Mediation in reputational bargaining

Jack Fanning[*†]

August 28, 2019

Abstract

This paper investigates the potential for mediation in a dynamic reputational bargaining model, with flexible rational agents and inflexible behavioral types. I first show how simple communication protocols used by professional mediators can improve outcomes. I then fully characterize equilibria with mediation under a "good faith" bargaining assumption, and identify a unique optimal protocol for symmetric games. Optimal mediation helps if there is curvature in the utility-possibility frontier, or agents' probability of inflexibility is small relative to demands. Outcomes differ markedly when a mechanism designer can impose agreement.

Keywords: Mediation, bargaining, reputation, behavioral types

1 Introduction

Mediators are third parties who help disputants reach *voluntary* bargaining agreements. Mediation is distinct from arbitration, another form of Alternative Dispute Resolution, which imposes agreement. It is widely used to help resolve conflicts ranging from international wars to labor union strikes to divorce. For instance, Dixon (1996) finds that mediation occurred in 13% of dispute phases of international conflicts between 1947-1982.[1] In a survey of general counsel for Fortune 1000 companies, Stipanowich and Lamare (2013) found that in each of commercial, employment and consumer disputes over 42% of companies "always" or "often" used mediation, and its use had increased in every category since 1997. By contrast binding arbitration was "always" or "often" used by less than 17% of companies in each category.

Attesting to mediation's benefits, Dixon found that mediated disputes were 47% less likely to escalate and 24% more likely to peacefully resolve compared to disputes with no conflict management.[2] More convincing evidence comes from Emery et al. (1991), who found that a treatment group, randomly selected to receive mediation services, settled 89% of contested

[*]Brown University. jack_fanning@brown.edu. Department of Economics, Robinson Hall, 64 Waterman Street, Brown University, Providence, RI 02912. See https://sites.google.com/a/brown.edu/jfanning for latest version.

[†]I am grateful to many people for their input on this project, especially Ennio Stacchetti, Mehmet Ekmekci, Oleg Semenov, Bobby Pakzad-Hurson, Larry Samuelson, and seminar audiences at Stanford GSB, UCSD, UCSB, Kellogg Northwestern, QMUL, Yale, MSU, Columbia and the BEET Conference.

[1]Dispute phases are distinguished by the level of conflict (e.g. threats of hostilities, open hostilities).

[2]Wilkenfeld et al. (2003) and Beardsley et al. (2006) also present positive empirical assessments of mediation's effectiveness in resolving conflict, while Fey and Ramsay (2010) is less positive.

custody cases out of court, compared to 28% of a control group. Mediation also halved the time spent reaching agreement and increased parties' satisfaction with the outcome.

Given that mediators are distinguished by having no formal power, why might mediation help? If a mediator is independently informed about the bargaining problem, then releasing her information can sometimes be beneficial (see Basak (2019)). However, typically mediators have no information beyond that already available to both parties (see Fey and Ramsay (2010)). Veteran mediator and former Secretary of Labor John Dunlop (Dunlop (1984), p16-24) describes the benefit of uninformed mediators in difficult "end-play" negotiations as follows: "The critical problem is that each side would prefer the other to move to avoid a further concession itself, and that any move may create the impression of being willing to move all the way to the position of the other side... In these circumstances a third party may greatly facilitate agreement. The separate conditional acceptance to the mediator by one side of the proposal does not prejudice the position of that side if there is no agreement. It is not unusual for a mediator to secure the separate acceptance of each side of a "package" of the mediator's design and then to bring the parties together to announce that, even if they do not know it, they have an agreement."

The claim is that mediators help, in part, by filtering agents' private information. Agents may resist proposing a compromise themselves for fear of being identified as a "weak" type, who is willing to concede entirely to her opponent's demand. The mediator can eliminate this fear by filtering the information that an agent is willing to compromise (e.g. releasing it only when an opponent is also willing to compromise), and so potentially encourage agreement.

Despite being intuitively appealing, it is far from obvious that such techniques do explain why mediation works; indeed, to the best of my knowledge, the existing theoretical literature has failed to identify any clear benefit from uninformed mediators in dynamic bargaining games. Uninformed mediators have been shown to usefully filter information in simple one-shot settings such as sender-receiver games (e.g. Goltsman et al. (2009)), however, the bargaining literature has shown that conclusions frequently don't extend from one-shot settings to dynamic ones. For instance, when a seller can make a take-it-or-leave-it offer to a buyer who has private information about her value, outcomes are inefficient with the seller getting her maximum possible profits, but in an infinite horizon, the Coase conjecture delivers an approximately efficient outcome with the buyer capturing all the gains from trade (Fudenberg et al. (1985) and Gul et al. (1986)). Such efficiency rules out Pareto improvements from mediation in this setting.[3]

In dynamic bargaining with two-sided incomplete information about values, it is hard to identify benefits from mediation given the vast set of equilibria even without mediation (e.g. see Ausubel et al. (2002)). Equilibrium multiplicity arises because agents can be harshly "punished with beliefs" for deviating from a given equilibrium path (i.e. identified as a weak type), giving them an arbitrarily small continuation payoff. This allows for both very efficient and very in-

[3]Clearly this is also true for canonical complete information models such as Rubinstein (1982).

efficient equilibria. As an example of the former, if the distributions of buyer and seller values are overlapping, independent, and have monotonic hazard rates, Ausubel and Deneckere (1993) construct equilibria of an alternating offer game which approximate the ex-ante constrained efficient bounds of Myerson and Satterthwaite (1983). If we believe that such efficient equilibria will arise absent mediation, there is again little scope for mediation. If we don't (strategies are nonintuitive: low cost sellers pool on demands with high cost sellers, but not medium cost sellers), then which of the many other equilibria should we compare mediation to?

I do, however, identify Pareto improvements from uninformed mediators in a dynamic reputational bargaining model with two-sided incomplete information. The reputational model has a unique equilibrium without mediation, which serves as a clear benchmark against which to assess mediation's effects. The model is adapted from the canonical reputational bargaining model of Abreu and Gul (2000) (henceforth AG).[4] Two agents must divide a dollar. They can make frequent offers over the course of an infinite horizon.[5] With positive probability each agent is an inflexible behavioral type who always demands a fixed share of the dollar and accepts nothing less, otherwise the agent is flexible (rational). The unique equilibrium resembles a war of attrition: flexible agents initially imitate inflexible demands and then concede slowly to their opponent.[6] If an agent is ever identified as flexible, she must immediately concede. Uniqueness arises because inflexible types cannot be punished with beliefs (they make their demand anyway). However, the model shares many tensions with other private information models, and can explain key features of unmediated bargaining highlighted by Dunlop, such as inefficient delay and negotiators' justified fear that small concessions will lead to larger ones.

I show that adding noise to the simple mediation (communication) protocol outlined by Dunlop can improve payoffs compared to unmediated bargaining, if agents are likely to be flexible. Dunlop proposed that mediators immediately suggest a compromise if both parties privately accept its terms. The added noise takes the form of the mediator failing to announce a deal with (possibly small) positive probability even when both parties privately accept it, which may occur because agents' messages sometimes go astray, or are misinterpreted.

Adding noise turns out to be essential: I also show that Dunlop's mediation protocol is necessarily ineffective without it. Without noise a flexible agent who accepts the mediator's compromise learns that her opponent didn't accept (and so is likely inflexible) when no deal is announced. This increases her incentive to subsequently concede, which destroys a flexible opponent's incentive to compromise in the first place. Adding noise makes agents less pessimistic about an opponent's type if no deal is announced (perhaps the mediator is at fault). The two results high-

[4] Myerson (1991) first analyzed reputational bargaining with one-sided inflexiblity. More recent contributions to the literature include: Kambe (1999), Abreu and Pearce (2007), Wolitzky (2012), Atakan and Ekmekci (2013).

[5] This might seem inconsistent with many mediation settings, where an impending trial implies a finite horizon (deadline). However, Fanning (2016) shows that infinite horizon and deadline reputational models are very similar.

[6] AG allow for discrete time bargaining, but show that outcomes converge to those of a unique continuous time equilibrium, regardless of the fine details of the bargaining protocol.

light the desirability of commitment for mediators, and suggest they shouldn't be incentivized to reach early agreements. Sometimes failing to agree spurs agreement.[7]

To more deeply understand why and when mediation works, the paper's main analysis adopts a mechanism design approach. Allowing for all possible mediator strategies, I fully characterize outcomes in what I call *good faith* equilibria, in which flexible agents never demand more than inflexible types.[8] This represents all equilibria if agents can recall an opponent's previous offer, and initially make an inflexible type's demand. The name derives from the National Labor Relations Act which requires that firms and unions bargain in "good faith" and not withdraw from provisions that they have previously agreed to.

My main result then characterizes an essentially unique good faith equilibrium that maximizes the sum of flexible agents' payoffs for any symmetric bargaining game.[9] Optimal mediation is symmetric, with each rational agent facing the same distribution of agreement times, and getting half the dollar when facing a rational opponent. The mediator delays agreements between two flexible agents by less than agreements between a flexible and inflexible type, to ensure that flexible types have incentives to truthfully reveal themselves. Mediation improves on unmediated payoffs if and only if agents are risk averse, or inflexible demands are larger than the probability of inflexibility.

A first interesting feature of this result is that curvature of the utility-possibility frontier can be both necessary and sufficient for beneficial mediation. In unmediated bargaining, flexible agents sometimes concede and are sometimes conceded to, which is inefficient if they are risk averse (an inefficiency that has received much less attention in the theoretical literature than delay). By replacing dispersed agreement terms with a single average agreement the mediator can always improve payoffs. Indeed, doing so creates strict incentives for flexible agents to reveal their type to the mediator and so she can then also reduce delay while preserving incentives.

The special importance of mediation when there is curvature in the utility-possibility frontier obviously extends beyond risk aversion,[10] and in particular applies to bargaining with multiple issues. Goldberg et al. (2012) emphasizes the special role for a mediator with multiple issues, because these allow for "integrative" bargaining solutions, where the mediator convinces parties to sacrifice low value issues in return for higher value concessions elsewhere. The mediator's role in devising an appropriate agreement "package" is also highlighted in Dunlop's quotation. The simplest illustration of how multiple issues create curvature is when agents must divide two continuous pies (with no transferable utility), where agent i values fraction $x_{i,j}$

[7] This may help explain why mediators are typically paid by the hour (see Velikonja (2009)) and is consistent with the advice of veteran mediators to not overly prioritize reaching an agreement (e.g. see Brazil (2007)).

[8] This restriction is with some loss of generality. It is discussed in Section 4.

[9] Symmetric agents have equal impatience, behavioral probability, behavioral demands, and utility functions.

[10] For example, warring parties (or divorcing parents) may strictly prefer a 50/50 split of disputed land (parenting time) to a 50/50 chance of all the land (sole custody) or none.

of pie j at $x_{i,j}v_{i,j}$ and $v_{i,i} > v_{i,-i} > 0$.[11] Giving each agent i all of pie i, therefore, provides higher payoffs than giving each a 50% chance of their first best agreement (all of both pies).

A second interesting feature of the result is that mediation may be most beneficial for significant disputes, where parties' demands are far apart. This is exemplified by the need for demands to be larger than the probability of inflexible types for mediation to improve outcomes for risk neutral agents. It is somewhat counterintuitive, as one might think parties have to give up too much for compromise to work in this case, however, the smaller payoff from conceding (to a large demand) increases a flexible agent's incentive to truthfully reveal her type to the mediator.

In addition to unmediated bargaining, a second important benchmark with which to compare mediation is a mechanism design problem where the designer can impose agreements, and disagreement between two reportedly inflexible types.[12] Not surprisingly, this benchmark always does better than mediation. In fact, the difference is dramatic. When inflexible types are likely or inflexible demands are moderate (close to 1/2) it achieves full efficiency, with payoffs matching those under complete information.[13] By contrast, the mediator was unable to improve on unmediated bargaining for risk neutral agents in exactly those circumstances. The very different predictions show that a mediator is constrained far more by agents' freedom to ignore her instructions, than by the informational problem alone.

Finally, it is worth emphasizing that the paper adopts the perspective that inflexible behavior is not vanishingly unlikely. When it is, AG show that for generic (asymmetric) bargaining games, bargainers agree immediately with probability approaching one, even without a mediator. Without curvature in the utility-possibility frontier, mediation may not be beneficial when *flexible* agents are vanishingly unlikely, however, I show it can still deliver substantially higher payoffs in such generic games for intermediate likelihoods of commitment.

The rest of the paper is arranged as follows. The remainder of this section discusses additional related literature; Section 2 outlines the model; Section 3 analyzes the model both without mediation and with simple mediation protocols inspired by Dunlop; Section 4 characterizes good faith equilibria and identifies an optimal mediation protocol, before comparing it to a mechanism design benchmark; Section 5 is a conclusion. All proofs are in Appendix B.

Further related literature

There is fairly extensive literature on uninformed mediation in simple one-shot games, which share some similarities with my results, despite important differences to dynamic bargaining.

[11] If values are private, a mediator may also filter this information. However, Jackson et al. (2015) show that with many privately valued issues and no reputational concerns, bargaining is almost efficient without a mediator.

[12] The difficulty of imposing disagreement may explain why such mechanisms aren't used instead of mediation.

[13] This contrasts with the necessary inefficiency of Myerson and Satterthwaite (1983) with continuous overlapping value distributions. The reputational model, however, is closer to a discrete value distribution setting.

Goltsman et al. (2009) characterizes the extent to which mediation, arbitration and negotiation (finitely many rounds of communication but with no discounting) can improve receiver payoffs in a sender-receiver game. Arbitration is (generically) more effective than mediation, while mediation is only sometimes more effective than communication. Both mediators and arbitrators filter information, but mediators must add noise. This is in line with my finding that noise is necessary to make Dunlop's simple protocol effective; however, the importance of noise isn't restricted to a mediator's communication. In a simple sender-receiver game with no mediator, Myerson (1991) highlights how noisy communication is informative, but noiseless communication isn't. In a one-shot conflict game, Fey and Ramsay (2010) shows that mediation cannot improve on unmediated communication. In a slight variant of this setting, however, Hörner et al. (2015) shows that arbitration and mediation are equally effective at deterring conflict, and these outperform communication when the intensity of conflict is high, or asymmetric information is large. Meirowitz et al. (2019) shows that unmediated peace talks increase the incentive to militarize and so increase eventual conflict, but mediated peace talks reduce militarization and conflict. The importance of mediator commitment for my results is in line with literature investigating the incentives of informed mediators. In particular, Kydd (2001) finds that mediators who only want to avoid war are ineffective, but biased mediators can help.

Jarque et al. (2003) is an important paper considering mediation in dynamic bargaining with two-sided private information. Time is continuous, and agents have reservation values drawn from a continuous distribution. A war of attrition equilibrium always exists, where agents only ever make two demands. A mediator adopts a simple version of Dunlop's protocol that is ineffective in the reputational model: she immediately announces a compromise when both parties accept it in private. When fundamentals are symmetric, there is an equilibrium with mediation if and only if the fraction of types willing to concede in the war of attrition is sufficiently small. Ex-ante efficiency can be higher than in the war of attrition. The reason Dunlop's simple protocol "works" here but not in the reputational model is that the mediator facilitates agreement between types who would never otherwise agree. For example, if buyers demand a price of 1 and sellers a price of 4 in a war of attrition, then a buyer with value 3 and a seller with cost 2 will never agree, but they can at the mediator's compromise price of 2.5. The extra payoffs from such agreements can add enough grease to the system to overcome the inherent difficulties of mediation. In the reputational model, by contrast, introducing a compromise agreement does not expand the set of types who ultimately agree. Čopič and Ponsatí (2008) extends this model to allow for a continuum of possible compromise agreements, and illustrates the existence of a mediator supported equilibrium, where all compatible types eventually agree.

Despite identifying an equilibrium with mediation, it is unclear whether mediation is actually beneficial in Jarque et al. (2003), because multiple equilibria can exist without mediation (the result only compares mediation to war of attrition equilibria). Without mediation, if *only* three alternative agreements are possible and strategies are Markov, then Ponsati (1997) shows

there must be a war of attrition in the reservation value model. However, she also constructs non-Markov equilibria, in which compromise agreements are used and shows these are more efficient than the war of attrition. Of course, it is also unclear why agents should only have three alternative agreements, when they could also seemingly agree to choose each alternative with positive probability (making the agreement set convex).

My equilibria with mediation are effectively communication equilibria in the sense of Myerson (1986). The fact that communication equilibria take the game being played as fixed distinguishes this from a typical mechanism design exercise. The presence of a fixed game is also shared with information design; indeed, the mediator's problem can be thought of as information design with an uninformed designer. Information design assumes that the designer has her own independent source of information. These links imply similarities between my optimal mediation problem and the literature on dynamic information design (e.g. see Ely (2017)), which is similarly concerned with the optimal release of information over time to affect behavior. The problem of how to optimally release information over time in order to affect behavior is also addressed by the literature on motivational ratings (e.g. see Hörner and Lambert (2016), and also Ekmekci (2011) for motivational ratings with behavioral types).

Basak (2019) considers a model very similar to reputational bargaining with a form of static information design. An informed mediator has access to a signal about the likelihood that an agent is committed to her bargaining demand. If the signal is perfectly informative then its release makes bargaining efficient (eliminating delay when at least one party is not committed); however, if the signal is only moderately informative, then its release may reduce payoffs.

Finally, it is important to acknowledge that mediation has many other reputed benefits beyond those considered in this paper (e.g. see Goldberg et al. (2012)). These include the mediator's acknowledgement of each side's grievances, her ability to create a less confrontational atmosphere for negotiation, and her ability to establish commonly accepted facts.

2 The model

The model presented below encompass all the mediation protocols I consider in a consistent way. The setup adapts the discrete-continuous time bargaining protocol of Abreu and Pearce (2007), although for much of the analysis time can be treated as completely continuous.

Two bargainers, $i = 1, 2$, must agree on how to divide a dollar and face an infinite horizon. Bargainers are either rational or behavioral types. I follow Abreu and Pearce (2007) in using the terms rational and behavioral in my formal analysis (rather than flexible and inflexible), which highlights behavioral types' lack of preferences. If rational bargainer i obtains a share $x_i \in [0, 1]$ of the dollar at real time t then her utility is $e^{-r_i t} u_i(x_i)$ where her discount rate is r_i and her utility function u_i is strictly increasing and concave with $u_i(0) = 0$. Behavioral types have

no preferences, but mechanically implement an exogenously defined strategy. A third player is a mediator, $i = 3$, who implements an exogenously fixed communication strategy (i.e. she is a behavioral type). There is no fee for mediation services.

Time is discrete-continuous to allow multiple events to occur at the same time in a sequential order. Each positive real time $t \in [0, \infty)$ is divided into five different discrete times t^1, t^2, t^3, t^4, t^5. Time follows a lexicographic ordering so that $t^k < t^{k+1}$, and $t^k < s^l$ whenever $t < s$. The set of discrete continuous times is $DC = [0, \infty) \times \{1, 2, ..., 5\} \cup \{\infty\}$. There is no discounting of payoffs *within* each time t. The bargaining protocol is as follows: At time 0^1 each bargainer i simultaneously announces a demand $\alpha_i(0^1) \in [0, 1]$; at $t^1 > 0^1$ each bargainer can concede to her opponent's existing demand (accept the share $(1 - \alpha_j(t^1))$), ending the game; at any t^2 each bargainer can send a private message to the mediator $\theta \in \Theta$, where typically this will simply indicate that she is rational; at t^3 the mediator can send a public message $\theta' \in \Theta$ to the agents, where typically this will simply involve a suggested dollar division $\theta' = (m_1, m_2)$ where $m_1 = 1 - m_2 \in [0, 1]$ (the message space is arbitrary but sufficiently rich that $\Theta \supseteq [0, 1]^2$); at t^4 each bargainer can simultaneously change her demand to $\alpha_i(t^4)$; at t^5 each bargainer can concede to her opponent's (possibly new) existing demand. If both bargainers concede at the same time then each proposal is selected with probability $\frac{1}{2}$.

At every $t^k > 0^1$ each bargainer is associated with an existing demand. If bargainer i changes her demand at t^4 then she cannot change her demand again until time $(t + \Delta)^4$ for some $\Delta > 0$. That is, if $\alpha_i(t^3) \neq \alpha_i(t^4)$ then i's existing demand at s^k is $\alpha_i(s^k) = \alpha_i(t^4)$ for $t^4 \leq s^k < (t + \Delta)^4$. Notice that agents can, however, change their demand from their initial demand at $(\Delta/2)^4$ because initial demand announcements at 0^1 are not counted as a change. Similarly, if agent i sent a message at t^k, then she cannot send another message until $(t + \Delta)^k$. These restrictions ensure that the bargaining environment is relatively stable, and so strategies and outcomes are more easily defined, but the fact that $\Delta > 0$ is not used to argue for any result.

The above setup can be extended to allow the mediator and agents to send both private and public messages. This does not affect any results, but makes describing agents' information and strategies considerably more cumbersome. The setup can also be adjusted to ensure that there is always a "first" time at which agents can change their demands after a change in the bargaining environment.[14] Again, this does not affect any results.

Bargainer i is a behavioral type with probability $z_i \in (0, 1)$, and is otherwise rational. A behavioral type for bargainer i initially demands a share $\alpha_i(0^1) = \alpha_i \in (0, 1)$ and never changes this.[15] She concedes to her opponent's demand at $t^k \in \{t^1, t^5\}$ if and only if $(1 - \alpha_j(t^k)) \geq \alpha_i$. She

[14] Let t be divided into $t^1, ..., t^{12}$. Times $t^1, ..., t^4$ are as before. At t^5 (at t^8) [at t^{11}] an agent can message the mediator if she observed an action at t^3 or t^5 (at t^6 or t^7) [at t^8 or t^9]. At t^6 (at t^9) the mediator can message agents if she observed an action at t^4 (at t^7). At t^7 (at t^{10}) agents can change demands if they observed an action at t^3, t^4 or t^6 (at t^6, t^7 or t^9). At t^{12} agents can concede. If an agent changes her demand at t^k, she can't do so again before $(t + \Delta)^1$. If an agent/the mediator sends a message at t^k, she can't do so again before $(t + \Delta)^1$.

[15] In AG, agents can imitate multiple behavioral types. I discuss this extension in the Conclusion.

never sends a message to the mediator, and so any message indicates rationality. (Alternatively, I could assume there is a message $\theta^* \in \Theta$ which a behavioral type never sends.) Because of this, I say that an agent who sends a message to the mediator *confesses* rationality, and is a *confessing* agent, otherwise she is a *non-confessing* agent. The behavioral demands of the two bargainers are incompatible, $\alpha^1 + \alpha^2 > 1$.

We can describe an explicit extensive form and strategies for this game by using stopping times.[16] An agent has a new information set (private history) only when she observes a change in her bargaining environment. At each of her information sets, she chooses an *action plan*, comprised of a planned future action and future action (stopping) time. A behavior strategy selects (potentially randomly) an action plan for every possible private history. Further details of this game form are laid out in Appendix A.

A perfect Bayesian equilibrium requires that at each of bargainer i's possible private histories, her behavior strategy maximizes her continuation payoff at that time, given others' strategies and her beliefs. Beliefs are determined by Bayes rule where possible.

3 Unmediated bargaining and simple mediation protocols

In this section, I first highlight the unique equilibrium of the model without mediation. I then examine the simplest version of the communication protocol suggested by Dunlop, in which the mediator seeks a specific compromise and immediately suggests that agreement if and only if both agents provisionally accept it. Finally, I add noise to that simple protocol.

3.1 A Baseline Without Mediation

I call bargaining without mediation the *Baseline* model. Without a mediator we can ignore times t^2 and t^3. AG's results (Lemma 1) imply that if agent i is revealed to be rational at time t^k in equilibrium (i.e. after i makes a non-behavioral demand), while agent j may be behavioral, then i must immediately concede. This relies only on continuation strategies being optimal at t^k. It mirrors the logic of the Coase conjecture in that one-sided asymmetric information implies an immediate agreement favorable to the informed party.

Given this immediate concession after revealing rationality, it is without loss of generality to assume that rational agents always imitate behavioral types and then simply choose when to concede. We can, therefore, describe (the on equilibrium path part of) agent j's strategy in continuous time, with a cumulative distribution function, $F_j \in [0,1]^{[-\infty,\infty]}$, where $F_j(t)$ is the *total* probability that agent j (who may be behavioral) has conceded before extended real time t. We then have $F_j(t) = 0$ for $t < 0$, while never conceding corresponds to a concession time of

[16] By using stopping times we can avoid many of the pathologies of continuous time games.

∞, so that $F_j(\infty) = 1$. Agent j's reputation for being behavioral at t is, therefore, $\bar{z}_j(t) = \frac{z_j}{1-F_j(t)}$. Given agent j's strategy, agent i's expected payoff to conceding at t is:[17,18]

$$U_i(t) = \int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j(s) + (1-F_j(t)) e^{-r_i t} u_i(1-\alpha_j) + \left(F_j(t) - \sup_{s<t} F_j(s)\right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j))$$

The unique equilibrium of this model is characterized by three properties: (i) at most one agent concedes with positive probability at time zero; (ii) both agents reach a probability one reputation at the same time, $T^* < \infty$; and (iii) agents are indifferent to conceding at any time on $(0, T^*]$. This third indifference condition implies that agent j must concede on the interval $(0, T^*]$ at the constant rate:

$$\frac{f_j(t)}{1 - F_j(t)} = \lambda_j = \frac{r_i u_i(1 - \alpha_j)}{u_i(\alpha_i) - u_i(1 - \alpha_j)} \tag{1}$$

This implies that $1 - F_j(t) = (1 - F_j(0))e^{-\lambda_j t}$. Next define rational agent j's *exhaustion time*, $T_j = -\frac{1}{\lambda_j} ln(z_j)$, as the time by which she must have conceded even if she did not concede at time zero (so $1 = z_j e^{\lambda_j T_j}$). Condition (i) and (ii) then imply $T^* = \min\{T_1, T_2\}$, and finally:

$$1 - F_j(0) = z_j e^{\lambda_j T^*} = \min\left\{1, z_j z_i^{-\frac{\lambda_j}{\lambda_i}}\right\} \tag{2}$$

Proposition 1 (AG, Proposition 1)**.** *The Baseline model has a unique distribution of equilibrium outcomes, characterized by equations (1) and (2).*

The fact that $T^* > 0$ implies that rational agents sometimes inefficiently delay agreement. This offers scope for mediation to improve outcomes. Baseline equilibrium payoffs are:

$$U_i^B = u_i(\alpha_i) F_j(0) + u_i(1 - \alpha_j)(1 - F_j(0))$$

3.2 Simple Dunlop (SD) mediation

I next consider the simplest version of the mediation protocol suggested by Dunlop. The mediator suggests an agreement (m_1, m_2) at 0^3 if and only if both agents confess their rationality at 0^2 and otherwise remains silent. I call this the *Simple Dunlop* (SD) mediation protocol, and show that it cannot improve on unmediated bargaining outcomes.

If the mediator *does* make an announcement at 0^3, then both agents are revealed to be rational. In this case, any dollar division or even perpetual delay is consistent with sequential rationality.[19] While this is different from discrete time models where there may be a unique

[17] Here and elsewhere, I suppress the explicit dependence of payoffs on strategies to minimize notation.

[18] Here and elsewhere, I assume (without loss of generality) that agents only ever concede at t^5 (and not at t^1).

[19] Agent i changes her demand to $\alpha_i(0^4) \in [0, 1]$ and subsequently plans not to concede unless j offers her more

continuation equilibrium when both agents are known to be rational, by revealing information on rationality sequentially in discrete time, an informed mediator would still have wide freedom to implement her preferred agreement.[20] Given the eventual negative result of SD mediation, it is without loss of generality to assume that agents do follow the mediator's suggestion (expected continuation payoffs must be weakly below some mediator proposal, and even those payoffs can't incentivize confession). If the mediator doesn't announce an agreement, but an agent subsequently reveals rationality, she must immediately concede.[21]

We can again simplify to a continuous time framework. Let agent i's (on equilibrium path) strategy be described as follows. Define $p_i^c \in [0,1]$ as the *total* probability that agent i confesses at 0^2 (c=confess). If both agents confess then agent i obtains the payoff $u_i(m_i)$; if not, she must choose when to concede. Agent i's concession choice is described by two cumulative distribution functions $F_i^c \in [0,1]^{[-\infty,\infty]}$ and $F_i^n \in [0,1]^{[0,\infty]}$. Let $F_i^c(t)$ be $F_i^c(t)$ be the probability that agent i has conceded to her opponent before extended real time t conditional on her confessing and no mediator suggestion. Similarly, let $F_i^n(t)$ be the probability that agent i has conceded before time t, conditional on her not confessing (n=not confess). Finally, let $F_i(t) = p_i^c F_i^c(t) + (1-p_i^c)F_i^n(t)$ be the probability that agent i has conceded by time t conditional on no mediator suggestion. Note that while I have not included additional subscripts or superscripts on F_i, it may be distinct from the distribution in the Baseline equilibrium.

Given agent j's strategy, rational agent i's utility from confessing and then conceding at time t if the mediator makes no suggestion, is:

$$\begin{aligned} U_i^c(t) =& p_j^c u_i(m_i) + (1-p_j^c)\Bigg(\int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j^n(s) + (1-F_j^n(t)) e^{-r_i t} u_i(1-\alpha_j) \\ &+ \left(F_j^n(t) - \sup_{s<t} F_j^n(s) \right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \Bigg) \end{aligned}$$

Alternatively, rational i's utility if she does not confess and concedes at time t is:

$$\begin{aligned} U_i^n(t) =& \int_{s<t} e^{-r_i s} u_i(\alpha_i) dF_j(s) + (1-F_j(t)) e^{-r_i t} u_i(1-\alpha_j) \\ &+ \left(F_j(t) - \sup_{s<t} F_j(s) \right) e^{-r_i t} \frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \end{aligned}$$

It is clear that the Baseline model's equilibrium can still be an equilibrium here; indeed this

than that. The claim is formally established in Lemma 3 in the Appendix.

[20] For instance, consider an alternating offer model. With complete information agents immediately agree to a division $(\alpha_1^R, 1-\alpha_1^R)$. Suppose an informed mediator wants a division (m_1, m_2) between rational agents where $m_1 \in (1-\alpha_2, \min\{\alpha_1^R, \alpha_1\})$ and for rational agents to concede to behavioral types. The mediator immediately announces whether agent 1 is behavioral. If agent 1 initially demands m_1, the mediator will reveal whether agent 2 is rational in period 2 and otherwise remain silent. Accepting m_1 is optimal for a rational agent 2 given $m_1 \leq \alpha_1^R$. Demanding m_1 is optimal for rational agent 1 for small Δ, because otherwise the resulting game of one-sided private information can only give her marginally more than $u_1(1-\alpha_2)$, due to the Coase conjecture (see AG).

[21] This again follows immediately from AG, Lemma 1.

is the case in all mediation protocols considered. If agent j does not confess with positive probability then agent i has no incentive to do so either.

It is also clear that there can be no equilibrium with $m_i \in (1-\alpha_j, \alpha_i)$ in which rational agent i *always* confesses and j does so with positive probability. If there was, then a confessing agent j would learn for sure that i was behavioral if the mediator made no announcement, and so would subsequently concede immediately. Knowing this, a rational agent i would optimally choose not to confess because this would give her a larger payoff, $\alpha_i > m_i$. Any equilibrium with mediation and $m_i \in (1-\alpha_j, \alpha_i)$, therefore, must involve both rational agents mixing between confessing and not.

The next proposition shows that there is no equilibrium where mediation ever improves on unmediated bargaining. While there can sometimes be an equilibrium where both parties confess with positive probability if $m_i \in \{1-\alpha_j, \alpha_i\}$, outcomes in this case remain identical to those in the Baseline equilibrium.

Proposition 2. *The distribution of outcomes in any equilibrium where the mediator adopts the Simple Dunlop mediation protocol is identical to that in the unique Baseline equilibrium.*

The explanation for this result is similar to why it is impossible for rational agents to always confess. I prove that if the mediator does not suggest an agreement (at 0^3), then at least one confessing agent, say j, must immediately concede with probability one ($F_j^c(0) = 1$). Such concession then destroys the incentive for her opponent to confess in the first place.

To understand why a confessing agent must immediately concede if there is no agreement suggested we must consider beliefs. If there is no agreement before time t, then agent i believes she faces a behavioral type with probability $\overline{z}_j^c(t) = \frac{z_j}{(1-p_j^c)(1-F_j^n(t))}$ if she did confess, but with probability $\overline{z}_j^n(t) = \frac{z_j}{1-F_j(t)}$ if she didn't confess.[22] She is, therefore, more pessimistic about her opponent's type if she confessed, $\overline{z}_j^c(t) \geq \overline{z}_j^n(t)$ and so, other things equal, has a greater incentive to concede earlier. This possibly small difference in beliefs ultimately has a big equilibrium effect. By adapting standard war of attrition arguments, I shows that if neither confessing agent has always conceded by time t ($F_j^c(t) < 1$, so $\overline{z}_j^c(t) > \overline{z}_j^n(t)$), then such agents must concede at a constant rate at t ($\frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$). But this constant concession rate would mean that they never concede with probability one in finite time, which cannot be optimal for a rational agent facing a possibly behavioral opponent.

The result is robust to allowing agents to confess continuously over time. In the Online Appendix, I extend the above model to allow agents to privately confess rationality to the mediator over the infinite horizon with the mediator suggesting an agreement as soon as both have confessed. This is the mediation protocol studied by Jarque et al. (2003) when agents have private

[22] There are potentially non-degenerate higher order beliefs in this game. If j did not confess, then she believes that i's beliefs about her likelihood of being behavioral are $\overline{z}_j^c(t)$ with probability $\frac{p_i^c(1-F_i^c(t))}{1-F_i(t)}$ and $\overline{z}_j^n(t)$ otherwise.

reservation values. I call it the *Ongoing Dunlop (OD)* mediation protocol and again show that it cannot improve on the Baseline equilibrium.

3.3 Noisy Dunlop (*ND*) mediation

I next consider a *Noisy Dunlop* (*ND*) mediation protocol which adds noise to the mediator's strategy in the *SD* protocol. The noise takes the form of the mediator failing to suggest an agreement even when both parties confess with positive probability. I show that this adapted protocol can improve on unmediated outcomes if behavioral types are unlikely.

The significance of the result is that it represents the first clear theoretical demonstration of the benefits of mediation in dynamic bargaining (to the best of my knowledge), and moreover, the mediation protocol used is close to one actually employed by professional mediators. In conjunction with the failure of the *SD* protocol (Proposition 2) it highlights the desirability of commitment for mediators: it is important for mediation to sometimes fail, in order to succeed.

In the *ND* protocol, if both agents confess at 0^2 the mediator suggests the agreement (m_1, m_2) at 0^3 with probability $b \in (0,1)$, and otherwise remains silent. This noise can also be interpreted as each agent's message going astray or being misinterpreted by the mediator with probability $1 - \sqrt{b}$ (with the mediator always announcing agreement when she knows both parties have confessed). I focus attention on what I call *ND equilibria*, which are equilibria where the mediator adopts the *ND* protocol, while rational agents always confess at 0^2 and subsequently immediately implement any mediator suggestion.

If the mediator makes no suggestion at 0^3, then rational agents must decide when to concede. We can then describe (on path) continuation strategies using the cumulative distribution function $H_i^c \in [0,1]^{[-\infty,\infty]}$, where $H_i^c(t)$ is the probability that a *rational* agent i has conceded before extended real time t conditional on confessing and the mediator making no suggestion. In an *ND* equilibrium, agent i's utility if she confesses and concedes at time t is then:

$$\begin{aligned} U_i^c(t) =&(1-z_j)\left(bu_i(m_i) + (1-b)\int_{s<t} e^{-r_i s}u_i(\alpha_i)dH_j^c(s)\right) + \left((1-z_j)(1-b)(1-H_j^c(t)) + z_j\right)e^{-r_i t}u_i(1-\alpha_j) \\ &+ (1-z_j)(1-b)\left(H_j^c(t) - \sup_{s<t} H_j^c(s)\right)e^{-r_i t}\frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \end{aligned} \tag{3}$$

Agent i's utility if she does not confess and then concedes at time t is:

$$\begin{aligned} U_i^n(t) =&(1-z_j)\int_{s<t} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) + \left((1-z_j)(1-H_j^c(t)) + z_j\right)e^{-r_i t}u_i(1-\alpha_j) \\ &+ (1-z_j)\left(H_j^c(t) - \sup_{s<t} H_j^c(s)\right)e^{-r_i t}\frac{1}{2}(u_i(\alpha_i) + u_i(1-\alpha_j)) \end{aligned}$$

In an *ND* equilibrium, if the mediator does not make a suggestion at 0^3, then rational agent j must believe that her opponent is behavioral with probability $\bar{z}_i = \frac{z_i}{1-(1-z_i)b}$. In this case, behavior in the continuation game must resemble that of the Baseline model but with initial reputations

$\bar{z}_i$ instead of z_i. As noted previously, this equilibrium is characterized by three conditions: (i) at most one agent concedes with positive probability at time zero; (ii) both agents reach a probability one reputation at the same time, $T^* < \infty$; and (iii) agents are indifferent to conceding at any time on $(0, T^*]$.

Let $F_j(t) = (1-\bar{z}_j)H_j^c(t)$ be the probability that a confessing agent i believes that j will concede before t conditional on no mediator announcement. Condition (iii) then implies that agent i must expect j to concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on $(0, T^*]$. Agent j's exhaustion time is now $T_j = -\frac{1}{\lambda_j} ln(\bar{z}_j)$. To ensure conditions (i) and (ii) are satisfied, we must have $T^* = min\{T_1, T_2\}$ and $1-F_j(0) = \min\left\{1, \bar{z}_j \bar{z}_i^{-\frac{\lambda_j}{\lambda_i}}\right\}$. More generally for $t \leq T^*$ we must have $1-F_j(t) = (1-F_j(0))e^{-\lambda_j t}$.

Given such behavior, a rational agent i who did not confess (not her equilibrium strategy) will subsequently find it in her interest to wait until T^* and then concede. This is because conditional on no mediator suggestion, the rate at which i expects j to concede is larger if she did not confess than if she did. That is, $\frac{(1-z_j)h_j^c(t)}{(1-z_j)(1-H_j^c(t))+z_j} > \frac{(1-\bar{z}_j)h_j^c(t)}{(1-\bar{z}_j)(1-H_j^c(t))+\bar{z}_j}$ on $(0, T^*)$, which implies $\frac{U_i^n(t)}{dt} > \frac{U_i^c(t)}{dt} = 0$. Hence, her continuation payoff, U_i^{*n}, after not confessing is:

$$U_i^{*n} = \max_t U_i^n(t) = (1-z_j)\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s) + z_j e^{-r_i T^*} u_i(1-\alpha_j)$$

When agent i does confess, she is subsequently indifferent to conceding at any $t \in (0, T^*]$. Her equilibrium payoff, U_i^{*c}, can therefore be written as:

$$U_i^{*c} = \max_t U_i^c(t) = U_i^c(T^*) = (1-z_j)\left(bu_i(m_i) + (1-b)\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)\right) + z_j e^{-r_i T^*} u_i(1-\alpha_j) \tag{4}$$

This payoff suggests that confessing leads to a lottery giving $u_i(m_i)$ with probability $(1-z_j)b$, $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$ with probability $(1-z_j)(1-b)$, and $e^{-r_i T^*} u_i(1-\alpha_j)$ with probability z_j. By contrast, not confessing implies a lottery giving $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$ with probability $(1-z_j)$, and $e^{-r_i T^*} u_i(1-\alpha_j)$ with probability z_j. Comparing these lotteries therefore reduces to comparing $u_i(m_i)$, the agent's payoff from a mediated agreement, and $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$, the stream of payoffs from a known rational agent's concession. That is, a necessary and sufficient condition for an *ND* equilibrium to exist is for $i = 1, 2$:

$$Q_i = \frac{U_i^{*c} - U_i^{*n}}{(1-z_j)b} = u_i(m_i) - \int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s) \geq 0 \tag{5}$$

The paper's first positive result (Proposition 3, below) shows that when agents' reputations are sufficiently small, an *ND* equilibrium always exists that (strictly) Pareto dominates the equilibrium of the Baseline model. The result allows b to be chosen arbitrarily close to one, and so we can guarantee efficiency as behavioral types become vanishingly unlikely.

Proposition 3. *For any given r_i, u_i, α_i for $i = 1, 2$, $b \in (0, 1)$ and fixed $K \geq 1$, there exists $\underline{z} > 0$ such that whenever $z_i \leq \underline{z}$ and $K \geq \frac{z_1}{z_2} \geq \frac{1}{K}$, an ND equilibrium exists, which both rational agents strictly prefer to the Baseline equilibrium.*

The reason equation (5) holds when behavioral probabilities are small is because in this case, the stream of payoffs from a rational opponent's concession, $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$, comes fairly slowly. For any $b \in (0, 1)$, if z_j is small then so is $\bar{z}_j$, and so the probability that agent j concedes before time t is very close to the probability that rational agent j concedes, $(1 - \bar{z}_j)H_j^c(t) \approx H_j^c(t)$. Hence the value of the stream of payoffs $\int_{s<T^*} e^{-r_i s} u_i(\alpha_i) dH_j^c(s)$, is approximately the same as her payoff in a Baseline equilibrium with reputations $\bar{z}_j$. Because the Baseline equilibrium is inefficient, however, it is possible to choose an m_i to satisfy equation (5) for both agents.

It is illuminating to consider the special case of symmetric bargaining games with $u_i = u$, $r_i = r$, $\alpha_i = \alpha$, $z_i = z$. In this case, we have $H^c(t) = (1 - e^{-\lambda t})/(1 - \bar{z})$. This implies $\int_{s<T^*} e^{-rs} u(\alpha) dH^c(s) = \frac{1-\bar{z}^{\frac{u(\alpha)}{u(1-\alpha)}}}{1-\bar{z}} u(1 - \alpha)$, which converges to $u(1 - \alpha)$ as $\bar{z} \to 0$ and $u(\alpha)$ as $\bar{z} \to 1$. We can, therefore, satisfy equation (5) for both agents with any $m_1 = 1 - m_2 \in (1 - \alpha, \alpha)$ as $\bar{z} \to 0$, although it is easiest to satisfy with $m_i = 0.5$. Given that $Q_i(1 - \bar{z})$ is concave in $\bar{z}$, there is a unique $\hat{z} \in (0, 1)$ such that equation (5) is satisfied with equality. We then have an *ND* equilibrium if and only if $\bar{z} \leq \hat{z}$, where the maximum feasible b is $b = \frac{\hat{z}-z}{\hat{z}-z\hat{z}}$. As an example, suppose agents are risk neutral and $\alpha = 0.75$, then we have $\hat{z} = 0.62$. Even a fairly large behavioral probability such as $z = 0.5$, then allows for $b = 0.38$ and payoffs of $U_i^{ND} = 0.30$. This compares to Baseline payoffs of $U_i^B = 0.25$, and a complete information upper bound on payoffs, when rational agents immediately split the dollar and concede to behavioral opponents, of $U_i^{CI} = 0.38$. A smaller behavioral probability $z = 0.25$, allows for $b = 0.79$ and payoffs of $U_i^{ND} = 0.40$ (compared to $U_i^B = 0.25$ and $U_i^{CI} = 0.44$).

It might seem strange that adding noise makes a difference to the success of mediation; after all, agents could use mixed strategies in the *SD* protocol (a particular form of noise). Indeed, if both agents confessed with probability b under the *SD* protocol then conditional on agent i confessing and hearing no mediator announcement, she will believe that j is behavioral with probability $\bar{z}_j$. These situations are quite distinct, however. In particular, when agents mix under the *SD* protocol, continuation play after time zero must provide incentives for both a confessing and non-confessing agent to concede continuously, whereas an *ND* equilibrium only needs to provide dynamic incentives for a confessing agent. As mentioned previously, the need for noise in effective communication is in line with the existing literature (e.g. see Myerson (1991)).

One nice feature of the *ND* mediation protocol is that it is fairly robust, in the sense that the mediator needn't know anything about the underlying parameters about the model. While it was convenient to assume that she suggested agreement terms (m_1, m_2), she could just send an arbitrary message θ' with probability $b < 1$ if both agents messaged her. When the probabil-

ity of behavioral types is small enough, there is an equilibrium where rational agents always confess and coordinate among themselves on an appropriate (m_1, m_2) after observing θ'.

We saw above that mediation cannot improve on unmediated outcomes in symmetric bargaining games when behavioral types are sufficiently likely. Proposition 4 shows that this is also true in asymmetric games. The reason can be readily ascertained, by reexamining the inequality in equation (5). By confessing, an agent gains an immediate payoff of $u_i(m_i)$ but loses a delayed payoff of $u_i(\alpha_i)$, when she faces a rational opponent. When even one agent is likely to be behavioral, however, there is very little delay (i.e. $T^* \approx 0$) and so the mediator must propose $m_i \approx \alpha_i$ to incentivize confession. Clearly she can't do this for both agents.

Proposition 4. *For any given r_i, u_i, α_i for $i = 1, 2$ there exists $\underline{z} < 1$ such that if $z_1 \geq \underline{z}$, then no ND equilibrium exists.*

This result does *not* imply that *ND* mediation is unable to deliver substantial benefits outside of (non-generic) symmetric bargaining games. AG shows that when $\lambda_i \neq \lambda_j$, agents agree immediately with probability approaching one even without mediation, as the probability of behavioral types becomes vanishingly small. However, Proposition 4 merely says that mediation can't improve outcomes when the probability of behavioral types is close to one. Computed examples reveal that *ND* mediation does deliver substantial benefits in asymmetric games for intermediate behavioral probabilities. For instance, if agents are risk neutral, $\alpha_1 = 0.75$, $\alpha_2 = 0.7$, $r_i = 1$ then $\lambda_1/\lambda_2 = 5/6$. If $z_i = z$ then mediation is beneficial whenever $z < 0.50$. If $z = 0.25$ the mediator can suggest immediate agreement between rational agents with probability $b = 0.66$ when $m_1 = 1 - m_2 = 0.50$, which delivers payoffs of $U_i^{ND} = 0.40$ to each agent. This compares with Baseline equilibrium payoffs of $U_1^B = 0.3$ and $U_2^B = 0.34$, and a complete information upper bound on payoffs of $U_1^{CI} = 0.45$ and $U_2^{CI} = 0.44$.

4 Good faith equilibria and optimal mediation

The previous analysis has restricted the mediator to very particular strategies. We would like to understand more generally what mediators can and cannot do by characterizing the set of equilibria when allowing for all possible mediator strategies. Moreover, having done this we would like to say something substansive about "optimal" mediation. This section makes progress towards these goals by adopting a mechanism design approach. I characterize the set of what I call good faith equilibria and within this set identify an essentially unique optimal equilibrium for symmetric bargaining games.

I define a *good faith* equilibrium to be one in which rational agents never demand more than behavioral types (on the equilibrium path). This is a large class of equilibria which includes the unmediated Baseline equilibrium and the *ND* equilibria of the last section, as special cases.

As mentioned in the Introduction, the name derives from the requirement in the National Labor Relations Act, that parties negotiate in good faith without withdrawing from previously agreed upon provisions. Good faith equilibria encompass all equilibria in which agents initially make behavioral demands, if agents can recall past offers. Mediators typically do intervene only after initial demands are in conflict, and there are clearly no equilibria where rational agents always make non-behavioral initial demands.[23] The ability to recall past offers is explicitly present in some bargaining situations, and implicitly present in many others, for instance because agents' reference points change so that they become committed to not accepting less than their opponent's most generous previous offer.[24] A more basic justification for good faith equilibria is that the idea of mediators encouraging more extreme demands does not seem plausible. Nonetheless, the restriction to good faith equilibria is with some loss of generality. In the Online Appendix I show that for symmetric bargaining games with risk averse agents and a sufficiently small probability of behavioral types, a bad faith equilibrium exists which delivers higher expected payoffs than any good faith equilibrium.[25]

We can describe the distribution of outcomes in any equilibrium as follows. Let $G^R \in [0,1]^{[-\infty,\infty]}$ be the cumulative distribution of equilibrium agreement times conditional on two rational agents (R=rational), so that $G^R(t)$ is the probability of agreement before extended real time t (with $t = \infty$ corresponding to no agreement). Likewise let $G_j^Z \in [0,1]^{[-\infty,\infty]}$ be the cumulative distribution function of agreement times conditional on agent j being behavioral (Z=behavioral) and i being rational, so that $G_j^Z(t)$ is the probability of agreement before extended real time t. The terms of any such rational-behavioral equilibrium agreement must be $(\alpha_j, 1-\alpha_j)$, because behavioral agent j always demands α_j allowing i to guarantee the dollar share $1-\alpha_j$ (and j never accepts less than α_j). Let $M_i^t \in [0,1]^{(-\infty,\infty)}$ be the cumulative distribution function of agent i's share conditional on an agreement between two rational agents at time t, so that $M_i^t(m)$ is the probability of agent i obtaining a share less than m. Feasibility implies $M_1^t(m) = 1 - \sup_{l>m} M_2^t(1-l)$ for all $m \in [0,1]$. The entire set of such distributions is described by the function $M_i : [0,\infty) \to [0,1]^{(-\infty,\infty)}$ such that $M_i(t) = M_i^t$. Finally define $T^R = \min\{t : G^R(t) = 1\}$ and $T_j^Z = \min\{t : G_j^Z(t) = 1\}$ on the extended real line.

We are interested in what constraints must hold in good faith equilibria. Given any such equilibrium, consider the following global deviation for rational agent i: act consistent with her equilibrium strategy up to time t^5 (this sometimes involves reaching agreement at $s^k \leq t^5$),

[23] If there were, then deviating to a behavioral demand would convince a rational opponent to immediately concede and so guarantee rational agent $i = 1, 2$ an equilibrium payoff of at least $(1-z_j)u_i(\alpha_i) + z_j u_i(1-\alpha_j)$, but that is infeasible given behavioral types' commitment and $\alpha_1 + \alpha_2 > 1$.

[24] This is assumed in Fershtman and Seidmann (1993). In fact any positive possibility of becoming committed to an opponent's previous offer will imply the implicit recall of such offers, due to the Coase conjecture.

[25] At time zero the mediator randomly selects one rational agent, say i, to demand the entire dollar. If she faces a behavior opponent, the mediator eventually tells i to concede and otherwise eventually tells $j \neq i$ to concede. This eliminates j's option of getting a positive payoff by conceding and so improves some incentive constraints. It shows that limiting agents' opportunity to strike a deal without the mediator can be useful.

but always concede an instant after that time.[26] Agent i's expected payoff if she adopts this deviation and obtains exactly the share $1-\alpha_j$ from conceding an instant after t^5 is:

$$\begin{aligned} U_i^c(t) =&(1-z_j)\int_{s\leq t} e^{-r_i s}\int u_i(m)dM_i^s(m)dG^R(s) + z_j\int_{s\leq t} e^{-r_i s}u_i(1-\alpha_j)dG_j^Z(s) \\ &+ e^{-r_i t}u_i(1-\alpha_j)\Big((1-z_j)(1-G^R(t)) + z_j(1-G_j^Z(t))\Big) \end{aligned}$$

In reality, agent i may obtain a larger share than $1-\alpha_j$ when she concedes an instant after t^5, but not less, given that agent j always demands less than her behavioral type. Notice that agent i's expected equilibrium payoff is $U_i^c(\max\{T^R, T_j^Z\})$. For agent i's equilibrium strategy to be optimal, she must not want to make the global deviation, and so we must have:

$$U_i^c(\max\{T^R, T_j^Z\}) = \max_t U_i^c(t) \quad \text{(Dynamic IC)} \tag{6}$$

I call this the *dynamic incentive constraint*. An immediate observation is that for this constraint to be satisfied we must have $T_j^Z \leq T^R$. If this were not true, $T^R < T_j^Z$, then agent i would realize at T^R that she faced a behavioral opponent j and would profitably concede.

Rational agent i also has the option of making an alternative global deviation in which she mimics a behavioral type prior to time t^5 (i.e. always demands α_i and never messages the mediator) and concedes an instant after that time. Agent i's expected payoff from this deviation, again assuming that she obtains the share $1-\alpha_j$ from conceding an instant after t^5 is:

$$U_i^n(t) = (1-z_j)\int_{s\leq t} e^{-r_i s}u_i(\alpha_i)dG_i^Z(s) + e^{-r_i t}u_i(1-\alpha_j)\Big((1-z_j)(1-G_i^Z(t)) + z_j\Big)$$

In this case, agent i makes exactly the same agreements as a behavioral type at $s \leq t^5$ (giving her a share α_i). For rational agent i not to want to make this deviation, we must have:

$$U_i^c(T^R) \geq \sup_t U_i^n(t) \quad \text{(Type IC)} \tag{7}$$

I call this the *type incentive constraint*.[27]

The dynamic and type incentive constraints are not only necessary for an equilibrium, they are also sufficient. This claim is formalized below in Theorem 1. It follows from the fact that any distribution of outcomes satisfying both constraints can be obtained in what I call a *direct mediation equilibrium*.

[26] This deviation isn't well defined in the sense that there is no $s^* = \min\{s > t^5\}$; however, this is not important. It is equivalent to the lack of a best deviation in a Betrand competition game.

[27] These incentive constraints are somewhat similar to veto incentive constraints in mechanism design (e.g. see Forges (1999)) in that agents can (mis)report their type, observe the proposed outcome, and then walk away. However, my agents' outside options are an endogenous outcome of a dynamic game in which agents can strike their own agreements, and so the mediator doesn't immediately inform agents of their proposed outcome.

In a *direct mediation protocol*, if both agents initially demand $\alpha_i(0^4) = \alpha_i$ and message the mediator at 0^2, she sends a single message back to them at some random time t^3 suggesting terms for an agreement (if neither has changed demand prior to t^3). If both agents initially demand $\alpha_i(0^4) = \alpha_i$, but only agent i messages the mediator at 0^2, the mediator sends a single message to the agents suggesting that i concede to j at some random time t^3 (if neither has changed demand prior to t^3). If neither agent messages the mediator at 0^2, or some agent demands $\alpha_i(t^4) \neq \alpha_i$, the mediator is silent for the rest of the game. A *direct mediation equilibrium* is then an equilibrium in which the mediator adopts a direct mediation protocol, rational agents always message the mediator at 0^2, immediately implement the mediator's suggestions, and demand α_i prior to such a suggestion.

Theorem 1. *Any distribution of outcomes $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ satisfies the dynamic and type incentive constraints (equations (6) and (7)) if and only if it can be obtained in some good faith equilibrium, and if and only if it can be obtained in a direct mediation equilibrium.*

Given the necessity of the dynamic and type incentive constraints in any good faith equilibrium and the fact that a direct mediation equilibrium is a good faith equilibrium, to establish the result it only remains to show that if a distribution of outcomes $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ satisfies both constraints, then it can be obtained in a direct mediation equilibrium. It is, however, straightforward to construct such an equilibrium where the mediator's direct protocol announcements are described by $(G^R, G_1^Z, G_2^Z, M_1, M_2)$.

The ability to restrict attention to direct mediation equilibria represents a strong version of the revelation principle. The dynamic and type incentive constraints are clearly necessary, regardless of whether agents or the mediator can make public or private messages, or the number of those messages. It is sufficient, however, to allow agents to make a single private message to the mediator at time zero, and for the mediator to send a single public message back. In general for multistage games, Myerson (1986) shows that in a communication equilibrium, a mediator must collect information and privately recommend actions to agents in each stage.

Direct mediation equilibria are fully described by their agreement distributions. Given this, I henceforth treat all mediation protocols with the same distribution of equilibrium agreements as an equivalence class, and use the term mediation protocol interchangeably with the distribution of equilibrium outcomes which arise from that protocol.

Because the mediator reveals no information prior to a suggested agreement in a direct mediation equilibrium, when there has been no agreement before time t, agent i believes that her opponent j is behavioral with probability $\bar{z}_j^c(t)$ if she confessed, and with probability $\bar{z}_j^n(t)$ if she did not, where:

$$\bar{z}_j^c(t) = \frac{z_j(1 - G_j^Z(t))}{z_j(1 - G_j^Z(t)) + (1 - z_j)(1 - G^R(t))}, \quad \bar{z}_j^n(t) = \frac{z_j}{z_j + (1 - z_j)(1 - G_i^Z(t))}.$$

So far, the only direct mediation equilibrium which we know to always exist is the unmediated Baseline equilibrium. The distribution of agreement times in that case is

$$1 - G_j^Z(t) = \frac{z_i}{1 - z_i}(e^{\lambda_i(T^R - t)} - 1) \qquad \text{and} \qquad 1 - G^R(t) = (1 - G_i^Z(t))(1 - G_j^Z(t)),$$

for $t \leq T^R = T_1^Z = T_2^Z = \min\{-\frac{1}{\lambda_1}ln(z_1), -\frac{1}{\lambda_2}ln(z_2)\}$. The distribution of dollar shares conditional on an agreement at $t \leq T^R$ is:

$$M_i^t(m) = \begin{cases} 0 & \text{if } m < 1 - \alpha_j \\ \frac{\frac{g_j^Z(t)}{1-G_j^Z(t)}}{\frac{g_j^Z(t)}{1-G_j^Z(t)} + \frac{g_i^Z(t)}{1-G_i^Z(t)}} & \text{if } m \in [1 - \alpha_j, \alpha_i) \text{ and } t \in (0, T^R] \\ \frac{G_j^Z(0)}{G_i^Z(0)+G_j^Z(0)} & \text{if } m \in [1 - \alpha_j, \alpha_i),\ G_i^Z(0) + G_j^Z(0) > 0 \text{ and } t = 0 \\ 1 & \text{if } m \geq \alpha_i \end{cases}$$

These distributions cause incentive constraints to bind, so that $U_i^c(t) = U_i^n(t)$ for all $t \in [0, T^R]$.

For risk averse agents, there is an obvious way for the mediator to improve these bargaining outcomes: to reduce dispersed agreement shares with an average agreement. That is, let rational agent i get $\hat{m}_i(t) = \int m dM_i^t(m)$ in any agreement with a rational opponent at each time t. Concavity ensures $u_i(\hat{m}_i(t)) \geq \int u_i(m) dM_i^t$, with a strict inequality at $t \in (0, T^R]$ if the agent is risk averse. It is easily verified that leaving the distribution of agreement *times* unchanged then ensures that both incentive constraints are satisfied.

Lemma 1. *If agents are risk averse,*[28] *then there is a good faith equilibrium which delivers strictly higher payoffs for both rational agents than the Baseline equilibrium.*

While this observation is technically trivial, it may be at least as important in explaining the benefits of mediation as reduced delay. Moreover, the slackness of the dynamic and type incentive constraints after eliminating dispersed outcomes, makes it clear that the mediator can *also* then reduce delay compared to the Baseline equilibrium (i.e. choose some $\hat{G}^R(t) > G^R(t)$ and $\hat{G}_i^Z(t) > G_i^Z(t)$ for $t < T^R$). In fact, Theorem 2, below, shows that risk aversion is necessary (and not just sufficient) for mediation to be beneficial when the probability of behavioral agents is larger than their demands and agents are symmetric. As discussed in the Introduction, the importance of curvature in the utility-possibility frontier for mediation extends beyond the case of risk aversion. In particular, this can explain why mediation is particularly effective for bargaining over multiple issues.

We want to understand more generally how a mediator should behave given the above incentive constraints. Proposition 3 showed that the mediator can improve even risk neutral agents'

[28] So that $u_1\left(0.5(\alpha_1 + 1 - \alpha_2)\right) > 0.5(u_1(\alpha_1) + u_1(1 - \alpha_2))$.

payoffs when the probability of behavioral types is sufficiently small. However, identifying an optimal mediation protocol for general asymmetric bargaining games is extremely challenging. It is clear that there are many moving parts and a great number of constraints (the type and dynamic incentive constraints really represent four sets of infinitely many non-independent constraints). Moreover, it is not at all obvious what objective function should be maximized to appropriately take account of any asymmetry between the agents.

The optimal mediation problem simplifies considerably, however, if the bargaining game is symmetric in the sense that $u_i = u$, $r_i = r$, $z_i = z$, and $\alpha_i = \alpha$ (which also implies $\lambda_i = \lambda$). For the rest of the paper I will assume this symmetry, and will not explicitly highlight it in remaining results. In this symmetric setting, I maximize the sum of rational agents' payoffs. The optimal protocol I identify, however, necessarily implies symmetric payoffs, and so also maximizes the Nash product of payoffs.[29]

I formally identify an optimal mediation protocol for symmetric bargaining games as a solution to the following problem:

$$\arg\max_{G^R, G_1^Z, G_2^Z, M_1, M_2} U_1^c(T^R) + U_2^c(T^R) \text{ subject to equations (6) and (7)}$$

We can immediately simplify this maximization by focussing on what I call strongly symmetric bargaining protocols. A mediation protocol is *symmetric* if $G_i^Z = G^Z$ and $M_i^t = M^t$ for all t and $i = 1, 2$, and is *strongly symmetric* if it is symmetric and additionally $M^t(0.5) = 1$ (i.e. the mediator always suggests a 50/50 division between rational agents). Given any equilibrium, it is easy to verify that there is an equilibrium of a strongly symmetric mediation protocol which obtains a weakly higher objective. This is formalized in Lemma 2.

Lemma 2. *If $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ describes a good faith equilibrium, then the strongly symmetric mediation protocol $(G^R, \hat{G}^Z)$ with $\hat{G}^Z = 0.5(G_1^Z + G_2^Z)$ describes a good faith equilibrium with a weakly higher objective, $U_1^c(T^R) + U_2^c(T^R)$.*

Optimal Strongly Symmetric Mediation Protocol (OSSMP)

Given the restriction to symmetric games and Lemma 2, we can restrict attention to searching for an *Optimal Strongly Symmetric Mediation Protocol* (OSSMP), described by (G^{R*}, G^{Z*}). I first define two special classes of agreement distributions, which will help describe an OSSMP.

Given any distribution of rational-behavioral agreements G^Z, define $G^R_{G^Z}$ as the distribution of rational-rational agreements, which keeps an agent who confessed indifferent between conceding on the interval $[0, T^Z]$, where $T^Z = T^R$.[30] The indifference condition $U^c(t) = U^c(T^Z)$

[29] While it might also seem reasonable to incorporate outcomes for behavioral types into the objective, it is not obvious how this should be done, making any such exercise highly speculative.

[30] This is well defined whenever G^Z can be combined with some G^R to satisfy both incentive constraints.

implies a linear ODE:

$$g^R_{G^Z}(t) = \lambda^m \left((1 - G^R_{G^Z}(t)) + \frac{z}{1-z}(1 - G^Z(t)) \right).$$

where

$$\lambda^m = \frac{ru(1-\alpha)}{u(0.5) - u(1-\alpha)}$$

Combining this with the boundary condition $G^R_{G^Z}(T^Z) = 1$ gives:

$$G^R_{G^Z}(t) = 1 - e^{-\lambda^m t} \int_t^{T^Z} \lambda^m e^{\lambda^m s} \frac{z}{1-z}(1 - G^Z(s))ds \tag{8}$$

Notice that a larger $G^Z(t)$ increases $G^R_{G^Z}(0)$, or more precisely, if $\tilde{G}^Z$ first order stochastically dominates G^Z, then $G^R_{\tilde{G}^Z}(0) > G^R_{G^Z}(0)$. The reason is that more rational-behavioral agreements before time t reduce a (confessing) agent's belief that she faces a behavioral opponent, $\bar{z}(t) = \frac{z(1-G^Z(t))}{(1-z)(1-G^R(t))+z(1-G^Z(t))}$, and so give her more incentive to wait (for $u(0.5)$ instead of $u(1-\alpha)$). This ultimately means that fewer rational-rational agreements are needed after time t to keep the agent indifferent to conceding, and so more agreements can occur at time zero.

Given the distribution of agreements $G^R_{G^Z}$, the expected payoff of a confessing agent is clearly:

$$U^c(T^Z) = U^c(0) = (1-z)G^R_{G^Z}(0)(u(0.5) - u(1-\alpha)) + u(1-\alpha). \tag{9}$$

We next turn to rational-behavioral agreements. Given any T^Z and $\check{t} \in [0, T^Z]$, define the distribution $G^Z_{\check{t},T^Z}$ as one which keeps an agent who didn't confess indifferent between conceding on the interval $[\check{t}, T^Z]$ and satisfies $G^Z_{\check{t},T^Z}(t) = 0$ for $t < \check{t}$.[31] The indifference condition $U^n(t) = U^n(T^Z)$ for $t \in [\check{t}, T^Z]$ implies

$$g^Z_{\check{t},T^Z}(t) = \lambda \left(1 - G^Z_{\check{t},T^Z}(t) + \frac{z}{1-z} \right)$$

Combining this with the boundary condition $G^Z_{\check{t},T^Z}(T^Z) = 1$ gives:

$$G^Z_{\check{t},T^Z}(t) = \begin{cases} 0 \text{ for } t < \check{t} \\ \frac{1-ze^{\lambda(T^Z-t)}}{1-z} \text{ for } t \in [\check{t}, T^Z] \end{cases} \tag{10}$$

Notice that for $t \geq \check{t}$ this distribution corresponds to the Baseline equilibrium distribution after accounting for the different times at which agreements are completed (i.e. T^Z). That is, $G^Z_{\check{t},T^Z}(t) = G^{Z_B}(t - ln(z)/\lambda - T^Z)$ for $t \geq \check{t}$, where G^{Z_B} is the distribution of rational-behavioral agreement times in the Baseline equilibrium which are completed by time $-ln(z)/\lambda$. When agents are risk neutral, we similarly have $G^R_{G^Z_{\check{t},T^Z}}(t) = G^{R_B}(t - ln(z)/\lambda - T^Z)$ for $t \geq \check{t}$, where G^{R_B}

[31] This is certainly well defined when $T^Z \leq -\frac{1}{\lambda} ln(z)$.

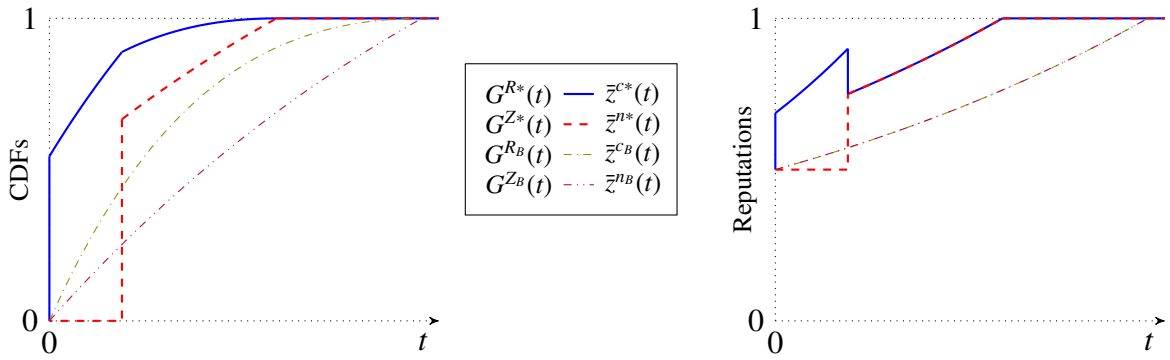

Figure 1. *Left:* Agreement time distributions in OSSMP (G^{R*}, G^{Z*}) and Baseline equilibrium (G^{R_B}, G^{Z_B}) when $\alpha = 0.75$, $z = 0.5$, $u(x) = x$. *Right:* Corresponding reputations $(\bar{z}^{c*}, \bar{z}^{n*})$ and $(\bar{z}^{c_B}, \bar{z}^{n_B})$. Payoffs: OSSMP=0.32, Baseline=0.25, Complete Information=0.38.

is the distribution of rational-rational agreement times in the Baseline equilibrium.

Having defined these distributions we are ready to state the paper's main result, Theorem 2. It establishes the existence of a unique OSSMP, whose agreement distributions have the above form, as well as precisely identifying when mediation is beneficial. These optimal distributions are illustrated and compared to the Baseline equilibrium distributions on the left hand side of Figure 1. The right hand side of the figure displays agents' implied beliefs (reputations).

Theorem 2. *A unique OSSMP exists. It satisfies* $G^{R*} = G^R_{G^*_Z}$ *and* $G^{Z*} = G^Z_{\check{t},T^Z}$ *for some* $\check{t} < T^Z$ *and implies* $U^c(T^Z) = U^n(T^Z)$. *For risk neutral agents it delivers higher payoffs than the Baseline equilibrium if and only if* $z < \alpha$, *where this also implies* $\bar{z}^c(\check{t}) = \bar{z}^n(\check{t}) = \alpha$ *and* $\check{t} > 0$. *For risk averse agents it always delivers higher payoffs than the Baseline equilibrium, and if behavioral types are sufficiently unlikely or make sufficiently large demands then* $\check{t} > 0$.[32]

The need for the optimal distribution of rational-rational agreements to have the specified form is fairly intuitive. It frontloads agreement as much as possible, while having the dynamic incentive constraint bind at all times (i.e. $U^c(t) = U^c(T^R)$ for all t). If the dynamic incentive constraint didn't bind at some time ($U^c(t) < U^c(T^R)$) then we could always move some rational-rational agreements forward in time, while maintaining incentives not to concede before T^R.[33]

The optimal distribution of rational-behavioral agreements is more interesting, particularly the presence of an initial interval with no agreement ($\check{t} > 0$) whenever behavioral types are unlikely or make large demands.[34] The reason that the mediator delays such initial agreements is that it keeps rational agents honest (i.e. it relaxes the type incentive constraint). Such agreements are worth only $u(1-\alpha)$ to an agent who confessed, but $u(\alpha)$ to an agent who claimed to be

[32] Lemma 10 states this final claim more precisely. Agents are risk averse if $u(0.5) > 0.5(u(\alpha) + u(1-\alpha))$.

[33] Such a change clearly also relaxes the type incentive constraint as $U^n(t)$ doesn't change.

[34] It is also interesting that G^{Z*} is always non-degenerate ($\check{t} < T^Z$), because this implies that the mediator is needed to help rational agents back down against behavioral opponents at the right time, and not just broker compromises. If $\check{t} = T^Z$ then a confessing agent could simply concede at T^Z without the need for direction.

behavioral. The discontinuity of G^{Z*} at $\check{t} > 0$[35] is also interesting because it implies that reputations are non-monotonic over time. For instance, in Figure 1's example, a rational risk neutral agent's belief about her opponent's commitment, $\bar{z}^{c*}(t)$, jumps from $z = 0.5$ to 0.69 at time zero, before continuously increasing to 0.91 and then jumping back down to $\alpha = 0.75$ at $\check{t}$. By contrast, reputations in the Baseline equilibrium increase monotonically with $\bar{z}^{c_B}(t) = \bar{z}^{n_B}(t)$.

The ability of the OSSMP to improve on the Baseline equilibrium for risk neutral agents if and only if the probability of behavioral types is smaller than their demand, $z < \alpha$, is perhaps the most interesting part of Theorem 2. The fact that there is more scope for mediation when agents are less likely to be committed to their demand (z small) is fairly intuitive, but the fact that this is also true when parties' bargaining positions are further apart (α large) is less obvious. Some intuition comes from the fact that the Baseline equilibrium is less efficient for more extreme demands (payoffs of $1 - \alpha$), and so it is easier for the mediator to improve outcomes; however, a more precise explanation must extend the discussion from the last paragraph.

The only way to benefit risk neutral agents is to reduce delay. To reduce delay while preserving incentives to confess rationality, rational-behavioral agreements must be delayed by more than rational-rational agreements (in particular, $G^R_{G^Z}(0) > G^Z(0)$ given $\alpha > 0.5$). Rational-behavioral agreements give a confessing agent a dollar share $1-\alpha$ with probability z, and a non-confessing agent α with probability $1 - z$. Delaying these agreements, therefore, hurts a non-confessing agent more than a confessing agent (improving the type incentive constraint) if and only if $z(1 - \alpha) < (1 - z)\alpha$, or equivalently $z < \alpha$. This also helps explain why reputations jump to $\alpha = \bar{z}^c(\check{t}) = \bar{z}^n(\check{t})$ at time $\check{t} > 0$ when $z < \alpha$. If $\bar{z}^c(\check{t}) = \bar{z}^n(\check{t}) < \alpha$, we could further relax the incentive constraint by delaying rational-behavioral agreements at $\check{t}$ until just after, while if $\bar{z}^c(\check{t}) = \bar{z}^n(\check{t}) > \alpha$ we could relax the constraint by bringing agreements forward to just before $\check{t}$.

Even though mediation is only beneficial for risk neutral agents when the probability of behavioral types is smaller than their demand, $z < \alpha$, mediation can have a significant effect even when the probability of behavioral types is quite large. For instance, in the example presented in Figure 1 we have $z = 0.5$, $\alpha = 0.75$ and risk neutrality. Mediated payoffs are $U^{OSSMP} = 0.32$, compared to $U^B = 0.25$ in the Baseline equilibrium, and a (symmetric) complete information upper bound on payoffs under complete information of $U^{CI} = 0.38$. In fact, the cutoff $z < \alpha$, clearly implies that optimal mediation *always* has a positive impact when less than 50% of agents are committed to their demand.

It is worthwhile to compare the result to those in subsection 3.3, concerning the *ND* mediation protocol. Those previous results extended to non-symmetric games, but made no claim of optimality. In symmetric games, *ND* mediation could improve on unmediated bargaining if and only if $z < \hat{z}$ for some $\hat{z} < 1$. It is simple to show that $\hat{z} < \alpha$ when agents are risk neutral.[36]

[35] If G^{Z*} was continuous at $\check{t}$, we couldn't have $U^c(T^Z) = U^n(T^Z) = U^n(\check{t}) > u(1 - \alpha)$.

[36] We showed that an *ND* equilibrium required $\frac{1-z\frac{u(\alpha)}{u(1-\alpha)}}{1-z}u(1 - \alpha) < u(0.5)$. When $u(x) = x$ and $z = \alpha$, this inequality can be transformed into the requirement that $\hat{Q}(\alpha) = \frac{\alpha}{1-\alpha}ln(\alpha) - ln(0.5) > 0$. Because $\hat{Q}$ is decreasing

For example with $\alpha = 0.75$ we had $\hat{z} = 0.62$, and when $z = 0.5$, payoffs were only $U^{ND} = 0.30$.

More insight into the characterization of Theorem 2 comes from examining the *time t type incentive constraint* of $IC_{G^Z}(t) = U^c(0) - U^n(t) \geq 0$ for an arbitrary G^Z, after substituting in for the distribution of rational-rational agreements, $G^R_{G^Z}$. Integrating by parts gives:

$$\begin{aligned} IC_{G^Z}(t) =& u(1-\alpha) + (u(0.5) - u(1-\alpha))(1-z)\left(1 - \int_0^{T^Z} \lambda^m e^{\lambda^m s} \frac{z}{1-z}(1 - G^Z(s))ds\right) \\ & - (1-z)\int_{s\leq t} e^{-rs}u(\alpha)dG^Z(s) - e^{-rt}u(1-\alpha)\left((1-z)(1-G^Z(t)) + z\right) \\ =& u(1-\alpha)(1-e^{-rt}) + (u(0.5) - u(1-\alpha))(1-z)\left(1 - \int_0^{T^Z} \lambda^m e^{\lambda^m s}\frac{z}{1-z}ds\right) \\ & - e^{-rt}G^Z(t)(1-z)\left(u(\alpha) - u(1-\alpha)\right) + \int_0^{T^Z} G^Z(s)r\left(u(1-\alpha)e^{\lambda^m s}z - \mathbb{1}_{[s\leq t]}u(\alpha)e^{-rs}(1-z)\right)ds \end{aligned} \tag{11}$$

In particular, consider the final integrand of $IC_{G^Z}(t)$, which is linear in $G^Z(s)$. This integrand captures the time t incentive costs and benefits of rational-behavioral agreements before time s (subject to not affecting $G^Z(t)$). The cost is direct in that earlier agreements increase a non-confessing agent's payoff, $-u(\alpha)e^{-rs}(1-z)$ when $s < t$. It is decreasing in s because later payoffs are discounted. The benefit is indirect, through relaxing the dynamic incentive constraint and so allowing earlier rational-rational agreements, $u(1-\alpha)e^{\lambda^m s}z$. This is increasing in s, because a change in $G^Z(s)$ has a larger effect on the likelihood ratio that a confessing agent faces a behavioral type, $\frac{\bar{z}(s)}{1-\bar{z}(s)} = \frac{z(1-G^Z(s))}{(1-z)(1-G^R(s))}$ when $(1 - G^R(s))$ is larger, ultimately implying a larger effect on the concession rate that makes a confessing agent indifferent, $\frac{g^R_{G^Z}(s)}{1-G^R_{G^Z}(s)} = \lambda^m\left(1 + \frac{\bar{z}(s)}{1-\bar{z}(s)}\right)$, which translates into a bigger effect on $G^R_{G^Z}(0)$.

Because the incentive benefits minus costs of earlier rational-behavioral agreements, $u(1-\alpha)e^{\lambda^m s}z - u(\alpha)e^{-rs}(1-z)$, are increasing in s, it must be better for both the objective function (the benefits) and incentives to have a larger $G^Z(s)$ later on in bargaining (given a fixed T^Z). This corresponds exactly to the structure of G^{Z*} which has $G^{Z*}(t) = 0$ for $t < \check{t}$ and a binding type incentive constraint, $IC_{G^{Z*}}(t) = 0$, for $t \in [\check{t}, T^Z]$; in other words $G^{Z*}(t)$ is as large as possible later on. When $u(1-\alpha)z - u(\alpha)(1-z) \geq 0$ (equivalently $z \geq \alpha$ for risk neutral agents), these incentive benefits minus costs of rational-behavioral agreements are necessarily positive for all s, which naturally results in $\check{t} = 0$ in the optimal distribution G^{Z*}.

Having identified the unique OSSMP, we can extend Lemma 2's claim (that strongly symmetric mediation protocols do at least weakly better than other protocols), to establish that any optimal protocol must be symmetric with the same distribution of agreement times as the OSSMP, and furthermore that the OSSMP is uniquely optimal when agents are risk averse. This result is established in Proposition 5, and is interesting because it suggests that a mediator *must* treat agents fairly, and not pick favorites if she wants to maximize total payoffs.[37]

in α and $\hat{Q}(0.5) = 0$, however, we must have $\hat{Q}(\alpha) < 0$ for $\alpha > 0.5$.

[37] This contrasts with the need for biased mediators in Kydd (2001), although the settings are quite different.

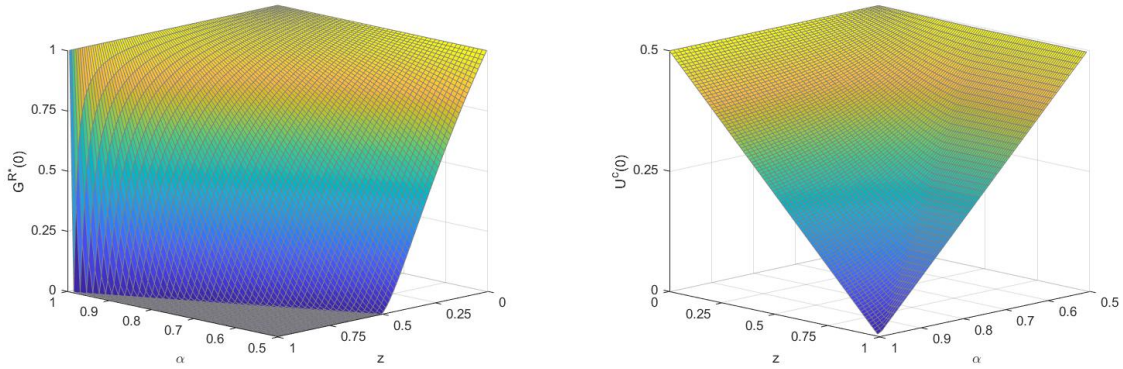

Figure 2. Numerical calculations of $G^{R*}(0)$ (*Left*) and $U^c(0)$ (*Right*) in OSSMP when $u(x) = x$.

Proposition 5. *Any optimal mediation protocol is symmetric with the same distribution of agreement times as in the unique OSSMP. If agents are risk averse,*[38] *then the unique optimal protocol is strongly symmetric.*

My final result on optimal mediation is Proposition 6, which provides limiting comparative statics on behavioral demands, behavioral probabilities, and utility functions. It shows that if the probability of behavioral types is arbitrarily small, or behavioral demands are arbitrarily large, or agents are arbitrarily risk averse, then mediation is approximately efficient in the sense that rational agents almost always reach agreement immediately, $G^{R*}(0) \approx 1$, to give payoffs that approximately match those possible under complete information $U^{CI} = (1 - z)u(0.5) + zu(1 - \alpha)$. These findings are in line with Theorem 2, in that mediation is again easier with risk averse agents, when behavioral demands are larger and behavioral probabilities are smaller. They are fairly intuitive given the previous analysis, and indeed can be established by using a lower bound on $G^{R*}(0)$ provided by the (non-optimal) *ND* mediation protocol.

Proposition 6. *Consider sequences of bargaining games,* $B^n = (\alpha, z^n, u, r)$ *with* $\lim_n z^n = 0$, $\check{B}^n = (\alpha^n, z, u, r)$ *with* $\lim_n \alpha^n = 1$, *and* $\hat{B}^n = (\alpha, z, u^n, r)$ *with* $\lim_n u^n(\alpha) = \lim_n u^n(0.5) > \lim_n u^n(1 - \alpha)$. *In the associated sequences of OSSMPs* $\lim_n G^{R*}(0) = 1$.

Establishing comparative statics more generally is difficult because the characterization in Theorem 2 does not provide a precise closed form solution for the OSSMP.[39] However, Figure 2 graphically illustrates numerical calculations of the probability of immediate rational-rational agreement, $G^{R*}(0)$, when agents are risk neutral. The probability is always decreasing in agents' probability of commitment, z, and increasing in their demands, α. Payoffs $U^c(0)$, which are an affine rescaling of $G^{R*}(0)$ (see equation (9)), are decreasing in z, but also decreasing in α (on net: larger demands are still bad).

[38] So that $u(0.5) > 0.5(u(x) + u(1 - x))$ for all $x \in [\alpha, 0.5)$.

[39] This is unsurprising given that the objective was maximized with respect to infinitely many constraints.

4.1 A mechanism design benchmark

In this subsection, I compare optimal mediation to a mechanism design benchmark where the mechanism designer can impose agreement and (perpetual) disagreement between agents. The exercise is somewhat speculative in that it is not entirely clear how the designer should accommodate behavioral types, but the contrast helps illustrate the mediator's problem more clearly.

The setup is as follows: Agents report their types to the designer, who chooses an agreement time and terms, or perpetual disagreement, based on the reported types. Behavioral agents always report their true type to the designer, but rational agents may lie. I require that the designer never imposes an agreement on a (reported) behavioral type which gives her less than her demand α. This can be thought of as a behavioral type's ex-post participation constraint (where her alternative to the designer's proposed agreement is perpetual disagreement). The designer must provide incentives for rational agents to reveal their type. I discuss imposing additional participation constraints below. I continue to restrict attention to symmetric bargaining games and have the designer maximize the sum of rational agents' payoffs.

Clearly, the designer should give a behavioral type exactly α in any rational-behavioral agreement, as giving her more can only decrease a rational agent's payoff and increase her incentive to imitate a behavioral type. For the same reasons as in Lemma 2 we can restrict attention to strongly symmetric mechanisms with $G_i^Z = G^Z$ and $M_i^t(0.5) = 1$ (transforming an arbitrary mechanism into a strongly symmetric one improves incentives and payoffs). I call an optimal mechanism of this kind $(G^{R\dagger}, G^{Z\dagger})$ an *Optimal Strongly Symmetric Delegation Mechanism* (OSSDM), because agents delegate their subsequent decision making power. It must solve:

$$\begin{aligned}\max_{G^R,G^Z} U^c =&(1-z)\int e^{-rs}u(0.5)dG^R(s) + z\int e^{-rs}u(1-\alpha)dG^Z(s)\\ \text{s.t. } U^c \geq U^n =&(1-z)\int e^{-rs}u(\alpha)dG^Z(s) \qquad \text{(Type IC}^\dagger\text{)}\end{aligned}$$

The delegation mechanism problem is much simpler than the optimal mediation problem. Increasing $G^R(0)$ strictly improves the objective function and the type incentive constraint, and so we must have $G^{R\dagger}(0) = 1$ in any solution. On the other hand, increasing $\int e^{-rs}dG^Z(s)$ strictly improves the objective function, but worsens the incentive constraint. If $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$, the incentive constraint can be satisfied even with $G^{Z\dagger}(0) = 1$, and otherwise the constraint must bind, so that $\int e^{-rs}dG^{Z\dagger}(s) = \frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$. There are many distributions which can make the constraint bind; one such distribution implies agents either agree immediately with probability $G^{Z\dagger}(0) = \frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$ or never agree, while other distributions imply eventual agreement.[40] This characterization of OSSDM is summarized in the following proposition.

[40]For instance $G^{Z\dagger}(t) = 0$ if $t < -\frac{1}{r}ln\left(\frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}\right)$ and $G^{Z\dagger}(t) = 1$ otherwise. Interesting, as $z^n \to 0$ or $\alpha^n \to 1$, this distribution converges to a point mass at $-\frac{1}{r}ln\left(\frac{u(0.5)}{u(\lim_n \alpha^n)}\right)$, which is also the limiting distribution of G^{Z*} in an OSSMP, while we also have $G^{R*}(0) \to 1 = G^{R\dagger}(0)$ in the limit.

Proposition 7. *An OSSDM exists. Any OSSDM must satisfy* $G^{R\dagger}(0) = 1$ *with* $G^{Z\dagger}(0) = 1$ *if* $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ *and* $\int e^{-rs} dG^{Z\dagger}(s) = \frac{(1-z)u(0.5)}{(1-z)u(\alpha)-zu(1-\alpha)}$ *otherwise.*

The characterization shows that a mediator is constrained much more by agents' freedom to ignore her suggestions, than by the informational problem alone. It is perhaps unsurprising that an OSSDM always achieves a strictly higher payoff than mediation.[41] However, not only do rational agent pairs reach immediate agreement, $G^{R\dagger}(0) = 1$, but because the type incentive constraint is less strict than under mediation (agents can't pretend to be behavioral and then subsequently concede) we can have an efficient outcome without any delay, $G^{R\dagger}(0) = G^{Z\dagger}(0) = 1$, whenever behavioral types are likely (z large) or make moderate demands (α small). For risk neutral agents, the cutoff for efficiency $z \geq \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ reduces to $z \geq 2\alpha - 1$. This holds, for instance, in the example from Figure 1 (with $\alpha = 0.75$, $z = 0.5$), so that OSSDM payoffs match those under complete information of 0.38 (compared to mediation payoffs of 0.32 and Baseline payoffs of 0.25). Of course, we also have an efficient outcome when $z \geq \alpha > 2\alpha - 1$, a situation where mediation was unable to improve on Baseline equilibrium payoffs at all!

More generally, comparative statics appear to be quite different to those established under mediation. Mediation payoffs (OSSMP) were $G^{R*}(0)(1-z)(u(0.5) - u(1-\alpha)) + u(1-\alpha)$, where $G^{R*}(0)$ was increasing in α and decreasing in z for risk neutral agents (see Figure 2). OSSDM payoffs are $(1-z)u(0.5) + G^{Z\dagger}(0)zu(1-\alpha)$, where $G^{Z\dagger}(0)$ is decreasing in α and increasing in z. Perhaps a more consistent way to compare payoffs is using $e(U) = \frac{U-U^B}{U^{CI}-U^B}$ as a measure of efficiency of payoff U, where $U^{CI} = (1-z)u(0.5) + zu(1-\alpha)$ is the complete information payoff and $U^B = u(1-\alpha)$ is the Baseline equilibrium payoff. Efficiency is then 100% in an OSSDM when behavioral types are likely or make moderate demands, but strictly smaller otherwise. For risk neutral agents, efficiency is 0% in an OSSMP when behavioral types are likely and make moderate demands, but strictly larger otherwise. In all cases, however, efficiency approaches 100% when behavioral types are very unlikely ($z \approx 0$) or make very large demands ($\alpha \approx 1$).

The OSSDM problem above lacks interim participation constraints, which might be thought to constrain its efficiency. Why should agents delegate their decision making power? This wasn't a problem for mediation because agents were free to (privately) talk to the mediator, or ignore her, and all agreements were voluntary. A natural participation constraint for delegation is that agents do better than in the Baseline equilibrium. This is certainly true for rational agents, as $U^c > u(1-\alpha)$. One plausible way to create a participation constraint for behavioral types is to assume that they have the same discount rate and same utility function as rational agents for dollar shares greater than α but obtain $-D$ for any smaller share, for D large. A weak improvement on the Baseline equilibrium then translates into a constraint, $U^n \geq u(1-\alpha)(1 - e^{rT^Z}z)$, where $T^Z = -\frac{ln(z)}{\lambda}$ (behavioral types expect a payoff $e^{-rT^Z}u(1-\alpha)z$ less than a rational agent who concedes at T^Z). Clearly this constraint is also satisfied because $U^n = U^c > u(1-\alpha)$

[41] The OSSMP distributions G^{R*} and G^{Z*} satisfy the OSSDM type incentive constraint strictly. Setting $G^R(0) = 1$ instead, strictly increases payoffs while preserving incentives.

if $z < \frac{u(\alpha)-u(0.5)}{u(1-\alpha)+u(\alpha)-u(0.5)}$ and $U^n = (1-z)u(\alpha)$ otherwise.[42] In fact, these participation constraints are typically also satisfied for mediation (OSSMP).[43]

While an OSSDM achieves a strictly higher objective than mediation, the credibility of this mechanism seems slightly dubious. It requires agents to fully delegate future decision making to the designer, who can then maintain perpetual disagreement between two (reported) behavioral types. However, courts don't typically enforce contracts in the absence of a harmed party (i.e. there is no party with standing to enforce a contract that constrains future agreements). Seemingly, therefore, a rational agent should always have the option to pretend to be behavioral and then change her mind and accept her (behavioral) opponent's demand.

An OSSDM isn't particularly close to the typical practice of arbitration (as a form of Alternative Dispute Resolution), because the designer sometimes imposes perpetual disagreement. An arbitrator who always immediately imposes some dollar division even between behavioral types, however, would seem likely to face fierce opposition. Using the assumptions about behavioral type "preferences" outlined above, this would necessarily give at least one behavioral type a payoff of less than $u(1) - D\frac{z}{2}$, which is worse than perpetual disagreement for large D. The designer would, therefore, seem unable to satisfy any reasonable interim participation constraint for such types. This illustrates an important point. The knowledge that a mediator will never impose an agreement that an agent strongly dislikes may be an important selling point of mediation compared to arbitration and can help explain its greater popularity in Stipanowich and Lamare (2013)'s survey. This problem of imposed agreements is in line with McEwen and Maiman (1981)'s finding that defendants in small claims courts are twice as likely to comply with mediated settlements as court imposed ones.

5 Conclusion

It is reassuring that economic theory can justify the effectiveness of mediation strategies similar to those used by professional mediators. The simple mediation protocol highlighted by Dunlop, of immediately announcing a deal when both parties agreed to its terms in private, is effective after adding noise (consistent with messages sometimes going astray) even if it is ineffective without noise. The reputational bargaining setting features two-sided incomplete information about agents' willingness to reach a deal. This may not be a perfect model for all situations where mediation is used; however, the tractability of this form of incomplete information allowed me to identify clear benefits from uninformed mediators, something that hasn't been possible in dynamic bargaining models with incomplete information about values.

[42] We can now imagine an extended game, where agents can agree to participate in an OSSDM at time zero, with unmediated bargaining occurring if some agent refuses to participate. Because all agents prefer the OSSDM to the Baseline equilibrium, beliefs can remain unchanged if some agent refuses to participate.

[43] In particular if $z \approx 0$, so a behavioral type's payoff in an OSSMP would be $U^n(T^Z) - ze^{-rT^Z}u(1-\alpha) \approx u(0.5)$.

The analysis highlighted two key ways in which mediation works. First, mediators can take advantage of curvature in the utility-possibility frontier, by bringing together flexible (rational) agents in agreements that are better than the average of the extreme demands proposed by inflexible (behavioral) types. This can explain why mediation may be especially important in multi-issue bargaining (which exhibits curvature). Second, by wisely choosing when to suggest agreements, a mediator can reduce delay more for flexible agents than for inflexible types, and thereby incentivize agents to reveal flexibility.

The analysis also revealed clear limits to what mediation can achieve. For symmetric games with no curvature in the utility-possibility frontier, mediation is beneficial if and only if agents are likely to be flexible, or inflexible types make large demands. Interestingly, this shows that mediation may be more effective when parties are initially further apart. The finding that a mechanism designer who can impose outcomes achieves full efficiency in settings where mediation is entirely ineffective, illustrates that getting agents to follow the mediator's suggestions constrains behavior much more than the informational problem alone.

One obvious direction for future work is to extend the characterization of optimal mediation to non-symmetric games. While a worthy goal, this also appears quite challenging. The ability to restrict attention to strongly symmetric protocols massively simplified the problem. It would obviously be even better to extend the analysis still further and allow agents to imitate many different (asymmetric) inflexible types, something which is possible in AG. However, clearly this makes the design of an optimal mediation protocol even more complex because mediation will affect flexible agents' demand choices in potentially perverse ways. In fact, in the Online Appendix I show that with many inflexible types, providing better mediation when agents make large demands can cause agents to make those demands more frequently, and so actually lower *both* flexible agents' payoffs compared to no mediation. These perverse effects of mediation are somewhat similar to Manzini and Ponsati (2006)'s finding that a third party with additional resources and an interest in agreement can actually increase bargaining delays. It is also similar to the detrimental effect of (unmediated) peace talks in Meirowitz et al. (2019).

References

Abreu, D. and F. Gul (2000). Bargaining and reputation. *Econometrica 68*(1), pp. 85–117. 3

Abreu, D. and D. Pearce (2007). Bargaining, reputation, and equilibrium selection in repeated games with contracts. *Econometrica 75*(3), 653–710. 3, 7

Atakan, A. E. and M. Ekmekci (2013, 08). Bargaining and reputation in search markets. *The Review of Economic Studies 81*(1), 1–29. 3

Ausubel, L. M., P. Cramton, and R. J. Deneckere (2002). Bargaining with incomplete information. In R. J. Aumann and S. Hart (Eds.), *Handbook of Game Theory with Economic Applications*, Volume 3, Chapter 50. Elsevier. 2

Ausubel, L. M. and R. J. Deneckere (1993, 04). Efficient sequential bargaining. *The Review of Economic Studies 60*(2), 435–461. 3

Basak, D. (2019). Fact-finding and bargaining. *Unpublished manuscript*. 2, 7

Beardsley, K. C., D. M. Quinn, B. Biswas, and J. Wilkenfeld (2006). Mediation style and crisis outcomes. *Journal of Conflict Resolution 50*(1), 58–86. 1

Brazil, W. D. (2007). Hosting mediations as a representative of the system of civil justice. *Ohio State Journal on Dispute Resolution 22*(2), 227–276. 4

Čopič, J. and C. Ponsatí (2008). Robust bilateral trade and mediated bargaining. *Journal of the European Economic Association 6*(2-3), 570–580. 6

Dixon, W. J. (1996). Third-party techniques for preventing conflict escalation and promoting peaceful settlement. *International Organization 50*(04), 653–681. 1

Dunlop, J. T. (1984). *Dispute resolution: Negotiation and consensus building*. Greenwood Publishing Group. 2

Ekmekci, M. (2011). Sustainable reputations with rating systems. *Journal of Economic Theory 146*(2), 479–503. 7

Ely, J. C. (2017, January). Beeps. *American Economic Review 107*(1), 31–53. 7

Emery, R. E., S. G. Matthews, and M. M. Wyer (1991). Child custody mediation and litigation: Further evidence on the differing views of mothers and fathers. *Journal of Consulting and Clinical Psychology 59*(3), 410. 1

Fanning, J. (2016). Reputational bargaining and deadlines. *Econometrica 84*(3), 1131–1179. 3

Fershtman, C. and D. Seidmann (1993). Deadline effects and inefficient delay in bargaining with endogenous commitment. *Journal of Economic Theory 60*, 306–321. 17

Fey, M. and K. W. Ramsay (2010). When is shuttle diplomacy worth the commute? Information sharing through mediation. *World Politics 62*(4), 529–560. 1, 2, 6

Forges, F. (1999). Ex post individually rational trading mechanisms. In *Current Trends in Economics*, pp. 157–175. Springer. 18

Fudenberg, D., D. Levine, and J. Tirole (1985). Infinite horizon models of bargaining with one-sided uncertainty. In *Game Theoretic Models of Bargaining*, Volume 73, pp. 79. Cambridge University Press. 2

Goldberg, S. B., F. E. Sander, N. H. Rogers, and S. Rudolph Cole (2012). *Dispute Resolution: Negotiation, Mediation, Arbitration, and Other Processes* (6 ed.). Wolters Kluwer. 4, 7

Goltsman, M., J. Hörner, G. Pavlov, and F. Squintani (2009). Mediation, arbitration and negotiation. *Journal of Economic Theory 144*(4), 1397–1420. 2, 6

Gul, F., H. Sonnenschein, and R. Wilson (1986, June). Foundations of dynamic monopoly and the Coase conjecture. *Journal of Economic Theory 39*(1), 155–190. 2

Hörner, J. and N. S. Lambert (2016). Motivational ratings. *Unpublished Manuscript*. 7

Hörner, J., M. Morelli, and F. Squintani (2015). Mediation and peace. *The Review of Economic Studies 82*(4), 1483–1501. 6

Jackson, M. O., H. Sonnenschein, and Y. Xing (2015). A theory of efficient negotiations. *Unpublished manuscript*. 5

Jarque, X., C. Ponsatı, and J. Sákovics (2003). Mediation: Incomplete information bargaining with filtered communication. *Journal of Mathematical Economics 39*(7), 803–830. 6, 12

Kambe, S. (1999, August). Bargaining with imperfect commitment. *Games and Economic Behavior 28*(2), pp. 217–237. 3

Kydd, A. (2001). Which side are you on? Mediation as cheap talk. *American Journal of Political Science 47*(3), 596–611. 6, 25

Manzini, P. and C. Ponsati (2006). Stakeholder bargaining games. *International Journal of Game Theory 34*(1), 67–77. 30

McEwen, C. A. and R. J. Maiman (1981). Small claims mediation in Maine. *Maine Law Review 33*, 237–268. 29

Meirowitz, A., M. Morelli, K. W. Ramsay, and F. Squintani (2019). Dispute resolution institutions and strategic militarization. *Journal of Political Economy 127*(1), 378–418. 6, 30

Myerson, R. (1991). *Game theory: Analysis of conflict*. Cambridge, Massachusetts: Harvard University Press. 3, 6, 15

Myerson, R. B. (1986). Multistage games with communication. *Econometrica*, 323–358. 7, 19

Myerson, R. B. and M. A. Satterthwaite (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory 29*(2), 265 – 281. 3, 5

Ponsati, C. (1997). Compromise vs. capitulation in bargaining with incomplete information. *Annales d'Economie et de Statistique*, 191–210. 6

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica 50*(1), pp. 97–109. 2, 58

Stipanowich, T. and J. R. Lamare (2013). Living with 'ADR': Evolving perceptions and use of mediation, arbitration and conflict management in Fortune 1,000 corporations. *Arbitration and Conflict Management in Fortune 1*. 1, 29

Velikonja, U. (2009). Making peace and making money: Economic analysis of the market for mediators in private practice. *Alb. L. Rev. 72*, 257. 4

Wilkenfeld, J., K. Young, V. Asal, and D. Quinn (2003). Mediating international crises: Cross-national and experimental perspectives. *Journal of Conflict Resolution 47*(3), 279–301. 1

Wolitzky, A. (2012, September). Reputational bargaining with minimal knowledge of rationality. *Econometrica 80*(5), 2047–2087. 3

Appendix A: The extensive form and strategies

At each *private history* h_i for agent i (an information set), she chooses an *action plan* $a_i(h_i)$. An action plan $a_i(h_i) = (\tau_i(h_i), x_i(h_i), i)$ for agent i consists of three parts: a *future* time to take action, $\tau_i(h_i)$; an action to take at that time, $x_i(h_i)$; and a marker for agent i. She can plan to never take a future action by setting $\tau_i(h_i) = \infty$. Let any actions planned for t^1 or t^5 be denoted $x_i(h_i) = C$ (C=Concession). Let the "action" planned for time ∞ be denoted $x_i(h_i) = \infty$. Actions at 0^1 or t^4 must specify a dollar division. Actions at t^2 and t^3 specify a message. The set of i's possible action plans is then a subset of $A_i = DC \times \{C\} \cup \Theta \cup [0,1] \times \{i\}$. A private history for agent i is then composed of a finite sequence of action plans which she has observed (of herself and others).

Which private histories are ultimately realized is determined inductively as follows. The initial *realized private history* is the null set, $h_i^1 = \emptyset$. Subsequent realized private histories, h_i^{k+1} (where $k \in \mathbb{N} \cup \infty$), are determined by the *joint realized history* $h^k = (h_1^k, h_2^k, h_3^k)$ and agents' action plans at realized private histories, $a_j(h_j^k)$. Let the time of the first action planned given h^k be $\tilde{\tau}(h^k) = \min\{\tau_1(h_i^k), \tau_2(h_i^k), \tau_3(h_i^k)\}$. Given h^k, let the set of agents whose actions i observes at $\tilde{\tau}(h^k)$ be $J_i(h^k)$. If $J_i(h^k) = \emptyset$ then let $h_i^{k+1} = h_i^k$. If $J_i(h^k) = \{1,2\}$ then $h_i^{k+1} = (h_i^k, (a_i(h_i^k), a_1(h_1^k), a_2(h_2^k)))$. If $J_i(h^k) = \{j\}$ then $h_i^{k+1} = (h_i^k, (a_i(h_i^k), a_j(h_j^k)))$. The game ends if ever $\tilde{\tau}(h^k) \in \{t^1, t^5\}$ (one agent concedes to a well defined existing demand) or else there is no agreement (agents get a zero payoff). If $h_i^{k+1} \neq h_i^k$ then define the *reference time* of realized private history h_i^{k+1} as $\check{\tau}(h_i^{k+1}) = \tilde{\tau}(h^k)$ and let $\check{\tau}(\emptyset) = 0^1$. This is the time at which the history h_i^{k+1} occurs.

An example will help clarify this structure. At time 0^1, bargainers make initial behavioral demands, so bargainer i's action plan at $h_i^1 = \emptyset$ is $a_i(h_i^1) = (\tau_i(h_i^1), x_i(h_i^1), i)$ where $\tau_i(\emptyset) = 0^1$, $x_i(\emptyset) = \alpha_i$. The mediator intends to send no message until she receives messages from both agents, and so $a_3(\emptyset) = (\infty, \infty, 3)$. The minimum time $\tilde{\tau}(h^1)$ in the joint realized history $h^1 = (\emptyset, \emptyset, \emptyset)$ is therefore 0^1. As all agents observe these demand announcements, the next realized private history for agent i is $h_i^2 = (h_i^1, (a_1(h_1^1), a_2(h_2^1), a_i(h_i^1)))$. Given h_1^2, agent 1 plans to message the mediator at 0^2, $a_1(h_1^2) = (0^2, \theta, 1)$, agent 2 plans to concede at time t^5, $a_2(h_1^2) = (t^5, C, 2)$, and the mediator plans to say nothing, $a_3(h_1^2) = (\infty, \infty, 3)$. This means $\tilde{\tau}(h^2) = 0^2$. Agent 1's action at 0^2 is observed by agent 1 and the mediator but not agent 2, so that $h_1^3 = (h_1^2, a_1(h_1^2))$, $h_3^3 = (h_3^2, (a_1(h_1^2), a_3(h_3^2)))$, and $h_2^3 = h_2^2$. Given h_1^3, agent 1 plans to change her demand to the entire dollar at $s^4 > t^5$ so that $a_1(h_1^3) = (s^4, 1, 1)$, the mediator plans to say nothing $a_3(h_3^3) = (\infty, \infty, 3)$. And so, $\tilde{\tau}(h^3) = t^5$, at which point the game ends with agent 2 conceding to agent 1's existing (initial, behavioral) demand α_1.

Let the set of possible joint histories be H and the set of agent i's possible private histories be H_i . A behavior strategy for agent i randomizes over her possible action plans at each of these private histories, $\sigma_i : H_i \to \Delta(A_i)$. A belief for bargainer i is $\mu_i : H_i \to \Delta(\{Z, R\} \times H)$. This describes both her belief about her opponent's type and her belief about the joint history at each of her possible private histories.

A perfect Bayesian equilibrium requires that at each of bargainer i's possible private histories, h_i, her behavior strategy maximizes her continuation payoff at reference time $\check{\tau}(h_i)$, given others' strategies and her beliefs. Beliefs are determined by Bayes rule where possible.

Appendix B:Proofs

Lemma 3. *If both agents are revealed as rational at time t^3, then there is a continuation equilibrium where agents agree at time t^5 on shares (m_1, m_2) for any $m_1 = 1 - m_2 \in [0, 1]$, and also an equilibrium with perpetual disagreement. If agent 1 is revealed as rational at time t^3 or t^4, but agent 2 has not been, then there is a continuation equilibrium where agent 1 immediately concedes to 2's behavioral demand at t^5.*

Proof. Continuation strategies are as follows: If both agents are revealed as rational at t^3 each flexible agent i plans to change her demands to $\alpha_i(t^4) = m_i$ at t^4 unless we want to implement perpetual disagreement in that case both agents demand $\alpha_i(t^4) = 1$. If agent 1 alone is revealed as rational at t^3 then she demands some $\alpha_1(t^4) \geq 1 - \alpha_2$ at t^4, while agent 2 plans never to concede or change her demand. If agents make compatible demands at t^4, then both agents plan to concede immediately at t^5. If agent 2 has still not been revealed as behavioral at t^4 (and demands are incompatible), then agent 1 plans to concede at t^5, while 2 plans never to concede or change her demand. If both agents demand $\alpha_i(t^4) = 1$ when we want to implement perpetual disagreement, then both plan never to concede. The mediator plans to say nothing in all circumstances. If agents follow strategies consistent with this, we get the desired outcomes. We fill in the rest of agents' continuation strategies below.

Let $s^4 \geq t^4$ be the first time some agent, call her i, observably deviates from this proposed equilibrium by either demanding: $\alpha_i(s^4) \neq m_i$, or $\alpha_i(s^4) \neq 1$ when we want to implement perpetual disagreement, or $\alpha_2(s^4) \neq \alpha_2$ when $i = 2$ hadn't been revealed as rational before t^4. This creates a new information set, where we designate agent i as the "loser", and j as the "winner" if only agent i deviated at s^4. If both agents deviate at s^4 then agent 1 is the loser. Neither the winner nor the loser ever plans to change her demands. So long as the winner doesn't change her demand the loser concedes to the winner immediately after any history at which she gets the chance, while the winner doesn't plan to ever concede except if demands are compatible. If the winner subsequently changes her demand, while the loser doesn't (at some $\tau^4 > s^k$), then roles are reversed, with the former winner (loser) becoming the new loser (winner). Clearly, given the winner's strategy it is optimal for the loser to concede immediately (delaying her concession cannot bring any benefit). Given the loser's strategy it is optimal for the winner to never concede unless demands are compatible. There is also clearly no benefit from deviating at t^4 (or any later time) as this will ensure that the agent becomes the loser (and so must then concede). □

Proof of Proposition 2

Suppose there is an equilibrium $\sigma = (\sigma_1, \sigma_2)$ with $p_i^c p_j^c > 0$. Let $A_i^c = \{t : U_i^c(t) = \max_s U_i^c(s)\}$ and $A_i^n = \{t : U_i^n(t) = \max_s U_i^n(s)\}$. Since σ is an equilibrium, $A_i^n \neq \emptyset \neq A_i^c$. Define $T_i^c = \inf\{t : F_i^c(t) = 1\}$, as the final time by which a confessing agent i concedes to her opponent. Similarly, define $T_i^n = \inf\{t : (1 - p_i^c)(1 - F_i^n(t)) = z_i\}$ as the final time a rational, non-confessing agent i concedes to her opponent. Finally, define $T^* = \max\{T_j^c, T_j^n, T_i^c, T_i^n\}$ and $T^c = \min\{T_i^c, T_j^c\}$. I next prove a series of claims, which help establish the result.

(a) *We must have $T_j^c \leq T_i^n < \infty$.* To establish $T_i^n \geq T_j^c$ suppose instead that $T_i^n < T_j^c$ then after time T_i^n a confessing agent j knows that she faces a behavioral opponent, and so would prefer to concede immediately rather than wait until T_j^c.
To establish $T_i^n < \infty$, let π_j^t be the conditional probability that agent j continues to act consistent with a behavioral type on the interval $[s, s + t)$ for arbitrary s. For agent i not to concede at s it must be that:

$$u_i(1 - \alpha_j) \leq (1 - \pi_j^t)u_i(1) + \pi_j^t e^{-r_i t} u_i(1)$$
$$\pi_j^t \leq \frac{u_i(1) - u_i(1 - \alpha_j)}{(1 - e^{-r_i t})u_i(1)}$$

where the second line simply rearranges the first. Fix $\delta \in \left(\frac{u_i(1)-u_i(1-\alpha_j)}{u_i(1)}, 1\right)$, and consider K such that $\delta^K < z_i$ and t' such that $\delta = \frac{u_i(1)-u_i(1-\alpha_j)}{(1-e^{-r_i t'})u_i(1)}$. Suppose agent i does not concede on the interval $[0, t'K)$ then it must be that the probability j acts consistent with a behavioral type on that interval is less than $(\pi_j^{t'})^K \leq \delta^K < z_i$, but this contradicts the fact that a behavioral type acts like itself. And so rational agent i will always concede by $T_i^n \leq t'K$

(b) *We must have* $\max\{T_j^c, T_j^n\} = T^* < \infty$. I first claim that $T_i^n \leq \max\{T_j^c, T_j^n\}$. Suppose not, so that $T_i^n > \max\{T_j^c, T_j^n\}$. Then after time $\max\{T_j^c, T_j^n\}$ a non-confessing rational agent i knows that she faces a behavioral opponent, and so would prefer to concede immediately rather than wait until T_i^n. By claim (a) we already know that $T_i^c \leq T_j^n \leq \max\{T_j^c, T_j^n\}$, hence $\max\{T_i^c, T_i^n\} \leq \max\{T_j^c, T_j^n\}$. Reversing the labelling we also have $\max\{T_i^c, T_i^n\} \geq \max\{T_j^c, T_j^n\}$, which establishes $\max\{T_j^c, T_j^n\} = T^*$. Claim (a) implies $\max\{T_i^c, T_j^c\} \leq \max\{T_j^n, T_i^n\} < \infty$, so that $T^* < \infty$.

(c) *There is no jump in* F_i^c *at* $t \in (0, T^*]$. *Furthermore, if* $F_j^n(0) > 0$ *then* $F_i^c(0) = 0$. Suppose F_i^c jumped at $t \in (0, T^*]$, then F_j^n must be constant on $[t-\varepsilon, t]$ for some $\varepsilon > 0$, as non-confessing agent j would prefer instead to concede an instant after t rather than on the interval $[t-\varepsilon, t]$. But in that case, a confessing agent i would prefer to concede at $t-\varepsilon$ rather than wait until t. Finally, if $F_j^n(0) > 0$ then a confessing agent i would strictly prefer to concede an instant after zero rather than at zero, so that $F_i^c(0) = 0$.

(d) *There is no jump in* F_i^n *at* $t \in (0, T^*]$. *Furthermore, if* $F_j(0) > 0$ *then* $F_i^n(0) = 0$. Suppose that F_i^n did jump at $t \in (0, T^*]$, then F_j is constant on $[t-\varepsilon, t]$ for some $\varepsilon > 0$, as a rational agent j would prefer instead to concede an instant after t rather than on the interval $[t-\varepsilon, t]$. But in that case, a non-confessing agent i would prefer to concede at $t-\varepsilon$ rather than wait until t. Finally, if $F_j(0) > 0$ then a non-confessing agent i would strictly prefer to concede an instant after zero rather than at zero, so that $F_i^n(0) = 0$.

(e) *If* F_i^n *is continuous at* t *then so is* U_j^c. *If* F_i *is continuous at* t *then so is* U_j^n. This follows from the definitions.

(f) *If* $T^* \geq t'' > t'$ *then* $F_i(t'') > F_i(t')$. Suppose not, then let $t_i^* = \sup\{t : F_i(t) = F_i(t')\} \in [t'', T^*]$. First, notice that no rational agent j can concede at $s \in (t', t_i^*)$ because this is strictly worse than conceding slightly earlier (e.g. at $\frac{s+t'}{2}$). Combining this with the continuity of F_j, U_i^c and U_i^n on $(0, T^*]$, established in claims (c), (d) and (e), implies that rational agent i (whether she confessed or not) would strictly prefer to concede at some early point in (t', t_i^*), such as $\frac{t_i^*+t'}{2}$, than wait to concede at or just after t_i^*. This, however, contradicts the definition $t_i^* \leq T^* < \infty$.

(g) *If* $T_j^c \geq t'' > t'$ *then* $F_i^n(t'') > F_i^n(t')$. Suppose not, then let $t_i^{*n} = \sup\{t : F_i^n(t) = F_i^n(t')\} \in [t'', T^*]$. First, notice that a confessing agent j will not concede at $s \in (t', t_i^{*n})$ because this is strictly worse than conceding slightly earlier (e.g. at $\frac{t'+s}{2}$). This ensures that $T_j^c \geq t_i^{*n}$. When combined with claim (f), we must have that F_j^n and F_i^c are strictly increasing on the interval (t', t_i^{*n}). Because F_i^c is strictly increasing on (t', t_i^{*n}), we must have that A_i^c is dense on that interval. By claims (d) and (e) U_i^c is continuous and hence constant on (t', t_i^{*n}). In turn that ensures that U_i^c is differentiable on (t', t_i^{*n}) with $\frac{dU_i^c(t)}{dt} = 0$, so that a non-confessing agent j must be conceding at rate $\frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$. Notice, however, that because $p_j^c(1-F_j^c(t)) > 0$ for $t < T_j^c$ where $T_j^c \geq t_i^{*n}$ for $t \in (t', t_i^{*n})$ we must have:

$$\frac{f_j(t)}{1-F_j(t)} = \frac{(1-p_j^c)f_j^n(t)}{(1-p_j^c)(1-F_j^n(t)) + p_j^c(1-F_j^c(t))} < \frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$$

A concession rate of exactly $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ would make a non-confessing agent i indifferent between conceding at any $t \in (t', t_i^{*n})$ and so a smaller concession rate, $\frac{f_j(t)}{1-F_j(t)} < \lambda_j$, means that she would strictly prefer to concede earlier rather than later on the interval (t', t_i^{*n}). The continuity of F_j and hence U_i^n on $(0, T^*]$, established in (c) and (d), then means that a non-confessing agent i must get a strictly lower payoff when conceding at or just after t_i^{*n} than if she conceded earlier (e.g. at $\frac{t'+t_i^{*n}}{2}$). This means that t_i^{*n} cannot be the supremum, a contradiction.

(h) *If $T_j^c > 0$, then agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ rate on $(0, T_j^c)$.* If $T_j^c > 0$, then claim (g) implies that F_i^n is strictly increasing on $[0, T_j^c]$, and so A_i^n is dense in $[0, T_j^c]$. From claims (c), (d), and (e) it follows that U_i^n is continuous. Hence, U_i^n is constant on this interval, and so differentiable with $\frac{dU_i^n(t)}{dt} = 0$, which implies that agent j concedes at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$.

(i) *If $T_j^c < T^*$, then $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$ on $(T_j^c, T^*]$.* First, suppose that $T_j^c \geq T_i^c$, then by claim (f) we must have that F_i^n is strictly increasing on $[T_i^c, T^*]$, and so A_i^n is dense in $[T_i^c, T^*]$. From claims (c), (d), and (e) it follows that U_i^n is continuous and hence is also constant on $(T_i^c, T^*]$. In turn that implies that U_i^n is differentiable on (T_i^c, T^*) with $\frac{dU_i^n(t)}{dt} = 0$, and so agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$. Next, suppose that $T_j^c < T_i^c$, so that at $t > T_j^c$, a confessing and non-confessing agent i have the same beliefs (in particular, both are certain that they face a non-confessing opponent j). This implies $A_i^n \cap (T_j^c, T^*] = A_i^c \cap (T_j^c, T^*]$. By claim (f) we know that F_i is strictly increasing on $(T_j^c, T^*]$, which implies that $A_i^n \cup A_i^c$ is dense in $(T_j^c, T^*]$, and so A_i^n is also dense on that interval. From claims (c), (d), and (e) it follows that U_i^n is continuous and hence constant on $(T_j^c, T^*]$. In turn that implies that U_i^n is differentiable on (T_j^c, T^*) with $\frac{dU_i^n(t)}{dt} = 0$, and so agent j must concede at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$. Finally, notice that for $t \geq T_j^c$ we have $1 - F_j(t) = (1 - p_j^c)(1 - F_j^n(t))$ and so $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)}$.

(j) *If $T^c \geq t'' > t'$, and $F_j^c(t'') = F_j^c(t')$ then $F_j^c(t'') = F_i^c(t'') = 0$.* I first claim that a confessing agent i will not concede with positive probability on the interval $[\max\{t' - \varepsilon, 0\}, t'']$ for some $\varepsilon > 0$. To see this, notice that given $F_j^c(t'') = F_j^c(t')$, to ensure that agent j on average concedes at rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on (t', t'') as required by claim (h), a non-confessing agent j must concede at rate:

$$\frac{f_j^n(t)}{1 - F_j^n(t)} = \lambda_j \left(1 + \frac{p_j^c(1 - F_j^c(t))}{(1 - p_j^c)(1 - F_j^n(t))}\right) \tag{12}$$

For $t < T^c$, however, $p_j^c(1 - F_j^c(t)) > 0$ and so this rate is strictly greater than λ_j, which implies that a confessing agent i would strictly prefer to concede at t'' rather than on the interval $[\max\{t' - \varepsilon, 0\}, t'')$ for some $\varepsilon > 0$. Next, define $t_i^{**} = \inf\{s : F_i^c(s) = F_i^c(t')\} \leq t'$. The previous argument with the roles of i and j reversed, implies $t_i^{**} = t_j^{**} = 0$ and $F_i^c(0) = F_j^c(0) = 0$. The continuity of F_i^c on $(0, T^*]$, established in claim (c), then implies $F_i^c(t'') = F_j^c(t'') = 0$.

(k) *Suppose $T^c > 0$, let $t^{*c} = \inf\{t : F_1^c(t) > 0$ or $F_2^c(t) > 0\}$, and suppose $t^{*c} \leq t' < t'' \leq T^c$, then $F_i^c(t'') > F_i^c(t')$.* Suppose not, so that $F_i^c(t'') = F_i^c(t')$. Claim (j) then implies that $F_1^c(t'') = F_2^c(t'') = 0$. Because $t^{*c} < t''$, however, we must have either $F_1^c(t'') > 0$ or $F_2^c(t'') > 0$ given $t^{*c} < t''$, a contradiction.

(l) *If $T^c > 0$ then $\frac{f_j^n(t)}{1-F_j^n(t)} = \frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$ on $(t^{*c}, T^c]$, where t^{*c} is defined in claim (k).* First notice that by (k) A_i^c must be dense in $[t^{*c}, T^c]$. From claims (d) and (e) U_i^c is continuous on $(0, T^c]$ and hence constant. In turn this implies that U_i^c is differentiable on (t^{*c}, T^c) with $\frac{dU_i^c(t)}{dt} = 0$, which implies that a non-confessing agent j concedes at rate $\frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$. By claim (h) we must also have a total concession rate $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ on (t^{*c}, T^c). If both these concession rates hold, then:

$$\lambda_j = \frac{f_j(t)}{1 - F_j(t)} = \frac{p_j^c f_j^c(t) + (1 - p_j^c) f_j^n(t)}{p_j^c(1 - F_j^c(t)) + (1 - p_j^c)(1 - F_j^n(t))} = \frac{p_j^c f_j^c(t) + (1 - p_j^c)(1 - F_j^n(t))\lambda_j}{p_j^c(1 - F_j^c(t)) + (1 - p_j^c)(1 - F_j^n(t))},$$

which rearranges to give $\frac{f_j^c(t)}{1-F_j^c(t)} = \lambda_j$.

(m) *We must have $T^c = 0$.* Suppose not, and so $T^c = T_j^c > 0$ for some agent j. This clearly implies $F_j^c(0) < 1$. Notice that the continuity of F_i^c on $(0, T^*]$ implies that either $t^{*c} = 0$ or $F_i^c(t^{*c}) = F_i^c(0) = 0$ (where t^{*c} is defined in claim (k)). Claim (l) then implies that if $t \in [t^{*c}, T^c]$, then $F_j^c(t) = 1 - (1 - F_j^c(0))e^{-\lambda_j t} < 1$, however, this contradicts $T_j^c < \infty$.

We are almost done. Suppose that $T_j^c = 0$ so that $F_j^c(0) = 1$ (and so $F_i^n(0) = 0$ by claim (d)). Claim (i) then

implies that $\frac{f_j(t)}{1-F_j(t)} = \frac{f_j^n(t)}{1-F_j^n(t)} = \lambda_j$ on $(0, T^*]$. This implies that $(0, T^*] \subseteq A_i^c = A_i^n$, and so a confessing agent i who concedes at $t \in A_i^c$ must get the payoff:

$$U_i^c(t) = p_j^c u_i(m_i) + (1 - p_j^c)\left(F_j^n(0)u_i(\alpha_i) + (1 - F_j^n(0))u_i(1 - \alpha_j)\right)$$

Whereas a non-confessing agent i who concedes at $t \in A_i^n$ must get the payoff:

$$U_i^n(t) = p_j^c u_i(\alpha_i) + (1 - p_j^c)\left(F_j^n(0)u_i(\alpha_i) + (1 - F_j^n(0))u_i(1 - \alpha_j)\right)$$

Therefore, if $p_j^c > 0$ we must have $m_i \geq \alpha_i$, or agent i would not find it optimal to confess. Clearly we cannot have $m_j = 1 - m_i < 1 - \alpha_i$, or confessing would deliver agent j a payoff of $p_i^c u_j(m_i) + (1 - p_i^c)u_j(1 - \alpha_j)$, which is strictly less than the payoff $u_j(1 - \alpha_i)$ which she could guarantee by not confessing and then conceding (recall that $F_j^c(0) = 1$ and $p_i^c > 0$).

Suppose finally that $m_j = 1 - \alpha_i$. In this case, we must have $U_j^c(t) \leq u_i(1 - \alpha)$ for all t (or we could not have $F_j^c(0) = 1$) and so we must similarly have $U_j^n(t) \leq u_i(1 - \alpha)$ for all t, which in particular implies $F_i(0) = 0$. Given $T_j^c = 0$, claim (b) implies that $T_j^n = T^*$.

Analogous to the requirement that both agents reach a probability one reputation at the same time in the Baseline model, we must have $T_j^n = T^* = \max\{T_i^c, T_i^n\}$. If $T_j^n < \max\{T_i^c, T_i^n\}$ then because $T_j^c = 0$, any rational agent i would know she faced a behavioral type at T_j^n and would concede at most an instant after. Similarly if $T_j^n > \max\{T_i^c, T_i^n\}$, then a non-confessing agent j would know she faced a behavioral type at $\max\{T_i^c, T_i^n\}$ and would concede at most an instant after.

Claims (h) and (i) then imply that agents must concede at rates $\frac{f_j(t)}{1-F_j(t)} = \lambda_j$ and $\frac{f_i(t)}{1-F_i(t)} = \lambda_i$ on $(0, T^*]$. Combined with the fact that $F_i(0) = 0$, this implies $1-F_i(t) = e^{-\lambda_i t}$ for $t \leq T^*$. The boundary conditions $(1-p_i^c)(1-F_i^n(T^*)) = z_i$ and $(1 - F_i^c(T^*)) = 0$, therefore imply $1 - F_i(T^*) = e^{-\lambda_i T^*} = z_i$ or $T^* = -\frac{1}{\lambda_i} ln(z_i)$. For agent j, these concession rates as well as $F_j^c(0) = 1$ imply that $(1 - p_j^c)(1 - F_j^n(t)) = (1 - F_j(t)) = (1 - F_j(0))e^{-\lambda_j t}$. The boundary condition $(1 - p_j^c)(1 - F_j^n(T^*)) = z_j$ then implies $(1 - F_j(0))e^{-\lambda_j T^*} = z_j$. Clearly if $T_j^* = -\frac{ln(z_j)}{\lambda^j} < -\frac{ln(z_i)}{\lambda^i} = T_i^* = T^*$, we have an immediate contradiction. Otherwise, $(1 - F_j(0)) = z_i e^{\lambda_j T^*} = z_i z_j^{-\frac{\lambda_j}{\lambda_i}}$. But in that case, any such equilibrium has exactly the same distribution of outcomes as the Baseline equilibrium. Such an equilibrium "involving" mediation exists whenever $T_i^* \neq T_j^*$ (e.g. let $p_i^c = 1 - z_i$ and $p_j^c \in \left(0, 1 - z_i z_j^{-\frac{\lambda_j}{\lambda_i}}\right]$ when $T_i^* < T_j^*$). □

Proof of Proposition 3

Agent i is indifferent to conceding at any $t \in (0, T^*]$. In particular, because it is optimal to concede an instant after time zero we must have:

$$U_i^{*c} = \max_t U_i^c(t) = (1 - z_j)\left(bu_i(m_i) + (1 - b)H_j^c(0)u_i(\alpha_i)\right) + \left(z_j + (1 - z_j)(1 - b)(1 - H_j^c(0))\right)u_i(1 - \alpha_j)$$

Setting this equal to the expression for $U_i^{*c} = U_i^c(T^*)$ in the main text (equation (4)) and rearranging gives:

$$\int_{s<T^*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) = u_i(\alpha_i) - (1 - H_j^c(0))(u_i(\alpha_i) - u_i(1 - \alpha_j)) + \frac{z_j(1 - e^{-r_i T^*})u_i(1 - \alpha_j)}{(1 - z_j)(1 - b)}. \tag{13}$$

And so, Q_i reduces to:

$$Q_i = u_i(m_i) - \left(u_i(\alpha_i) - (1 - H_j^c(0))(u_i(\alpha_i) - u_i(1 - \alpha_j)) + \frac{z_j(1 - e^{-r_i T^*})u_i(1 - \alpha_j)}{(1 - z_j)(1 - b)})\right)$$

Suppose that $T^* = T_j = -\frac{1}{\lambda_j} ln(\bar{z}_j) \leq T_i$, so that $1 - H_j^c(0) = 1$ and $1 - H_i^c(0) = \frac{\bar{z}_i}{1-\bar{z}_i}\left(\bar{z}_j^{-\frac{\lambda_i}{\lambda_j}} - 1\right)$. Substituting in for

these equalities gives:

$$Q_i = u_i(m_i) - u_i(1-\alpha_j) - \frac{z_j\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)u_i(1-\alpha_j)}{(1-z_j)(1-b)}$$

$$Q_j = u_j(m_j) - u_j(\alpha_j) - \frac{z_i\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}\right)u_j(1-\alpha_i)}{(1-z_i)(1-b)} + (u_j(\alpha_j) - u_j(1-\alpha_i))\frac{z_i\left(\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}-1\right)}{(1-z_i)(1-b)}$$

Define $\underline{m_i} < \alpha_i$ as the mediation share that causes $Q_i = 0$, that is:

$$\underline{m_i} = u_i^{-1}\left(u_i(1-\alpha_j) + \frac{z_j\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)u_i(1-\alpha_j)}{(1-z_j)(1-b)}\right)$$

Notice that $\underline{m_i} \to 1-\alpha_j$ as $z_j \to 0$. Setting $m_i = 1-m_j = \underline{m_i}$ we then have:

$$\frac{Q_j}{z_i} = \frac{u_j(1-\underline{m_i}) - u_j(\alpha_j)}{z_i} - \frac{\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}\right)u_j(1-\alpha_i)}{(1-z_j)(1-b)} + (u_i(\alpha_i) - u_i(1-\alpha_j))\frac{\left(\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}-1\right)}{(1-z_i)(1-b)} \tag{14}$$

We are interested in the limit of $\frac{Q_j}{z_i}$ as $z_j \to 0$. It is clear that $\lim_{z_j\to 0}\frac{1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_j}{\lambda_j}}}{(1-z_j)} = 1$ while $\frac{\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}-1}{(1-z_i)} = \infty$. By assumption we have $K \geq \frac{z_j}{z_i} \geq \frac{1}{K}$. We can then use l'Hopital's rule and the inverse function to show that:

$$\lim_{z_j\to 0}\frac{u_j(1-\underline{m_i}) - u_j(\alpha_j)}{z_j} = -\frac{u_j'(\alpha_j)u_i(1-\alpha_j)}{u_i'(1-\alpha_j)(1-b)} > -\infty$$

where this uses the fact that $\left.\frac{\partial\left(1-\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{r_i}{\lambda_j}}\right)z_j(1-z_j)^{-1}}{\partial z_j}\right|_{z_j=0} = 1$, and where $u_j'(\alpha_j)$ is a left derivative and $u_i'(1-\alpha_j)$ is a right derivative. And so we must have $\lim_{z_j\to 0}\frac{Q_j}{z_i} = \infty$. This ensures the existence of some $\underline{z}' > 0$ such that if $z_j \leq \underline{z}'$ we must have $Q_j \geq 0$ and $Q_i \geq 0$ and so an ND equilibrium exists.

It remains to show that such an equilibrium can strictly improve the payoff of both agents. If $\lambda_j = \lambda_i$ and $z_j \geq z_i$ then clearly we must have $T_i \geq T_j$ in any ND equilibrium and in the Baseline equilibrium. Alternatively, suppose that $\lambda_j > \lambda_i$ (and possibly $z_j < z_i$). In this case let $\overline{z}'' > 0$ be such that for $z_j \leq \overline{z}''$ we have $\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{\lambda_i}{\lambda_j}-1} \geq K$. This implies:

$$\left(\frac{z_j}{1-(1-z_j)b}\right)^{\frac{\lambda_i}{\lambda_j}} \geq \frac{Kz_j}{1-(1-z_j)b} \geq \frac{Kz_j}{1-(1-Kz_j)b} \geq \frac{z_i}{1-(1-z_i)b}$$

$$T_j = -\frac{1}{\lambda_j}ln\left(\frac{z_j}{1-(1-z_j)b}\right) \leq -\frac{1}{\lambda_i}ln\left(\frac{z_i}{1-(1-z_i)b}\right) = T_i$$

The first inequality on the first line is directly implied, the second follows because $K \geq 1$, the third because $Kz_j \geq z_i$. The second line is then simply a rearrangement of the inequality of the first and final term from the first line, and implies that $T_j \leq T_i$ in an ND equilibrium. The bound also ensures that $z_j^{\frac{\lambda_i}{\lambda_j}} \geq Kz_j \geq z_i$ so that in the Baseline equilibrium we must also have $T_j \leq T_i$.

Let $z_j \leq \min\{\overline{z}', \overline{z}'', \frac{1}{2}\}$, then the payoff to agent i in the Baseline equilibrium is $u_i(1-\alpha_j)$. Given that $U_i^{*n} > u_i(1-\alpha_j)$ in any ND equilibrium it is clear that we must also have $U_i^{*c} > u_i(1-\alpha_j)$. In the Baseline equilibrium agent j's

payoff is $U_j^B = u_j(\alpha_i) - (u_j(\alpha_j) - u_j(1-\alpha_i))z_i z_j^{-\frac{\lambda_i}{\lambda_j}}$. We need to compare this to her payoff in an ND equilibrium, which can be expressed as:

$$U_j^{*c} = (1-z_i)bu_j(m_j) + (1-b(1-z_i))u_i(\alpha_j) - (u_i(\alpha_i) - u_i(1-\alpha_j))z_i\left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}$$

The best possible ND equilibrium for agent j (consistent with a fixed b) has $m_j = 1 - \underline{m}_i$. For that equilibrium we have:

$$\begin{aligned}\frac{U_j^{*c} - U_j^B}{z_i} =&(1-z_i)b\frac{u_j(1-\underline{m}_i) - u_i(\alpha_j)}{z_i} + (u_i(\alpha_i) - u_i(1-\alpha_j))\left(z_j^{-\frac{\lambda_i}{\lambda_j}} - \left(\frac{z_j}{1-(1-z_j)b}\right)^{-\frac{\lambda_i}{\lambda_j}}\right)\\ \geq& b(K-z_j)\frac{u_j(1-\underline{m}_i) - u_i(\alpha_j)}{z_j} + (u_i(\alpha_i) - u_i(1-\alpha_j))z_j^{-\frac{\lambda_i}{\lambda_j}}\left(1 - \left(\frac{2}{2-b}\right)^{-\frac{\lambda_i}{\lambda_j}}\right)\end{aligned}$$

where the second line follows from the assumption $K \geq \frac{z_i}{z_j} \geq \frac{1}{K}$ and $\frac{1}{1-(1-z_j)b} \geq \frac{2}{2-b}$ when $z_j \leq \frac{1}{2}$ (this is equivalent to $2 - b \geq 2(1-(1-z_j)b)$).

We previously established that the $\lim_{z_j\to 0}\frac{u_j(1-\underline{m}_i)-u_i(\alpha_j)}{z_j} > -\infty$. Additionally noticing that $\left(\frac{2}{2-b}\right)^{-\frac{\lambda_i}{\lambda_j}} < 1$ and that $\lim_{z_j\to 0} z_j^{-\frac{\lambda_i}{\lambda_j}} = \infty$ it is clear that $\lim_{z_j\to 0}\frac{U_j^{*c}-U_j^B}{z_i} = \infty$. This implies that there exists $\underline{z} > 0$ such that for $z_j \leq \underline{z}$ we have an ND equilibrium with $m_i = \underline{m}_i$ where both rational agents' expected payoffs exceed their payoff in the Baseline equilibrium. □

Proof of Proposition 4

Suppose this were not true, then there must exist some sequence of games $(r_i, u_i, \alpha_i, z_i^n, m^n, b^n)$ with $z_1^n \to 1$ and a sequence of ND equilibria in each. I first claim that any (sub)sequence of these ND equilibria must satisfy $\lim_n T^* = 0$. This follows immediately from the fact that $T^* \leq T_1 = -\frac{1}{\lambda^1} ln\left(\frac{z_1^n}{(1-z_1^n)(1-b)+z_1^n}\right)$.

Notice that $\int_{s<T^*} e^{-r_i s}u_i(\alpha_i)dH_j^c(s) \geq e^{-r_i T^*}u_i(\alpha_i)$, hence for any $\varepsilon > 0$, for all sufficiently large n we need $m_i > \alpha_i - \varepsilon$ for $i = 1, 2$, in order to have $Q_i \geq 0$. Choosing $\varepsilon = \frac{\alpha_1+\alpha_2-1}{2}$ we have $m_1 + m_2 > \alpha_1 + \alpha_2 - 2\varepsilon = 1$, a contradiction.□

Proof of Theorem 1

Following the arguments in the text it remains only to show that when a distribution of outcomes $(G^R, G_1^Z, G_2^Z, M_1, M_2)$ satisfies the dynamic and type incentive constraints, there is a direct mediation equilibrium with that distribution of outcomes. We consider the direct mediation protocol where the mediator suggests an agreement before time t[3] with probability $G^R(t)$ if both agents message her at 0^2, and with probability $G_i^Z(t)$ if only j messages her. In the former case, the mediator suggests that i gets a share less than m with probability $M_i^t(m)$ in a time t agreement. If neither agent messages the mediator at 0^2, or agent i changes her demand from α_i without being suggested to do so by the mediator, then the mediator is silent for the rest of the game. On the equilibrium path, rational agent i always messages the mediator at 0^2, demands α_i prior to the mediator's suggestion, and immediately follows the mediator's suggestions.

We know from Lemma 3 that when both agents are revealed to be rational by the mediator's announcement, it is possible to implement any agreement share (including the one suggested by the mediator) in the continuation game. Furthermore, if agent i alone is revealed to be rational, by the mediator suggesting that i concedes to j or by i changing her demand before any mediator suggestion, then it is a continuation equilibrium outcome for i to immediately concede to j's behavioral demand. After initially confessing to the mediator, we know each agent cannot profitably deviate by conceding before the mediator makes a suggestion, because the dynamic incentive constraint is satisfied. Furthermore, the type incentive constraint ensures that the agent optimally confesses her rationality at 0^2, regardless of any continuation concession strategy. It only remains to show that a rational agent

has a well defined optimal continuation concession strategy after she deviates by failing to confess rationality at 0^2. This is described by a stopping time. Consider a sequence t_m such that $U_i^n(t_m) \to \sup_t U_i^n(t)$. The extended real line is compact so we can consider a subsequence, converging to $\underline{t}$. The right continuity of G_i^Z then ensures that $U_i^n(\underline{t}) = \sup_t U_i^n(t)$. Moreover, planning to concede at $\underline{t}^5$ (if there is no mediator announcement by t^3) ensures exactly this payoff (because there is no concession by j at t^5 unless the mediator suggests it at t^3). □

Proof of Lemma 1

Let $U_i^c(t)$ and $U_i^n(t)$ be the utilities under the Baseline equilibrium, and let $\hat{U}_i^c(t)$ and $\hat{U}_i^n(t)$ be utilities when the mediator suggests a time t agreement between rational agents, which gives agent 1, $\hat{m}_1(t) = 1 - \hat{m}_2(t) = \int m dM_1^t(m) - \varepsilon$ for $t \in [0, \varepsilon]$ and $\hat{m}_1(t) = 1 - \hat{m}_2(t) = \int m dM_1^t(m)$ for $t > \varepsilon$. For $\varepsilon = 0$ we have $\hat{U}_i^c(T^R) \geq U_i^c(t) = U_i^n(t) = \hat{U}_i^n(t)$ with a strict inequality for agent 1 for $t < T^R$. By continuity, for all small enough $\varepsilon > 0$ this inequality must be strict inequality for both agents, so that the type incentive constraint is satisfied. Moreover, for $t \leq T^R$ we have

$$\hat{U}_i^c(T^R) - \hat{U}_i^c(t) = \hat{U}_i^c(T^R) - \hat{U}_i^c(t) - U_i^c(T^R) + U_i^c(t) = (1 - z_j) \int_{t<s\leq T^R} e^{-r_i s} \left(u(\hat{m}_i(t)) - \int u_i(m) dM_i^s(m) \right) dG^R(s),$$

where the first inequality follows because $U_i^c(T^R) = U_i^c(t)$ for $t \leq T^R$. Clearly, $\hat{U}_i^c(T^R) - \hat{U}_i^c(t) \geq 0$ for $\varepsilon = 0$ and strictly so for agent 1, for all $t < T^R$. For all small enough $\varepsilon > 0$, therefore, we have a strict inequality for both agents for $t \in [0, \varepsilon)$ and at least a weak inequality for $t \in [\varepsilon, T^R]$. □

Proof of Lemma 2

First consider the associated symmetric protocol $(G^R, \check{G}^Z, \check{M})$ where $\check{G}^Z = \hat{G}^Z = 0.5(G_1^Z + G_2^Z)$ and $\check{M} = 0.5(M_1 + M_2)$. Let the utilities in the original equilibrium be $U_i^c(t)$ and $U_i^n(t)$ and the utilities in the symmetric protocol be $\check{U}^c(t)$, $\check{U}^n(t)$. Symmetry then implies:

$$\begin{aligned}
\check{U}^c(t) =& (1-z) \int_{s\leq t} e^{-rs} \int u(m) 0.5(dM_1^s(m) + dM_2^s(m)) dG^R(s) + z \int_{s\leq t} e^{-rs} u(1-\alpha) 0.5(dG_1^Z(s) + dG_2^Z(s)) \\
&+ e^{-rt} u(1-\alpha) \left((1-z)(1 - G^R(t)) + z(1 - 0.5(G_1^Z(t) + G_2^Z(t))) \right) = 0.5(U_1^c(t) + U_2^c(t)), \\
\check{U}^n(t) =& (1-z) \int_{s\leq t} e^{-rs} u(\alpha) 0.5(dG_1^Z(s) + dG_1^Z(s)) + e^{-rt} u(1-\alpha) \left((1-z)(1 - 0.5(G_1^Z(t) + G_2^Z(t))) + z \right) \\
=& 0.5(U_1^n(t) + U_2^n(t)).
\end{aligned}$$

This immediately ensures that the symmetric protocol obtains the same objective. Moreover, because the original equilibrium satisfied $U_i^c(T^R) = \max_t U_i^c(t) \geq \sup_t U_i^n(t)$, it is clear that $\check{U}^c(T^R) = \max_t \check{U}^c(t) \geq \sup_t \check{U}^n(t)$.

The proof of Lemma 1 then shows that moving from this symmetric protocol to a strongly symmetric one preserves incentive compatibility and weakly increases the objective, because $u(0.5) \geq \int u(m) d\check{M}^t(m)$. □

Proof of Theorem 2

We establish Theorem 2 by proving each of its claims in a series of lemmas. I present all of these lemmas, before turning to their proofs. The first lemma establishes that given any distribution G^Z, the distribution $G^R_{G^Z}$ maximizes the objective, $U^c(T^R)$.

Lemma 4. *Given any distribution G^Z, if the set of distributions G^R which satisfy both incentive constraints is non-empty, then $G^R_{G^Z}$ uniquely maximizes $U^c(T^R)$ in this set.*

The proofs of Lemma 4 and other lemmas making substantive claims from Theorem 2 use the following intermediate result. It shows that when we can adjust distributions (G^R or G^Z) in a monotonic way to improve the objective function, $U^c(T^R)$, there exists some distribution which cannot be improved in that way. For the statement of this

result, and elsewhere, it is helpful to define the set $\Delta_T \subset [0,1]^{[-\infty,\infty]}$ of cumulative distribution functions on the extended real line such that $G \in \Delta_T$ if $G(T) = 1$ and $G(t) = 0$ for $t < 0$.

Lemma 5. *For fixed $T \leq \infty$. Consider functions of the form:*

$$v(T,G) = \int_{s\leq T} e^{-rs} A_1(s) dG(s) + A_2(T)$$
$$w(t,G) = \int_{s\leq t} A_3(s) dG(s) + \int_{s\leq T} e^{-rs} A_4(s) dG(s) + A_5(t) + e^{-rt} A_6(t) G(t)$$

where each A_k is a continuous bounded function on $[0,T]$ with A_5 is continuously differentiable and $G \in \Delta_T$. Let $X = \{G \in \Delta_T : w(t,G) \geq 0, \forall t \in [0,T]\}$ and $X' = \{G \in X : w(T,G) = 0\}$.

(a) Define the partial order $\gtrsim$ on X by $G \sim G$, and $\tilde{G} \succ G$ if $v(T,\tilde{G}) > v(T,G)$ and $\tilde{G}(t) \geq G(t)$ for all t. If there exists some $\tilde{G} \succ G$, then there exists some $\overline{G} \succ G$ for which there is no $\hat{G}$ such that $\hat{G} \succ \overline{G}$.

(b) For any $G \in X'$ define $\bar{t}_G = \inf\{t : w(s,G) = 0 \text{ for } s \in [t,T]\}$. For any $\tilde{t} < \infty$, define the partial order $\gtrsim_{\tilde{t}}$ on X' by $G \sim_{\tilde{t}} G$, and $\tilde{G} \succ_{\tilde{t}} G$, if $v(T,\tilde{G}) > v(T,G)$, $\bar{t}_{\tilde{G}} \leq \bar{t}_G$, $\tilde{G}(t) \geq G(t)$ for $t \geq \tilde{t}$, and $\tilde{G}(s) \leq G(s)$ for $s < \tilde{t}$. If there exists some $\tilde{G} \succ_{\tilde{t}} G$, then there exists some $\overline{G} \succ_{\tilde{t}} G$ for which there is no $\hat{G}$ such that $\hat{G} \succ_{\tilde{t}} \overline{G}$.

The next lemma shows that type incentive constraint must bind at T^Z in an OSSMP and also at any $t \geq \hat{t}$ where:

$$\hat{t} = \frac{1}{\lambda^m + r} ln\left(\frac{(1-z)u(\alpha)}{zu(1-\alpha)}\right)$$

Lemma 6. *Consider any distribution G^Z such that $\inf_t IC_{G^Z}(t) \geq 0$. If $IC_{G^Z}(t) > 0$ for some $t \in \left[\min\{\hat{t}, T^Z\}, T^Z\right]$ then there is an alternative distribution $\tilde{G}^Z$ such that $\min_s IC_{\tilde{G}^Z}(s) = 0 = IC_{\tilde{G}^Z}(t)$ for $t \in \left[\min\{\hat{t}, T^Z\}, T^Z\right]$, and $G^R_{\tilde{G}^Z}(0) > G^R_{G^Z}(0)$.*

Given this lemma, when $\hat{t} \leq 0$ the optimal distribution of rational-behavioral agreements must have the form G^Z_{0,T^Z}, and so the OSSMP problem can be reduced to finding the minimum T^Z such that $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$. It is then fairly simple to establish that mediation cannot be beneficial for risk neutral agents in this case (where $\hat{t} \leq 0$ if and only if $z \geq \alpha$).

Lemma 7. *If agents are risk neutral and $z \geq \alpha$, then the distribution of agreement times and payoffs in the unique OSSMP are identical to those in the Baseline equilibrium.*

The next lemma establishes the converse, that mediation is beneficial for risk neutral agents when $z < \alpha$.

Lemma 8. *If $z < \alpha$, then an OSSMP delivers higher payoffs than the Baseline equilibrium.*

The negative final integrand of equation (11) for $s < \min\{t, \hat{t}\}$, implies that $G^Z_{\min\{T^Z,\hat{t}\},T^Z}$ maximizes $IC_{G^Z}(T^Z)$ among all distributions G^Z with the same T^Z. This means that T^Z is consistent with the incentive constraints if and only if $IC_{G^Z_{\min\{T^Z,\hat{t}\},T^Z}}(T^Z) \geq 0$. In this case, we can define:

$$\check{t}(T^Z) = \min\{\check{t} \geq 0 : IC_{G^Z_{\check{t},T^Z}}(T^Z) \geq 0\}$$

The next lemma establishes that an optimal distribution G^{Z*} must be of the form $G^Z_{\check{t}(T^Z),T^Z}$.

Lemma 9. *For any distribution $G^Z \neq G^Z_{\check{t}(T^Z),T^Z}$ with $\min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$, we have $G^R_{G^Z_{\check{t}(T^Z),T^Z}}(0) > G^R_{G^Z}(0)$.*

The ability to restrict attention to distributions of the form $G^Z_{\check{t},T^Z}$ allows me to rapidly establish four more facts about an OSSMP, which complete the characterization of Theorem 2. First, an equilibrium exists. Second, it is unique. Third, it features $\check{t} < T^Z$, so that the optimal distribution G^{Z*} is non-degenerate. Fourth, it features $\check{t} > 0$ when the probability of behavioral types is small or behavioral demands are large (in particular, $z < \alpha$ for risk neutral agents), so that there is a non-degenerate interval with no rational-behavioral agreements. Finally, for risk-neutral agents we have $\bar{z}^c(\check{t}) = \bar{z}^n(\check{t}) = \alpha$.

Lemma 10. *A unique OSSMP exists. The optimal distribution $G^{Z*} = G^Z_{\check{t},T^Z}$ satisfies $\check{t} < T^Z$. Moreover, there exists $\underline{z}(\alpha,u) > 0$ and $\underline{\alpha}(z,u) < 1$ such that if $z < \underline{z}(\alpha,u)$ or $\alpha > \underline{\alpha}(z,u)$ then $\check{t} > 0$. If agents are risk neutral and $z < \alpha$ then $\check{t} > 0$ and $\bar{z}^c(\check{t}) = \bar{z}^n(\check{t}) = \alpha$.*

I now turn to the proof of these lemmas.

Proof of Lemma 4. Throughout this proof I hold G^Z fixed and consider G^R such that both incentive constraints are satisfied. Suppose that such a distribution G^R, implies $T^R > T^Z$. In that case, the alternative distribution $\tilde{G}^R$ with $\tilde{G}^R(t) = G^R(t)$ for $t < T^Z$ and $\tilde{G}^R(T^Z) = 1$ so that $T^R = T^Z$ strictly increases $U^c(T^R)$, while relaxing both incentive constraints. It is, therefore, without loss of generality to focus on G^R that imply $T^R = T^Z$.

Suppose that G^R implies that the dynamic incentive does not bind, in the sense that $U^c(T^Z) - U^c(\bar{t}_1) = \delta > 0$ for some real valued time $\bar{t}_1 < T^Z$. We next want to show that in this case there must exist some alternative distribution $\check{G}^R$ with $\check{G}^R(t) \geq G^R(t)$ satisfying both constraints, which increases $U^c(T^Z)$. To that end define $\bar{t}_2 = T^Z$ if $T^Z < \infty$ and $\bar{t}_2 = \min\{t : e^{-rt}u(1) \leq \frac{\delta}{2}\}$ otherwise. Notice that we must have $U^c(T^Z) - U^c(\bar{t}_2) \leq e^{-r\bar{t}_2}u(1) \leq \frac{\delta}{2}$, so that $U^c(\bar{t}_2) - U^c(\bar{t}_1) \geq \frac{\delta}{2} > 0$ and $\bar{t}_2 > \bar{t}_1$.

Next define $\bar{t}_3 = \min\{t \in [\bar{t}_1, \bar{t}_2] : U^c(\bar{t}_2) - U^c(t) \leq (\bar{t}_2 - t)\frac{\delta}{4(\bar{t}_2 - \bar{t}_1)}\}$. This is well defined because the right continuity of G^z and G^R ensures that $U^c(t)$ is right continuous also. By construction $\bar{t}_3 > \bar{t}_1$, and $U^c(\bar{t}_3) - U^c(t) = [U^c(\bar{t}_2) - U^c(t)] - [U^c(\bar{t}_2) - U^c(\bar{t}_3)] > \frac{\delta(\bar{t}_3 - t)}{4(\bar{t}_2 - \bar{t}_1)}$ for all $t \in [\bar{t}_1, \bar{t}_3)$. For such t we have:

$$\begin{aligned}
U^c(\bar{t}_3) - U^c(t) =& (1-z)\int_{s\in(t,\bar{t}_3]} e^{-rs}u(0.5)dG^R(s) + e^{-r\bar{t}_3}u(1-\alpha)\Big((1-z)(1-G^R(\bar{t}_3)) + z(1-G^Z(\bar{t}_3))\Big) \\
&+ z\int_{s\in(t,\bar{t}_3]} e^{-rs}u(1-\alpha)dG^Z(s) - e^{-rt}u(1-\alpha)\Big((1-z)(1-G^R(t)) + z(1-G^Z(t))\Big) \\
\leq& (1-z)e^{-rt}(G^R(\bar{t}_3) - G^R(t))\left(u(0.5) - u(1-\alpha)\right) \\
&+ (e^{-r\bar{t}_3} - e^{-rt})\Big((1-z)(1-G^R(\bar{t}_3)) + z(1-G^Z(\bar{t}_3))\Big)u(1-\alpha),
\end{aligned}$$

where the inequality follows from the fact that the integrals in the first two lines are respectively smaller than $(1-z)e^{-rt}u(0.5)(G^R(\bar{t}_3) - G^R(t))$ and $ze^{-rt}u(1-\alpha)(G^Z(\bar{t}_3) - G^Z(t))$, and some rearrangement. Let $\varepsilon = \frac{\delta}{4(\bar{t}_2 - \bar{t}_1)}$ so that $U^c(\bar{t}_3) - U^c(t) > \varepsilon(\bar{t}_3 - t)$ for $t \in [\bar{t}_1, \bar{t}_3)$. By dividing the right hand side of the above inequality by $(\bar{t}_3 - t)$ and taking its limit infimum as $t \to \bar{t}_3$ gives:

$$e^{-r\bar{t}_3}\left(u(0.5) - u(1-\alpha)\right)(1-z)lim_{t\to\bar{t}_3}\inf_{s\in[t,t_3)}\frac{G^R(\bar{t}_3) - G^R(s)}{\bar{t}_3 - s} - ru(1-\alpha)e^{-r\bar{t}_3}\Big((1-z)(1-G^R(\bar{t}_3)) + z(1-G^Z(\bar{t}_3))\Big) \geq \varepsilon$$

This in turn implies that there exists $\varepsilon' > 0$ and $\bar{t}_4 < \bar{t}_3$ such that for all $t \in [\bar{t}_4, \bar{t}_3]$,

$$G^R(\bar{t}_3) - G^R(t) \geq \left((1 - G^R(\bar{t}_3)) + \frac{z}{1-z}(1 - G^Z(\bar{t}_3))\right)\lambda^m(\bar{t}_3 - t) + \varepsilon'(\bar{t}_3 - t).$$

Consider then an alternative distribution, $\hat{G}^R$. This is defined by $\hat{G}^R(t) = G^R(t)$ for $t \geq \bar{t}_3$, and satisfies the indifference condition $U^c_{\hat{G}^R}(t) = U^c_{\hat{G}^R}(\bar{t}_3)$ for $t \leq \bar{t}_3$ (where $U^c_{\hat{G}^R}(t)$ is the utility of conceding at t given $\tilde{G}^R$). This indifference condition implies that $\hat{G}^R(t)$ is differentiable on $(0, \bar{t}_3)$ with $\hat{g}^R(t) = \left((1 - \hat{G}^R(t)) + \frac{z}{1-z}(1 - G^Z(t))\right)\lambda^m$.

It is clear that this implies the existence of some $\bar{t}_5 < \bar{t}_3$ such that for all $t \in [\bar{t}_5, \bar{t}_3]$,

$$\hat{G}^R(\bar{t}_3) - \hat{G}^R(t) \leq \left((1 - \hat{G}^R(\bar{t}_3)) + \frac{z}{1-z}(1 - G^Z(\bar{t}_3))\right)\lambda^m(\bar{t}_3 - t) + \frac{\varepsilon'}{2}(\bar{t}_3 - t).$$

Letting $\bar{t}_6 = \max\{\bar{t}_4, \bar{t}_5\}$ we must then have $\hat{G}^R(t) > G^R(t)$ for all $t \in [\bar{t}_6, \bar{t}_3)$. We can now define $\check{G}^R(t) = \hat{G}^R(t)$ for $t \geq \bar{t}_6$ and $\check{G}^R(t) = G^R(t)$ elsewhere. This distribution implies $\check{G}^R(t) \geq G^R(t)$ for all t, and $\check{G}^R(t) > G^R(t)$ for $t \in [\bar{t}_6, \bar{t}_3)$. This ensures that $U^c_{\check{G}^R}(t) > U^c_{G^R}(t)$ for all $t \in [\bar{t}_6, T^Z]$.

I claim that $\check{G}^R$ must satisfy the Dynamic IC constraint. For $t \geq \bar{t}_3$ we have $U^c_{\check{G}^R}(T^Z) - U^c_{\check{G}^R}(t) = U^c_{G^R}(T^Z) - U^c_{G^R}(t) \geq 0$. For $t \in [\bar{t}_6, \bar{t}_3]$ we have $U^c_{\check{G}^R}(T^Z) - U^c_{\check{G}^R}(t) = U^c_{\check{G}^R}(T^Z) - U^c_{\check{G}^R}(\bar{t}_3) \geq 0$ (recall that $U^c_{\check{G}^R}(t) = U^c_{\check{G}^R}(\bar{t}_3)$). Finally for $t < \bar{t}_6$ we have $U^c_{\check{G}^R}(t) = U^c_{G^R}(t)$ and so $U^c_{\check{G}^R}(T^Z) > U^c_{\check{G}^R}(t)$. The new distribution $\check{G}^R$ must certainly also satisfy the type IC constraint, because G^Z is unchanged and therefore so is $U^n(t)$.

For arbitrary cumulative distribution function G^R on $[0, T^Z]$ with $T^Z = T^R$ let $T = T^Z$, as well as $v(T, G^R) = U^c_{G^R}(T)$, $w(t, G^R) = U^c_{G^R}(T) - U^c_{G^R}(t)$. The proof above establishes that for any G^R satisfying both constraints, if the dynamic incentive constraint doesn't bind ($w(t, G^R) > 0$ for some $t < T$), then there is some alternative distribution $\check{G}^R$ on $[0, T]$ with $\check{G}^R(t) \geq G^R(t)$ which improves rational payoffs but still satisfies both incentive constraints ($v(T, \check{G}^R) > v(T, G^R) \geq U^n(t)$ and $w(t, \check{G}^R) \geq 0$ for $t \in [0, T]$). We can then apply Lemma 5, which in this case establishes the existence of some $\overline{G}^R$ which satisfies both incentive constraints and delivers a higher time T payoff than G^R (i.e. $v(T, \overline{G}^R) > v(T, G^R) \geq U^n(t)$ and $w(t, \overline{G}^R) \geq 0$ for $t \in [0, T]$), for which there is no other distribution on $[0, T]$ with $\tilde{G}^R(t) \geq \overline{G}^R(t)$. This implies that the dynamic incentive constraint must bind for $\overline{G}^R$ (i.e. $w(t, \overline{G}^R) = 0$ for $t \in [0, T]$). □

Proof of Lemma 5. Define $\overline{u}(\hat{G}) = \sup_{\tilde{G} \in X}\{v(T, \tilde{G}) : \tilde{G} \gtrsim \hat{G}\}$ (respectively $\overline{u}_{\tilde{t}}(\hat{G}) = \sup_{\tilde{G} \in X}\{v(T, \tilde{G}) : \tilde{G} \gtrsim_{\tilde{t}} \hat{G}\}$). Let $G^0 = G$ and choose $G^{k+1} \gtrsim G^k$ (respectively $G^{k+1} \gtrsim_{\tilde{t}} G^k$) such that $v(T, G^{k+1}) \geq \frac{\overline{u}(G) + v(T, G^k)}{2}$ (respectively $v(T, G^{k+1}) \geq \frac{\overline{u}_{\tilde{t}}(G) + v(T, G^k)}{2}$). Let $\underline{G}(t) = \lim_k G^k(t)$ and then define the cumulative distribution function $\overline{G}$ by $\overline{G}(t) = \inf\{\underline{G}(s) : s > t\}$. Clearly we have G^k weakly converging to $\overline{G}$ ($G^k \underset{w}{\to} \overline{G}$).

Given $G^k(T) = \overline{G}(T) = 1$ and the weak convergence of G^k, we clearly have $\lim_k \int_{s \leq T} A_k(s) dG^k(s) = \int_{s \leq T} A_k(s) d\overline{G}(s)$ and so ultimately $\lim_k v(T, G^k) = v(T, \overline{G})$ and $\lim_k w(T, G^k) = w(T, \overline{G}) \geq 0$. Noticing that $\overline{G}$ is continuous almost everywhere, let $Y = \{t : \overline{G} \text{ is continuous at } t\}$. For $t \in Y$ we have $\lim_k G^k(t) = \overline{G}(t)$. Define the cumulative distribution functions $G^{k,t}$ on $[0, t]$ by $G^{k,t}(s) = \frac{G^k(s)}{G^k(t)}$ and $\overline{G}^t(s) = \frac{\overline{G}(s)}{\overline{G}(t)}$, then $G^{k,t} \underset{w}{\to} \overline{G}^t$. This ensures $\lim_k \int_{s \leq t} A_k(s) dG^k(s) = \int_{s \leq t} A_k(s) d\overline{G}(s)$ and so ultimately for $t \in Y$ we have $\lim_k w(t, G^k) = w(t, \overline{G}) \geq 0$. For $t \notin Y$, the right continuity of $\overline{G}$ implies that $w(t, \overline{G}) \geq \sup_{\nu > t} \inf_{s \in (t, \nu] \cap Y} w(s, \overline{G}) \geq 0$. This establishes $\overline{G} \in X$.

If $G^{k+1} \gtrsim G^k$ then $\overline{G}(t) \geq G^k(t)$ and $v(T, \overline{G}) \geq v(T, G^k) > v(T, G^0)$ so that $\overline{G} \gtrsim G^k \succ G^0$. If $G^{k+1} \gtrsim_{\tilde{t}} G^k$ we have $\bar{t}_{\overline{G}} \leq \bar{t}_{G^k}$ (indeed, $G^k(t) = \overline{G}(t)$ for $t \geq \bar{t}_{G^k}$), $\overline{G}(t) \geq G^k(t)$ for $t \geq \tilde{t}$ and $\overline{G}(t) \leq G^k(t)$ for $t \leq \tilde{t}$ and $v(T, \overline{G}) \geq v(T, G^k) > v(T, G^0)$ so that $\overline{G} \gtrsim_{\tilde{t}} G^k \succ_{\tilde{t}} G^0$.

Suppose then that there exists $\hat{G} \in X$ such that $\hat{G} \succ \overline{G}$ (respectively $\hat{G} \succ_{\tilde{t}} \overline{G}$), then define $\varepsilon = v(T, \hat{G}) - v(T, \overline{G}) > 0$. Clearly, $\hat{G} \succ G^k$ so that $\overline{u}(G^k) \geq v(T, G^k) + \varepsilon$ (respectively $\hat{G} \succ \overline{G}$ so that $\overline{u}_{\tilde{t}}(G^k) \geq v(T, G^k) + \varepsilon$. That in turn implies $v(T, G^k) \geq v(T, G^0) + k\frac{\varepsilon}{2}$ and so $v(T, \overline{G}) = \infty$, which contradicts the fact that $v(T, \overline{G})$ must be bounded. □

Proof of Lemma 6. We are given that $IC_{G^Z}(t) > 0$ for some $t \in [\min\{\hat{t}, T^Z\}, T^Z]$. We first want to find an alternative distribution $\hat{G}^Z$ with $IC_{\hat{G}^Z}(t) \geq 0$ and $\hat{G}^Z(t) \geq G^Z(t)$ for all t so that $G^R_{\hat{G}^Z}(0) > G^R_{G^Z}(0)$. Initially suppose that $IC_{G^Z}(T^Z) = \delta > 0$. If $T^Z < \infty$ then let $t' = T^Z$. If $T^Z = \infty$ then let $e^{-rt'}u(1 - \alpha) = \frac{\delta}{3}$ so that $IC_{G^Z}(t) \geq \frac{2\delta}{3}$ for $t \geq t'$. Given the right continuity of G^Z, we must have $IC_{G^Z}(t) \geq \frac{\delta}{3}$ for $t \geq t' - \varepsilon$ for some $\varepsilon > 0$. Consider the alternative distribution $\hat{G}^Z$, such that $\hat{G}^Z(t) = G^Z(t)$ for $t < T^Z - \varepsilon$ and $\hat{G}^Z(t) = \min\{G^Z(t) + \varepsilon', 1\}$ for $t \geq t' - \varepsilon$ and some $\varepsilon' > 0$. Notice that for $t < t' - \varepsilon$ we must have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t)$. For all $t \geq t' - \varepsilon$ we have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t) - \varepsilon' u(\alpha)$. Given $IC_{G^Z}(t) \geq \frac{\delta}{3}$, by selecting $\varepsilon' > 0$ sufficiently small, we must have $IC_{\hat{G}^Z}(t) \geq 0$ for all $t \geq t' - \varepsilon$.

Next, suppose that $IC_{G^Z}(T^Z) = 0$ but $IC_{G^Z}(t') = \delta > 0$ for some $t' \in [\hat{t}, T^Z)$. Let $\hat{G}^Z(t) = G^Z(t)$ if $t < t'$ and $\hat{G}^Z(t) = \max\{G^Z(t') + \varepsilon, G^Z(t)\}$ otherwise, for some $\varepsilon > 0$. Clearly, $IC_{\hat{G}^Z}(t') \geq IC_{G^Z}(t') - \varepsilon u(\alpha)$ so that for $\varepsilon \leq \frac{\delta}{2u(\alpha)}$ we have $IC_{\hat{G}^Z}(t') \geq \frac{\delta}{2}$. If $\hat{G}^Z(t) = G^Z(t)$ then because the final integrand in equation (11) is always positive for $s \geq t'$ and $\hat{G}^Z(s) \geq G^Z(s)$, we must have $IC_{\hat{G}^Z}(t) \geq IC_{G^Z}(t) \geq 0$. If $\hat{G}^Z(t) = \hat{G}^Z(t') > G^Z(t)$, however, then $U^n_{\hat{G}^Z}(t) \leq U^n_{\hat{G}^Z}(t')$ and so $IC_{\hat{G}^Z}(t) \geq IC_{\hat{G}^Z}(t') > 0$ (a larger t simply delays a non-confessing agent's payoff from concession).

For arbitrary $\tilde{G}^Z$ define $v(T, \tilde{G}^Z) = G^R_{\tilde{G}^Z}(0)$ and $w(t, \tilde{G}^Z) = IC_{\tilde{G}^Z}(t)$, as well as $T = \infty$. The proof above establishes that if the time t type incentive constraint doesn't bind for some $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ (i.e. $w(t, \tilde{G}^Z) > 0$) then there is some alternative incentive compatible $\hat{G}^Z$ ($w(t, \hat{G}^Z) \geq 0$ for all $t \leq T$) delivering higher payoffs ($v(T, \hat{G}^Z) > v(T, \tilde{G}^Z)$) with $\hat{G}^Z(s) \geq \tilde{G}^Z(s)$. Invoking Lemma 5, this implies the existence of some $\overline{G}^Z$ with $\overline{G}^Z(t) \geq \tilde{G}^Z(t)$, $v(T, \overline{G}^Z) > v(T, \tilde{G}^Z)$ and $w(t, \overline{G}^Z) \geq 0$ for all $t \leq T$ such that there is no alternative $\check{G}^Z$ with $\check{G}^Z(t) \geq \overline{G}^Z(t)$, $v(T, \check{G}^Z) > v(T, \bar{G}^Z)$ and $w(t, \check{G}^Z) \geq 0$. And so, this incentive compatible distribution delivers higher payoffs and must satisfy $IC_{\overline{G}^Z}(t) = 0$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$. □

Proof of Lemma 7. By Lemma 6, we can restrict attention to distributions of the form G^Z_{0,T^Z}, and so maximizing $U^c(T^R)$ reduces to the problem of minimizing T^Z such that $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$ where:

$$\begin{aligned} IC_{G^Z_{\check{t},T^Z}}(T^Z) =& u(1-\alpha) + (1-z)G^R_{G^Z_{\check{t},T^Z}}(0)(u(0.5) - u(1-\alpha)) - e^{-r\check{t}}\left(u(1-\alpha) + (1-z)G^Z_{\check{t},T^Z}(\check{t})(u(\alpha) - u(1-\alpha))\right) \\ =& (1-z)(u(0.5) - u(1-\alpha))\left(1 - \frac{z\lambda^m}{1-z}\int_0^{\check{t}} e^{\lambda^m s}ds - \frac{z^2\lambda^m}{(1-z)^2}\int_{\check{t}}^{T^Z} e^{\lambda T^Z + (\lambda^m - \lambda)s} - e^{\lambda^m s}ds\right) \\ &+ u(1-\alpha)(1 - e^{r\check{t}}) - e^{-r\check{t}}(u(\alpha) - u(1-\alpha))(1 - e^{\lambda(T^Z - \check{t})}z) \end{aligned} \tag{15}$$

For risk neutral agents we have $IC_{G^Z_{0,T^Z}}(T^Z) = 0$ when $T^Z = -\frac{1}{\lambda}ln(z)$. Taking the derivative of $IC_{G^Z_{0,T^Z}}(T^Z)$ we get

$$\begin{aligned} \frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z} &= -\frac{z^2}{1-z}ru(1-\alpha)\int_0^{T^Z} \lambda e^{\lambda T^Z + (\lambda^m - \lambda)s}ds + zru(1-\alpha)e^{\lambda T^Z} \\ &= -\frac{z^2}{1-z}ru(1-\alpha)\frac{\lambda}{\lambda^m - \lambda}(e^{\lambda^m T^Z} - e^{\lambda T^Z}) + zru(1-\alpha)e^{\lambda T^Z} \end{aligned}$$

Notice that $u(x) = x$ implies $\lambda^m = 2\lambda$ and so $\frac{\lambda}{\lambda^m - \lambda} = 1$. In turn, this implies $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z}\Big|_{T^Z = -\frac{1}{\lambda}ln(z)} = 0$. Finally, notice that $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z}e^{-\lambda T^Z}$ is strictly decreasing in T^Z and so $\frac{dIC_{G^Z_{0,T^Z}}(T^Z)}{dT^Z} > 0$ for $T^Z < -\frac{1}{\lambda}ln(z)$. Hence, we must have $IC_{G^Z_{0,T^Z}}(T^Z) < 0$ whenever $T^Z < -\frac{1}{\lambda}ln(z)$. In the OSSMP, therefore, we must have $T^Z = -\frac{1}{\lambda}ln(z)$, so that the optimal distributions G^{Z*} and G^{R*} correspond exactly to the Baseline equilibrium. □

Proof of Lemma 8. Given Lemma 1, we only need to contend with the risk neutral case. This proof derives some expressions in greater detail than is needed, but which are used in later proofs. Consider a distribution of the form $G^Z_{\check{t},T^Z(\check{t},\kappa)}$ where:

$$T^Z(\check{t}, \kappa) = \check{t} + \frac{1}{\lambda}ln\left(\frac{u(\alpha) - \kappa e^{r\check{t}}}{z(u(\alpha) - u(1-\alpha))}\right) \tag{16}$$

is defined to ensure $U^n(t) = \kappa$ for $t \in [\check{t}, T^Z]$. Such a distribution implies that:

$$G^R_{G^Z_{\check{t},T^Z(\check{t},\kappa)}}(0) = 1 - \frac{z}{1-z}\int_0^{\check{t}} \lambda^m e^{\lambda^m s}ds - \left(\frac{z}{1-z}\right)^2 \lambda^m \int_{\check{t}}^{T^Z} e^{\lambda T^Z + (\lambda^m - \lambda)s} - e^{\lambda^m s}ds$$

and so

$$\frac{dG^R_{G^Z_{\check{t},T^Z(\check{t},\kappa)}}(0)}{d\check{t}} = \frac{z\lambda^m}{(1-z)^2}\left(ze^{\lambda T^Z+(\lambda^m-\lambda)\check{t}} - e^{\lambda^m\check{t}}\right) - \frac{dT^Z(\check{t},\kappa)}{d\check{t}}\frac{z^2}{(1-z)^2}\lambda\lambda^m\int_{\check{t}}^{T^Z} e^{\lambda T^Z+(\lambda^m-\lambda)s}ds \tag{17}$$

which has the same sign as

$$\begin{aligned} Y(w) =& \frac{dG^R_{G^Z_{\check{t},T^Z(\check{t},\kappa)}}(0)}{d\check{t}}\frac{(1-z)^2}{z\lambda^m e^{\lambda^m\check{t}}} = -1 + ze^{\lambda(T^Z-\check{t})} - \frac{dT^Z(\check{t},\kappa)}{d\check{t}}\frac{z\lambda}{\lambda^m-\lambda}e^{\lambda(T^Z-\check{t})}(e^{(\lambda^m-\lambda)(T^Z-\check{t})}-1) \\ =& -1 + \left(\frac{u(\alpha)-w}{u(\alpha)-u(1-\alpha)}\right) - \frac{\lambda u(\alpha)}{(\lambda^m-\lambda)u(1-\alpha)}\left(\frac{u(1-\alpha)-w}{u(\alpha)-u(1-\alpha)}\right)\left(\left(\frac{u(\alpha)-w}{z(u(\alpha)-u(1-\alpha))}\right)^{\frac{\lambda^m-\lambda}{\lambda}} - 1\right) \end{aligned} \tag{18}$$

where $w = \kappa e^{r\check{t}}$. The first line evaluates the integral and the second imposes

$$\frac{dT^Z(\check{t},\kappa)}{d\check{t}} = \frac{u(\alpha)\lambda - (r+\lambda)\kappa e^{r\check{t}}}{\lambda(u(\alpha)-\kappa e^{r\check{t}})} = \frac{u(\alpha)}{u(1-\alpha)}\frac{u(1-\alpha)-\kappa e^{r\check{t}}}{u(\alpha)-\kappa e^{r\check{t}}}$$

where this uses $\frac{r+\lambda}{\lambda} = \frac{u(\alpha)}{u(1-\alpha)}$. Clearly $Y(u(1-\alpha)) = 0$, as occurs when $\check{t} = 0$ and $\kappa = u(1-\alpha)$. Let $V(w) = \frac{dY(w)}{dw}(u(\alpha)-u(1-\alpha))$ then:

$$V(w) = -1 + \frac{u(\alpha)\lambda}{u(1-\alpha)(\lambda^m-\lambda)}\left(\left(\frac{u(\alpha)-w}{z(u(\alpha)-u(1-\alpha))}\right)^{\frac{\lambda^m-\lambda}{\lambda}} - 1 + (u(1-\alpha)-w)\frac{\lambda^m-\lambda}{\lambda}\frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}}\right) \tag{19}$$

Imposing risk neutrality, $u(x) = x$, so that $\lambda^m = 2\lambda$, and evaluating $V(w)$ at $w = (1-\alpha)$ gives:

$$V(1-\alpha) = -1 + \frac{\alpha}{1-\alpha}\left(\frac{1}{z} - 1\right) = \frac{\alpha - z}{z(1-\alpha)}$$

This is clearly positive whenever $\alpha > z$, which implies $Y(w) > 0$ when w is slightly greater than $(1-\alpha)$, and so $G^R_{G^Z_{\check{t},T^Z(\check{t},\kappa)}}(0) > 0$ when $\kappa = (1-\alpha)$ and $\check{t}$ is slightly greater than 0. Recalling that $U^n(t) = \kappa$ for $t \in [\check{t}, T^Z(\check{t},\kappa)]$ implies $IC_{G^Z_{\check{t},T^Z(\check{t},\kappa)}}(t) > 0$ for all t. □

Proof of Lemma 9. First notice that $IC_{G^Z_{\check{t},T^Z}}(T^Z)$ is continuous and strictly increasing in $\check{t}$ for $\check{t} \leq \min\{T^Z, \hat{t}\}$ (see equation (15)), and so $\check{t}(T^Z)$ is well defined given $IC_{G^Z_{\min\{T^Z,\hat{t}\},T^Z}}(T^Z) \geq 0$. We are given the existence of some $\check{G}^Z$ with $\min_s IC_{\check{G}^Z}(s) = 0 = IC_{\check{G}^Z}(t)$ for $t \in [\min\{\hat{t}, T^Z\}, T^Z]$ and $\check{G}^Z \neq G^Z_{\check{t}(T^Z),T^Z}$. Let the dependence of T^Z on G^Z be explicit, so that $T^Z_{G^Z} = \min\{t : G^Z(t) = 1\}$, and then define:

$$\check{X} = \{G^Z \in \Delta_\infty : T^Z_{G^Z} = T^Z_{\check{G}^Z} \text{ and } \min_s IC_{G^Z}(s) = 0 = IC_{G^Z}(t) \text{ for } t \in [\min\{\hat{t}, T^Z\}, T^Z]\}.$$

recalling that Δ_∞ is the set of cumulative distribution functions with $G \in \Delta_\infty$ if $G(t) = 0$ for $t < 0$. For the rest of the proof, I restrict attention to $G^Z \in \check{X}$. Let $\underline{t}_{G^Z} = \inf\{t : G^Z(t) > 0\}$ and $\bar{t}_{G^Z} = \min\{t : IC_{G^Z}(s) = 0 \text{ for } s \in [t, T^Z]\}$. Clearly, we have, $\underline{t}_{\check{G}^Z} \leq \bar{t}_{\check{G}^Z} \leq \min\{\hat{t}, T^Z\}$. Moreover, notice that we must have $G^Z(t) = G^Z_{\check{t}(T^Z),T^Z}(t)$ for $t \geq \bar{t}_{\check{G}^Z}$.

Given $\check{G}^Z \neq G^Z_{\check{t}(T^Z),T^Z}$, I claim that $\bar{t}_{\check{G}^Z} > \check{t}(T^Z)$. Suppose not, then $\check{G}^Z(t) \geq G^Z_{\check{t}(T^Z),T^Z}(t)$ for all t and $IC_{G^Z_{\check{t}(T^Z),T^Z}}(T^Z) = 0$. Because the final integrand in the expression for $IC_{G^Z}(T^Z)$ in equation (11) is negative for all $s < \min\{\hat{t}, T^Z\}$, we must then have $IC_{\check{G}^Z}(T^Z) - IC_{G^Z_{\check{t}(T^Z),T^Z}}(T^Z) < 0$, a contradiction. I further claim that $\underline{t}_{\check{G}^Z} < \bar{t}_{\check{G}^Z}$. Suppose not, so that $\underline{t}_{\check{G}^Z} = \bar{t}_{\check{G}^Z} > \check{t}(T^Z)$. In this case $\check{G}^Z(t) \leq G^Z_{\check{t}(T^Z),T^Z}(t)$ for all t, which implies $IC_{\check{G}^Z}(T^Z) - IC_{G^Z_{\check{t}(T^Z),T^Z}}(T^Z) > 0$. Given that by assumption $IC_{\check{G}^Z}(T^Z) = 0$ we again have a contradiction.

Given some fixed $\tilde{t}$, define the partial order $\succsim^*_{\tilde{t}}$ over $\check{X}$ as follows. Let $G^Z \sim^*_{\tilde{t}} G^Z$, and let $G^Z \succ^*_{\tilde{t}} \tilde{G}^Z$ if $\underline{t}_{\tilde{G}^Z} \leq \underline{t}_{G^Z}$, $\bar{t}_{\tilde{G}^Z} \geq \bar{t}_{G^Z}$, $G^R_{G^Z}(0) > G^R_{\tilde{G}^Z}(0)$ and either (a) $\underline{t}_{G^Z} = \tilde{t}$ and $G^Z(t) \geq \tilde{G}^Z(t)$ for $t \geq \tilde{t}$, or (b) $\bar{t}_{G^Z} = \tilde{t}$ and $G^Z(t) \leq \tilde{G}^Z(t)$ for $t < \tilde{t}$.

Given $\check{G}^Z \neq G^Z_{\check{t}(T^Z),T^Z}$, I claim that for any $\tilde{t} \in (\underline{t}_{\check{G}^Z}, \bar{t}_{\check{G}^Z})$, there exists a distribution $\bar{G}^Z_{\tilde{t}}$ such that $\bar{G}^Z_{\tilde{t}} \succ^*_{\tilde{t}} \check{G}^Z$, and there is no distribution G^Z for which $G^Z \succ^*_{\tilde{t}} G^Z_{\tilde{t}}$.

I first simply look to find an alternative distribution $\hat{G}^Z$, which generates higher payoffs than $\check{G}^Z$. For some $\varepsilon \in (0, \tilde{t} - \underline{t}_{\check{G}^Z})$ and $\varepsilon' \geq 0$, let $\hat{G}^Z(t) = \min\{\check{G}^Z(t) - \varepsilon', 0\}$ for $t \leq \underline{t}_{\check{G}^Z} + \varepsilon$. Suppose that $\check{G}^Z$ is discontinuous at some $t' \in (\tilde{t}, \bar{t}_{\check{G}^Z}]$ so that $\sup_{s<t'} \check{G}(s) < \check{G}(t')$. In this case let $\hat{G}^Z(t) = \sup_{s<t'} \check{G}(s) + \varepsilon''$ for $t \in [t' - \varepsilon, t')$, where $\varepsilon'' \geq 0$ is still to be defined and we additionally restrict attention to $\varepsilon < t' - \tilde{t}$. If on the other hand $\check{G}(t)$ is continuous on $(\tilde{t}, \bar{t}]$ then there must exist some $t' \in (\tilde{t}, \bar{t}]$ such that $IC_{G^Z}(t') > 0$ and $G^Z(t') > G^Z(t)$ for all $t < t'$. In this case define $\hat{G}^Z(t) = \max\{\check{G}(t) + \varepsilon'', \check{G}(t')\}$ for $t \in [t' - \varepsilon, t')$. Let $\hat{G}^Z(t) = \check{G}^Z(t)$ elsewhere. We approximately have, $\hat{G}^Z - \check{G}^Z(s) \approx -\varepsilon' \leq 0$ for $s \in [\underline{t}_{\check{G}^Z}, \underline{t}_{\check{G}^Z} + \varepsilon]$, whereas $\hat{G}^Z - \check{G}^Z(s) \approx \varepsilon'' \geq 0$ for $s \in [t' - \varepsilon, t')$ and $\hat{G}^Z = \check{G}^Z(s)$ elsewhere.

For $t \geq t'$ we have:

$$IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = \int_{\underline{t}_{\check{G}^Z}}^{\underline{t}_{\check{G}^Z}+\varepsilon} (\hat{G}^Z - \check{G}^Z(s)) r\left(u(1-\alpha)e^{\lambda^m s} z - u(\alpha)e^{-rs}(1-z)\right) ds + \int_{t'-\varepsilon}^{t'} (\hat{G}^Z - \check{G}^Z(s)) r\left(u(1-\alpha)e^{\lambda^m s} z - u(\alpha)e^{-rs}(1-z)\right) ds$$

This difference is continuous and strictly increasing in ε' and $-\varepsilon''$, is positive for $\varepsilon'' = 0$ and negative for $\varepsilon' = 0$. For all sufficiently small ε' therefore, there is a uniquely defined ε'' such that $IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = 0$ for $t \geq t'$. This leaves $\hat{G}^Z$ as a function of ε' and ε. For $IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t) = 0$ as first $\varepsilon' \to 0$ and then $\varepsilon \to 0$ we must have:

$$\lim_{\varepsilon\to 0}\lim_{\varepsilon'\to 0} \frac{IC_{\hat{G}^Z}(t) - IC_{\check{G}^Z}(t)}{\varepsilon\varepsilon'} = 0 = \lim_{\varepsilon\to 0}\lim_{\varepsilon'\to 0} \frac{\varepsilon''}{\varepsilon'} r\left(u(1-\alpha)e^{\lambda^m t'} z - u(\alpha)e^{-rt'}(1-z)\right) - r\left(u(1-\alpha)e^{\lambda^m \underline{t}_{\check{G}^Z}} z - u(\alpha)e^{-r\underline{t}_{\check{G}^Z}}(1-z)\right)$$

Notice that for sufficiently small ε, we have $IC_{\check{G}^Z}(s) \geq \delta$ for $s \in [t' - \varepsilon, t')$ and some $\delta > 0$ and so for such s, $IC_{\hat{G}^Z}(s) > 0$ for all sufficiently small ε'. For $s < t' - \varepsilon$ we have $U^n_{\check{G}^Z}(s) \leq U^n_{\hat{G}^Z}(s)$, hence, so long as we can show $G^R_{\hat{G}^Z}(0) > G^R_{\check{G}^Z}(0)$ then all time t type incentive constraints will be satisfied for $\hat{G}^Z$. To that end, notice that:

$$G^R_{\hat{G}^Z}(0) - G^R_{\check{G}^Z}(0) = \int_0^{T^Z} \lambda^m e^{\lambda^m s} \frac{z}{1-z} (\hat{G}^Z(s) - \check{G}^Z(s)) ds$$

$$\begin{aligned}
\lim_{\varepsilon\to 0}\lim_{\varepsilon'\to 0} \frac{G^R_{\hat{G}^Z}(0) - G^R_{\check{G}^Z}(0)}{\varepsilon\varepsilon'} \frac{1-z}{z\lambda^m} &= e^{\lambda^m t'} \lim_{\varepsilon'\to 0}\lim_{\varepsilon\to 0} \frac{\varepsilon''}{\varepsilon'} - e^{\lambda^m \underline{t}_{\check{G}^Z}} \\
&= \frac{e^{\lambda^m t'}\left(u(1-\alpha)e^{\lambda^m \underline{t}_{\check{G}^Z}} z - u(\alpha)e^{-r\underline{t}_{\check{G}^Z}}(1-z)\right) - e^{\lambda^m \underline{t}_{\check{G}^Z}}\left(u(1-\alpha)e^{\lambda^m t'} z - u(\alpha)e^{-rt'}(1-z)\right)}{u(1-\alpha)e^{\lambda^m t'} z - u(\alpha)e^{-rt'}(1-z)} \\
&= \frac{u(\alpha)(1-z)e^{(\lambda^m - r)\underline{t}_{\check{G}^Z}}(e^{-r(t' - \underline{t}_{\check{G}^Z})} - e^{\lambda^m(t' - \underline{t}_{\check{G}^Z})})}{u(1-\alpha)e^{\lambda^m t'} z - u(\alpha)e^{-rt'}(1-z)} > 0,
\end{aligned}$$

where the final two equalities hold for $t' < \hat{t}$ and the inequality in the final line follows because the denominator is negative for $t' < \hat{t}$, and the numerator is also negative given $t' > \underline{t}_{\check{G}^Z}$. If $t' = \hat{t}$, on the other hand then we must have $\lim_{\varepsilon\to 0} \lim_{\varepsilon'\to 0} \frac{\varepsilon''}{\varepsilon'} = \infty$ and so the second line must certainly be strictly positive. The implication of this is that for sufficiently small ε and ε', $G^R_{\hat{G}^Z}(0) > G^R_{\check{G}^Z}(0)$.

Fix $\tilde{t} \in (\underline{t}_{\check{G}^Z}, \bar{t}_{\check{G}^Z})$ and $T = T^Z_{\check{G}^Z}$. For arbitrary G^Z define $v(T, G^Z) = G^R_{G^Z}(0)$ and $w(t, G^Z) = IC_{G^Z}(t)$. We established above that there exists some $\hat{G}^Z$ such that $\bar{t}_{\hat{G}^Z} \leq \bar{t}_{\check{G}^Z}$, $\hat{G}^Z(t) \geq \check{G}^Z(t)$ for $t \geq \tilde{t}$, $\hat{G}^Z(s) \leq \check{G}^Z(s)$ for $s < \tilde{t}$, $v(T, \hat{G}^Z) > v(T, \check{G}^Z)$ and $w(t, \hat{G}^Z) \geq 0$ for all $t \leq T$. Invoking Lemma 5 (part (b)), therefore, there exists $\overline{G}^Z_{\tilde{t}} \in \check{X}$ such that $IC_{\overline{G}^Z_{\tilde{t}}}(t) \geq 0$, $G^R_{\overline{G}^Z_{\tilde{t}}}(0) > G^R_{\check{G}^Z}(0)$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \leq \bar{t}_{\check{G}^Z}$, $\bar{G}^Z_{\tilde{t}}(t) \geq \check{G}^Z(t)$ for $t \geq \tilde{t}$, $\bar{G}^Z_{\tilde{t}}(s) \leq \check{G}^Z(s)$ for $s < \tilde{t}$ (implying $\underline{t}_{\bar{G}^Z_{\tilde{t}}} \leq \underline{t}_{\check{G}^Z}$), and furthermore, there is no alternative incentive compatible G^Z with $G^Z(t) \geq \overline{G}^Z_{\tilde{t}}(t)$ for $t \geq \tilde{t}$ and $G^Z(s) \leq \overline{G}^Z_{\tilde{t}}(s)$ for $s < \tilde{t}$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \geq \bar{t}_{G^Z}$ such that $G^R_{G^Z}(0) > G^R_{\overline{G}^Z_{\tilde{t}}}(0)$.

I claim that $\bar{G}^Z_{\tilde{t}} \succ^*_{\tilde{t}} \check{G}^Z$. Given the comparison of the descriptions above, for this not to be true requires that both $\bar{t}_{\bar{G}^Z_{\tilde{t}}} < \tilde{t}$ and $\bar{t}_{\bar{G}^Z_{\tilde{t}}} > \tilde{t}$. But in this case, we know from the previous construction that there must exist an incentive compatible distribution $\tilde{G}^Z$ with $\tilde{G}^Z(t) \geq \overline{G}^Z_{\tilde{t}}(t)$ for $t \geq \tilde{t}$ and $\tilde{G}^Z(t) \leq \overline{G}^Z_{\tilde{t}}(t)$ for $t < \tilde{t}$, $\bar{t}_{\bar{G}^Z_{\tilde{t}}} \geq \bar{t}_{\tilde{G}^Z}$ such that $G^R_{\tilde{G}^Z}(0) > G^R_{\bar{G}^Z_{\tilde{t}}}(0)$, a contradiction.

I next define a new partial order $\gtrsim^*$ on $\check{X}$ as follows. Let $G^Z \sim^* G^Z$, and let $G^Z \succ^* G^Z$ if there exists some $\tilde{t}$ such that $G^Z \succ^*_{\tilde{t}} G^Z$. Let $\overline{u}^*(G^Z) = \sup_{\tilde{G}^Z \in \check{X}}\{G^R_{\tilde{G}^Z}(0) : \tilde{G}^Z \gtrsim^* G^Z\}$. Now define a sequence of distribution functions by $G^0 = \check{G}^Z \neq G^Z_{\check{t}(T^Z),T^Z}$, and $G^{k+1} \gtrsim^* G^k$ such that $G^R_{G^{k+1}}(0) \geq \frac{\overline{u}^*(G^k)+G^R_{G^k}(0)}{2}$. Clearly, if $G^k = G^Z_{\check{t}(T^Z),T^Z}$ for some k then the proof is complete (as then $G^R_{G^Z_{\check{t}(T^Z),T^Z}}(0) \geq G^R_{\check{G}^Z}(0)$), so suppose not and $G^{k+1} \succ^* G^k$ for all k.

We want to establish that G^k weakly converges to $G^Z_{\check{t}(T^Z),T^Z}$ ($G^k \overrightarrow{w} G^Z_{\check{t}(T^Z),T^Z}$), which if shown completes the proof as then $G^R_{G^Z_{\check{t}(T^Z),T^Z}}(0) > G^R_{\check{G}^Z}(0)$. Suppose not. Let $\tilde{t}^k \in [\underline{t}_{G^0}, \bar{t}_{G^0}]$ be such that $G^{k+1} \succ^*_{\tilde{t}^k} G^k$, and (taking a subsequence if necessary) let $\tilde{t}^k \to \tilde{t}^*$.

I claim that $\underline{G}(t) = \lim_k G^k(t)$ is well defined except possibly at $t = \tilde{t}^*$, where we can let it be defined by some arbitrary subsequence if necessary. This is clearly the case if $\underline{t}_{G^k} \to \tilde{t}^*$ and $\bar{t}_{G^k} \to \tilde{t}^*$ (notice that $\underline{t}_{G^k}$ and $-\bar{t}_{G^k}$ must be increasing in k). If $\underline{t}_{G^k} \not\to \tilde{t}^*$, then for some $t < \tilde{t}^*$ we have $\underline{t}_{G^k} < t$ for all k, (i.e. $G^k(t) > 0$). For all sufficiently large k we must then have $\tilde{t}^{k-1} > t$, and so $IC_{G^k}(s) = 0$ for $s \in [\tilde{t}^{k-1}, T^Z]$ and $G^k(v) \leq G^{k-1}(v)$ for $v \leq \tilde{t}^{k-1}$. But in that case $\underline{G}(t)$ is defined for all t. Suppose instead that $\bar{t}_{G^k} \not\to \tilde{t}^*$, then for some $t > \tilde{t}^*$ we have $\bar{t}_{G^k} \geq t$ for all k. For all sufficiently large k we must have $\tilde{t}^{k-1} < t$, and so $G^k(s) = 0$ for $s \leq \tilde{t}^{k-1}$ and $G^k(s) \geq G^{k-1}(s)$ for $s \geq \tilde{t}^{k-1}$. But in that case $\underline{G}(t)$ is again defined for all t.

We can now define the cumulative distribution function $\overline{G}$ by $\overline{G}(t) = \inf\{\underline{G}(s) : s > t\}$, so that $G^k \overrightarrow{w} \overline{G}$. We clearly have $\bar{G} \in \check{X}$ and $\overline{G} \gtrsim^* G^k$ (the proof of Lemma 5 provides explicit reasoning). If $\overline{G} \neq G^Z_{\check{t}(T^Z),T^Z}$, then there must exist some $\tilde{G}^Z \succ^* \overline{G}$. But in that case $\tilde{G}^Z \succ^* G^k$ and so $G^R_{G^{k+1}}(0) - G^R_{G^k}(0) \geq \frac{\overline{G}^R_{\tilde{G}^Z}(0) - G^R_{\overline{G}}(0)}{2} > 0$, but this must then contradict $G^R_{G^k}(0) \leq 1$ for sufficiently large k, and so imply $\overline{G} = G^Z_{\check{t}(T^Z),T^Z}$. □

Proof of Lemma 10. Suppose $z \geq \frac{u(\alpha)}{u(\alpha)+u(1-\alpha)}$ and so $\hat{t} \leq 0$. By Lemma 6, therefore, we can restrict attention to G^Z of the form G^Z_{0,T^Z}. To maximize $U^c(T^R)$, therefore, we simply need to minimize T^Z among such distributions, subject to $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$. Equation (15), shows that $IC_{G^Z_{0,T^Z}}(T^Z)$ is continuous in T^Z. Because G^Z_{0,T^Z} corresponds to the Baseline equilibrium distribution when $T^Z = -\frac{1}{\lambda}ln(z)$ and so $IC_{G^Z_{0,T^Z}}(T^Z) \geq 0$ we have a non-empty set over which to minimize T^Z. Clearly, therefore, an optimal distribution G^{Z*} must not only exist but be unique in this case.

Now suppose $z < \frac{u(\alpha)}{u(\alpha)+u(1-\alpha)}$. By Lemma 9 we can restrict attention to G^Z of the form $G^Z_{\check{t}(T^Z),T^Z}$. To establish existence, first notice $IC_{G^Z_{\check{t},T^Z}}(T^Z)$ is continuous and strictly increasing in $\check{t}$ and continuous in T^Z (again, see equation (15)), and so the set of T^Z for which $\check{t}(T^Z)$ is defined, is closed. The fact that incentive constraints are satisfied under the Baseline distribution $G^Z_{0,-\frac{1}{\lambda}ln(z)}$ implies that this set is also non-empty. Moreover, $\check{t}(T^Z)$ is continuous on that closed set. Hence, the OSSMP problem can be reduced to maximizing a continuous function of T^Z, $G^R_{G^Z_{\check{t}(T^Z),T^Z}}(0)$ on a compact set $\{T^Z : IC_{G^Z_{\min\{T^Z,\hat{t}\},T^Z}}(T^Z) \geq 0\} \cap \left[0, -\frac{1}{\lambda}ln(z)\right]$. We can rule out any $T^Z > -\frac{1}{\lambda}ln(z)$ because this must be worse than the Baseline equilibrium in the sense that $G^Z_{\check{t},T^Z}(t) \leq G^Z_{0,T^{Z'}}(t)$ where $T^{Z'} = -\frac{1}{\lambda}ln(z)$ and $IC_{G^Z_{0,T^{Z'}}}(t) \geq 0$ (and so $G^R_{G^Z_{0,T^{Z'}}}(0) > G^R_{G^Z_{\check{t},T^Z}}(0)$ if $G^R_{G^Z_{\check{t},T^Z}}$ can even be defined).

We now turn to uniqueness. Let the maximized objective be $U^c(0) = \underline{u} > u(1-\alpha)$. Because $IC_{G^{Z*}}(t) = 0$ for $t \in [\check{t}(T^Z), T^Z]$, we must also have $U^n(t) = \overline{u}$ for such t. Knowing this, we can consider a reduced problem of maximizing $G^R_{G^Z_{\check{t},T^Z(\check{t},\underline{u})}}(0)$, with respect to $\check{t}$ where $T^Z(\check{t},\kappa)$ is defined in equation (16), to ensure that $U^n(T^Z) = \kappa$. This reduced problem must have the same maximizers (i.e. implied distribution function) as the original problem.

Equation (20) in the proof of Lemma 8 defines the variable $Y(w)$ which has the same sign as $\frac{dG^R_{G^Z_{\check{t},T^Z(\check{t},\underline{u})}}(0)}{d\check{t}}$ where $w = \underline{u}e^{r\check{t}}$. Equation (19) evaluates $\frac{dY(w)}{dw}$. Its second derivative is:

$$\frac{d^2Y(w)}{dw^2} = -\frac{u(\alpha)}{u(1-\alpha)(u(\alpha)-u(1-\alpha))}\left(\frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}} + \frac{(u(\alpha)-w)^{\frac{\lambda^m-2\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}} + (u(1-\alpha)-w)\frac{\lambda^m-2\lambda}{\lambda}\frac{(u(\alpha)-w)^{\frac{\lambda^m-3\lambda}{\lambda}}}{(z(u(\alpha)-u(1-\alpha)))^{\frac{\lambda^m-\lambda}{\lambda}}}\right)$$

Evaluating at $w = \overline{u}e^{r\check{t}} \in (u(1-\alpha), u(\alpha))$ and noticing that $\lambda^m \in (\lambda, 2\lambda]$, it is clear that this expression is strictly negative as each of the bracketed terms is positive, the first two strictly. The implication is that $G^R_{G^Z_{\check{t},T^Z(\check{t},\overline{u})}}$ is strictly quasiconcave in $\check{t}$, and so has a unique maximizer $\check{t}^*$. This completes the proof of uniqueness.

To establish that $\check{t} < T^Z$, notice that $\check{t} = T^Z(\check{t}, \underline{u})$ if and only if $w = \overline{u}e^{r\check{t}} = u(\alpha) - z(u(\alpha) - u(1-\alpha))$, but in this case $Y(w) = -(1-z) < 0$ and so decreasing $\check{t}$ slightly (which is certainly possible given $T^Z > 0$) would strictly increase $G^R_{G^Z_{\check{t},T^Z(\check{t},\underline{u})}}(0)$.

Turning to the claim that $\check{t} > 0$ for sufficiently small z or sufficiently large z, recall that Proposition 6 establishes $\lim_n G^{R*}(0) = 1$ if $z_n \to 0$ or $\alpha_n \to 1$, and so $\lim_n \overline{u} \to u(0.5)$ when $z_n \to 0$ and $\lim_n \overline{u} \to (1-z)u(0.5)$ when $\alpha_n \to 0$. Notice that:

$$\lim_{z_n \to 0} z_n^{\frac{\lambda^m-\lambda}{\lambda}} Y(w) = -\frac{\lambda u(\alpha)}{(\lambda^m-\lambda)u(1-\alpha)}\frac{u(1-\alpha)-\lim_n w}{(u(\alpha)-u(1-\alpha))}\left(\frac{u(\alpha)-\lim_n w}{u(\alpha)-u(1-\alpha)}\right)^{\frac{\lambda^m-\lambda}{\lambda}}.$$

This must either equal zero, or it must be negative while $\lim_n w = \lim_n \overline{u} = u(0.5)$ (because in an OSSMP, we either need $Y(w) = 0$ or $Y(w) \leq 0$ with $w = \overline{u}$). This clearly implies $\lim_n w = u(0.5)e^{r\lim_n \check{t}} = u(\alpha)$ (taking a subsequence if necessary to ensure convergence, and noting that $\lim_n w \geq u(0.5) > u(1-\alpha)$) and so $\lim_n \check{t} = \frac{1}{r}ln\left(\frac{u(\alpha)}{u(0.5)}\right) > 0$, ensuring $\check{t} > 0$ for all sufficiently small z.

Similarly, notice that:

$$\lim_{\alpha_n \to 0} Y(w)u(1-\alpha_n) = -\frac{u(0.5)u(1)}{(u(1)-u(0.5))}\left(\frac{-\lim_n w}{u(1)}\right)\left(\left(\frac{u(1)-\lim_n w}{z(u(1))}\right)^{\frac{u(1)-u(0.5)}{u(0.5)}} - 1\right) \tag{20}$$

Again, this must either equal zero, or it must be negative while $\lim_n w = \lim_n \overline{u} = u(0.5)$ This clearly implies $\lim_n w = (1-z)u(0.5)e^{r\lim_n \check{t}} = (1-z)u(1)$ and so again $\lim_n \check{t} = \frac{1}{r}ln\left(\frac{u(1)}{u(0.5)}\right) > 0$ ensuring $\check{t} > 0$ for all sufficiently large α. This establishes the existence of some $\underline{z}(u,\alpha) > 0$ and $\underline{\alpha}(u,z) < 1$ such that if $z < \underline{z}(u,\alpha)$ or $\alpha > \underline{\alpha}(u,z)$ then $\check{t} > 0$ in the OSSMP.

For risk neutral agents, Lemma 7 establishes that the distribution G^Z_{0,T^Z} can only satisfy both incentive constraints for risk neutral agents when it matches the Baseline equilibrium distribution (i.e. $T^Z = -\frac{1}{\lambda}ln(z)$), and so we must have $\check{t} > 0$ in the OSSMP when $z < \alpha$. For these risk neutral agents with $z < \alpha$, we finally turn to the claim that $\overline{z}^c(\check{t}) = \overline{z}^n(\check{t}) = \alpha$. We know from Lemma 9 that we must have

$$\underline{u} = U^c(T^Z) = U^n(\check{t}) = e^{-r\check{t}}\left((1-z)G^Z_{\check{t},T^Z}(\check{t})(0.5-(1-\alpha)) + (1-\alpha)\right) = e^{-r\check{t}}\left(\alpha - z(2\alpha-1)e^{\lambda(T^Z-\check{t})}\right)$$

Letting $\bar{w} = e^{\lambda(T^Z-\check{t})}$ we, therefore, have $w = e^{r\check{t}}\underline{u} = \alpha - z(2\alpha-1)\bar{w}$. Plugging this into $Y(w)$ we, get an equivalent function $\bar{Y}$ in terms of $\bar{w}$:

$$Y(\alpha - z(2\alpha-1)\bar{w}) = \bar{Y}(\bar{w}) = -1 + z\bar{w} - \frac{\alpha}{1-\alpha}(z\bar{w}-1)(\bar{w}-1)$$

The function $\bar{Y}$ is readily seen to be strictly concave in $\bar{w}$ (its quadratic). Moreover, we have $\bar{Y}(1/z) = \bar{Y}(1/\alpha) = 0$. Recalling that $Y(w) = \bar{Y}(\bar{w})$ has the same sign as $\frac{dG^R_{G^Z_{\check{t},T^Z}(\check{t},\underline{u})}(0)}{d\check{t}}$ we would be able to increase $G^R_{G^Z_{\check{t},T^Z}(\check{t},\underline{u})}(0)$ unless in the OSSMP $\check{t}$ and T^Z imply $\bar{w} = e^{\lambda(T^Z-\check{t})} = 1/\alpha$ (by adjusting $\check{t}$ and so $T^Z(\check{t},\underline{u})$ given our fixed $\underline{u}$), a contradiction. Finally, notice that for at $t \in [\check{t}, T^Z]$ we must have:

$$\bar{z}^c(t) = \frac{z(1-G^Z_{\check{t},T^Z}(t))}{z(1-G^Z_{\check{t},T^Z}(t)) + (1-z)(1-G^R_{G^Z_{\check{t},T^Z}}(t))} = e^{-\lambda(T^Z-t)} = \frac{z}{z+(1-z)(1-G^Z_{\check{t},T^Z}(t))} = \bar{z}^n(t)$$

where we have just established that $e^{-\lambda(T^Z-\check{t})} = \alpha$ in the OSSMP. □

This completes the proof of Theorem 2. □

Proof of Proposition 6

Notice that payoffs in an OSSMP and under a strongly symmetric ND protocol are of the form $G^R(0)(1-z)(u(0.5)-u(1-\alpha))+u(1-\alpha)$ where $G^R(0) = b$ in the latter. Hence, in an OSSMP we must have $G^R_{G^{Z*}}(0) \geq b$, for any b which is part of an ND equilibrium. Let $b \in (0,1)$ arbitrary, then the condition for an ND equilibrium to exist as highlighted by the proof of Proposition 3 is that

$$Q = u(0.5) - u(1-\alpha) - \frac{z\left(1-\left(\frac{z}{1-(1-z)b}\right)^{\frac{r}{\bar{\lambda}}}\right)u(1-\alpha)}{(1-z)(1-b)} \geq 0$$

Suppose first that $B^n = (\alpha, z^n, u, r)$ with $\lim_n z^n = 0$. It is clear that the final expression vanishes so that $\lim_n Q = u(0.5) - u(1-\alpha) > 0$. Suppose next that $\check{B}^n = (\alpha^n, z, u, r)$ with $\lim \alpha^n = 1$. In this case $\lim u(1-\alpha^n) = 0$ and $\lim \frac{r}{\bar{\lambda}} = \infty$ so that $\lim_n Q = u(0.5) > 0$.

Finally, suppose that $\hat{B}^n = (\alpha, z, u^n, r)$ with $\lim_n u^n(\alpha) = \lim_n u^n(0.5) > \lim_n u^n(1-\alpha)$, and without loss of generality normalize $u^n(1-\alpha) = \underline{u}(1-\alpha) > 0$ for all n. Evaluating the limit of Q and rescaling gives

$$\hat{Q}(b) = \lim_n \frac{Q(1-b)(1-z)}{\underline{u}(1-\alpha)} = \frac{r}{\bar{\lambda}}(1-b)(1-z) - z\left(1-\left(\frac{z}{1-(1-z)b}\right)^{\frac{r}{\bar{\lambda}}}\right)$$

where $\frac{r}{\bar{\lambda}} = \frac{\lim_n u^n(\alpha) - \underline{u}(1-\alpha)}{\underline{u}(1-\alpha)}$. Moreover,

$$\frac{d\hat{Q}(b)}{db} = -\frac{r}{\bar{\lambda}}(1-z) + \frac{r}{\bar{\lambda}}(1-z)z^{\frac{r+\bar{\lambda}}{\bar{\lambda}}}(1-(1-z)b)^{-\frac{r+\bar{\lambda}}{\bar{\lambda}}}$$
$$\frac{d^2\hat{Q}(b)}{db^2} = \frac{r}{\bar{\lambda}}\frac{r+\bar{\lambda}}{\bar{\lambda}}(1-z)^2 z^{\frac{r+\bar{\lambda}}{\bar{\lambda}}}(1-(1-z)b)^{-\frac{r+2\bar{\lambda}}{\bar{\lambda}}} > 0$$

Evaluating at $b = 1$ we get $\hat{Q}(1) = 0$ and $\frac{d\hat{Q}(b)}{db}\Big|_{b=1} = 0$, so that $\hat{Q}(b) > 0$ for any $b < 1$. □

Proof of Proposition 5

Suppose there is some optimal mediation protocol which is not symmetric, $(G^R, G^Z_1, G^Z_2, M_1, M_2)$. The proof of Lemma 2 highlighted that the strongly symmetric mediation protocol $(G^R, \check{G}^Z)$ with $\check{G}^Z = 0.5(G^Z_1 + G^Z_2)$ implied an equilibrium with utilities $\check{U}^c(t) \geq 0.5(U^c_1(t) + U^c_2(t))$ and $\check{U}^n(t) = 0.5(U^n_1(t) + U^n_2(t))$. Clearly, if this is not the OSSMP, then the non-symmetric protocol is not optimal. Suppose that it is an OSSMP, then we must have $\check{U}^c(T^R) = \check{U}^c(t)$ for $t \leq T^R = T^Z$, $\check{U}^c(T^R) = \check{U}^n(t)$ for $t \in [t^*, T^R]$, and $\check{G}^Z(t) = 0$ for $t < t^*$. This immediately implies $G^Z_i(t) = 0$ for $t < t^*$ for $i = 1, 2$.

Because the non-symmetric protocol is an equilibrium, we must have $U^c_i(T^R) \geq U^c_i(t)$ and $U^c_i(T^R) \geq U^n_i(t)$ for

all t. Suppose that $U_i^c(T^R) > U_i^n(t)$ for some $i \in \{1,2\}$ and some $t \in [t^*, T^R]$, then clearly $\check{U}^c(T^R) > U^n(t)$, a contradiction. However, if we have $U_i^c(T^R) = U_i^n(t)$ for $t \in [t^*, T^R]$ then on this interval we must have $G_i^Z(t) = \frac{1-ze^{\lambda(T^R-t)}}{1-z} = \check{G}^Z$ and so the original protocol must in fact be symmetric. Finally, notice that when $u(0.5) < \frac{u(1-x)+u(x)}{2}$ for $x < 0.5$, then $u(0.5) < \int u(m)d\check{M}^t(m)$ for any symmetric $\check{M}^t$ where $\check{M}^t(0.5) < 1$. Hence, we must have $\int \mathbb{1}_{[t:\check{M}^t(0.5)=1]} dG^R(t) = 1$. □

For online publication: Appendix C

Ongoing Dunlop (*OD*) mediation

Below, I extend the Simple Dunlop mediation protocol by allowing the mediator to respond to agents who confess rationality continuously over the infinite horizon, in what I call the *Ongoing Dunlop (OD)* mediation protocol.

In the *OD* protocol, if agent i confesses at time t^2 and agent j confesses at $s^2 \geq t^2$, then at s^3 the mediator suggests the agreement (m_1, m_2). For tractability reasons, I focus on what I call *OD equilibria* in which rational agents follow the mediator's suggestion by changing their demands to (m_1, m_2) at s^4. Focussing on such equilibria entails some loss of generality because this imposes the same continuation payoffs for an agent after a mediator announcement, regardless of the time at which she confessed and whether she confessed before or after her opponent. Given this assumption, however, it is then without loss of generality to assume $m_i \in (1-\alpha_j, \alpha_i)$, because if $m_i \geq \alpha_i$ then rational i would confess with probability one at 0^2 if this had any chance of affecting the outcome.

In an *OD* equilibrium, agent i's strategy reduces to choosing a time to confess and a time to concede (to her opponent's behavioral demand). It is without loss of generality to assume that an agent never concedes at t^1 but only at t^5, and confesses before she concedes (because doing so strictly increases her payoff whenever it affects the game's outcome). We can again, therefore, analyze the game in continuous time. Agent i's strategy is described by two cumulative distribution functions, $F_i^c \in [0,1]^{[-\infty,\infty]}$ and $F_i^d \in [0,1]^{[0,\infty]}$. Let $F_i^c(t)$ be the total probability that agent i has confessed before time t, and $F_i^d(t)$ be the total probability that agent i has conceded before time t (c=confessed, d=defeated) where $F_i^c(t) \geq F_i^d(t)$. Given j's equilibrium strategy, rational strategy, rational agent i's expected utility from confessing at time s and conceding at time $t \geq s$ is:

$$\begin{aligned} U_i(s,t) = &\int_{v<s} e^{-r_i v} u_i(\alpha_i) dF_j^d(v) + \int_{v\in(s,t]} e^{-r_i v} u_i(m_i) dF_j^c(v) \\ &+ (1-F_j^c(t)) e^{-r_i t} u_i(1-\alpha_j) + (F_j^c(s) - \sup_{v<s} F_j^d(v)) e^{-r_i s} u_i(m_i) \end{aligned}$$

Of course, the Baseline equilibrium is still an *OD* equilibrium, where $F_i^c(t) = F_i^d(t)$ for all t. The next proposition establishes that this is the only *OD* equilibrium.

Proposition 8. *The distribution of outcomes in any OD equilibrium is identical to that in the unique Baseline equilibrium.*

The idea of the proof is similar to that of Proposition 2 in that unless behavior matches the Baseline equilibrium with $F_i^c(t) = F_i^d(t)$, then indifference conditions for confessing and non-confessing agents imply a contradiction to the fact that rational agents must concede within finite time. However, it is somewhat more involved.

Proof of Proposition 8. Suppose there is an equilibrium $\sigma = (\sigma_1, \sigma_2)$. In this setup, I refer to agent j who has confessed but not yet conceded, as a confessing agent. Let $A_i = \{(s,t) : U_i(s,t) = \max_{v,w} U_i(v,w)\}$. Since σ is an equilibrium, $A_i \neq \emptyset$. Finally, define $T_i^d = \inf\{t : F_i^d(t) = 1 - z_i\}$ and $T^* = \max\{T_1^d, T_2^d\}$.

(a) *We must have* $T_i^d = T^* < \infty$. This follows for the reasons as outlined in the proof of Proposition 2, claim (a). We must have $T_i^d = T_j^d$, because if a rational agent knows she faces a behavioral opponent she will concede immediately. We must have $T_j^d < \infty$ because if a rational agent j does not concede at some t to get $u_j(1-\alpha_i) > 0$, she must expect her opponent to stop acting like a behavioral type soon and therefore must eventually become convinced that her opponent is behavioral.

(b) *If* F_i^d *jumps at* $t \in [0, T^*]$ *then* F_j^c *is constant on* $[t-\varepsilon, t]$ *for some* $\varepsilon > 0$. This follows because if agent j has not confessed before $t-\varepsilon$, she would strictly increase her payoff by confessing an instant after t compared

to slightly before as this would give her $u_j(\alpha_j)$ rather than $u_j(m_j)$ with positive probability (at least $F_i^d(t) - \sup_s F_i^d(s) > 0$).

(c) *If F_i^c jumps at $t \in (0, T^*]$ then F_j^d is constant on $[t-\varepsilon, t)$ for some $\varepsilon > 0$.* This follows because agent j would prefer to concede an instant after t rather than slightly before as this would give her $u_j(m_j)$ rather than $u_j(1-\alpha_i)$ with positive probability (at least $F_i^c(t) - \sup_{s<t} F_i^c(t) > 0$).

(d) *Let $t' \leq t'' < t''' \leq T^*$. If $F_i^c(t''') = F_i^c(t')$ and $F_j^c(t'') > F_j^d(t'')$ then $F_j^d(t'') = F_j^d(t''')$.* If this is not true so that $F_j^d(t'') < F_j^d(t''')$, then there must exist some $s \leq t''$ and some $t \in (t'', t''']$ such that $(s,t) \in A_j$. However, given that $F_i^c(t'') = F_i^c(t''')$ the alternative strategy of conceding slightly earlier (e.g. at $\frac{1}{2}(t''+t)$) while still confessing at s is strictly more profitable as it moves the concession payoff $u_j(1-\alpha_i)$ forward in time (with probability greater than $z_i > 0$).

(e) *Let $t' < t''' \leq T^*$. If $F_i^c(t''') = F_i^c(t')$ then either $F_j^d(t') = F_j^d(t''')$ or for all $t \in [t', t''')$ we have $F_j^c(t) = F_j^d(t)$.* Suppose not, then for some $t'' \in [t', t''')$ we have $F_j^d(t'') < F_j^c(t'')$ and $F_j^d(t') < F_j^d(t''')$. Define $\check{t}_i = \sup\{t : F_i^c(t) = F_i^c(t')\}$. By claim (d) we have $F_j^d(t'') = \sup_{s<\check{t}_i} F_j^d(s)$ and $F_j^c(t'') > F_j^d(t'')$. This implies that F_i^c must be continuous at $\check{t}_i$, i.e. $F_i^c(\check{t}_i) = \sup_{s<\check{t}_i} F_i^c(s)$. To see this, notice that confessing at $\check{t}_i$ and conceding at some later date t must give i a strictly lower payoff than confessing slightly earlier (e.g. at $\frac{1}{2}(\check{t}_i + t'')$ and still conceding at t (with probability $F_j^c(t'') - F_j^d(t'') > 0$ she receives the payoff $u_i(m_i)$ earlier). By claim (d), therefore, we must have $F_j^d(t''') = F_j^d(\check{t}_i)$. But in that case any strategy in which agent i confesses an instant after $\check{t}_i$ cannot be optimal either, contradicting the definition of the supremum $\check{t}_i$.

(f) *Let $T^* \geq t'' > t'$. If $F_i^d(t'') = F_i^d(t')$ and $F_i^c(t') > F_i^d(t')$ then $F_j^c(t') = F_j^c(t'')$* Suppose not so that $F_j^c(t') < F_j^c(t'')$. Then there exists $(s,t) \in A_j$ such that $s \in (t', t'']$. However, given $F_i^d(t'') = F_i^d(t')$, the alternative plan of confessing slightly earlier (e.g. at $\hat{s} = \frac{1}{2}(t'+s)$) while still conceding at t would be strictly better for j as this gives her the payoff $u_j(m_j)$ with positive probability at an earlier time (at least $(F_i^c(t') - F_i^d(t')) > 0$).

(g) *There is no jump in F_i^d at $t \in (0, T^*]$.* Suppose not, then by claim (b) F_j^c is constant on $[t-\varepsilon, t]$ for some $\varepsilon > 0$. Hence, by claim (e) either $F_i^d(t) = F_i^d(t-\varepsilon)$ (a direct contradiction) or $F_i^c(s) = F_i^d(s)$ for $s \in [t-\varepsilon, t)$. It must then be that F_i^c also jumps at t, because we must have $\sup_{s<t} F_i^c(s) = \sup_{s<t} F_i^d(s) < F_i^d(t) \leq F_i^c(t)$. Hence by claim (c), F_j^d is constant on $[t-\varepsilon, t)$ for some $\varepsilon > 0$ (assume the same ε without loss of generality). Given that F_i^c and F_i^d jump at t, we must have $(t,t) \in A_i$. However, the alternative strategy for i of both confessing and immediately conceding slightly earlier (e.g. at $t - \frac{\varepsilon}{2}$) delivers strictly higher expected profits as she gets the payoffs $(F_j^c(t-\varepsilon) - F_j^d(t-\varepsilon))u_i(m_i)$ and $(1 - F_j^c(t-\varepsilon))u_i(1-\alpha_j) > 0$ at an earlier date, without affecting other payoffs.

(h) *If F_i^d is continuous at $s \leq t$ then $U_i(s,t)$ is continuous at s, and if F_i^c is continuous at t then $U_i(s,t)$ is continuous at t.* This follows from how $U_i(s,t)$ is defined.

For claims (i)-(m) suppose that $F_1^c(t') > F_1^d(t')$ for some $t' \in [0, \infty)$ (symmetric arguments apply if $F_2^c(t') > F_2^d(t')$). Define $\bar{t}_1 = \inf\{t \geq t' : F_1^c(t) = F_1^d(t)\}$ and $\underline{t}_1 = \inf\{t : F_1^c(s) > F_1^d(s)\ \forall s \in [t, t']\}$. Notice that by claim (g), the continuity of F_1^d, we have $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. Also note that $\bar{t}_1 > t' \geq \underline{t}_1$ and $F_1^c(t) > F_1^d(t)$ for all $t \in (\underline{t}_1, \bar{t}_1)$. Let $\bar{t}_1 \geq t''' > t'' > \underline{t}_1$.

(i) *We must have $F_2^c(t''') > F_2^c(t'')$.* Suppose not, and so let $\check{t}_2 = \sup\{t : F_2^c(t) = F_2^c(t'')\} \geq t'''$. I first establish the subclaim (i') that this must imply either $F_1^d(t'') = F_1^d(\check{t}_2)$ or $F_1^c(t) = F_1^d(t)$ for $t \in [t'', \check{t}_2)$. Suppose not (again), then $F_1^d(t'') < F_1^d(\check{t}_2)$ and there is some $t \in [t'', \check{t}_2)$ such that $F_1^d(t) < F_1^c(t)$. By claim (g), the continuity of F_1^d, we must have $F_1^d(t'') < F_1^d(\check{t}_2 - \varepsilon)$ for all $\varepsilon > 0$ sufficiently small. Choose such an appropriately small $\varepsilon < \check{t}_2 - t$, then we have $F_2^c(t'') = F_2^c(\check{t}_2 - \varepsilon)$, $F_1^d(t) < F_1^c(t)$ for some $t \in [t'', \check{t}_2 - \varepsilon)$ and $F_1^d(t'') < F_1^d(\check{t}_2 - \varepsilon)$, which contradicts claim (e).
By assumption we have $F_1^c(t) > F_1^d(t)$ for all $t \in [t'', \bar{t}_1)$ so that subclaim (i') in fact implies $F_1^d(t'') = F_1^d(\check{t}_2)$.

This in turn ensures $\bar{t}_1 > \check{t}_2$ because $F_1^d(\check{t}_2) = F_1^d(t'') < F_1^c(t'') \leq F_1^c(\check{t}_2)$ whereas $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. I next claim that it can't be optimal for agent 2 to confess at $\check{t}_2$ while conceding at some $t \geq \check{t}_2$. To see this, notice that agent 2 would do strictly better confessing slightly earlier (e.g. at $\frac{1}{2}(\check{t}_2 + t'')$) while still conceding at t as this would bring forward the payoff $u_2(m_2)$ with positive probability (at least $F_1^c(t'') - F_1^d(t'') > 0$), without affecting other payoffs. Given claim (g), the continuity of F_1^d, this argument similarly also implies that confessing an instant after $\check{t}_2$ is strictly worse than confessing at $\frac{1}{2}(\check{t}_2 + t'')$. This contradicts the definition of the supremum $\check{t}_2$.

(j) *We must have* $F_1^d(t''') > F_1^d(t'')$. Suppose not, then let $\check{t}_1 = \sup\{t : F_1^d(t) = F_1^d(t'')\} \geq t'''$. Given claim (g), the continuity of $F - i^d$, we have $F_1^d(\check{t}_1) = F_1^d(t'')$. Given $F_1^d(\check{t}_1) = F_1^d(t'') < F_1^c(t'') \leq F_1^c(\check{t}_1)$ we must have $\check{t}_1 < \bar{t}_1$. By claim (f) we must then have $F_2^c(\check{t}_1) = F_2^c(t'')$ which contradicts claim (i), that F_2^c is increasing on $(\underline{t}_1, \bar{t}_1]$.

(k) *We must have* $F_2^d(t''') > F_2^d(t'')$. Suppose not so that $F_2^d(t''') = F_2^d(t'')$. Given that F_2^c is increasing on the interval $[t'', t''']$ by claim (i), we must have $F_2^c(t) > F_2^d(t)$ for $t \in (t'', t''']$. Define $\bar{t}_2 = \inf\{t \geq t''' : F_2^c(t) = F_2^d(t)\}$ and $\underline{t}_2 = \inf\{t : F_2^c(s) > F_2^d(s)\ \forall s \in [t, t''']\}$, then switching the labelling for 1 and 2, claim (i) implies $F_1^c(t''') > F_1^c(t'')$ and claim (j) implies $F_2^d(t''') > F_2^d(t'')$, a contradiction.

(l) *We must have* $F_1^c(t''') > F_1^c(t'')$. Suppose not, and so $F_1^c(t''') = F_1^c(t'')$. Let $\check{t}_1 = \inf\{t : F_1^c(t) = F_1^c(t'')\}$. The right continuity of F_1^c ensures that $F_1^c(\check{t}_1) = F_1^c(t'')$. Clearly, we have $\check{t}_1 \geq \underline{t}_1$ (if $\check{t}_1 < \underline{t}_1$ then certainly at some $t \in (\check{t}_1, t'']$ we must have $F_1^c(t) = F_1^d(t) = F_1^c(t'') \geq F_1^d(t'') \geq F_1^d(t)$, which contradicts $F_1^c(t'') > F_1^d(t'')$). By claim (e), we then have either $F_2^d(t''') = F_2^d(\check{t}_1)$, which contradicts claim (k), or $F_2^c(t) = F_2^d(t)$ for all $t \in [\check{t}_1, t''')$. Notice that because F_1^d is strictly increasing on $[\check{t}_1, t''')$ by claim (j) while F_1^c is by assumption constant, for some $s \leq \check{t}_1$ and some $t \in (\check{t}_1, t''')$ we must have $(s, t) \in A_1$. Furthermore, if $(s', t') \in A_1$ where $s' \in [s, \check{t}_1]$ then $(s', t) \in A_1$. This is simply because at time s' an agent who confessed at s and another who previously confessed at s' have the same incentives to concede thereafter. I claim, however, that $(\check{t}_1, t) \notin A_1$. To see this, notice that such a strategy is strictly worse than both confessing and conceding at t, which gives agent 1 the higher payoff of $u_1(\alpha_1)$ instead of $u_1(m_1)$ from the positive concession of agent 2 on the interval $[\check{t}_1, t)$. That is:

$$\begin{aligned} U_1(t,t) - U_1(\check{t}_1, t) &\geq \int_{\check{t}_1 \leq v \leq t} (u_1(\alpha_1) - u_1(m_1)) e^{-r_1 v} dF_2^c(v) \\ &\geq e^{-r_1 t}(u_1(\alpha_1) - u_1(m_1))(\sup_{v<t} F_2^c(v) - F_2^c(\check{t}_1)) > 0 \end{aligned}$$

where the first inequality follows from $F_2^c(t) = F_2^d(t)$ on $[\check{t}_1, t''')$, the second from $t \geq v \in [\check{t}_1, t]$ and the third from claim (i). For the same reason, confessing an instant before $\check{t}_1$ and conceding at t cannot be optimal either. This either contradicts the definition of $\check{t}_1$ as an infimum or implies $\check{t}_1 = 0$ and $F_2^c(0) = 0$. The latter possibility, however, clearly contradicts $F_1^c(v) > F_1^d(v)$ for all $v \in (\check{t}_1, t''')$.

(m) F_i^c *is continuous on* $(\underline{t}_1, \bar{t}_1]$. If F_i^c did jump at $t \in (\underline{t}_1, \bar{t}_1]$ then by (c), F_j^d is constant on $(t - \varepsilon, t)$ for some $\varepsilon > 0$, contradicting either claim (j) or (k).

We are almost done. Because F_1^c, F_1^d are increasing on $(\underline{t}_1, \bar{t}_1)$, established in claims (j) and (l), while by assumption $F_1^d(t) < F_1^c(t)$ on this interval, it follows that there is some $s' \in (\underline{t}_1, \bar{t}_1)$ such that A_1 is dense in the set $\{(s', t) : t \in [s', \bar{t}_1]\}$. Notice that regardless of whether agent 1 confesses at s' or $s \in (s', \bar{t}_1)$, she faces the same incentives to concede after s if she has not already done so. Notice also, that there is always a positive probability that agent 1 has confessed before s but has not conceded. From the continuity of F_2^c on $(\underline{t}_1, \bar{t}_1]$ it follows that $U_1(s', t)$ is constant on $[s', \bar{t}_1]$, and hence differentiable with respect to t with zero partial derivative, $\frac{\partial U_1(s',t)}{\partial t} = 0$. This implies:

$$\frac{f_2^c(t)}{1 - F_2^c(t)} = \lambda_2^c = \frac{r_1 u_1(1 - \alpha_2)}{u_1(m_1) - u_1(1 - \alpha_2)}$$

for $t \in [\underline{t}_1, \bar{t}_1]$. Solving this linear ODE gives $(1 - F_2^c(s)) = (1 - F_j^c(\underline{t}_1)) e^{-\lambda_2^c (s - \underline{t}_1)}$.

By the same reasoning there must be some $s'' \in (\underline{t}_1, \bar{t}_1)$ such that A_1 is dense in the set $\{(s, s'') : s, \in [\underline{t}_1, s'']\}$. The continuity of F_2^d on $(\underline{t}_1, \bar{t}_1]$ then implies that $U_1(s, s'')$ is constant on $(\underline{t}_1, s'']$, and hence differentiable with respect to s with zero partial derivative, $\frac{\partial U_1(s,s'')}{\partial s} = 0$. Rearranging this zero derivative condition gives:

$$\frac{f_2^d(s)}{F_2^c(s) - F_2^d(s)} = \lambda_2^d = \frac{r_1 u_1(m_1)}{u_1(\alpha_1) - u_1(m_1)}$$

This should already suggest a problem. When $F_2^c(s) - F_2^d(s)$ becomes arbitrarily small $f_2^d(s)$ must be similarly small. However, $f_2^c(t) \geq \lambda_2^c(1 - F_2^c(t)) \geq \lambda_2^c z_2$ is bounded above zero, implying $F_2^c(t) - F_2^d(t) > 0$ on $(\underline{t}_1, \bar{t}_1]$. To be more precise, the above linear ODE is solved to give:

$$(1 - F_2^d(s)) = \begin{cases} \phi_2^d e^{-\lambda_2^d(s-\underline{t}_1)} + \phi_2^c \psi_2 (e^{-\lambda_2^c(s-\underline{t}_1)} - e^{-\lambda_2^d(s-\underline{t}_1)}) \text{ if } \lambda_2^d \neq \lambda_2^c \\ (\phi_2^d + \lambda_2^d \phi_2^c (s - \underline{t}_1)) e^{-\lambda_2^d(s-\underline{t}_1)} \text{ if } \lambda_2^d = \lambda_2^c \end{cases}$$

where, $\psi_2 = \frac{\lambda_2^d}{\lambda_2^d - \lambda_2^c}$ and $\phi_2^d = (1 - F_2^d(\underline{t}_1)) \geq (1 - F_2^c(\underline{t}_1)) = \phi_2^c$. Define the gap between F_2^c and F_2^d as $d_2(s) = F_2^c(s) - F_2^d(s)$, and consider the following transformations of this gap:

$$\begin{aligned} d_2(s) \frac{e^{\lambda_2^d(s-\underline{t}_1)}}{\psi_2 - 1} &= \frac{\phi_2^d - \psi_2 \phi_2^c}{\psi_2 - 1} + e^{(\lambda_2^d - \lambda_2^c)(s-\underline{t}_1)} \quad \text{if} \quad \lambda_2^d > \lambda_2^c \\ d_2(s) \frac{e^{\lambda_2^c(s-\underline{t}_1)}}{\phi_2^d - \psi_2 \phi_2^c} &= e^{(\lambda_2^c - \lambda_2^d)(s-\underline{t}_1)} + \frac{\psi_2 - 1}{\phi_2^d - \psi_2 \phi_2^c} \quad \text{if} \quad \lambda_2^d < \lambda_2^c \\ d_2(s) e^{\lambda_2^d(s-\underline{t}_1)} &= \phi_2^d + \lambda_2^d \phi_2^c (s - \underline{t}_1) - \phi_2^c \quad \text{if} \quad \lambda_2^d = \lambda_2^c \end{aligned}$$

I claim that each of these transformations is positive. Notice that $\psi_2 - 1 = \frac{\lambda_2^c}{\lambda_2^d - \lambda_2^c} > 0$ when $\lambda_2^d > \lambda_2^c$. Similarly $\phi_2^d - \psi_2 \phi_2^c \geq -\phi_2^c \frac{\lambda_2^c}{\lambda_2^d - \lambda_2^c} > 0$ when $\lambda_2^d < \lambda_2^c$, where the first inequality follows from $\phi_2^d \geq \phi_2^c$. Each of the transformed gaps is strictly increasing in s, implying that $d_2(s) > 0$ for $s \in (\underline{t}_1, \bar{t}_1]$. Recall that we must have $\bar{t}_1 \leq T^* < \infty$, and $F_1^c(\bar{t}_1) = F_1^d(\bar{t}_1)$. Now define $\bar{t}_2 = \inf\{t > \underline{t}_1 : F_2^c(t) = F_2^d(t)\} \leq T^* < \infty$, where this is consistent with the definition of $\bar{t}_2$ in the proof of claim (k). We can now repeat the above arguments with the roles of agent 1 and 2 reversed to find that $d_1(s) > 0$ for $s \in (\underline{t}_1, \bar{t}_2]$ and $F_2^c(\bar{t}_2) = F_2^d(\bar{t}_2)$. Let $\bar{t} = \min\{\bar{t}_1, \bar{t}_2\}$. For some i we must have $\bar{t} = \bar{t}_i$, but that implies both $F_i^c(\bar{t}_i) = F_i^d(\bar{t}_i)$ and $d_i(\bar{t}_i) = F_i^c(\bar{t}_i) - F_i^d(\bar{t}_i) > 0$, a contradiction. We must, therefore, have $F_i^c(t) = F_i^d(t)$ for $t \in [0, \infty)$. Given this, the unique equilibrium must match that of the Baseline model by standard arguments (see AG). □

Bad faith equilibrium

As discussed in the main text, the restriction to good faith equilibria is with some loss of generality. I show, below, that when the probability of behavioral types is sufficiently small, the bargaining problem is symmetric, and agents are risk neutral, a bad faith equilibrium exists which delivers higher payoffs to rational agents than the OSSMP (optimal good faith equilibrium).

The equilibrium takes the following form: Rational agents make a behavioral demand α at 0^1 and (conditional on both doing so) confess rationality to the mediator at time 0^2. If both agents confess rationality then at time 0^3, with probability 0.5 the mediator (publicly) tells one agent i to demand $\alpha_i(0^4) = 1 > \alpha$ at 0^4 (revealing i's rationality), and otherwise tells agent $j \neq i$ to demand $\alpha_j(0^4) = 1$. In the former (latter) case, the mediator suggests that agent i (j) gets the whole dollar in any subsequent agreement. If only agent i confesses rationality, then the mediator always tells her to demand $\alpha_i(0^4) = 1$ at 0^4. In this case agent i obtains the share $(1 - \alpha)$ in all subsequent agreements. If neither agent confesses, then the mediator says nothing. If an agent fails to demand α at time 0^1

or fails to follow the mediator's suggestion at some time, then the mediator subsequently says nothing and the continuation equilibrium specifies that she should concede to her opponent immediately.

Conditional on agent i (upon instruction) demanding $\alpha_i(0^4) = 1$ at 0^4, let the cumulative distributions of her agreement times with a rational and behavioral opponent be $G^{R,o}$ and $G^{Z,o}$ respectively. Let $T^{R,o} = \min_t\{t : G^{R,0}(t) = 1\}$ and $T^{Z,o} = \min_t\{t : G^{Z,0}(t) = 1\}$. The conditional probability that agent i faces a behavioral opponent after her instruction is $\bar{z} = \frac{2z}{1+z}$. If she subsequently concedes at t^5 (without being instructed to by the mediator), then she obtains the expected utility:

$$\begin{aligned}U^{c,o}(t) =&(1-\bar{z})\int_{s\leq t} e^{-rs}u(1)dG^{R,o}(s) + \bar{z}\int_{s\leq t} e^{-rs}u(1-\alpha)dG^{Z,o}(s)\\ &+ e^{-rt}u(1-\alpha)\Big((1-\bar{z})(1-G^{R,o}(t)) + \bar{z}(1-G^{Z,o}(t))\Big)\end{aligned}$$

which leads to a new dynamic incentive constraint

$$U^{c,o}(T^{R,o}) = \max_t U^{c,o}(t).$$

An agent j who confessed rationality but is not told to change her demand, can obtain $u(0) = 0$ from conceding and also gets a continuation payoff of zero, and so is indifferent to subsequently following the mediator's instructions.

If rational agent i does not confess and the mediator announces nothing at 0^3, then the agent realizes she must face a behavioral opponent and subsequently immediately concedes. If the mediator instead tells her opponent at 0^3 to demand $\alpha_j(0^4) = 1$, then because i obtains at most $u_i(0) = 0$ from conceding or revealing rationality, her continuation payoff is $\int_{t\leq T^R} e^{-rs}u(\alpha)dG^{Z,o}(s)$. In sum, her expected payoff to not confessing is:

$$U^{n,o} = zu(1-\alpha) + (1-z)\int_{s\leq T^{R,o}} e^{-rs}u(\alpha)dG^{Z,o}(s)$$

The new type incentive constraint is then simply:

$$\frac{(1+z)U^{c,o}(T^{R,o})}{2} \geq U^{n,o}.$$

Consider the distributions $G^{Z,o}_*(t) = 0$ for $t < T^{R,o} = T^{Z,o}$ and $1 - G^{R,o}_*(t) = \frac{\bar{z}}{1-\bar{z}}(e^{\bar{\lambda}(T^{R,o}-t)} - 1)$ where $\bar{\lambda} = \frac{ru(1-\alpha)}{u(1)-u(1-\alpha)}$. These ensure that the dynamic incentive constraint binds in the sense that $U^{c,o}(T^{R,o}) = U^{c,o}(t)$ for $t \leq T^{R,o}$. We then attempt to select the minimum $T^{R,o}$ such that the type incentive constraint binds. To that end define:

$$\begin{aligned}W(T^{R,o}) =&\frac{(1+z)U^{c,o}(T^{R,o})}{2} - U^n = \frac{1-z}{2}\left(1 - \frac{2z}{1-z}(e^{\bar{\lambda}T^{R,o}} - 1)\right)(u(1) - u(1-\alpha))\\ &+ \frac{1+z}{2}u(1-\alpha) - zu(1-\alpha) - (1-z)e^{-rT^{R,o}}u(\alpha)\end{aligned}$$

and let $T^{R,o}_* = \min\{T^{R,o} : W(T^{R,o}) \geq 0\}$. Notice that $W(T^{R,o})$ is strictly concave in $T^{R,o}$. For all sufficiently large z, $T^{R,o}_*$ is not well defined (this is certainly the case for $z \geq \alpha$ for risk neutral agents). However, it is well defined for sufficiently small z because as $z \to 0$ we have $W(T^{R,o}) \to 0.5u(1) - e^{-rT^{R,o}}u(\alpha)$, which also implies that $T^{R,o}_* \to \frac{1}{r}ln\left(\frac{2u(\alpha)}{u(1)}\right)$. For all sufficiently small z, therefore, there is a bad faith equilibrium. We are interested in payoffs under this protocol as compared to payoffs in the OSSMP as $z \to 0$ when agents are risk neutral. The difference between these payoffs is:

$$\begin{aligned}U^{c,o}(T^{R,o}) - U^c(T^R) =&\frac{1-z}{2}G^{R,o}_*(0)(u(1) - u(1-\alpha)) + \frac{1+z}{2}u(1-\alpha) - (1-z)G^R_{G^{Z*}}(0)(u(0.5) - u(1-\alpha)) - u(1-\alpha)\\ =&\frac{1-z}{2}\left(\alpha(G^{R,o}_*(0) - G^R_{G^{Z*}}(0)) - (1 - G^R_{G^{Z*}}(0))(1-\alpha)\right)\end{aligned}$$

where the second equality imposes $u(x) = x$. By rearranging it is clear that $U^{c,o}(T^{R,o}) - U^c(T^R) \geq 0$ if and only if

$$\frac{2\alpha - 1}{\alpha} \geq \frac{1 - G^{R,o}_*(0)}{1 - G^R_{G^{Z*}}(0)}.$$

The optimal distribution of rational-behavioral agreement times in the OSSMP satisfies $G^{Z*} = 0$ for $t < t^*$. This implies that $1 - G^R_{G^{Z*}}(0) \geq \frac{z}{1-z}(e^{\lambda^m t^*} - 1)$ whereas $1 - G^{R,o}_*(0) = \frac{2z}{1-z}(e^{\bar{\lambda} T^{R,o}_*} - 1)$. Therefore, if we can show that

$$\frac{2\alpha - 1}{\alpha} \geq \frac{2(e^{\bar{\lambda} T^{R,o}_*} - 1)}{e^{\lambda^m t^*} - 1}$$

then the low option equilibrium delivers higher payoffs than the OSSMP. Above we noted that as $z \to 0$ we have $T^{R,o}_* \to \frac{1}{r} ln\,(2\alpha)$. The proof of Lemma 10 likewise showed that $t^* \to \frac{1}{r} ln\,(2\alpha)$ in this case. And so the bad faith equilibrium gives higher payoffs than the OSSMP for all sufficiently small z if:

$$\frac{2\alpha - 1}{\alpha} > \lim_{z \to 0} \frac{2(e^{\bar{\lambda} T^{R,o}_*} - 1)}{e^{\lambda^m t^*} - 1} = \frac{2\left((2\alpha)^{\frac{1-\alpha}{\alpha}} - 1\right)}{(2\alpha)^{\frac{2-2\alpha}{2\alpha-1}} - 1}.$$

To show that this holds for $\alpha \approx 0.5$, define:

$$V(\alpha) = (2\alpha - 1)\left((2\alpha)^{\frac{2-2\alpha}{2\alpha-1}} - 1\right) - 2\alpha\left((2\alpha)^{\frac{1-\alpha}{\alpha}} - 1\right)$$

noticing that $\lim_{\alpha \to 0.5} (2\alpha)^{\frac{2-2\alpha}{2\alpha-1}} = e^{\lim_{\alpha \to 0.5} \frac{ln(2\alpha)}{2\alpha-1}} = e$ it is clear that $\lim_{\alpha \to 0.5} V(\alpha) = 0$. However,

$$\frac{dV(\alpha)}{d\alpha} = 2\left((2\alpha)^{\frac{2-2\alpha}{2\alpha-1}} - 1\right) + (2\alpha)^{\frac{2-2\alpha}{2\alpha-1}}\left(\frac{2 - 2\alpha}{\alpha} - \frac{2ln(2\alpha)}{2\alpha - 1}\right) + 2\left((2\alpha)^{\frac{1-\alpha}{\alpha}} - 1\right) - (2\alpha)^{\frac{1}{\alpha}} \frac{1 - \alpha - ln(2\alpha)}{\alpha^2}$$

and so $\lim_{\alpha \to 0.5} \frac{dV(\alpha)}{d\alpha} = 2e - 4 > 0$, implying that $V(\alpha) > 0$ for $\alpha \approx 0.5$.

Analytically showing that $V(\alpha)$ or a suitable rescaling is strictly positive more generally for $\alpha \in (0.5, 1)$ is tricky (for instance $V(\alpha)$ is not quasiconcave); however, it is easily verified numerically.

Lower payoffs from mediation

In the main text, I focussed on the question of whether mediation could deliver Pareto improvements on unmediated bargaining. Here, I will show how mediation can also lower agents' payoffs compared to unmediated outcomes.

The first result (Proposition 9) shows that when the probability of behavioral types is small and there is only a single behavioral type for each agent, there always exists a mediation protocol that gives each agent i a payoff $u_i(1 - \alpha_j)$, her payoff from conceding immediately. In the Baseline equilibrium, agent i's payoff was $F_j(0)u_i(\alpha_i) + (1 - F_j(0))u_i(1 - \alpha_j)$ where $F_j(0) > 0$ if and only if $T_i = -\frac{1}{\lambda_i} ln(z_i) < T_j$. Hence, generically (whenever $T_i \neq T_j$) such mediation is Pareto inferior to the Baseline equilibrium. Clearly, this only lowers payoffs for at most one of the two agents.

The second result (Proposition 10) highlights some of the particular difficulties for mediation when agents can imitate multiple behavioral types. I extend the model slightly to allow for multiple types, and so mediation can affect rational agents' initial demand choices. I show that when the probability of behavioral types is sufficiently small each agent obtains a payoff approximately equal to that which she would receive facing her most aggressive possible behavioral opponent for sure. This is strictly lower than the Baseline equilibrium payoff for both agents when the set of behavioral types is even moderately rich.

Proposition 9. *For any given r_i, u_i, α_i for $i = 1, 2$ and fixed $K \geq 1$, there exists $\underline{z} > 0$ such that whenever $z_i \leq \underline{z}$ and $K \geq \frac{z_1}{z_2} \geq \frac{1}{K}$, there is an equilibrium with mediation where each agent i's payoff is exactly $u_i(1-\alpha_j)$.*

Proof. I prove this by construction. Following the notation of Section 4 in the main text, consider the distributions $G_j^Z(t) = 0$ for $t < T^{Z_1} = T^{Z_2} = T^R$. We would like to find some T^R, some mediation proposals $m_i : [0, T^R] \to [0, 1]$ and a distribution G^R such that *both* agents are indifferent to conceding on $[0, T^R]$ and $G^R(0) = 0$. For arbitrary m_i and T^R the fraction of remaining j agents at $t < T^R$ is $1 - F_j(t) = (1 - G^R(t))(1 - z_j) + z_j$. For confessing agent i to be indifferent to concession on $(0, T^R]$ we must have

$$\frac{f_j(t)}{1 - F_j(t)} = \lambda_j^m(t) = \frac{r_i u_i(1-\alpha_j)}{u_i(m_i(t)) - u_i(1-\alpha_j)}$$

Imposing the boundary condition $1 - F_j(T^R) = z_j$ and solving the linear ODE, we get

$$1 - G^R(t) = 1 - \frac{F_j(t)}{1 - z_j} = \frac{z_j}{1 - z_j}\left(exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) - 1\right) \tag{21}$$

We want this equation to hold for both agents i and j for all t and so we get:

$$P(t) = \frac{g^R(t) - g^R(t)}{\lambda_j^m(t)\lambda_i^m(t)} = \frac{z_j}{(1-z_j)\lambda_i^m(t)} exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) - \frac{z_i}{(1-z_i)\lambda_j^m(t)} exp\left(\int_t^{T^R} \lambda_i^m(s)ds\right) = 0 \tag{22}$$

We can immediately identify $m_j(T^R)$ as the unique value which solves this at T^R:

$$\frac{z_j(1-z_i)}{z_i(1-z_j)} = \frac{\lambda_i^m(T^R)}{\lambda_j^m(T^R)} = \frac{u_i(1-m_j(T^R)) - u_i(1-\alpha_j)}{u_j(m_j(T^R)) - u_j(1-\alpha_i)} \frac{r_j u_j(1-\alpha_i)}{r_i u_i(1-\alpha_j)}. \tag{23}$$

More generally, imposing $m_i(t) = 1 - m_j(t)$ and differentiating gives:

$$\begin{aligned}\frac{dP(t)}{dt} = 0 = &\frac{z_j}{1-z_j}\left(\frac{u_j'(m_j(t))m_j'(t)}{r_j u_j(1-\alpha_i)} - \frac{\lambda_j^m(t)}{\lambda_i^m(t)}\right) exp\left(\int_t^{T^R} \lambda_j^m(s)ds\right) \\ &+ \frac{z_i}{1-z_i}\left(\frac{u_i'(1-m_j(t))m_j'(t)}{r_i u_i(1-\alpha_j)} + \frac{\lambda_i^m(t)}{\lambda_j^m(t)}\right) exp\left(\int_t^{T^R} \lambda_i^m(s)ds\right)\end{aligned}$$

Combining this with equation (22) and solving for $m_j'(t)$ gives:

$$m_j'(t) = \frac{\lambda_j^m(t) - \lambda_i^m(t)}{\frac{\lambda_i^m(t)u_j'(m_j(t))}{r_j u_j(1-\alpha_j)} + \frac{\lambda_j^m(t)u_i'(1-m_j(t))}{r_i u_i(1-\alpha_i)}} = \frac{r_j u_j(1-\alpha_i)(u_j(m_j(t)) - u_j(1-\alpha_i)) - r_i u_i(1-\alpha_j)(u_i(1-m_j(t)) - u_i(1-\alpha_j))}{(u_i(1-m_j(t)) - u_i(1-\alpha_j))u_j'(m_j(t)) + (u_j(m_j(t)) - u_j(1-\alpha_i))u_i'(1-m_i(t))}$$

This is uniformly Lipschitz continuous in $m_j(t)$ for $t \in [0, T^R]$ (its derivative is continuous) and is continuous in t, hence by Picard's Theorem it has a unique solution. Define $\bar{m}_j$ as the unique value which solves the equality:

$$1 = \frac{u_j(\bar{m}_j) - u_j(1-\alpha_i)}{u_i(1-\bar{m}_j) - u_i(1-\alpha_j)} \frac{r_i u_i(1-\alpha_j)}{r_j u_j(1-\alpha_i)}$$

and $\bar{\lambda}^m = \frac{r_j u_j(1-\alpha_i)}{u_j(\bar{m}_j) - u_j(1-\alpha_i)}$. When $z_i = z_j$, it is clear that we must have $m_j(t) = \bar{m}_j$ and $\lambda_j^m(t) = \lambda_i^m(t) = \bar{\lambda}^m$ for all t. More generally, it is clear that $m_j(t)$ (respectively $\lambda_j(t)$) is a convex combination of $\bar{m}_j$ and $m_j(T^R)$ (respectively $\bar{\lambda}^m$ and $\lambda_j^m(T^R)$) given that $m_j'(t) > 0$ when $\lambda_j^m(t) > \lambda_i^m(t)$ (which decreases $\frac{\lambda_j^m(t)}{\lambda_i^m(t)}$).

Let the solution be indexed by T^R, $m_j^{T^R}$, with associated agreement time distribution $G_{T^R}^R$. Clearly $m_j^{T^R}(T^R - t)$ is independent of T^R and so $G_{T^R}^R(0)$ is continuous and strictly decreasing in T^R (see equation (21)). This ensures that there is a unique value of T^R such that $G_{T^R}^R(0) = 0$; call this $T^{R,0}$. Clearly the distribution $G_{T^{R,0}}^R$ ensures that a

confessing agent i obtains the payoff $u_i(1-\alpha_j)$.

Now consider the payoff of an agent i who doesn't confess. Given that $G_i^Z(t) = 0$ for $t < T^R$, conceding at $t \in (0, T^R)$ is strictly worse than conceding at 0 for a payoff of $u_i(1-\alpha_j)$. Conceding after T^R gives at most $e^{-r_i T^R} u_i(\alpha_i)$. If $T^R \geq \bar{T}^R = \max_{i\in 1,2} \left\{-\frac{1}{r_i} ln\left(\frac{u_i(1-\alpha_j)}{u_i(\alpha_i)}\right)\right\}$, therefore, agent i who doesn't confess obtains a payoff of exactly $u_i(1-\alpha_j)$.

Assume $z_i, z_j \leq 1-\varepsilon$ for any $\varepsilon > 0$ and fix u_i, r_i and α_i. Examining equation (23) it is clear that the bound $\frac{z_i}{z_j} \in \left[\frac{1}{K}, K\right]$ implies that we can uniformly bound $\lambda_j(T^R)$ and hence $\lambda_j(t)$, so that $\lambda_j(t) \in \left[\frac{1}{L}, L\right]$ for some $L \geq 1$. Given this bound, equation (21) shows that in order to have $G^R_{T^{R,0}}(0) = 0$ as $z_j \to 0$, we must have $T^{R,0} \to \infty$. And so, there exists $\bar{z} > 0$ such that if ever $z_i \leq \bar{z}$ then $T^{R,0} \geq \bar{T}^R$. This completes the proof. □

Next I consider a generalization of the model, following AG, in which agents make their demand announcements sequentially and agents can imitate multiple different behavioral types. To do this I introduce a new time 0^0 at which agent 1 makes her initial demand announcement. Agent 2 can then either immediately concede at 0^1 or announce a counterdemand. For each agent i there is a finite set of behavioral type demands E_i, where the conditional probability of a behavioral agent i being of type α_i is $\pi_i(\alpha_i)$. If a behavioral type has $\alpha_2 < 1-\alpha_1$ then she immediately concedes at 0^1. Assume that $\max E_i > 1 - \min E_j$. Let rational agent 1's demand choice be described by a probability distribution μ_1 on E_1, and rational 2's choice after observing α_1, be described by a probability distribution $\mu_1^{\alpha_1}$ on $E_2 \cup Q$ where Q indicates immediate concession. Reputations after demand choices are:

$$\bar{z}_1(\alpha_1) = \frac{z_1\pi_1(\alpha_1)}{z_1\pi_1(\alpha_1) + (1-z_1)\mu_1(\alpha_1)} \qquad \bar{z}_2^{\alpha_1}(\alpha_2) = \frac{z_2\pi_2(\alpha_2)}{z_2\pi_2(\alpha_2) + (1-z_2)\mu_2^{\alpha_1}(\alpha_2)}$$

AG establish a unique equilibrium of this game without mediation, characterized by the condition that each type a rational agent imitates must give her the same expected continuation payoff. After time zero, behavior matches the Baseline equilibrium described in the main text but with z_i replaced by $\bar{z}_i$. Let $\lambda_j^{\alpha_j,\alpha_i} = \frac{r_i u_i(1-\alpha_j)}{u_i(\alpha_i)-u_i(1-\alpha_j)}$ be the concession rate by j that would keep i indifferent to conceding after time zero without a mediator and demands α_j, α_i. AG show that as the fraction of behavioral types becomes small ($z_i \to 0$ and $\frac{z_1}{z_2} \in \left[\frac{1}{K}, K\right]$ for some $K \geq 1$) then bargaining becomes arbitrarily efficient so long as $\lambda_j^{\alpha_j,\alpha_i} \neq \lambda_i^{\alpha_i,\alpha_j}$ for each pair of incompatible demands α_i, α_j. Moreover, let $\alpha_i^R = \arg\max_{\alpha_i} u_i(\alpha_i)^{\frac{r_j}{r_i+r_j}} u_j(1-\alpha_i)^{\frac{r_i}{r_i+r_j}}$. This is the complete information alternating offers demand (Rubinstein (1982)) when the time between offers converges to zero. If agent i can imitate some type $\alpha_i' \leq \alpha_i^R$, then AG show she must obtain an equilibrium payoff greater than $u_i(\alpha_i')$. This holds because $\alpha_i' \leq \alpha_i^R$ implies $\lambda_j^{\alpha_j,\alpha_i'} < \lambda_i^{\alpha_i',\alpha_j}$ for any $\alpha_j > \alpha_i'$ so that agent i builds reputation exponentially more quickly than j. If agents imitate the demands (α_i', α_j) with positive limit probability as $z_i \to 0$, therefore, agent j must concede with probability approaching one at 0^4 to ensure that both agents reach a probability one reputation at the same time.

The next result, by contrast, shows that as the fraction of behavioral types becomes small, there exist equilibria with mediation which give rational agents' payoffs arbitrarily close to $u_i(1-\max E_j)$. The mediation protocol used for each incompatible demand pair is the same as in Proposition 9. Demand choices can then be distorted so that both agents almost exclusively start imitating their maximum demand type. Clearly, whenever the type space is rich enough that agent i can imitate a type $\alpha_i' \in (1-\max E_j, \alpha_i^R]$, then this implies strictly lower payoff under mediation than without.

Proposition 10. *Consider a sequence of bargaining games $B^n = \{u_i, r_i, E_i, \pi_i, z_i^n\}$ such that $\lim_n z_i^n = 0$ and $\frac{z_1^n}{z_2^n} \in \left[\frac{1}{K}, K\right]$ for some constant $K \geq 1$. Then there is a sequence of equilibria with mediation such that the limit of agent i's equilibrium payoffs is $\lim_n U_i^n = u_i(1-\max E_j)$ for $i = 1, 2$, $i \neq j$.*

Proof. I first construct an equilibrium which will hold for all arbitrarily large n. Given any demand α_1 suppose that whenever agent 2 makes counterdemand $\alpha_2 > 1-\alpha_1$ then agent 1's continuation payoff is $u_1(1-\alpha_2)$. In this

case agent 1's expected payoff from demanding α_1 is:

$$\underline{U}_1(\alpha_1) = u_1(\alpha_1)\left((1-z_2)\mu^{\alpha_1}(Q) + \sum_{\alpha_2 \leq 1-\alpha_1} z_2\pi_2(\alpha_2)\right) + \sum_{\alpha_2 > 1-\alpha_1} u_1(1-\alpha_2)\left(z_2\pi_2(\alpha_2) + (1-z_2)\mu_2^{\alpha_1}(\alpha_2)\right)$$

Whenever $|E_1| = 1$ then let $\mu_1(\max E_1) = 1$. Otherwise, consider any $\varepsilon \in (0,1)$ and define $\mu_1(\max E_1) = 1-\varepsilon$ and $\mu_1(\alpha_1) = \frac{\varepsilon}{|E_1|-1}$ for $\alpha_1 < \max E_1$. If $|E_2| = 1$ let $\mu_2^{\alpha_1}(\max E_2) = 1$, in this case because $\max E_2 > 1 - \min E_1$, we have that agent 1's continuation payoff is $u_1(1-\max E_2)$ for all α_1. Suppose then $|E_2| > 1$. Let $\mu_2^{\max E_1}(\max E_2) = 1-\varepsilon$ and $\mu_2^{\max E_1}(\alpha_2) = \frac{\varepsilon}{|E_2|-1}$ if $\alpha_2 < \max E_2$. If $\alpha_1 < \max E_1$ then $\mu_2^{\alpha_1}(\max E_2) = 1 - \varepsilon^{\alpha_1}$ and $\mu_2^{\alpha_1}(\alpha_2) = \frac{\varepsilon^{\alpha_1}}{|E_2|-1}$ if $\alpha_2 \in D_2(\alpha_1) = \{\alpha_2 \in (1-\alpha_1, \max E_2)\}$ and $\mu_2^{\alpha_1}(Q) = \frac{\varepsilon^{\alpha_1}(|E_2|-1-|D_2|)}{|E_2|-1}$, where ε^{α_1} is defined to ensure that $\underline{U}_1(\alpha_1) = \underline{U}_1(\max E_1)$. To check that ε^{α_1} is well defined, consider:

$$\begin{aligned}\underline{U}_1(\max E_1) - \underline{U}_1(\alpha_1) = &\sum_{\alpha_2 \leq 1-\alpha_1}\left((u_1(1-\alpha_1) - u(\alpha_1))\, z_2\pi_2(\alpha_2) + \frac{(1-z_2)}{|E_2|-1}\left(\varepsilon u_1(1-\alpha_2) - \varepsilon^{\alpha_1}u_1(\alpha_1)\right)\right)\\ &+ (\varepsilon^{\alpha_1} - \varepsilon)(1-z_2)\left(u_1(1-\max E_2) - \sum_{\alpha_2 > 1-\alpha_1}\frac{u_1(1-\alpha_2)}{|E_2|-1}\right)\end{aligned}$$

Given $1 - \max E_2 < \min E_1$, this expression is continuous and strictly increasing in ε and decreasing in ε^{α_1}. Choose $\bar{\varepsilon} > 0$ such that

$$\sum_{\alpha_2 \leq 1-\alpha_1}\frac{\bar{\varepsilon}u_1(1-\alpha_2) - u_1(\alpha_1)}{|E_2|-1} + (1-\bar{\varepsilon})\left(u_1(1-\max E_2) - \sum_{\alpha_2 > 1-\alpha_1}\frac{u_1(1-\alpha_2)}{|E_2|-1}\right) < 0.$$

for $\alpha_1 = \min E_2$. It is clear that for all $\varepsilon \leq \bar{\varepsilon}$ and for all $\alpha_1 \in E_1$ there must exist some $\underline{z}_2$ such that if $z_2 \leq \underline{z}_2$, then $\underline{U}_1(\max E_1) - \underline{U}_1(\alpha_1) < 0$ if $\varepsilon^{\alpha_1} = 1$, hence $\varepsilon^{\alpha_1} \geq \varepsilon$ is well defined and indeed, bounded away from 1.

Now consider the sequence of bargaining games B^n and choose N sufficiently large that $z_2^n \leq \underline{z}_2$ for all $n \geq N$. Suppose that agents play the demand choice strategies above for all such n. Given that agent 1 imitates all her types with probability bounded away from 0, and agent 2 likewise imitates all possible incompatible counterdemands with probability bounded away from 0, it is clear that $z_i^n \to 0$ implies $\bar{z}_1(\alpha_1) \to 0$ and moreover there exists $L \geq 1$ such that $\frac{\bar{z}_1(\alpha_1)}{\bar{z}_1^{\alpha_1}(\alpha_2)} \in \left[\frac{1}{L}, L\right]$ for all incompatible behavioral demand pairs. By Proposition 9, therefore, there exists some $N'_{\alpha_1,\alpha_2} \geq N$ such that if $n \geq N'_{\alpha_1,\alpha_2}$, we can find an equilibrium with mediation for the continuation game with incompatible demand pair (α_1, α_2) such that continuation payoffs for each agent i are exactly $u_i(1-\alpha_j)$. For all $n \geq N' = \max_{\alpha,\alpha_2}\{N'_{\alpha_1,\alpha_2}\}$, by construction agent 1 is indifferent between all her demand choices and agent 2 is indifferent between all her incompatible counterdemands and conceding immediately.

The above equilibrium with mediation gives agents an expected utility of at most $\lim_n U_i \leq (1-\varepsilon)u_i(1-\max E_j) + \varepsilon u_i(\max E_i)$. Given that $\varepsilon \in (0, \bar{\varepsilon}]$ was arbitrary it is clear that we can choose a sequence $\varepsilon^n \to 0$ such that there is an equilibrium with mediation of the above form for all $n \geq N'$, and so i's payoff converges to $u_i(1-\max E_j)$. □