The Honest Broker:

Mediation and Mistrust[1]

Andrew Kydd

February 24, 2004

[1]Presented at the Mershon Center, Ohio State University, March 5, 2004. Please send comments to akydd@wcfia.harvard.edu

Abstract

Mediators are often said to facilitate conflict resolution or mutually beneficial exchange by building trust between suspicious adversaries or potential trading partners. This paper examines the conditions under which it is possible for mediators to build trust through information provision, rather than providing physical guarantees. In one shot interactions unbiased mediators who wish to encourage cooperation will not be credible. However if there are future opportunities for mediation, there can be reputational incentives for unbiased mediators to be credible, and in fact unbiased mediators will be preferable in these circumstances. In order to ensure honesty on the part of the mediator, the payoff in the case of successful cooperation and fee for serving as a mediator (unconditional on the outcome) must be kept in careful balance.

Mediators are often thought to provide reassurance in situations of mutual mistrust. In a conflict situation, mediators can reassure each side that the other is genuinely interested in peace, and not attempting to deceive and exploit them. For instance, 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). The fact that they were able to develop this degree of trust enabled the parties to cooperate, at least for a while.

Where two parties are contemplating an exchange, a mediator can vouch for the trustworthiness of each side of a risky transaction. 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 behaved honorably 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 their promises.

If trustbuilding is part of the mediator's job description, we need to know when a mediator will be effective at this task. Does it have to do with the characteristics of the meditator, her motivations and biases, or with the structure of the mediation? Is unbiasedness conducive to trustbuilding, and what is the role of the mediator's reputation for honesty?

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 them. The mediator has some private information about the trustworthiness of the two sides, and can share this information with the parties in an effort to reassure them. The central results of the model are that mediators in one shot situations have an incentive to lie and hence will be ineffective at fostering reassurance. However, if the game is repeated, the mediator can acquire a reputational incentive to be honest, which overcomes short term incentives to lie. Bias proves to be harmful in this situation, unbiased mediators make the best trustbuilders.

I will first discuss how the need for reassurance arises in various settings, and the role of mediation in reassurance The next section develops the model and summarizes the main results. I conclude with some discussion of broader implications on the role of mediation in reassurance.

1 Mediation and Mistrust

The need for reassurance, and the possibility that a mediator might provide it, arises in at least two distinct but related social contexts, the problem of conflict resolution and the problem of exchange.

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). John 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). Barbara 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 trust-building, and that this could be a role for mediators in the conflict resolution or prevention process.

Another social context in which trust is important is the problem of exchange. 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. Milgrom, North and Weingast argue that these sorts of problems hampered medieval trade. Merchants had to travel long distances and trade in unfamiliar environments in a weakly institutionalized setting where centralized enforcement of contracts was weak or non-existent. In these conditions, institutions are necessary to facilitate the collection and transmission of information on who has cheated and to encourage punishment of the guilty

parties.

The dominant theoretical model of these two problems, cooperation and exchange, is the Prisoner's Dilemma. The Prisoner's Dilemma seems like a natural model for these problems because each side has a temptation to exploit the other side, but if the situation is repeated, the threat of future retaliation may suffice to keep them honest (Axelrod 1984). Considerable work has gone into investigating what strategies can maintain cooperation under varying conditions. Bendor and colleagues and Signorino investigate the robustness of cooperation to noise, or uncertainty about what the other side has done (Bendor, Kramer, and Stout 1991; Signorino 1996; Bendor and Swistak 1997). Kandori and Ellis investigate situations, like the Milgrom, North and Weingast setup, where participants are matched randomly from anonymous populations, and show that cooperation can be sustained if the players fear a general collapse of social cooperation consequent to their own defection (Kandori 1992; Ellis 1994).

The Prisoner's Dilemma framework, while generating many insights, is not fully adequate to study trust and mistrust. Trust is a matter of uncertainty about whether the other side is inclined to reciprocate cooperation or not (Hardin 2002). To trust someone is to think that they prefer to reciprocate cooperation, to mistrust someone is to think they would prefer to exploit your cooperation. In the Prisoner's Dilemma and repeated PD framework, there is no such uncertainty. In the one shot game there can be no trust because the players have a dominant strategy to defect. In the repeated game, either there are sufficient incentives to guarantee cooperation or there are not, in neither case is there any uncertainty about whether the other side prefers to reciprocate cooperation or exploit it.[1] Thus trust must be modeled

[1]In the models featuring uncertainty about what players have done there can be uncertainty about

using incomplete information games in which there is uncertainty over the preferences of the actors (Kydd 2000).

The literature on mediation has not ignored the issues of trust and reassurance (for a recent review see (Wall, Stark, and Standifer 2001).) 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). 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).[2] 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. 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

whether the other side will reciprocate cooperation, but that is not really mistrust, since everyone knows that in equilibrium everyone intends to cooperate so that instances of non-cooperation are errors.

[2]Game theory, alas, has not yet reached the stage where the use of humor in dispute resolution can be adequately modeled.

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 of necessity biased (Princen 1991; Princen 1992).

Drawing on the cheap talk literature in economics (Farrell and Rabin 1996), Kydd has argued that mediators must be biased to be effective (Kydd 2003). Consider a mediator who is trying to convince one side to make a concession to the other by arguing that without the concession an agreement is unlikely. A mediator who is unbiased will not be believed when she makes such a statement, because such a mediator would have an incentive to say this even if she did not believe it, because anything that convinces the parties to make concessions makes an agreement more likely, which is what the mediator wants. A mediator who is biased towards one party, however, can credibly tell them that the adversary will not make peace without the concession, because the mediator would not urge a concession on the party unless the mediator thought it was truly necessary.

This result, however, may not hold in other contexts and for other tasks that the mediator might wish to perform. In the context of trust building, as we will see below, there is a similar result in the one shot game. Unbiased mediators are unable to build trust in a one shot situation because an unbiased mediator would say anything to raise the likelihood of peace, (or successful exchange), in the absence of any penalty for being caught in a lie. Repetition can provide just such a penalty, however, if the mediator's chance of future employment as a mediator hinges on not vouching for the trustworthiness of a party who proves to be untrustworthy. A betrayal of trust is a very visible event, and the parties can easily punish a mediator who vouches for an untrustworthy actor. Indeed, as the model below demonstrates,

bias can be problematic in the repeated game, because it can give the mediator an incentive to vouch for the side towards whom it is biased, even if that side is not so likely to be trustworthy. Too much bias can therefore make the mediator unable to serve as a credible source of information, preventing it from providing reassurance. The model illuminates the conditions under which this is the case.

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 while its opponent receives $-a$. There are two types of player, trustworthy types for whom $b_t \in (0, 1)$ and untrustworthy types for whom $b_u > 1$. In terms of the familiar two by two games, trustworthy types have Assurance payoffs and untrustworthy types have Prisoner's Dilemma payoffs. Nature starts the game by determining the types based on a t_i likelihood that player i is trustworthy and a corresponding $1 - t_i$ likelihood that player i is untrustworthy.

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 player 1 unilaterally defects, the mediator receives β and if player 2 unilaterally defects, the mediator receives $-\beta$. Thus β is a measure of how biased in favor of player 1 the mediator is. If $\beta > 0$ the mediator is biased towards player 1, if $\beta < 0$

Table 1: The Mediation Game

		Player 2: Cooperate	Player 2: Defect
Player 1	Cooperate	1, 1, ρ	$-a$, b, $-\beta$
	Defect	b, $-a$, β	0, 0, 0

the mediator is biased towards player 2, and if $\beta = 0$ the mediator is unbiased. If both sides defect the mediator receives zero.

After Nature chooses the players' types, Nature sends the mediator a signal about the type of each of the players which is not observed by the players. The mediator, therefore, has additional information about how likely each player is to be trustworthy. The signal is either T for trustworthy or U for untrustworthy so Nature's strategy space is $\{T, U\} \times \{T, U\}$. The likelihood that the signal about player i is accurate is $1 - \epsilon_i$, while the likelihood that it is in error is ϵ_i.

After receiving the T signal, the mediator's posterior beliefs that player i is trustworthy is derived from Bayes rule and is

$$p(iT|T) = \frac{t_i(1-\epsilon_i)}{t_i(1-\epsilon_i) + (1-t_i)\epsilon_i}$$

while after receiving the U signal, it is

$$p(iT|U) = \frac{t_i\epsilon_i}{t_i\epsilon_i + (1-t_i)(1-\epsilon_i)}.$$

Since $p(iT|T) > t_i > p(iT|U)$, the mediator becomes more trusting after receiving the T signal, and more suspicious after receiving the U signal.

The mediator then makes a public announcement about the messages received from Nature. The mediator's strategy space is $\{T, U, N\} \times \{T, U, N\}$ where T stands for trustworthy, U stands for untrustworthy and N stands for no comment, to allow the mediator to avoid comment on a player who is trustworthy enough for the other side to cooperate without additional reassurance. T, T means that both sides are trustworthy, T, U means that player 1 is trustworthy but player 2 is not, etc.

After the mediator makes an announcement, if the mediator is believed to be telling the truth, the parties' beliefs will shift to mirror the mediator's. That is, an honest report from the mediator is informationally equivalent to directly observing the signal from Nature. The mediator will of course take this into account when deciding whether or not to be honest. Alternatively, if the mediator send the no comment signal regardless of the signal received from Nature, the parties' beliefs will remain unchanged because the mediator's communication is uninformative.

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 One Round Game

I solve for perfect Bayesian equilibria. Untrustworthy types have a dominant strategy to defect and do so in any equilibrium. Trustworthy types can cooperate if both are sufficiently trusting. I am interested in the existence of equilibria in which the mediator tells the truth, if she comments on the players types, and the trustworthy types trust each other enough to

cooperate.[3]

3.1 When Mediation is Useful

Mediation will only be necessary and feasible if the trustworthy types are not too pessimistic nor to optimistic. If the players are very trusting, mediation will be superfluous, they can cooperate without reassurance. If the players are too suspicious, even reassurance from an honest mediator will not be sufficient, and they will fail to cooperate.

To derive the upper bound, note that if trustworthy types are expected to cooperate, the payoff for cooperation for player i is $t_j 1 + (1 - t_j)(-a)$ while the payoff for defection is $t_j b_t + (1 - t_j)0$. Cooperation beats defection if the level of trust exceeds a minimum trust threshold, $\overline{m}$.

$$t_j \geq \overline{m} \equiv \frac{a}{1 - b_t + a}$$

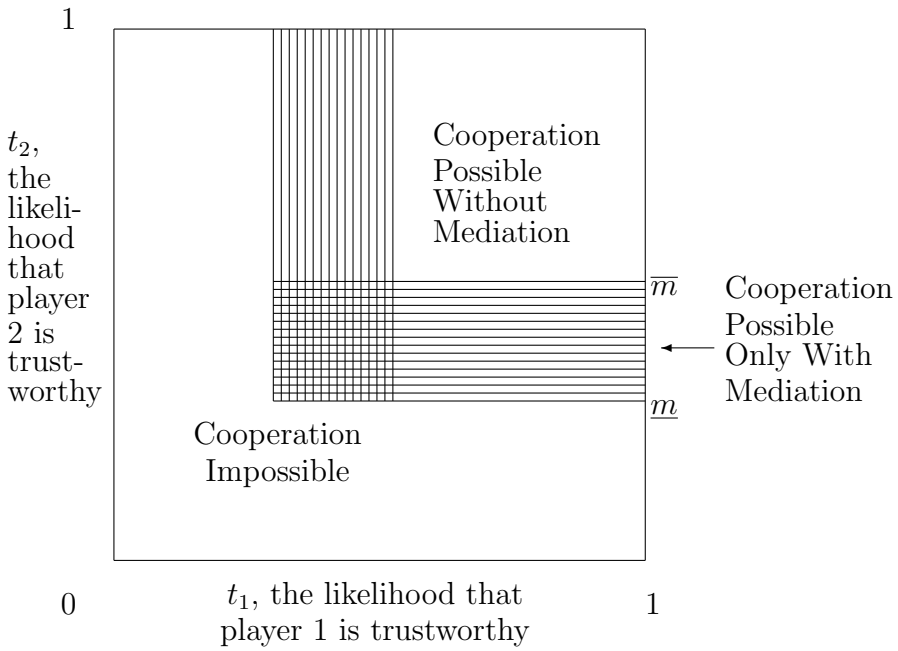
If the prior beliefs exceed this threshold, then mediation is unnecessary and an equilibrium exists in which the mediator makes the N, N statement regardless of the information received from Nature, which provides no information to the players, and trustworthy types cooperate nonetheless.

To derive the lower bound, note that cooperation would be possible if the posterior beliefs exceeded the minimum trust threshold, which would be the case if player i received the T signal and believed it, provided that $p(jT|T) \geq \overline{m}$ or $\frac{t_j(1-\epsilon_j)}{t_j(1-\epsilon_j)+(1-t_j)\epsilon_j} \geq \frac{a}{1-b_t+a}$. Solving for the prior beliefs we get

$$t_j \geq \underline{m} \equiv \frac{a\epsilon_j}{a\epsilon_j + (1 - b_t)(1 - \epsilon_j)}$$

[3]Off equilibrium path beliefs are discussed in the Appendix.

Figure 1: Where Mediation is Useful



Thus if $t_j \in [0, \underline{m})$, the prior beliefs are so pessimistic that even on receiveing the T signal and believing it player i will not cooperate. Cooperation is impossible here with or without mediation. If $t_j \in [\underline{m}, \overline{m})$ cooperation would be possible between trustworthy types if and only if the mediator were truthful, and both players received the T message. These constraints are illustrated in Figure 1. There are three regions where mediation matters, where player 1 is trusting enough to cooperate without new information but player 2 is not (shaded vertically), where player 2 will cooperate without new information but player 1 will not (shaded horizontally), and where neither side will cooperate without reassurance (the cross hatched region).

3.2 When the Mediator is Honest

When does the mediator have an incentive to be truthful? Consider the symmetrical case where both sides need reassurance to cooperate, $t_i \in [\underline{m}, \overline{m})$.[4] Is there a truthtelling equilibrium in which the players cooperate if they are trustworthy? We must consider each of the possible signals from Nature and make sure the mediator has the incentives to tell the truth in each case.

If the mediator gets the U, U signal, the payoff for telling the truth by sending the U, U signal is 0, as is the payoff for sending any other signal but T, T. The payoff for lying by sending the T, T signal, which will encourage the trustworthy types to cooperate, is

$$p(1T|U)p(2T|U)\rho + p(1T|U)(1 - p(2T|U))(-\beta) + (1 - p(1T|U))p(2T|U)(\beta)$$

Telling the truth beats lying if

$$0 > p(1T|U)p(2T|U)\rho + (p(2T|U) - p(1T|U))\beta$$

Constraints for the U, T, T, U and T, T signals are derived similarly, and the conditions are summarized below. The fourth condition has a reversed sign because after receiving the T, T signal from Nature, honesty requires passing this on, which produces the risky payoff on the right, while lying in any way results in non-cooperation for sure, the payoff on the left.

$$0 \geq p(1T|U)p(2T|U)\rho + (p(2T|U) - p(1T|U))\beta \quad (1)$$

$$0 \geq p(1T|U)p(2T|T)\rho + (p(2T|T) - p(1T|U))\beta \quad (2)$$

$$0 \geq p(1T|T)p(2T|U)\rho + (p(2T|U) - p(1T|T))\beta \quad (3)$$

[4] I discuss the asymmetrical case where only one side needs reassurance in the Appendix.

$$0 \quad \leq \quad p(1T|T)p(2T|T)\rho + (p(2T|T) - p(1T|T))\beta \tag{4}$$

The fourth condition is easily satisfied, given that $\rho > 0$. That is, the mediator has every incentive to tell the parties they are trustworthy, if the mediator thinks this is the case, since it only makes cooperation more likely, which is what the mediator wants. The first three, where the mediator gets bad information about one or the other player, are more problematic. An unbiased mediator, with $\beta = 0$, would have an incentive to lie here. This is because lying encourages the trustworthy types to cooperate with each other, which is the unbiased mediator's only chance of earning a positive payoff, the reward for a successful agreement, ρ. If the mediator honestly reveals that she got bad information about one or both players, cooperation will not ensue and the mediator's payoff will be zero. Therefore the unbiased mediator in the one shot game has an incentive to lie, as in Kydd's treatment of mediation in bargaining. In fact, a truthtelling equilibrium is impossible in the one round game in the zone in which it is needed to produce cooperation, regardless of the mediator's bias. The result for the one shot game is summarized in the following theorem.

Theorem 1 *In the single round mediation game, if mediation is necessary to produce cooperation, $t_i \in [\underline{m}, \overline{m})$, and the mediator is rewarded for successful cooperation, $\rho > 0$, there is no truthtelling equilibrium.*

Proof: Given that $\rho > 0$, the second term in the first three conditions must be negative, which indicates that the mediator must be biased in favor of a player who is always more trustworthy than the other. Consider the case where $\beta > 0$ and $p(1T|U) > p(2T|T)$, the reverse case is symmetrical. For condition 2 to be satisfied, it must be the case that $p(2T|T)\rho \leq \beta\left(1 - \frac{p(2T|T)}{p(1T|U)}\right)$. However, for conditions 2 and 4 to be jointly satisfied it must

be that $p(1T|U)p(2T|T)\rho+(p(2T|T)-p(1T|U))\beta \leq p(1T|T)p(2T|T)\rho+(p(2T|T)-p(1T|T))\beta$ or $p(2T|T)\rho \geq \beta$ which produces a contradiction.

Q.E.D.

4 The Repeated Game

Now consider a repeated version of the 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 she has never said that a player is trustworthy when a player defects in a situation where the player is expected to cooperate if trustworthy. If the mediator does make this mistake, she is never hired again.[5] The mediator receives a payoff of ϕ as a fee for participating regardless of what happens in the game. For all rounds in which the mediator is not hired, she receives a payoff of zero.

I consider perfect Bayesian equilibria, so that the mediator's choice at each round must be optimal given her information. As in the one round game, untrustworthy types will defect in any equilibrium while trustworthy types can cooperate if they are trusting enough. If the players are too pessimistic, $t_i \in [0, \underline{m})$, then mediation cannot promote cooperation.[6] If the players are very optimistic, $t_i \in [\overline{m}, 1]$, then an equilibrium exists in which the mediator always makes the N, N statement, is never caught in a lie and so stays in the game forever, and

[5] We can imagine a pool of mediators available such that the parties can select a different mediator if the mediator in question has lied.

[6] The equilibrium payoff for the mediator is $\frac{\phi}{1-\delta}$. Equilibria exist in which the mediator is honest, dishonest in various ways, or acts at random.

the trustworthy types cooperate which maximizes the mediator's payoff.[7] More interesting is the case of middling beliefs, I consider the zone in which $t_i \in [\underline{m}, \overline{m})$ for both players.

4.1 The Equilibrium Payoff

The payoff for the equilibrium strategy is a function of the per round payoff and the likelihood of passing to the next round. Denote the per-round payoff for telling the truth π_t. The base payoff is ϕ, which is earned regardless of the outcome. In addition, in each round, if the mediator gets at least one U signal, she will convey it which will result in non-cooperation and a payoff of 0, and if she gets two T signals, she will convey them and trustworthy types will cooperate and untrustworthy types will not. The payoff is therefore composed of four components, one where the T, T signal was correct, $t_1(1-\epsilon_1)t_2(1-\epsilon_2)\rho$, and three were it was not $t_1(1-\epsilon_1)(1-t_2)\epsilon_2(-\beta)$, $(1-t_1)\epsilon_1 t_2(1-\epsilon_2)\beta$, and $(1-t_1)\epsilon_1(1-t_2)\epsilon_2(0)$. The payoff for the round is therefore the fee plus the sum of these.

$$\pi_t = \phi + t_1(1-\epsilon_1)t_2(1-\epsilon_2)\rho + [(1-t_1)\epsilon_1 t_2(1-\epsilon_2) - t_1(1-\epsilon_1)(1-t_2)\epsilon_2]\beta$$

Now consider the ex-ante likelihood of getting to the next round assuming the mediator follows the equilibrium strategy, denoted γ_t. If the mediator tells the truth, she will continue to the next round for sure following the U, U, U, T and T, U signals, because the players defect and the mediator cannot be caught in a lie. Following the T, T signal, the mediator will continue to the next round if both sides are actually trustworthy. Hence the likelihood of getting to the next round is one minus the likelihood of getting the T, T signal in error,

[7] Provided the mediator is not too biased.

or

$$\gamma_t = 1 - [t_1(1-\epsilon_1)(1-t_2)\epsilon_2 + (1-t_1)\epsilon_1 t_2(1-\epsilon_2) + (1-t_1)\epsilon_1(1-t_2)\epsilon_2]$$

With the payoff per round, π_t, and the likelihood of going into the next round, γ_t, the equilibrium payoff in the game is the following sum.

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

The discount factor is multiplied by the likelihood of getting into the next round, representing a total discounting of future rounds by $\delta\gamma_t$.

4.2 Equilibrium Conditions

Now we need to consider the payoffs associated with various deviations. First I will consider one round deviations. In general, the payoff for a one round deviation can be written as follows, where the one round payoff for deviating is π_d and the likelihood of continuing to the next round is γ_d

$$\pi_d + \delta\gamma_d\pi_t + \delta\gamma_d\delta\gamma_t\pi_t + \cdots$$

Following the equilibrium payoff will beat the deviation if

$$\pi_t + \delta\gamma_t\frac{\pi_t}{1-\delta\gamma_t} \geq \pi_d + \delta\gamma_d\frac{\pi_t}{1-\delta\gamma_t}$$

Now consider specific deviations, after the mediator is informed by Nature. Consider first if the mediator receives the U, U message. The equilibrium mandates honest communication followed by no cooperation, for a payoff of ϕ, then continuing on to the next round for sure. The equilibrium payoff is therefore

$$\phi + \delta\frac{\pi_t}{1-\delta\gamma_t}$$

Deviating to U,T or T,U will produce no change in payoff. Deviating to T,T will cause trustworthy types to cooperate. The likelihood of continuing to the next round will then be $p(1T|U)p(2T|U)$ rather than 1 as in the equilibrium. The payoff for the deviation is

$$\phi + p(1T|U)p(2T|U)\rho + [p(2T|U) - p(1T|U)]\beta + p(1T|U)p(2T|U)\delta\frac{\pi_t}{1-\delta\gamma_t}$$

The deviation will be unprofitable if

$$\begin{aligned}\phi \;\geq\; & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(1T|U)p(2T|U)}{1-p(1T|U)p(2T|U)} - t_1(1-\epsilon_1)t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(2T|U)-p(1T|U)}{1-p(1T|U)p(2T|U)} - [(1-t_1)\epsilon_1 t_2(1-\epsilon_2) - t_1(1-\epsilon_1)(1-t_2)\epsilon_2]\right)\beta\end{aligned} \tag{5}$$

While more complex than before, this condition is similar to the previous case and places a lower bound on the fee, ϕ in terms of the reward for successful cooperation, ρ and the mediator's bias β.

The conditions for the other cases are derived similarly and are the following.

$$\begin{aligned}\phi \;\geq\; & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(1T|U)p(2T|T)}{1-p(1T|U)p(2T|T)} - t_1(1-\epsilon_1)t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(2T|T)-p(1T|U)}{1-p(1T|U)p(2T|T)} - [(1-t_1)\epsilon_1 t_2(1-\epsilon_2) - t_1(1-\epsilon_1)(1-t_2)\epsilon_2]\right)\beta\end{aligned} \tag{6}$$

$$\begin{aligned}\phi \;\geq\; & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(1T|T)p(2T|U)}{1-p(1T|T)p(2T|U)} - t_1(1-\epsilon_1)t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(2T|U)-p(1T|T)}{1-p(1T|T)p(2T|U)} - [(1-t_1)\epsilon_1 t_2(1-\epsilon_2) - t_1(1-\epsilon_1)(1-t_2)\epsilon_2]\right)\beta\end{aligned} \tag{7}$$

$$\begin{aligned}\phi \;\leq\; & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(1T|T)p(2T|T)}{1-p(1T|T)p(2T|T)} - t_1(1-\epsilon_1)t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_t}{\delta}\frac{p(2T|T)-p(1T|T)}{1-p(1T|T)p(2T|T)} - [(1-t_1)\epsilon_1 t_2(1-\epsilon_2) - t_1(1-\epsilon_1)(1-t_2)\epsilon_2]\right)\beta\end{aligned} \tag{8}$$

If these four conditions are satisfied, no one round deviation is profitable. As for multiple round deviations, the following lemma is proved in the appendix.

Lemma 1 *If no one round deviation is profitable, no longer one will be either.*

Therefore, the conditions identified above are the ones governing whether a truthtelling equilibrium is possible.

4.3 The Role of Reputation

Examining the four inequalities several things can be deduced. On the left we have the per period fee for participation, ϕ. On the right hand side we have functions of ρ and β as before. If β is small enough, or ρ large enough, the right hand side in 8 is greater than that in the other three inequalities. If it is greater than zero, there will exist a value for ϕ such that it exceeds the first three right hand sides, but not the last one, satisfying the equilibrium constraints. This will definitely be the case for small enough δ, as reducing δ increases the right hand side. Thus we have the following theorem.

Theorem 2 *In the repeated mediation game, when mediation is necessary to produce cooperation, $t_i \in [\underline{m}, \overline{m})$, and the mediator is rewarded for successful cooperation, $\rho > 0$, truthtelling equilibria exist regardless of the degree of bias of the mediator.*

Proof: For large enough ρ, the right hand side of condition 8 will be greater than in the previous three conditions. For small enough δ, an appropriate level of ϕ can therefore be found to satisfy all four conditions.

Q.E.D.

Intuitively, the fact that the game is repeated can give the mediator an incentive to be honest. The mediator knows that if she is caught in a lie, saying a player is trustworthy who

then defects when a trustworthy player would have cooperated, she will never be hired again. Being hired again is valuable because of the non-contingent fee ϕ, as well as the opportunity to get more rewards ρ for successful cooperation. So the mediator may not wish to jeopardize her chances of getting to the next round by vouching for the players when she 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 trustworthy types will not cooperate, so the mediator cannot be caught in a lie.

For the mediator to be honest in all circumstances, however, the payoffs need to be carefully balanced. For instance, if the non-contingent fee ϕ is too great, then condition 8 will be violated. This means that even when the mediator gets information that the two sides are trustworthy, she will still say they are untrustworthy in order to maximize her chance of getting to the next round and getting another ϕ. Similarly, the reward for successful cooperation, ρ, cannot be too great, or conditions 5, 6, or 7 may be violated. If the payoff for getting the parties to cooperate is too great, the temptation to encourage them to cooperate by vouching for them even when they may not be trustworthy will be irresistible, as in the one shot game. Thus for the mediator to have the correct incentives to be honest in the repeated game, the non-contingent fee for mediation ϕ and the extra payoff in case of success, ρ must be kept in balance with each other. Holding one fixed, the other cannot be allowed to grow too much.

Under some conditions the parties might be able to influence or choose the mediator's payoffs. This raises the question of what the optimal payoff schedule for the mediator might be, from the perspective of parties aiming to minimize mediation costs. We can probably assume that the bias, β, is exogenously given by the previous relationship or sympathies of

the mediator with the parties.[8] If β is fixed, we can derive the minimum payoffs that will support an equilibrium by increasing ρ from zero until the right hand side in condition 8 is just larger than in the previous three conditions. This enables one to select the minimum ϕ such that it is just below the right hand side of condition 8 and just above the largest of the other three, satisfying the equilibrium. This pair will be the minimum payoffs for the mediator that will permit the truthtelling equilibrium to hold.

4.4 The Role of Bias

If ρ and ϕ are fixed, however, too much bias on the part of the mediator can make the truthtelling equilibrium break down, rendering the mediator incapable of fostering cooperation. This might be the case in an international mediation in which the parties are incapable of influencing the payoffs of the mediator in any serious way, as when the U.S. mediates between small countries. Thus in mistrust situations, mediator bias is generally a bad thing.

This is most easily seen in the case where the players are symmetrical, $t_1 = t_2$, $\epsilon_1 = \epsilon_2$. With symmetrical players, if we look at conditions 5 and 8 we can see that the bias term drops out, so the level of bias is unimportant when the mediator receives the same information about the two players, whether positive or negative.

Conditions 6 and 7, where the mediator receives good information about one player and bad information about the other still contain the bias term. However, they are simplified by the fact that the second term in parentheses multiplying β is zero, which makes the sign of the first term determinative of the sign of the effect of the bias. If we consider condition 6, the term multiplying β is positive, because the mediator got good information about player

[8]Players might try to bribe the mediator, however (Milgrom, North, and Weingast 1990, 16).

2 and bad information about player 1. If β is also positive, then the whole term is positive, so if it is too large, it will invalidate the inequality. Too much bias can therefore prevent the truthtelling equilibrium from working. If the mediator is biased towards player 2, $\beta < 0$, then the danger arises in condition 7, because the first term in parentheses is negative, so the product is positive, once again threatening the inequality.

The intuition behind the result is that mediators may be tempted to lie about a side towards whom they are biased. Imagine that the mediator learns that player 1 is less likely to be trustworthy but player 2 is more likely. If player 1 is less likely to be trustworthy than player 2 it means that the outcome in which player 1 exploits player 2 is more likely than the reverse outcome in which player 2 exploits player 1. The more biased in favor of player 1 the mediator is, the better off the mediator is when player 1 exploits player 2. Thus the mediator may be tempted to lie about the side towards whom it is biased, saying that player 1 is trustworthy even though she got bad information about player 1. This will encourage the two sides to cooperate which may produce mutual cooperation or an outcome in which player 1 exploits player 2, either one of which is not bad from the mediator's perspective.

In the end, if the mediator is biased towards one side, the mediator will be more likely to vouch for them despite lingering doubts. Being less sensitive to the possibility of player 2 being exploited, they will take a risk on encouraging cooperation. Excess bias will lead to the mediator not being honest about the side they are biased towards, rendering the mediator useless. This is summarized in the following theorem.

Theorem 3 *In the repeated mediation game, there is an upper limit on the mediator's bias beyond which no truthtelling equilibrium is possible.*

Proof: Hold ρ and ϕ fixed. At least one of the terms multiplying β in conditions 5, 6 and 7 is either positive or negative. If it is positive, making β too large and positive will violate the condition, if it is negative, making β too large and negative will violate the condition.

Q.E.D.

5 Discussion

The main implications of the model are as follows.

Hypothesis 1 *For a mediator to build trust between the parties, she must have a reputational incentive to be honest, such that she would tell the parties that she thought them to be untrustworthy if this was the case.*

Mediators must have an ongoing stake in a reputation for honesty. In the model this arises from a series of mediating opportunities with different parties, however, it could just as easily arise from different issues arising with the same parties, provided the parties types are not correlated across the issues. 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).

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. While the coding for this category is doubtless not equivalent to the trustbuilding statements envisioned by the model, it is likely that trustbuilding statements would more likely fall in this category than in the other two categories, "procedural," and "directive".

This condition is more likely to be met in the exchange scenario than in international conflict resolution. In the exchange scenario, 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, however, 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 ϕ and δ). When President Clinton mediated between Arafat and Barak, the reward for a successful settlement would have been tremendous, while Clinton's concern about future mediation scenarios could have been minimized by uncertainty that there would ever be a next time, at least of such magnitude.

Hypothesis 2 *Biased mediators will be less effective at building trust than unbiased mediators.*

The second hypothesis supports the common intuition that mediators should be unbiased. The more biased a mediator is, the greater danger that she will lie about the side towards whom she is biased, in order to encourage cooperation in spite of the risk that her 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 unbiased. Some small amount of bias may be acceptable, however, provided that the long term incentives for honesty are in place. Indeed, any amount of bias can be compensated for by an adequate adjustment of the other payoffs, as Theorem 2 shows.

Countless scholars and practitioners have noted the advantage of neutrality in mediation. 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). The phrase "honest broker" implies neutrality, as Secretary of State Haig noted when he argued, "the honest broker must above all be neutral" (Haig 1984, 226). 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).

Hypothesis 3 *To generate the proper incentives for honesty, the mediator's unconditional payoff for serving as mediator and the reward for successful cooperation must be balanced.*

The third hypothesis warns against letting either the fee or the reward for successful cooperation get too large. If the fee gets too large, the mediator will do whatever maximizes the likelihood of getting to the next round, in this case saying that the players are untrustworthy. If the reward for successful cooperation is too large, the mediator will vouch for the players even when they are not trustworthy, in an effort to maximize the likelihood that they cooperate. Another way of putting this is that the mediator cannot consider the current interaction too important, in comparison to reputational considerations and the possibility of future interactions. A mediator who wishes to build trust in the current interaction at all costs will not be credible, nor will one who just wants to minimize the chance of a mistake in the current round in order to pass on to the next.

Another way of conceiving of the proper relationship between the payoffs is suggested by the discussion following Theorem 2. Most mediators have some degree of bias, if it is only a matter of personally liking one side more than the other. To overcome this, there needs to be a reward for a successful agreement, ρ, so that the mediator is focused on obtaining a settlement or successful exchange, rather than simply indulging her biases by favoring one side rather than the other. However, this sets up an incentive to say anything to get an agreement, so to counter this there needs to be an incentive, ϕ, which can be obtained in the future with greater probability if the mediator is honest today. When these payoffs can be chosen at will, the parties can minimize the expense of mediation by choosing the smallest ρ that will convince the mediator to go for a settlement, and the smallest ϕ that will convince

her to nonetheless be honest about it.

6 Conclusion

When mediators attempt to build trust, they should be unbiased and have a reputational incentive to be honest. This will result in the mediator discouraging cooperation when she thinks the parties are untrusworthy, and encouraging cooperation only when she thinks they are trustworthy. The mediator should be compensated enough to give them a stake in the future, but not enough to be willing to do anything to avoid making a mistake. Thus the ideal trustbuilding mediator is neutral, in it for the long haul, and moderately compensated for each mediation and in case of success.

The appropriateness of the unbiased mediator for trustbuilding contrasts with the case of a mediator who attempts to get one side to make a concession, analyzed by Kydd. The one round game here exhibits a similar logic, in that in the absence of future rounds, the mediator will say anything to encourage cooperation, vouching for the players even when they are not trustworthy. A shadow of the future, however, provides the needed incentive for the unbiased mediator to be honest. This highlights the fact that the characteristics of the mediator that are conducive to success are highly situation and context dependent. Different tasks call for mediators with different attributes. Mediators attempting to prise out a concession in a one shot interaction need to be biased towards the side they are communicating with. Mediators attempting to build trust who have a stake in the future are more effective if they are unbiased.

Table 2: Notation in the Game

$-a$	Payoff for unilateral cooperation
b	Payoff for unilateral defection
ρ	Mediator's reward for successful agreement
β	Mediator's degree of bias
t_i	Likelihood that player i is trustworthy
$[\underline{m}, \overline{m})$	Zone where mediation is useful
δ	Mediator's discount factor
ϕ	Mediator's fee for participating in repeated game
π	Mediator's per round payoff in repeated game
γ	Mediator's likelihood of passing to the next round

Appendix

Notation in the game is summarized in Table 2.

Off Equilibrium Path Beliefs

The game can go off the equilibrium path if the mediator sends a signal that has probability zero. I assume that elements of the signal that have probability zero are perceived to be uncorrelated with the player's types. For instance, if the mediator is supposed to send the N signal about player 1, indicating no comment, if she instead sends the T or U signals they are taken to be uncorrelated with player 1's type as in a babbling equilibrium. This is always sustainable in equilibrium if the players do not condition their behavior on the signal

in this case.

Multiple Round Deviations

First consider a finite deviation, assuming that conditions are such that a one round deviation is not profitable. Any finite defection is composed of a sequence of deviations followed, eventually, by a return to equilibrium behavior, ie. truthtelling. If we let d designate a typical round in which a deviation occurs, and denote the one round payoff for round d as π_d and the associated likelihood of getting to the next round as γ_d, the payoff for an n round deviation is

$$\pi_1 + \delta\gamma_2\pi_2 + \delta^2 \prod_{d=2}^{3} \gamma_d\pi_3 + \cdots + \delta^{n-1} \prod_{d=2}^{n} \gamma_d\pi_n + \delta^n \prod_{d=2}^{n+1} \gamma_d\pi_t + \delta^{n+1} \prod_{d=2}^{n+1} \gamma_d\gamma_t\pi_t + \cdots$$

Now consider an identical deviation except that it is one round shorter so that the mediator returns to truthtelling one round earlier, in round n. The payoff is

$$\pi_1 + \delta\gamma_2\pi_2 + \delta^2 \prod_{d=2}^{3} \gamma_d\pi_3 + \cdots + \delta^{n-2} \prod_{d=2}^{n-1} \gamma_d\pi_{n-1} + \delta^{n-1} \prod_{d=2}^{n} \gamma_d\pi_t + \delta^n \prod_{d=2}^{n} \gamma_d\gamma_t\pi_t + \cdots$$

The payoff is identical in all rounds until n. The shorter deviation beats the longer if

$$\delta^{n-1} \prod_{d=2}^{n} \gamma_d\pi_t + \delta^n \prod_{d=2}^{n} \gamma_d\gamma_t\pi_t + \cdots > \delta^{n-1} \prod_{d=2}^{n} \gamma_d\pi_n + \delta^n \prod_{d=2}^{n+1} \gamma_d\pi_t + \cdots$$

or if

$$\pi_t + \delta\gamma_t \frac{\pi_t}{1-\delta\gamma_t} > \pi_n + \delta\gamma_{n+1} \frac{\pi_t}{1-\delta\gamma_t}$$

But this is just the condition making a one round deviation unprofitable. So if the one round deviation is unprofitable, then in the last round of the finite deviation, reverting to the equilibrium strategy would be preferable to waiting one more round to do so. Therefore

any finite deviation is beat by one in which the mediator reverts to the equilibrium strategy one round earlier. This zips back to the one round deviation that begins it, which in turn, by assumption, is worse than the equilibrium strategy. So if no one round deviation is profitable, neither is any longer but finite one.

Turning to infinite deviations, consider any infinite deviation, with payoff y, and the sequence of finite deviations that consists of following the infinite deviation for n rounds and then returning to the equilibrium strategy of telling the truth, each with payoff x_n and the equilibrium payoff, denoted z. The limit of the payoffs for the finite deviations will approach the payoff for the infinite deviation as $n \to \infty$,

$$\lim_{n=1}^{\infty} x_n = y$$

because the reversion to the equilibrium strategy will be postponed further and further, and hence be discounted to insignificance. By the logic of the finite case, we know that the sequence of payoffs for the finite deviations is decreasing, $x_n > x_{n-1}$. Given that the equilibrium payoff z beats the one round deviation, and the payoff decreases with the length of the deviation, and the payoff for the infinite deviation is the limit of the sequence, the equilibrium strategy also beats the infinite round deviation. That is, if $z > x_1$, $x_n > x_{n+1}$ and $\lim_{n=1}^{\infty} x_n = y$, then $y < x_i$, $\forall i$, so that $z > y$.

Thus if all one round deviations are unprofitable, no longer deviation will be, proving the lemma.

The Asymmetrical Case

Now consider the asymmetrical case, where $t_1 \geq \overline{m}$ and $t_2 \in [\underline{m}, \overline{m})$. In this case player 1 needs to be reassured about player 2, but player 1 is trustworthy enough for player 2 to cooperate without additional information. The equilibrium considered in the body may still hold under such conditions. However, another equilibrium may be possible in which the mediator sends the no comment signal, N, about player 1 regardless of the signal received from Nature, without preventing the trustworthy player 2 from cooperating. Note this equilibrium involves telling the truth about player 2 and preserving a tactful silence about player 1, so some might quibble with the label truthtelling. However, the mediator does provide information which enables cooperation where it would not be possible otherwise.

In the one round game, the only difference in the conditions arises in the second case, where the mediator gets the U, T signal from nature. Here the mediator sends the no comment signal about player 1 and is truthful about player 2, N, T, encouraging cooperation. The new condition 2 is the following.

$$0 \leq p(1T|U)p(2T|T)\rho + (p(2T|T) - p(1T|U))\beta$$

The new condition, like condition 4, is easily satisfied, given that $\rho > 0$. The first and third however, remain problematic as before. An unbiased mediator, with $\beta = 0$, would still have an incentive to lie, supporting the first theorem. As before, under certain conditions a truthtelling equilibrium is possible if the mediator is biased and the players are asymmetrical. For instance, if the mediator is biased in favor of player 1, $\beta > 0$, honesty could be induced by the fear that if the mediator encourages player 1 to cooperate after receiving bad information about player 2, there will be an increased risk that player 1 is exploited. If $\beta > 0$ then

if $p(2T|U) < p(1T|U)$, conditions 1 and 3 may be satisfied. This could be the case if the information about player 2 was better than that about player 1. For instance, if $t_1 = t_2 = 0.5$, $\epsilon_1 = 0.5$, $\epsilon_2 = 0.25$, and $\rho = \beta = 2$ then all four inequalities are satisfied.

The repeated game in the asymmetrical case is also similar. Posit an equilibrium strategy of saying N about player 1 in each round and telling the truth about player 2. The expected per round payoff will be the following.

$$\pi_{at} = \phi + t_1 t_2 (1 - \epsilon_2)\rho + [(1 - t_1)t_2(1 - \epsilon_2) - t_1(1 - t_2)\epsilon_2]\beta$$

Note it is no longer contingent on the information about player 1, since the mediator says the same thing regardless of the information about player 1. The mediator will get to the next round unless player 2 is untrustworthy but the mediator gets the wrong signal, or if player 1 is untrustworthy and player 2 is trustworthy and the mediator gets the correct signal, so the likelihood of getting into the next round is the following.

$$\gamma_{at} = 1 - [(1 - t_2)\epsilon_2 + (1 - t_1)t_2(1 - \epsilon_2)]$$

The equilibrium payoff is then

$$\frac{\pi_{at}}{1 - \delta\gamma_{at}}$$

The four equilibrium conditions are slightly different because of the difference in the equilibrium payoff, they are as follows.

$$\begin{aligned}
\phi \;\geq\; & \left(\frac{1 - \delta\gamma_{at}}{\delta}\frac{p(1T|U)p(2T|U)}{1 - p(1T|U)p(2T|U)} - t_1 t_2(1 - \epsilon_2)\right)\rho + \\
& \left(\frac{1 - \delta\gamma_{at}}{\delta}\frac{p(2T|U) - p(1T|U)}{1 - p(1T|U)p(2T|U)} - [(1 - t_1)t_2(1 - \epsilon_2) - t_1(1 - t_2)\epsilon_2]\right)\beta \\
\phi \;\leq\; & \left(\frac{1 - \delta\gamma_{at}}{\delta}\frac{p(1T|U)p(2T|T)}{1 - p(1T|U)p(2T|T)} - t_1 t_2(1 - \epsilon_2)\right)\rho + \\
& \left(\frac{1 - \delta\gamma_{at}}{\delta}\frac{p(2T|T) - p(1T|U)}{1 - p(1T|U)p(2T|T)} - [(1 - t_1)t_2(1 - \epsilon_2) - t_1(1 - t_2)\epsilon_2]\right)\beta
\end{aligned}$$

$$\begin{aligned}\phi \;\geq\; & \left(\frac{1-\delta\gamma_{at}}{\delta}\frac{p(1T|T)p(2T|U)}{1-p(1T|T)p(2T|U)} - t_1t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_{at}}{\delta}\frac{p(2T|U)-p(1T|T)}{1-p(1T|T)p(2T|U)} - [(1-t_1)t_2(1-\epsilon_2) - t_1(1-t_2)\epsilon_2]\right)\beta\end{aligned}$$

$$\begin{aligned}\phi \;\leq\; & \left(\frac{1-\delta\gamma_{at}}{\delta}\frac{p(1T|T)p(2T|T)}{1-p(1T|T)p(2T|T)} - t_1t_2(1-\epsilon_2)\right)\rho + \\ & \left(\frac{1-\delta\gamma_{at}}{\delta}\frac{p(2T|T)-p(1T|T)}{1-p(1T|T)p(2T|T)} - [(1-t_1)t_2(1-\epsilon_2) - t_1(1-t_2)\epsilon_2]\right)\beta\end{aligned}$$

The chief difference is that in the second case, after the U, T signal, the mediator must be willing to send the N, T signal and encourage cooperation, even though as before she is unwilling to do so if she gets the T, U signal. This would be facilitated if player 1 were more trustworthy than player 2, if $t_1 > t_2$, and especially if the information about player 1 was of lower quality, if $\epsilon_1 > \epsilon_2$. If the signal is not very informative about player 1 but very informative about player 2, the mediator will be willing to ignore the signal about player 1 and act only on that about player 2.

References

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Bendor, J., R. M. Kramer, and S. Stout (1991). When in Doubt . . . Cooperation in a noisy prisoner's dilemma. *Journal of Conflict Resolution 35*(4), 691–719.

Bendor, J. and P. Swistak (1997). The Evolutionary Stability of Cooperation. *American Political Science Review 91*(2), 290–307.

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.

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

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

Ellis, G. (1994). Cooperation in the Prisoner's Dilemma with Anonymous Random Matching. *The Review of Economic Studies 61*(3), 567–588.

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

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.

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

Haig, A. (1984). *Caveat: Realism, Reagan and Foreign Policy*. New York: MacMillan.

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

Herz, J. H. (1950). Idealist Internationalism and the Security Dilemma. *World Politics 2*(2), 157–180.

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

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.

Kandori, M. (1992). Social Norms and Community Enforcement. *The Review of Economic Studies 59*(1), 63–80.

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.

Kydd, A. (2000). Trust, Reassurance and Cooperation. *International Organization 54*(2), 325–357.

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.

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.

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

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.

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.

Signorino, C. S. (1996). Simulating International Cooperation under Uncertainty. *Journal of Conflict Resolution 40*(1), 152–205.

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.