

Princen, T. (1991). Camp David: Problem Solving or Power Politics As Usual. *Journal of Peace Research 28*(1), 57–69.

Princen, T. (1992). *Intermediaries in International Conflict*. Princeton: Princeton University Press.

Rauchhaus, R. W. (2005). Asymmetric Information, Mediation and Conflict Management.

Riley, J. G. (2001). Silver Signals: Twenty-Five Years of Screening and Signaling. *Journal of Economic Literature 39*(2), 432–478.

Ross, W. H. and C. Wieland (1996). Effects of Interpersonal Trust and Time Pressure on Managerial Mediation Strategy in a Simulated Organizational Dispute. *Journal of Applied Psychology 81*(3), 228–248.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica 50*(1), 97–109.

Schmidt, H. (2004). When (and Why) Do Brokers Have to be Honest? Impartiality and Third Party Support for Peace Implementation after Civil Wars 1945-1999. *manuscript*.

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

Spulber, D. F. (1996). Market Microstructure and Intermediation. *Journal of Economi Perspectives 10*(3), 135–152.

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

Touval, S. and I. W. Zartman (1989). Mediation in International Conflicts. In K. Kressel and D. G. Pruitt (Eds.), *Mediation Research: The Process and Effectiveness of Third Party Intervention*, pp. 115–137. Hoboken: Jossey-Bass.

Wall, J. A., J. B. Stark, and R. L. Standifer (2001). Mediation: A Current Review and Theory Development. *Journal of Conflict Resolution 45*(3), 370–391.

Walter, B. F. (2002). *Committing to Peace: the Successful Settlement of Civil Wars.* Princeton: Princeton University Press.

Walton, R. E. (1969). *Interpersonal Peacemaking: Confrontations and Third Party Consulation.* Reading, MA: Addison-Wesley.

Wehr, P. and J. P. Lederach (1991). Mediating Conflict in Central America. *Journal of Peace Research 28*(1), 85–98.

Young, O. R. (1967). *The Intermediaries: Third Parties in International Crises.* Princeton: Princeton University Press.