Reputational Bargaining*

Jack Fanning, Alexander Wolitzky

April 2020

1 Introduction

In their 1992 survey of noncooperative bargaining theory, Binmore, Osborne, and Rubinstein observe that "Schelling's (1960) view of bargaining as a 'struggle to establish commitments to favorable bargaining positions' remains largely unexplored as regards formal modeling," (Binmore et al. (1992), p. 200). A generation later, this is no longer true. One branch of the literature has modelled attempts to establish commitment as explicit moves in a complete-information bargaining game; it is surveyed in the current volume by Miettinen. A second branch considers simple bargaining games with only offers and accept/reject decisions, but introduces incomplete information about whether a bargainer is a "type" committed to obtaining a large share of the surplus; uncommitted bargainers then have incentives to imitate these types to develop a tough reputation. This is the literature on "reputational bargaining." Such models have proved remarkably tractable and often provide clear predictions that are independent of details such as the bargaining procedure and the distribution of commitment types. They have also delivered new insights in settings beyond bilateral bargaining, such as repeated games and search markets.

The predictions of complete-information bargaining models following Rubinstein (1982) have been criticized for depending on unobserved details of the extensive form, such as whether one party can make offers more frequently than the other, or whether offers are sequential or simultaneous (Wilson (1987), Kreps (1990b)).[1] The reputational bargaining literature's eschewal of explicit extensive-form modeling of commitment thus reflects an ambition to predict the outcome of negotiations on the basis of players' preferences and beliefs alone, rather than on how bargaining is assumed to proceed.

The first hint that such procedure-independent predictions might be possible comes from the

*We thank Mehmet Ekmekci, Deepal Basak, Drew Fudenberg, David Pearce, the editors, and an anonymous reviewer for helpful comments.

[1] Perry and Reny (1993) provide support for Rubinstein's predictions in a continuous-time model with endogenously timed offers. Cooperative bargaining solutions offer an alternative, procedure-free approach.

classical Coase conjecture (Fudenberg et al. (1985), Gul et al. (1986); see Ausubel et al. (2002) for a survey). Loosely speaking, this result states that in bargaining with one-sided private information about valuations for a good, the informed party's equilibrium payoff is no less than it would be if she were known to have her most favorable valuation (e.g. a seller with a known cost immediately proposes a price that all buyer value types accept). However, the extension of this model to two-sided private information about valuations produced less compelling predictions. An inevitable feature of such environments is that a player's offers signal her information, typically leading to vast equilibrium multiplicity: signalling allows a player to be "punished with beliefs" for deviating from a proposed equilibrium path (e.g., she is identified as the weakest possible type and given a low continuation payoff), and the threat of this punishment can support a wide variety of behavior, ranging from no-trade to Myerson and Satterthwaite (1983)'s constrained-efficient bounds (Ausubel and Deneckere (1993)). While attempts have been made to impose "reasonable" refinements on these equilibria, it is often hard to agree on what is reasonable, particularly when some natural refinements give paradoxical results such as no-trade (Ausubel et al. (2002)).

The reputational bargaining literature thus starts not from the classical Coase conjecture, but from a "reputational Coase conjecture" established by Myerson (1991). Myerson considers an infinite-horizon, alternating-offers bargaining game, where one player has a small probability of being a "commitment type" who always demands some exogenous, pre-specified share α of the surplus and never accepts less. Myerson shows that, when both players are patient, the possibly-committed player cannot receive an equilibrium payoff significantly below α, regardless of the players' relative costs of delay. Kreps (1990b) conjectured the same result, and predicted it would hold regardless of the details of the bargaining protocol.[2]

The seminal reputational bargaining model of Abreu and Gul (2000) (henceforth AG) vastly generalizes Myerson's result by introducing general bargaining protocols (rather than alternating offers), multiple commitment types (rather than Myerson's single "α-insistent type"), and two-sided reputation formation (i.e., commitment types on both sides). AG find a unique equilibrium that is independent of the details of the bargaining protocol, so long as both sides can make offers frequently. Punishing with beliefs does not arise despite two-sided incomplete-information, because commitment types are immune to belief punishments: they insistently make their pre-specified demands, forcing this behavior onto the equilibrium path. The equilibrium features a war of attrition structure, with uncommitted players on both sides mimicking commitment types before eventually conceding. This offers a good description of some real-world negotiations and links AG to earlier models of incomplete-information wars of attrition (e.g. Kreps and Wilson (1982), Milgrom and Roberts (1982)). AG provide especially clear predictions when commitment behavior is vanishingly unlikely: under some conditions, payoffs approximate those from complete-information, alternating-offers bargaining. Similarly clear

[2]See Chapter 5 of Kreps (1990b), and also Exercise 9 to Chapter 15 of Kreps (1990a).

predictions in the complete-information limit arise in many other reputational bargaining models, even those with multiple equilibria.

The rest of this chapter is arranged as follows. Section 2 describes AG's model and its predictions. Section 3 discusses extensions of the reputational bargaining framework. Section 4 presents applications to specific economic environments. Section 5 discusses experimental evidence. Section 6 concludes by highlighting some open questions.

2 The Abreu-Gul (AG) Reputational Bargaining Model

AG's paper has three parts. It first analyzes a simple concession game with a single commitment type on each side. The game is then generalized to allow multiple commitment types, with a focus on the complete-information limit. Finally, AG show that equilibria in a large class of discrete-time reputational bargaining games converge to the unique equilibrium of the concession game as offers become frequent.

The concession game with a single commitment type

Two players must divide a dollar at some point in continuous time. Each player $i \in \{1, 2\}$ is a "commitment type" with independent probability z_i (alternatively, a "behavioral," "inflexible," "insistent," "obstinate," or "irrational" type) and otherwise is rational. A commitment type always demands some fixed share $\alpha_i \in (0, 1)$, where the commitment demands are incompatible: $\alpha_1 + \alpha_2 > 1$. At any moment each player i can "concede" (accept), obtaining share $1 - \alpha_j$ and giving her opponent α_j. Commitment types never concede. Each player i discounts payoffs exponentially at rate $r_i > 0$, so that if she obtains share x_i at time t, she receives payoff $e^{-r_i t} x_i$.

Each player i's strategy is conveniently described by a distribution F_i over concession times, where $F_i(t)$ is the probability that player i concedes by time t. Because commitment types never concede, player i's reputation for being committed at time t (absent agreement) is

$$z_i(t) = \frac{z_i}{1 - F_i(t)}.$$

This concession game has a unique Nash equilibrium, in which play follows a war of attrition. It is characterized by three properties:

(a) Both players' reputations reach probability 1 at the same time T^*.

(b) At most one player concedes with positive probability at time 0.

(c) On the interval $(0, T^*)$, each player i concedes at the constant rate that keeps her opponent indifferent between waiting and conceding.

These properties are not difficult to establish. Property (a) must hold, because if rational player i were ever certain that she faced committed player j, she would concede immediately. If property (b) did not hold, a player could profitably wait until an instant after time 0 before conceding, to see if her opponent concedes first. By similar reasoning, concession must be continuous after time 0, which implies that each player must always be indifferent between waiting and conceding. For i to be indifferent, j must concede at the constant rate λ_j given by

$$r_i(1-\alpha_j) = \lambda_j(\alpha_i + \alpha_j - 1) \quad \Leftrightarrow \quad \lambda_j = \frac{r_i(1-\alpha_j)}{\alpha_i + \alpha_j - 1}. \tag{1}$$

This equates i's flow cost of delaying concession (the lost interest on j's offer, $1-\alpha_j$) and her flow benefit of delay (the probability that j concedes multiplied by i's payoff gain when that happens, $\alpha_i - (1-\alpha_j)$).

Equation (1) implies that for $t \leq T^*$, we have

$$1 - F_j(t) = (1 - F_j(0))e^{-\lambda_j t},$$

where $F_j(0)$ is the probability that j concedes at time 0. Since both players' reputations reach 1 at time T^*, we have

$$z_j(T^*) = \frac{z_j e^{\lambda_j T^*}}{1 - F_j(0)} = 1.$$

If i does not concede with positive probability at time 0, her reputation reaches 1 at time

$$T_i = -\frac{\ln z_i}{\lambda_i}.$$

Because at most one player concedes at time 0, the equilibrium characterization is completed by setting

$$T^* = \min\{T_1, T_2\} \quad \text{and} \quad F_j(0) = 1 - z_j e^{\lambda_j T^*} = \max\left\{0, 1 - z_j z_i^{-\lambda_j/\lambda_i}\right\}. \tag{2}$$

A useful way to understand the equilibrium is to first ask which player would win a "race" to reach reputation 1 absent concession at time 0: that is, which player has the smaller T_i. The losing player (the one with the larger T_i) must then concede at time 0 to give her reputation a sufficient "head start" to reach 1 at the same time as her opponent's. Note that because each player i is indifferent to conceding an instant after time 0, her equilibrium payoff is

$$F_j(0)\alpha_i + (1 - F_j(0))(1 - \alpha_j),$$

which exceeds her payoff from immediately conceding only if she wins the reputational race.

So, which player wins the race? Note that $T_i < T_j$ if and only if

$$\frac{\ln z_i}{\ln z_j}\frac{r_i}{r_j}\frac{1-\alpha_j}{1-\alpha_i} < 1.$$

Therefore, i is better-positioned to win the race when her initial reputation is larger, when she is more patient, and when her demand is smaller. This last comparison—smaller demands increase bargaining strength—plays a major role in the expanded model with multiple commitment types, because it incentivizes rational players to make moderate demands.

It is also important to note that players' initial reputations z_i, z_j enter bargaining strength through the ratio of their logarithms, unlike the concession rates λ_i, λ_j. This has dramatic implications for the complete-information limit, where commitment types become vanishingly unlikely. Consider a sequence of concesssion games where initial reputations converge to zero at the same rate, $z_i^n, z_j^n \to 0$ with $z_i^n/z_j^n \in [1/K, K]$ for some $K \geq 1$, with all other parameters are fixed. If $\lambda_i > \lambda_j$ then j must concede at time 0 with probability approaching 1 along the corresponding sequence of equilibria. This is immediate from examining equation (2), which shows that j's time 0 concession satisfies $F_j(0) \geq 1 - K z_i^{1-\lambda_j/\lambda_i}$ when $z_j \leq K z_i$, and noting that this lower bound on $F_j(0)$ is close to 1 when $z_i \approx 0$ and $\lambda_i > \lambda_j$. To see the intuition, notice that to reach a probability 1 reputation when the initial reputations are small, players must concede with probability close to 1. Since after time 0 players concede at constant rates (which are independent of the initial reputations), the reputational race must continue for a long time. During this long race i's reputation grows exponentially faster than j's, $(dz_i(t)/dt)/z_i(t) = \lambda_i > \lambda_j$, which overwhelms any fixed proportional advantage for j in the initial reputations.

The concession game with multiple commitment types

Now suppose for each player i there is a finite set of commitment types $C_i \subset (0,1)$, where each type is identified with its demand. The (exogenous) probability that i demands $\alpha_i \in C_i$ conditional on being committed is $\pi_i(\alpha_i)$, while the total probability that she is committed remains z_i. At time 0, first player 1 publicly announces a demand $\alpha_1 \in C_1$, and then player 2 announces a counterdemand $\alpha_2 \in C_2$, whence play continues into a concession game. Denote the (endogenous, equilibrium) probability that rational player 1 demands α_1 by $\mu_1(\alpha_1)$, and denote the probability that rational player 2 counterdemands α_2 by $\mu_2^{\alpha_1}(\alpha_2)$. Players' reputations at the start of the concession game with demands α_1, α_2 are

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

AG show there is still an essentially unique equilibrium after incorporating this demand-choice stage. The basic intuition is that when rational player i becomes more likely to mimic type

α_i, this reduces her posterior reputation after announcing α_i, which reduces her continuation payoff and makes mimicking α_i less appealing. This "strategic substitutability" pushes towards a unique equilibrium.

What happens in the complete-information limit of this richer game? AG show that, along a sequence of concesssion games where initial reputations converge to zero at the same rate, each player i can guarantee a limiting payoff of at least

$$\underline{\alpha}_i = \max\left\{\alpha_i \in C_i : \alpha_i \leq \frac{r_j}{r_i + r_j}\right\}.$$

To see this for $i = 1$, note that if player 2 counterdemands $\alpha_2 > 1 - \underline{\alpha}_1$, then $r_2/(r_1 + r_2) \geq \underline{\alpha}_1 > 1 - \alpha_2$, and therefore

$$r_1(1 - \alpha_2) < \frac{r_1 r_2}{r_1 + r_2} \leq r_2(1 - \underline{\alpha}_1),$$

so $\lambda_2 < \lambda_1$. As we have seen, this implies that player 2 concedes at time 0 with probability 1 in the complete-information limit.

Given the above result, if the space of commitment types is sufficiently rich, player i's payoff must be approximately $r_j/(r_i + r_j)$, which is also her payoff in Rubinstein (1982)'s complete-information, alternating-offers game when offers are frequent. Types who demand exactly $\alpha_i^* = r_j/(r_i + r_j)$ are sometimes called "cannonical" types. When they are present, we can precisely identify equilibrium outcomes in the complete-information limit. The independence of this prediction to the distribution of commitment types, π_i, is a crucial robustness property. It is similar to Fudenberg and Levine (1989)'s finding that, in the presence of a type that always plays a Stackelberg action, a patient long-run player facing short-run opponents obtains approximately her Stackelberg payoff.

Convergence of discrete-time bargaining to the concession game

The final part of AG's paper considers discrete-time reputational bargaining games. The only assumption made about the bargaining protocol is that each player can make at least one offer in every length-$\Delta > 0$ interval of real time. AG show that all perfect Bayesian equilibrium outcome distributions of all such games converge to the unique equilibrium of the concession game as offers become frequent ($\Delta \to 0$). This crucial result shows that the preceding analysis (of concession games where players cannot change their demands) applies equally to any bargaining game with frequent offers, independent of the details of the bargaining protocol. It is the basis of much of the subsequent literature, which often directly adopts AG's tractable continuous-time concession game structure.

To understand AG's convergence result, suppose we knew that if a player takes an action inconsistent with any of her commitment types ("reveals rationality") before her opponent does,

then she must immediately concede. We would then be back to a (discrete-time) concession game: after making her initial demand, each player's only remaining choice is whether to keep mimicking her chosen commitment type or to reveal rationality and concede. Convergence to continuous time would then be a technical exercise.

The key part of AG's result is therefore that, with frequent offers, revealing rationality is essentially the same as conceding, and in particular gives approximately the same continuation payoffs. This follows from a generalization of Myerson's reputational Coase conjecture, discussed above. It remains to explain the logic of Myerson's result.

Suppose player 1 is possibly committed, while player 2 is known to be rational. Note first that there exists a finite time T such that, if player 1 always demands α and never accepts less, then player 2 concedes by T. The argument is similar to ones in the literature on reputation in repeated games (e.g. Fudenberg and Levine (1989); see Mailath and Samuelson (2006) for a survey): If player 2 does not immediately accept, she must believe that 1 will cease commitment behavior soon with positive probability. So, if 1 does not cease, 2's belief that 1 is committed must increase. Iterating this argument, 2 must eventually become certain that 1 is committed, and so accept.

This argument implies that at any time $t < T$ rational player 1 can guarantee a continuation payoff of $e^{-r_1(T-t)}\alpha_1$ by insisting on α until T. To complete the proof, we argue that T converges to 0 as offers become frequent. Suppose towards a contradiction that T remains bounded away from 0, and suppose 1 insists on α_1 until time $T - \varepsilon$ for some small $\varepsilon > 0$. From this point forward player 2 can expect at most $1 - e^{-r_1\varepsilon}\alpha_1$ in any agreement. Fixing another small number $\eta > 0$, agreements reached after time $T - \eta\varepsilon$ must be worth even less to player 2 from the perspective of time $T - \varepsilon$: at most $e^{-r_2(1-\eta)\varepsilon}(1 - e^{-r_1\eta\varepsilon}\alpha_1) < 1 - \alpha_1$. Hence, for 2 to delay acceptance from $T - \varepsilon$ to $T - \eta\varepsilon$, she must believe 1 will cease commitment behavior before $T - \eta\varepsilon$ with high probability. Iterating this argument for $k \in \mathbb{N}$, if time $T - \eta^k\varepsilon$ is reached, 1 must cease commitment behavior before $T - \eta^{k+1}\varepsilon$ with high probability.[3] But these repeated expected deviations from commitment behavior eventually exhaust the probability that 1 is rational before time T, a contradiction.

3 Extensions

Endogenous commitment demands

In AG, the interpretation of the distribution over commitment types π_i is somewhat ambiguous. Certainly, real-world bargainers may not have a very precise sense of the probabilities with which their opponents can be committed to various bargaining positions. One of AG's key

[3] Frequent offers guarantee that 2 has an opportunity to accept 1's demand within each such interval.

messages is that the details of π_i are often irrelevant in the complete-information limit, but they do assume that the relative probabilities of different commitment types do not blow up, and π_i also matters away from the limit. These considerations have led some researchers to consider models where the distribution of commitment demands is endogenous.

The first paper in this area—written shortly after AG's paper was first circulated—is due to Kambe (1999). Kambe considers an elegant variant of AG, where each player i is initially rational for sure, but after making any initial demand α_i becomes committed to it with some probability z_i. (Thus, a player is parameterized by a single number z_i rather than a distribution π_i.) A player does not observe whether the opponent becomes committed; moreover, the initial demands cannot signal commitment, because they are made before commitment arises.[4] Once players make their initial demands (and potentially become committed), play proceeds as in AG's concession game.

Kambe shows that in equilibrium players make the unique just-compatible demands α_i, α_j that lead to a tie in AG's reputational race:[5] that is, demands satisfy the system of equations

$$\frac{\ln z_i}{\ln z_j}\frac{r_i}{r_j}\frac{1-\alpha_j}{1-\alpha_i} = 1, \quad \alpha_i + \alpha_j = 1,$$

which has solution

$$\alpha_i = \frac{r_j \ln z_i}{r_i \ln z_i + r_j \ln z_j}.$$

If player i's demand is more aggressive than this, she loses the reputational race and ends up conceding; while if she is less aggressive, she gets a smaller share when her opponent accepts.

Kambe's model thus predicts immediate agreement, even when z_i is large (unlike AG). However, $\alpha_i \to r_j/(r_i + r_j)$ when z_i and z_j go to 0 at the same rate, so Kambe's model coincides with AG in the complete-information limit. It can thus be viewed as a reinterpretation of AG where the exogenous commitment type distribution is replaced by endogenous bargaining postures.[6]

Non-stationary types and payoffs-as-you-go

In the models considered so far, the play of commitment types takes a very simple form: always demand some fixed share α_i, and never accept less. There is no obvious reason to restrict attention to such stationary types, and indeed it seems plausible that a player could be committed to richer behaviors, such as making tougher or weaker demands over time, or responding

[4] These assumptions mirror those of the complete-information bargaining model of Crawford (1982).

[5] More precisely, these demands arise in the unique equilibrium without randomization over initial demands, and payoffs in equilibria with randomization are similar.

[6] Sanktjohanser (2018) considers a hybrid of Kambe and AG, where each player knows at time 0 whether she is a "stubborn type," all types are free to make any initial demand, and stubborn types become committed to any initial demand they make. This model re-introduces signalling concerns, which allow almost any equilibrium payoffs; however, the paper also characterizes behavioral properties that hold across all symmetric equilibria.

aggressively to certain opposing actions.[7] Such *non-stationary* types are considered by Abreu and Pearce (2007) (henceforth AP). AP also analyze "bargaining with payoffs-as-you-go," a hybrid between the pure bargaining model considered so far and a repeated game: players first announce potentially non-stationary commitment types (where all commitment types announce truthfully), and then repeatedly play a stage game and receive payoffs, while simultaneously offering each other binding contracts to govern future play of the game.

In this rich and complex model, the authors establish a remarkable result: in the complete-information limit, payoffs converge to the Nash bargaining with threats (NBWT) payoffs identified by Nash (1953), so long as there is a type on each side that always plays the corresponding NBWT action and insistently demands the NBWT payoff. Thus, the stationary NBWT type is canonical, while non-stationary types have no effect.[8]

Intuition for the result comes from first generalizing AG's model with only stationary types to a setting with arbitrary flow/disagreement payoffs and feasible agreements corresponding to the stage game payoffs. With equal patience, a type demanding her Nash bargaining payoff is canonical: it ensures that a player concedes faster than her opponent in the war of attrition. Now allow players to first choose mixed actions in the stage game to determine the flow payoffs. Anticipating that they will agree on Nash bargaining payoffs relative to the flow payoffs, the players will choose their Nash threat actions. Finally, AP show that a war of attrition structure is preserved when non-stationary types are introduced, although now concession rates are determined by equilibrium continuation payoffs rather than current offers. While a player imitating the NBWT type may concede with lower probability than her opponent at certain times, she concedes with higher probability over the long run and so still wins the reputational race.[9]

Wolitzky (2011) notes a caveat to AP's powerful equilibrium selection result: it relies on the assumption that commitment types are "transparent," in that they truthfully announce their future behavior at the beginning of the game. Suppose there is instead a positive probability of a weak commitment type that initially claims to be the NBWT threat type and mimics its behavior for a long time, before eventually conceding to any demand. If this type is more likely than the true NBWT type, a player will wait when her opponent claims to be the NBWT type, hoping that he is actually the weak type. Thus, when commitment types are both non-stationary and non-transparent, equilibrium selection depends on the relative frequency of different types.[10]

[7] Richer types also let us avoid AG's somewhat counterfactual "no-haggling" prediction that a player who changes her offer immediately concedes.

[8] Recall that given a stage game with action sets A_i and utility functions u_i, the NBWT solution is the Nash equilibrium of the game where players choose "threats" $\beta_i \in \Delta(A_i)$ and payoffs are given by the Nash bargaining solution for the feasible payoff set of the stage game with disagreement point $u(\beta_1, \beta_2)$.

[9] A related paper by Atakan and Ekmekci (2013b) obtains a war of attrition structure in a class of repeated games with two-sided reputation. In recent work, Abreu and Pearce (2019) extend their NBWT prediction to settings without binding contracts, by imposing a form of renegotiation proofness.

[10] If all types are stationary then transparency is irrelevant, because initial play reveals the entire strategy.

Non-equilibrium analysis

The complexity of the equilibrium reasoning involved in reputational bargaining models raises the question of what predictions are robust to letting players hold more permissive, non-equilibrium beliefs about the opponent's behavior. Wolitzky (2012) investigates this issue in a bargaining model where players can announce any (potentially non-stationary) path of bargaining demands, before become committed to the announced path with probability z_i (as in Kambe's variant of AG). He asks what predictions can be made assuming only that players' strategies can be rationalized by some belief, and what path of demands a player must announce to guarantee her largest possible payoff. It turns out that a player with ex-ante commitment probability z_i can guarantee a "minmax" payoff of $\alpha_i^* = 1/(1 - \ln z_i)$ against an uncommitted opponent, which is substantial even for small z_i. The announcement which guarantees this payoff initially demands α_i^* and subsequently demands compensation for any delay: more precisely, it demands $\min\{e^{r_i t}\alpha_i^*, 1\}$ at each time t. The intuition is that a demand path that increases slower than this leaves the player with a payoff below α_i^* when the opponent accepts after some delay; while a path that increases faster fails to convince the opponent that the player is committed by the time her demand reaches 1, and thus could lead to a permanent impasse.

Non-stationary environments

Fanning (2016) extends AG's model to a non-stationary environment where players must agree before a random deadline that is continuously distributed on a finite interval $[0, T]$. When commitment is vanishingly unlikely, outcomes differ markedly depending on whether or not commitment types are stationary, unlike in AP. With a rich set of stationary types, players can approximately guarantee their Nash bargaining payoff regardless of their impatience. This occurs because small initial reputations cause bargaining to continue until close to T, when the cost of delay explodes. A Nash demand player concedes much faster than her opponent at that point, and so wins the reputational race. With non-stationary types, the type that adopts the time-varying, complete-information, alternating-offers strategy for this environment is canonical. The intuition is that alternating offers give players equal opportunities to use the threat of costly delay to extract surplus, so if agreement were ever delayed this would be equally costly to both players. Therefore, in reputational bargaining, a player who demands more than her alternating-offers share faces higher delay costs than an opponent imitating an alternating-offers type, and so concedes slower and loses the reputational race.

When commitment types are stationary, the model also predicts "deadline effects" similar to those observed empirically (e.g. Roth and Malouf (1979)). There is frequent agreement at time 0 and close to the deadline but not in between, and some disagreement. Here the time 0 agreements reflect initial concessions as in AG, while the subsequent lull in agreement followed

by a spike at the deadline occurs because war of attrition concession rates are proportional to delay costs.

Fanning (2018) considers a different non-stationary extension of AG, where now players' costs of delay can change at some "revelation time" $R > 0$, with both players initially uncertain about the direction of such changes. For instance, an election at time R may determine political parties' costs of resisting an agreement in a divided legislature. The main result shows that there is often delay, even in the complete-information limit with a rich set of stationary commitment types. Rational players make aggressive, incompatible demands and then wait until time R in the hope that the opponent will turn out to have a large delay cost, and so concede. Mutually beneficial compromises exist; however, a player who proposes one increases her opponent's option value of waiting, so the opponent still waits.

Incomplete information about preferences

A final extension combines commitment types with incomplete information about preferences. One interpretation of AG's results is that perturbing a complete-information bargaining model with a rich set of commitment types selects a unique equilibrium outcome. Abreu et al. (2015) ask the same question for the incomplete-information model of Rubinstein (1985), where one player's preferences are known while the other can have one of two possible discount rates. The authors show that perturbing this model with a rich set of stationary commitment types and taking the limit as those types become vanishingly unlikely supports a "Coasean" prediction: the outcome is the same as if the informed player were known to be patient. The intuition is that, since the patient rational type concedes at a slower rate in the war of attribution, in the limit the outcome of the reputational race is solely determined by this type's behavior. By contrast, allowing non-stationary types that can delay making their initial offer yields a non-Coasean equilibrium where the rational informed player no longer receives the payoff corresponding to her patient type. The problem is that the patient type has an incentive to separate by delaying her initial demand, which breaks the pooling equilibrium.[11]

[11] Peski (2019) studies multi-issue reputational bargaining with incomplete information about players' weights on different issues, where bargainers can offer menus of alternative agreements. With one-sided preference uncertainty, the uninformed party gets half the total surplus by offering a menu consisting of all allocations that give her that payoff.

4 Applications

Outside options and search markets

In an early critique of AG, Compte and Jehiel (2002) investigate the effect of outside options on reputational bargaining. Their main point can be seen when each player i has a single commitment type, which demands α_i. Assume an alternating-offers bargaining protocol, which gives complete information payoffs $v_1^* = 1 - v_2^* = (1 - \delta_2)/(1 - \delta_1\delta_2)$. Further, assume that each player can opt out of bargaining at any time, yielding payoffs v_i^{out}, v_j^{out}, and that player i prefers her commitment demand, over her complete-information payoff, over opting out, over a committed opponent's offer: that is, $\alpha_i > v_i^* > v_i^{out} > 1 - \alpha_j$.

Compte and Jehiel show that in the unique equilibrium play proceeds exactly as in the model without outside options or commitment types: the players immediately agree to the complete-information payoffs. The intuition is that since each player will opt out if she becomes convinced her opponent is committed, players have no incentive to build reputations, but instead reveal rationality and bargain under complete information.

This analysis suggests that much depends on whether outside options are sufficiently attractive relative to the commitment types' offers. With a rich set of commitment types (or in Kambe's endogenous demand model) players make moderate equilibrium demands, and in the complete-information limit we have $\alpha_i + \alpha_j = 1$, which violates the assumption $\alpha_i > v_i^* > v_i^{out} > 1 - \alpha_j$. Thus, Compte and Jehiel's critique is most significant when there is only a small number of relatively aggressive commitment types.

Atakan and Ekmekci (2013a) consider reputational bargaining with outside options endogenously determined by a search market. Firms and workers flow into the market and are randomly matched to bargain. They exit the market after reaching an agreement that generates a unit of surplus (or randomly dying). Players are rational or committed. Player i's single commitment type always demands α_i, but also stops bargaining and returns to the market if convinced that her opponent is committed. On returning to the market, players must wait time $\tau \geq 0$ before being rematched.

The paper derives several results concerning steady-state equilibria. A headline result is that when search costs are minimal ($\tau \approx 0$) and firms and workers enter the market at the same rate, bargaining involves no initial concessions, so the outcome is inefficient with total payoffs $2-\alpha_1-\alpha_2 < 1$ (in contrast with AG, where initial concessions lead to efficiency in the complete-information limit). The reason why initial concessions cannot occur is that this would give players on the other side of the market outside options that are greater than their payoffs from conceding, which is inconsistent with equilibrium. However, it is unclear whether a richer set of commitment types would constrain this inefficiency.

Endogenous outside options are also central to Özyurt (2015), who shows that even vanishingly small reputational concerns allow a wide range of prices in a Bertrand-like setting. This occurs because buyers who observe a seller undercutting her rival's posted price, use the lower price as an outside option in bargaining with the high price rival.[12]

Mediation

Fanning (2019) investigates how an uninformed mediator can improve efficiency in AG's model. He first considers a simple form of mediation often used by professional mediators: publicly suggesting a deal only when both parties accept it in private. This can be effective, but only if the mediator sometimes fails to suggest the deal even when both parties accept. Fanning then characterizes the equilibrium with mediation that maximizes rational players' payoffs in symmetric games. Mediation improves on unmediated bargaining if and only if players are risk averse, or commitment demands are larger than the probability of commitment types. Mediation works by (1) replacing dispersed agreement shares between rational players with an average agreement, and (2) reducing delay between two rational players more than between a rational player and a reported commitment type, which incentivizes truthful reporting of rationality.

5 Experimental evidence

Behavior resembling reputational bargaining was observed even in early unstructured bargaining experiments. For example, Roth and Malouf (1979) had subjects divide the probability of winning a monetary prize before a deadline by sending proposals and free-form messages over a computer. When the prize was worth three times as much to one player, agreements clustered around two focal points: equal probability of winning a prize (the Nash solution) and equal expected payoffs (75% probability for the low-prize subject). Agreements occurred close to the deadline, although some subjects never agreed. These focal divisions correspond to two notion of fairness that may motivate commitment behavior. By contrast, when each subject could win the same prize, they always split the probability equally, again suggesting that commitment to a demand may depend on its perceived fairness and/or focality.

One feature of behavior in these experiments that does not align with AG's predictions is that small demand changes do not precipitate immediate agreement. Nonetheless, Fanning and Kloosterman (2019) provide support for the basic Coasean underpinnings of reputational bargaining when there is only one fair/focal division: in an infinite-horizon bargaining experiment

[12] Özyurt (2014) introduces commitment types into Fearon (1994)'s "crisis bargaining" model, where in addition to waiting or conceding, bargainers can end the game by "attacking"; this is another type of outside option.

where one subject makes all the offers, outcomes were close to immediate agreement on an equal division (in contrast to relatively unequal divisions in one-shot ultimatum bargaining).

Other experiments have sought to test reputational bargaining predictions more directly. Embrey et al. (2014) allow a simultaneous initial demand stage followed by a continuous-time concession stage with fixed demands. When subjects faced either another subject or a computer committed to a fixed demand (without knowing which), they made the computer's demands more frequently than in a control treatment in which they always faced another subject. This suggests that subjects understood the benefit of mimicking a tough computer bargainer. However, many subjects still demanded an equal split instead of the computer demand, and there was more delay than predicted by AG. Heggedal et al. (2020) test Compte and Jehiel (2002)'s predictions by adding treatments with outside options to the above experimental setup. Outside options reduce imitation of aggressive computer demands but do not improve bargaining efficiency, because they are used too often.

6 Open questions

Foundations for commitment behavior

A key feature of reputational bargaining models is that commitment behavior is exogenous. This has the the advantage of limiting signalling and equilibrium multiplicity. But it also raises important questions of where commitment behavior comes from, and what forms of commitment behavior are most likely to be observed.

Abreu and Sethi (2003) address these questions using evolutionary game theory. They consider a population of commitment and rational types who are randomly matched and then bargain. All players have the same preferences over agreements, but rational types incur an extra cost reflecting their more sophisticated behavior. This cost ensures that commitment types always exist in every evolutionary stable equilibrium.[13] The main result is that whenever there is a commitment type demanding α there must also be a complementary type demanding $1-\alpha$. Complementary types ensure that invading types demanding more than α are incompatible with the complementary type, and so earn lower profits. For any $\alpha > 0.5$, an equilibrium exists with only two complementary types demanding α and $1-\alpha$ (in addition to rational types, when the cost of rationality is sufficiently small). The equilibrium is efficient when $\alpha \to 0.5$.

Basak (2019) provides a simple foundation for commitment in an alternating-offers model where players have private "reservation values". Each player's reservation value ω_i is drawn

[13] Abreu and Sethi's notion of evolutionary stability requires that all types in the population obtain the same expected payoffs, and obtain strictly higher payoffs than the population average after introducing a small fraction of invading types.

from a binary distribution, where the high value exceeds her complete-information payoff but is compatible with the opponent's low value. A player receives utility x_i for obtaining a dollar share $x_i \geq \omega_i$, but receives negative utility for a lower share. The unique equilibrium matches AG's war of attrition: high types always demand their reservation value, and low types imitate them before eventually conceding. Uniqueness arises because reservation values do not affect players' intertemporal preferences for dollar shares larger than that value.[14]

Weinstein and Yildiz (2016) show that any (stationary or non-stationary) commitment type's behavior in a repeated game is the unique rationalizable strategy of a utility-maximizing type with different payoffs and information about the stage game. Rational players face the same strategic situation as in the original game with commitment types, while the commitment-behavior types sometimes face types that were absent in the original game (in particular, values may be interdependent, so commitment-behavior types may not know their own payoffs). The permissiveness of this result provides some support for AP's approach of flooding the game with a wide variety of types when commitment is vanishingly unlikely.

Other directions

Further open questions include: What is a tractable model of multilateral reputational bargaining?[15] Can reputational bargaining's powerful equilibrium selection results be further extended in general dynamic games, such as repeated games?[16] How does repeated reputational bargaining unfold?[17] What are the effects of allowing players to gain and lose commitment over time? Does considering commitment types who randomize their behavior or behave non-transparently deliver new insights?

Finally, we began this survey by discussing the overarching ambition of reputational bargaining models to make predictions on the basis of putatively observably factors like players' beliefs about each other's commitment behavior, rather than the details of the bargaining protocol. A crucial question is thus whether these models can predict and explain bargaining field data better than competing models. Such empirical application of reputational bargaining models is currently a wide open area.

[14] Basak also considers the effect of releasing information about the reservation values. Fully informative signals ensure immediate agreement, but partially informative signals may reduce efficiency.

[15] Kambe (2019) analyzes a multilateral, incomplete-information war of attrition with some similarities to reputational bargaining. Ma (2020) shows that in majoritarian bargaining, an agent may benefit from having a lower reputation, because this leads to her inclusion in more winning coalitions.

[16] As in Abreu and Pearce (2007), Atakan and Ekmekci (2013b), and more broadly the literate on reputation in repeated games surveyed by Mailath and Samuelson (2006).

[17] Lee and Liu (2013) consider an incomplete-information repeated bargaining model with some features of reputational bargaining.

References

Abreu, D. and F. Gul (2000). Bargaining and reputation. *Econometrica 68*(1), pp. 85–117. 2

Abreu, D. and D. Pearce (2007). Bargaining, reputation, and equilibrium selection in repeated games with contracts. *Econometrica 75*(3), 653–710. 9, 15

Abreu, D. and D. Pearce (2019). Bargaining, reputation, and equilibrium selection in repeated games without contracts. *Unpublished Manuscript*. 9

Abreu, D., D. Pearce, and E. Stacchetti (2015). One sided uncertainty and delay in reputational bargaining. *Theoretical Economics 10*, 719–773. 11

Abreu, D. and R. Sethi (2003). Evolutionary stability in a reputational model of bargaining. *Games and Economic Behavior 44*(2), 195–216. 14

Atakan, A. E. and M. Ekmekci (2013a, 08). Bargaining and reputation in search markets. *The Review of Economic Studies 81*(1), 1–29. 12

Atakan, A. E. and M. Ekmekci (2013b). A two-sided reputation result with long-run players. *Journal of Economic Theory 148*(1), 376 – 392. 9, 15

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. 2

Basak, D. (2019). Fact-finding and bargaining. *Unpublished manuscript*. 14

Binmore, K., M. J. Osborne, and A. Rubinstein (1992). Noncooperative models of bargaining. *Handbook of game theory with economic applications 1*, 179–225. 1

Compte, O. and P. Jehiel (2002). On the role of outside options in bargaining with obstinate parties. *Econometrica 70*(4), 1477–1517. 12, 14

Crawford, V. P. (1982). A theory of disagreement in bargaining. *Econometrica 50*(3), 607–637. 8

Embrey, M., G. R. Frchette, and S. F. Lehrer (2014, 09). Bargaining and Reputation: An Experiment on Bargaining in the Presence of Behavioural Types. *The Review of Economic Studies 82*(2), 608–631. 14

Fanning, J. (2016). Reputational bargaining and deadlines. *Econometrica 84*(3), 1131–1179. 10

Fanning, J. (2018). No compromise: Uncertain costs in reputational bargaining. *Journal of Economic Theory 175*, 518 – 555. 11

Fanning, J. (2019). Mediation in reputational bargaining. *Unpublished Manuscript*. 13

Fanning, J. and A. Kloosterman (2019). A simple experimental test of the coase conjecture: fairness in dynamic bargaining. *Unpublished manuscript*. 13

Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review 88*(3), 577592. 13

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

Fudenberg, D. and D. K. Levine (1989, July). Reputation and Equilibrium Selection in Games with a Patient Player. *Econometrica: Journal of the Econometric Society 57*(4), 759–778. 6, 7

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

Heggedal, T.-R., L. Hellandy, and M. V. Knutsenz (2020). The power of outside options in the presence of obstinate types. *Unpublished Manuscript*. 14

Kambe, S. (1999, August). Bargaining with imperfect commitment. *Games and Economic Behavior 28*(2), pp. 217–237. 8

Kambe, S. (2019). An n-person war of attrition with the possibility of a non-compromising type. *Theoretical Economics 14*(3), 849–886. 15

Kreps, D. M. (1990a). *A Course in Microeconomic Theory*. Princeton University Press. 2

Kreps, D. M. (1990b). *Game Theory and Economic Modelling*. Clarendon Press. 1, 2

Kreps, D. M. and R. Wilson (1982). Reputation and imperfect information. *Journal of Economic Theory 27*(2), 253–279. 2

Lee, J. and Q. Liu (2013). Gambling reputation: Repeated bargaining with outside options. *Econometrica 81*(4), 1601–1672. 15

Ma, Z. (2020). Efficiency and surplus distribution in majoritarian reputational bargaining. *Unpublished manuscript*. 15

Mailath, G. J. and L. Samuelson (2006). *Repeated Games and Reputations: Long-Run Relationships*. Oxford University Press. 7, 15

Milgrom, P. and J. Roberts (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory 27*(2), 280–312. 2

Myerson, R. (1991). *Game Theory: Analysis of Conflict*. Cambridge, Massachusetts: Harvard University Press. 2

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

Nash, J. F. (1953, January). Two-Person Cooperative Games. *Econometrica 21*(1), 128–140. 9

Özyurt, S. (2014). Audience costs and reputation in crisis bargaining. *Games and Economic Behavior 88*, 250 – 259. 13

Özyurt, S. (2015). Bargaining, reputation and competition. *Journal of Economic Behavior & Organization 119*, 1 – 17. 13

Perry, M. and P. J. Reny (1993). A non-cooperative bargaining model with strategically timed offers. *Journal of Economic Theory 59*(1), 50 – 77. 1

Peski, M. (2019). Bargaining over hetrogenous good with structural uncertainty. *Unpublished manuscript*. 11

Roth, A. E. and M. W. Malouf (1979). Game-theoretic models and the role of information in bargaining. *Psychological Review 86*(6), 574–594. 10, 13

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica 50*(1), pp. 97–109. 1, 6

Rubinstein, A. (1985). Choice of conjectures in a bargaining game with incomplete information. In *Game-Theoretic Models of Bargaining*. Cambridge University Press. 11

Sanktjohanser, A. (2018). Optimally stubborn. *Unpublished Manuscript*. 8

Weinstein, J. and M. Yildiz (2016). Reputation without commitment in finitely repeated games. *Theoretical Economics 11*(1), 157–185. 15

Wilson, R. (1987). Game theoretic analysis of trading processes. *Advances in Economic Theory*. 1

Wolitzky, A. (2011). Indeterminacy of Reputation Effects in Repeated Games with Contracts. *Games and Economic Behavior 73*(2), 595–607. 9

Wolitzky, A. (2012, September). Reputational Bargaining with Minimal Knowledge of Rationality. *Econometrica 80*(5), 2047–2087. 10