

0.1 General Results with Noisy Evolution

Still, there may be concern that these results are specific to case where the noise in the evolutionary process is uniform or may be sensitive to other simplifying assumptions. In this section, we first show that the existence and uniqueness of a stable distribution of preferences where conflict occurs with positive probability and the comparative statics on the cost of conflict hold for a wide class of noise distributions, while maintaining that only the type with the highest fitness payoff reproduces. We then proceed to the more challenging case where multiple types reproduce, where we can still show that conflict must occur in any stable equilibrium.

We consider noise distributions with the following properties

Theorem 1. *For any noise distribution G such that $G(0) \in (0, 1)$ with a continuous and differentiable and single-peaked density g ,*

- i. there exists a unique single reproducer stable preference distribution where the probability of conflict is strictly positive, and*
- ii. if the noise distribution has an upper bound $\bar{\epsilon}$, then for k sufficiently large $\beta^* = k - \bar{\epsilon}/2$, and the probability of conflict is:*

$$p_c^* = \int_{\nu_1=0}^{\bar{\epsilon}} (1 - G(\bar{\epsilon} - \nu_1))g(\nu_1)d\nu_1 > 0 \quad (1)$$

Proof If the previous generation reproducer was type β_m , then the distribution of the current generation types is given by $G(\beta - \beta_m)$ with density $g(\beta - \beta_m)$. So, the expected payoff for being type β_j when β_{-j} is drawn from this distribution can be written:

$$\Pi(\beta_j; k, \beta_m) = \int_{-\infty}^{2k-\beta_j} \left(v + \frac{\beta_j - \beta_{-j}}{2} \right) g(\beta_{-j} - \beta_m) d\beta_{-j} + \int_{2k-\beta_j}^{\infty} (v - k) g(\beta_{-j} - \beta_m) d\beta_{-j}$$

We first consider the case where g has full support on \mathbb{R} . In this case Π is continuous with respect to β_j , and by Leibniz's rule the derivative is:

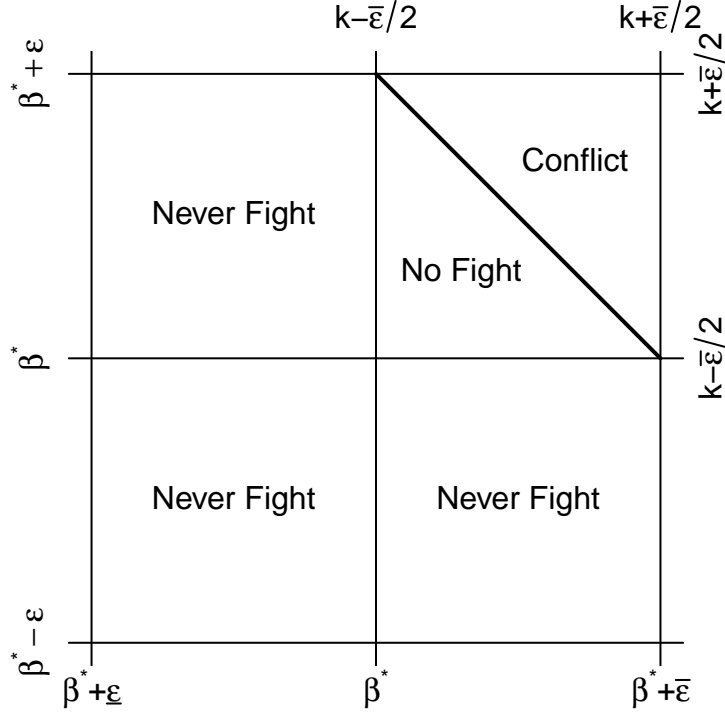
$$\frac{\partial \Pi}{\partial \beta_j} = g(2k - \beta_j - \beta_m) \left[(v - k) - \left(v + \frac{\beta_j - (2k - \beta_{-j})}{2} \right) \right]_{\beta_{-j}=2k-\beta_j} + F(2k - \beta_j)/2 \quad (2)$$

If g is bounded $2k - \beta_j$ is below the support of g (i.e., β_j is high enough to always fight), this reduces to $v - k$. If $2k - \beta_j$ lies above the support of g (i.e., a deal is always struck), this reduces to

Parts i provides a more general statement of our main results about the inevitability of conflict. The logic is the same as described above: in any distribution of toughness which is entirely peaceful, the toughest type is the one that reproduces, generating even tougher offspring.

Part ii says that if the noise distribution has an upper bound, not only must conflict occur, but there is a lower bound on the probability of conflict greater than zero. It would seem that adding an upper bound to how much tougher children can be from their parents would make it easier to sustain peace. However this intuition ignores how such an upper bound affects the long-run distribution

Figure 1: Illustration of the lower bound on conflict for proposition 1. Types below β^* on either axis never fight, while those above β^* fight if their partner is sufficiently tough.



of preferences. Without an upper bound on the noise, when conflict becomes extremely costly, the probability of conflict can go to zero.* However, when such an upper bound exists and conflict is costly, the ideal type to be is always the toughest one who never fights – which is impossible without an upper bound. When the preference distribution is centered around the toughest type that never fights, any actor that is tougher than their parent will *sometimes* fight.

Figure 1 gives a graphical illustration of this point for a noise distribution with bounds $[\underline{\epsilon}, \bar{\epsilon}]$ (the lower bound clarifies the illustration but is not necessary for the result). Each axis corresponds to the toughness distribution for the types, so the square corresponds to all possible pairs of toughness in the stable distribution. When conflict gets extremely costly, the stable preference distribution is always centered around $\beta^* = k - \bar{\epsilon}/2$, as this implies the toughest type is $\underline{\beta} = k - \bar{\epsilon}/2 + \bar{\epsilon} = k + \bar{\epsilon}/2$. So, when β^* is matched with the toughest type their aggregate toughness is exactly $2k$ and a deal is struck.

So, in the stable preference distribution, anyone who is tougher than β^* will fight the toughest types. In particular, the diagonal line in the top right corner corresponds to the line where $\beta_1 + \beta_2 = 2k$, so for any pair of actors above this line conflict will occur. This corner corresponds to $1/8$

*However, if there is no upper bound on the noise distribution, then the result that conflict must happen is trivial: no matter what the ideal toughness level, there are always arbitrarily tough offspring who fight.

of the entire square, illustrating why with a uniform distribution this is the lower bound on the probability of conflict. For non-uniform distributions, equation 1 measures the likelihood of having a sufficiently high sum of toughness, and hence the probability of conflict. Depending on the precise noise distribution there will be more or less density in this corner, but it will never be empty.

The fact that the probability of conflict can remain non-trivial even as fighting because arbitrarily costly has implications for how changes in military technology – in particular, the spread of nuclear weapons – affects the possibility and expected costs of interstate war (?). Even if nuclear weapons do make war less likely, this effect may be offset by the increased cost of conflict should war occur. Of course, if, for example, nuclear weapons make war so costly (i.e., when k is arbitrarily large) that the probability of occurring is zero, the expected costs from war will also be zero. However, our model suggests that when war is extremely costly, this probability of war never approaches zero, as the actors always continue to get tougher in their bargaining stances. So, the threat of extremely costly war may not be enough to ensure it does not occur if preferences are endogenous.

Finally, our most general result considers an evolutionary process where more than one type reproduces. Formally, there is a “weight function” $w : \mathbb{R} \rightarrow \mathbb{R}_+$ which determines how many offspring each player gets as a function of their fitness payoff (where a normalization ensures that the population size is constant). Again, offspring have a toughness equal to their parent’s plus a noise term ν drawn from density g with both positive and negative support (i.e., $G(0) \in (0, 1)$, where G is the corresponding cumulative density function). So, a distribution of preferences is stable if the density function resulting from this process is equal to the initial density f , or:

$$f^*(\beta) = \int f^*(\beta - \nu) \frac{w(\Pi(\beta; f^*, \sigma^*))}{\int w(\Pi(\beta; f^*, \sigma^*)) f^*(\beta) d\beta} g(\nu) d\nu \quad (3)$$

Our solution concept for evolution this general process is as follows:

Definition A density of types f^* and strategy profile σ constitute a *Stable Noisy Preferences-Subgame Perfect Equilibrium* (SNP-SPE) if:

- (1) σ^* is derived from lemma ??.
- (2) f^* and σ^* satisfy equation 3.

While it is difficult to derive conditions when such a stable distribution exists, we can easily show that or any cost of conflict, weight function, and noise distribution, conflict must occur with positive probability in any stable distribution:

Theorem 2. *There is no SNP-SPE equilibrium where conflict never happens.*

Proof See the appendix.

The intuition for this result is a generalization of the previous results about the inevitability of conflict. In any preference distribution such that conflict never occurs, the next generations must have some offspring from the toughest members of the population. Hence some of these offspring will have a higher level of toughness than anyone in the previous generation.

1 Extensions

While some aspects of the results in the previous section are quite general, we should admit caveats. First, the bargaining protocol used is very simple. Still, we expect the result will generalize: in a proposed world without conflict, there are evolutionary incentives to be more willing to fight, since without the realization of conflict there is no drawback to having such preferences. As long as the toughest type is best off in a proposed equilibrium with no conflict, such an equilibrium is impossible. (See ? and for conditions in a large class of bargaining games where this will hold, albeit without perfect information.)

Second, we only allow for preferences to diverge from objective payoffs in a specific way. Third, we assume that each player knows both their type (not too unrealistic) and the type of their adversary (much more problematic). We address both of these concerns in the extensions, first by allowing different levels of toughness based on being in the proposer or responder role, and second by allowing the types to only be sometimes observed.

1.1 More General Preferences

A strong simplifying assumption we make in the analysis in the previous section is to only allow the preferences to diverge from payoffs in a single way, by increasing the value attached to conflict regardless of whether in the proposer or responder role. However, since the deal made is given by the responder reservation point, there is no value to developing a taste for fighting in the proposer role. So, our first extension with more general preferences allows the toughness to vary based on whether the actor is in the proposer or responder role.

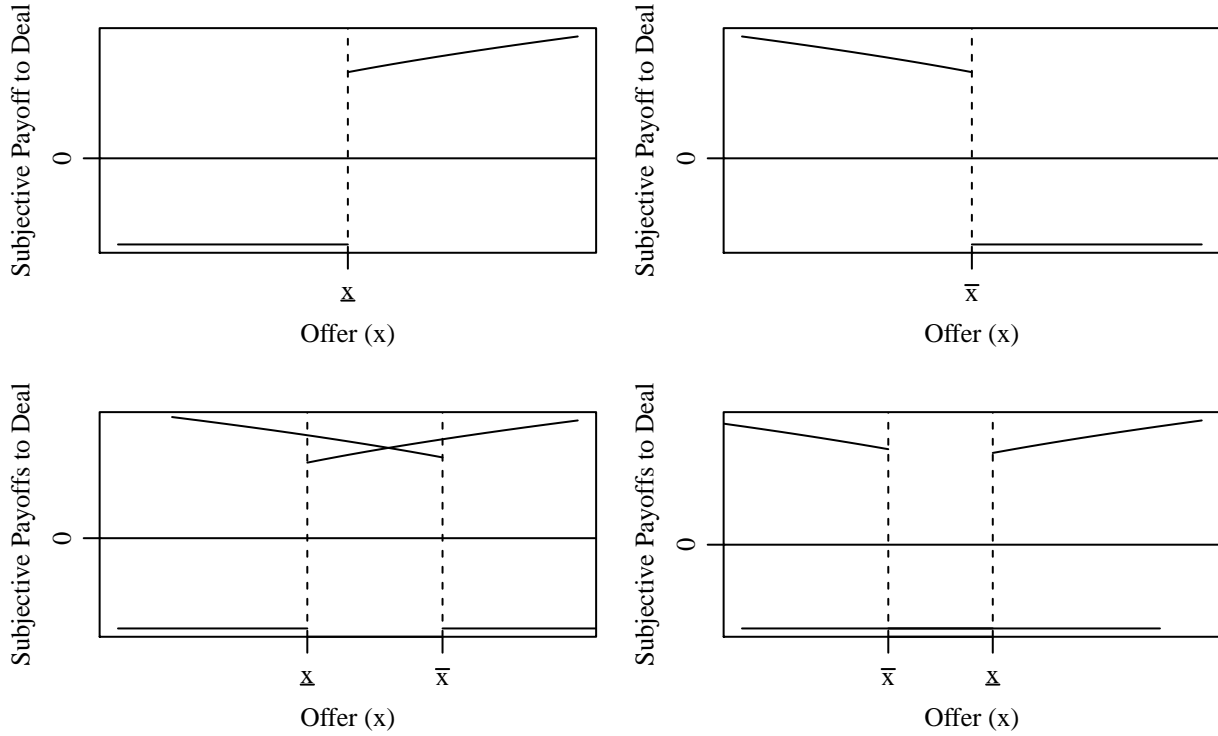
In particular, suppose the payoff to fighting when in the proposer role is $v - k + \beta^p$ and when in the responder role is $v - k + \beta^r$. We search for a stable distribution of preferences such that when β^i is uniformly distributed on $[\beta^{i,*} - \epsilon^i, \beta^{i,*} + \epsilon^i]$, for $i \in \{r, p\}$, a type with toughness $\beta^* = (\beta^{p,*}, \beta^{r,*})$ gets the highest fitness payoff.

As indicated above, having a toughness not equal to 0 in the proposer can only harm the fitness of an actor: if $\beta^p > 0$ they may end up fighting partners when reaching an agreement would be better, and if $\beta^p < 0$ they may end up making deals with partners when fighting would give a higher objective payoff. So, in any stable preference distribution, $\beta^{p,*} = 0$.

The tradeoffs for increasing the taste for fighting in the responder role are similar to the baseline model, but since proposers have toughness centered at 0 this allows responders to be even tougher. In fact, when the noise in the evolutionary process is the same for the proposer toughness and responder toughness (i.e., $\epsilon^r = \epsilon^p$), the equilibrium toughness in the responder role is exactly twice the equilibrium toughness in the main model, and the probability of conflict is unchanged.

In the stable preference distribution, there are some types that have positive toughness in the proposer role because of the noise in the evolutionary process, and this makes them more apt to fight. So, it may seem that conflict would be less likely if there were less noise in the toughness of proposers. Interestingly, when the cost of conflict is high, the opposite is true. In fact, when k is large and $\epsilon^p \rightarrow 0$, the probability of conflict is $1/2$, four times that in the baseline model. This is because when conflict is costly, the optimal toughness in the responder role is to be as tough as possible while never fighting. However, when $\epsilon^p \rightarrow 0$, the level of toughness in the responder

Figure 2: Illustration of “sacred value” preferences.



role which never fights approaches $2k$. Given this distribution of preferences, any type which is tougher than the parent will *always* fight.[†] So, not allowing preferences to diverge from fitness in a way that leads to “unnecessary” toughness actually leads to *more* conflict.

Next, we consider a very different type of deviation from the objective preferences, intended to capture the idea that the actors assign a “sacred value” to attaining a certain share of the prize. For example, assigning a value to fairness can be captured by assuming an actor faces a large negative shock to their preferences when accepting a deal that gives them less than v (that is, $x < v$ for the responder and $x > v$ for the proposer). If what is being bargained over includes a piece of land that both actors – here, perhaps ethnic groups – believe to be sacred, then they can assign an arbitrarily low subjective payoff to any accepted deal which does not give them all of this land. Similarly, ? finds that citizens may punish leaders for compromising on issues that are “moralized”, which could certainly apply to many issues where disagreement may lead to conflict.

Figure 2 illustrates how we formalize this idea, and the resulting behavior in the bargaining game. The top left panel illustrates the preferences for a responder with “ \underline{x} –sacred value preferences,” which we assume means they assign an arbitrarily low value to any accepted deal which gives them less than \underline{x} . We normalize the payoff to conflict to zero, so any negative value assigned

[†]Conversely, as $\epsilon^r \rightarrow 0$, the probability of conflict approaches 0 when k is high, see the appendix.

to deals less than \underline{x} results in equivalent behavior. Finally, the payoff to deals better than \underline{x} is increasing in x , though this effect can be qualitatively “small” relative to the importance of getting at least \underline{x} .

The right panel illustrates the analogous preferences for a proposer with \bar{x} –sacred value preferences. A player with these preferences assigns an arbitrarily low subjective preference to any offer greater than \bar{x} – equivalently, any offer which leaves them with less than $2v - \bar{x}$.

The bottom panels illustrate the equilibrium behavior for actors with these preferences. In the bottom left, the minimal value demanded by the responder (\underline{x}) is less than the maximal value the proposer is willing to give up (\bar{x}), and hence there is a range of mutually acceptable deals. So, by familiar logic, the proposer will offer \underline{x} which will be accepted. In the bottom right, $\underline{x} < \bar{x}$, indicating there is no deal which does not violate at least one players values. So, conflict will occur.

So, equilibrium behavior is determined only by where the discontinuity of the payoffs lie. As demonstrated in the appendix, identical results arise in a model where the \underline{x} and \bar{x} for each actor are guided by a noisy evolutionary process. That is, if \underline{x} and \bar{x} for each actor is equal to their parent’s thresholds plus uniform noise, the equilibrium probability of conflict is the same as the case where we parameterize deviations from fitness with changes in the conflict payoff.

Finally, this example illustrates that any set of preferences where the proposer has a maximally acceptable deal and the responder has a minimally acceptable deal result in the same equilibrium behavior. As shown in the appendix, for a general class of preferences and assumptions about how they evolve the general conclusions from the model with role-based toughness are unchanged.

These results suggest an interesting connection between several canonical findings in behavioral economics and the study of conflict. First, as argued in ?, the fact that actors develop different levels of toughness depending on their role in the bargaining process provides an explanation for the “endowment effect”, a strong empirical regularity that people place more value on objects that they own.[‡]

This interpretation interacts in an interesting way when modeling sacred value preferences. If groups adapt preferences that make striking a deal that forfeits a certain piece of land which is “theirs” unacceptable, this can be valuable in retaining this piece of land from potential invaders.[§] If multiple groups develop sacred values and aversion to territorial divisibility, then peace agreements may be more difficult if not impossible (?).

****[THIS DOES NOT MAKE A TON OF SENSE TO ME.....]*** Further, when these preferences form around objects or pieces of land which are at focal locations or not obviously divisible (a Temple, land between two particular rivers), it may make less sense to model noise in the evolutionary process that transmits these preferences to future generations. That is, if the sacred value preference is over controlling a well-demarcated piece of land or a building, it would seem odd for the next generation preferences to allow for controlling an island minus a tiny sliver or a building minus a section of the walls. So, the forces described here can endogenizing in a mostly rationalist framework many of the arguments for why bargaining over sacred land may be intractable in the

[‡]While being in the proposer or responder role does not directly correspond to being a buyer or seller, a more general model where one player owns a piece of land being bargained over could provide similar results to (?).

[§]This dynamic is seen in so-called “sons of the soil” conflicts (?).

long term.

1.2 Imperfect Observation of Types

One objection to the analysis above is that it assumes the types of the players are always common knowledge, which is particularly problematic given the key role that incomplete information about preferences plays in the literature on bargaining on conflict (e.g., ???).

Here we show that a central conclusion from this past work – that incomplete information is a key cause of conflict – becomes less straightforward when preferences are endogenous. This results from the fact that willingness to reject offers can only convey an evolutionary advantage if the proposer knows the responder’s taste for fighting. So, while making it harder to observe the type of the responder leads to more conflict *for fixed preferences*, it can also lead the actors to evolve less belligerent preferences, potentially leading to less conflict.

To simplify, we build on the model of role-based toughness, and assume that all players have standard preferences when in the proposer role (i.e., $\beta^p = 0$) but get subjective payoff $v - k + \beta^r$ when rejecting in the responder role. When two players are matched, the type of the responder is observed with probability q , and if not the proposer only knows the distribution of toughness in the population. So, the proposer faces a standard risk-reward tradeoff of reaching more bargains but getting worse bargains when making higher offers. As above, we assume the toughness in a given generation is equal to the toughness that attained the highest fitness payoff in the previous round plus uniform noise.

As demonstrated in the appendix, there are two possible stable preference distributions. When q is high, there is not enough of a chance that the type is unobserved for the stable preference to differ from the case of complete observation. So, when the type is observed there is sometimes conflict as above. When the type is unobserved, the proposer makes an offer which is never accepted.

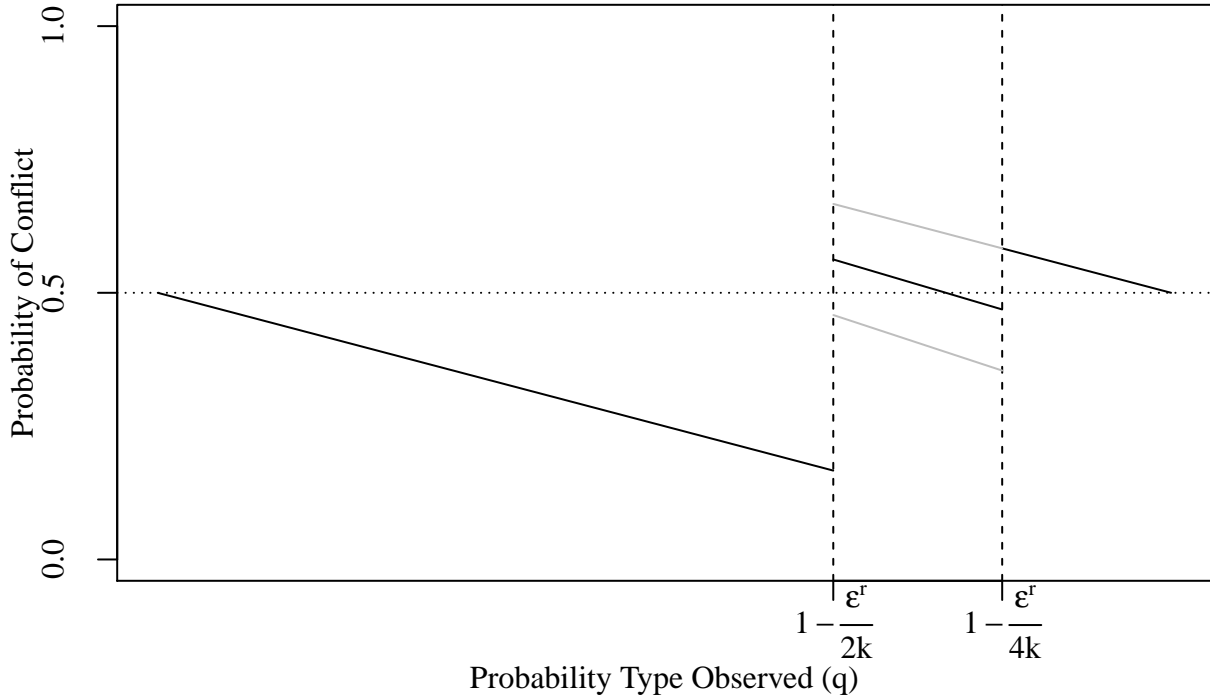
When the probability that the proposer observes the responder type (q) is low, the benefit to being tough decreases to the point where the players evolve less belligerent preferences. In particular, a bargain is always struck when the type is observed. When the type is unobserved, the proposer makes an offer which is sometimes accepted when the type is unobserved, and the type that attains the highest payoff is the toughest type that accepts this offer.

There is also an intermediate range of q where there is no stable preference distribution, but it is easy to characterize a “preference cycle” where alternating generations are relatively more or less belligerent.

Figure 3 shows how changing the probability of the responder type being observe affects the probability of conflict in the stable preference distribution. For the first and third interval, there is a unique stable preference distribution as described above. In the intermediate interval, generations alternate between a relatively tough and less tough average toughness, with the grey lines tracing the probability of conflict in these two cases. The black line in this interval is the average probability of conflict across generations.

Other than the points where the stable preference distribution changes, the probability of conflict is always decreasing in q , in line with standard arguments about how incomplete information fuels conflict (??). However, making preferences easier to observe also leads to discrete increases in the probability of conflict when the stable preference distribution does change. In fact, in the

Figure 3: Equilibrium Probability of Conflict with Partially Observed Preferences



“highest” range of q – corresponding to the case where the type is almost always observed and hence the type that gets the highest objective payoff is one who always fights when the type is unobserved but not when the type is observed – has a higher probability of conflict than any value of q where the highest fitness is attained by a type who does not fight when the type is unobserved.

In addition to the abstract contribution to the study of incomplete information and conflict, these results also have implications for how changes in technology affect the likelihood of conflict. For example, some have argued that international organizations or better intelligence gathering reduce uncertainty about rival state’s capabilities or willingness to fight, which can then reduce the possibility of conflict (???, e.g.,). If the evolutionary forces described here can act “quickly”, then the model in this section provides a potential counter to this: while making preferences more observable unambiguously leads to more conflict in the short term, it may lead to states or leaders (a point we discuss further in the following section) developing even tougher preferences in the long term, which can mitigate or even reverse the short-term effect.

Whether evolution on such a short time scale makes sense depends on the mechanisms of evolution. Biological evolution takes place on a wider time scale which is unlikely to react quickly to shifts in the strategic environment. However, if states learn to elect the kinds of leaders who are successful in the recent past, it may be plausible to conceptualize “generations” as turnover in leadership, which may be quite fast. So, if changes in the observability of preferences, cost of

conflict, or other factors that affect international negotiation change rapidly, it may only take a few generations for citizens to start selecting different kinds of leaders with different preferences.

2 Conclusion

The idea of a bargaining breakdown leading to conflict plays a particularly important role in the literature on interstate war. Yet most previous theories do not explicitly model the evolution of aggressive (versus non aggressive) types. Given recent theoretical and empirical insights of behavioral economics and psychology on variation in aggression, and political science on the importance of leader-specific characteristics influencing crisis behavior, this is an important gap. We provide a theoretical mechanism – evolution of preferences – that bridges these insights with standard political science models of interstate war. Finally, we argue that our results also provide a mechanism for understanding recent empirical work on the importance of the personalities of leaders (??). Our paper demonstrates the importance of environmental context for selecting certain traits (aggressiveness), and how these traits can subsequently influence the bargaining process, and likelihood of conflict.

This evolutionary process can be further fueled by the conflict itself, with individuals becoming “stuck” in the cycle of violence, and refusing to compromise ??.[¶] To pick a prominent example of the a dispute that plausibly lacks a peaceful solution (at least to date), the Israeli-Palestinian conflict, Hamas spokesman Fawzi Barhoum was quoted on July 8, 2014 about possible negotiations to stop fighting between Israel and Hamas in July-August of 2014:

“Today there is no intention of relaxation and calm. Palestinian blood has been spilled. There is no place for talking about peace with the Israeli occupation. If they want to protect their entity from Hamas’s missiles, they will have to put an Iron Dome (Israel’s missile defense shield) on every home in Israel.”

On the other side, hardline Israeli Economy Minister Naftali Bennett said of the recent Gaza conflict:

“Only a decisive victory (in Gaza) will prevent the next war. Radical Islam seeks to erase the Jewish state from the face of the earth. They do not seek a strip of land or a Palestinian state. Only our annihilation. (Ayatollah) Khamenei in Iran. Nasrallah in Lebanon. Haniyeh in Gaza. Meshaal in Qatar. All of the jihadist arms that are waiting to see how we respond.”

Both quotes highlight the possibility that (broadly construed) environments in which leaders operate can shape preferences that make leaders even more predisposed to belligerence. Thus violence may shift preferences of both the population (?), and leaders that precludes a peaceful settlement. These processes are not unique to the Israeli-Palestinian conflict, but characterize

[¶]We recognize that violence can affect strategic incentives and reveal information (??). However, here we are only referring to how the violence can shift preferences, and does not reveal strategic information. For example, see ? for an empirical example on how exposure to violence shifts risk and other preferences.

vendettas (i.e. Hatfield and McCoys (?)), other international disputes (India-Pakistan (?)), and civil wars (Iraqi and Syrian Civil Wars (?)).

More speculatively, our model shows that arguments about certain conflicts being “intractable” because there are no peaceful agreements that all sides prefer to fighting (setting aside information, commitment, and other problems) is less naive than the extant bargaining literature claims. Of course, the fact that many participants in conflict make claims that they are unwilling to compromise is not dispositive evidence that their preferences admit no peaceful agreement. However, the fact that we have theories where conflict occurs even though preferences *do* admit a peaceful agreement – and are also generally consistent with actors claiming to be less willing to compromise than they truly are – is not dispositive evidence that such a bargaining range must always exist. Exploring how the evolutionary forces we model here interact with these other explanations – as complements or substitutes – could provide a promising new avenue for conflict research.

Appendix

Let $\pi(\beta_j; \beta_{-j}, \sigma)$ be the expected objective fitness for a player of toughness β_j when matched with a partner with toughness β_{-j} and they use strategies $\sigma = (\sigma_1, \sigma_2)$ when in the subscripted role in the bargaining game.

As derived above, these strategy profiles depend on the types of the players, which for now we assume are common knowledge. So, a strategy for the proposer $\sigma_1(\beta_1, \beta_2) \mapsto \mathbb{R}$ specifies an offer made as a function of the types. A strategy for the responder $\sigma_2(\beta_1, \beta_2, x) \mapsto \{\text{accept, reject}\}$ specifies what offers are accepted as a function of the types.

If the distribution of toughness levels in the population is given by a distribution function F , then the expected fitness payoff for a player with toughness level β is:

$$\Pi(\beta; F, \sigma) = \int \pi(\beta; \beta_{-j}, \sigma) dF(\beta_{-j})$$

Our static equilibrium concept is:

Definition A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*)$ and preference distribution F comprise a *Stable Preferences-Subgame Perfect Equilibrium* (SP-SPE) if:

- (1) $(\sigma_1^*(\beta_1, \beta_2), \sigma_2^*(\beta_1, \beta_2))$ is a SPNE of the bargaining game for all $(\beta_1, \beta_2) \in \mathbb{R}^2$.
- (2) $\text{supp}(F) \in \arg \max_{\beta} \Pi(\beta; F, \sigma^*)$

The first part states that the outcome of the bargaining game is a SPNE given the subjective payoffs of the players, i.e., the players behave optimally given their preferences. The second part states that all feasible types (i.e., with positive probability or density) get the highest possible objective payoff when playing strategies meeting the first condition. That is, no “invader” type β' would get a higher fitness payoff when facing a population with distribution F , and all types in F get the same fitness payoff.

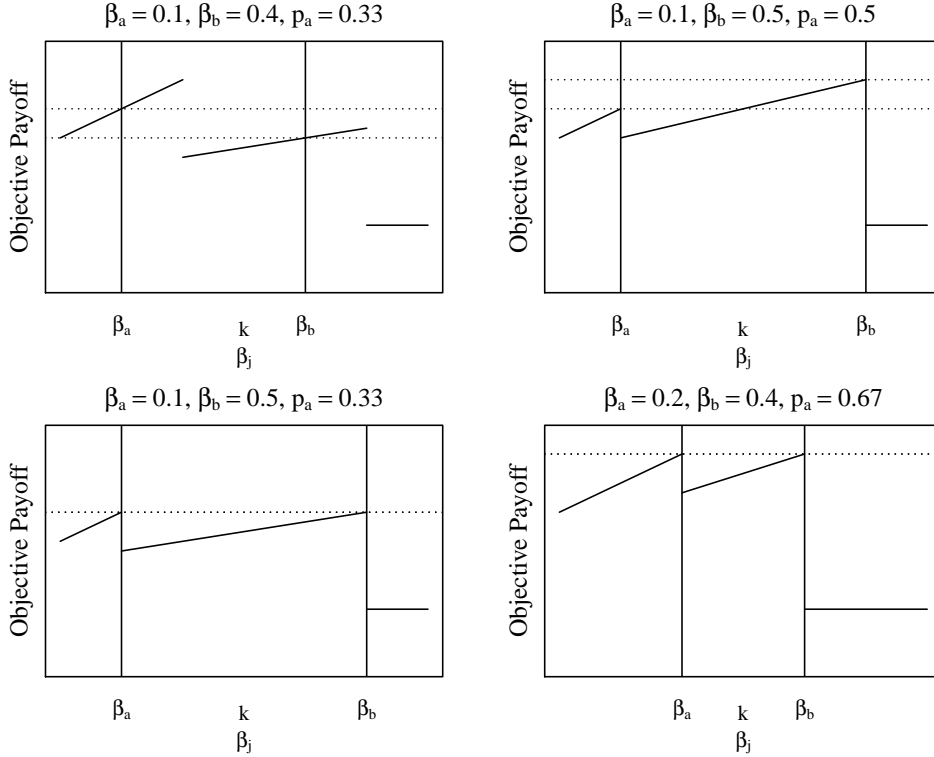
We first characterize a class of equilibria where the players have either one or two types in equilibrium. In particular, suppose the players are type β_h with probability $p_h \in [0, 1]$ and type $\beta_l < \beta_h$ with probability $1 - p_h$.^{||} That is, the “high” (h) or tough types get a higher subjective payoff from fighting than the “low” (l) or weak types.

Figure 4 illustrates the equilibrium condition for a two-type equilibrium (see the appendix for a formal derivation). Each panel plots the objective payoff for being type β' when population which has proportion p_h of the β_h types and proportion $1 - p_h$ of the β_l types. For this distribution of preferences to be a part of an SP-SPE, it must be the case that both β_l and β_h (the vertical lines) are global maxima of the objective payoff function. So, the type distributions in the top two panels cannot comprise an SP-SPE, while the bottom two panels do.

The top left panel is not an SP-SPE because neither β_l or β_h correspond to local maxima. This is because being marginally tougher than a β_l or β_h type leads to better bargains without increase the chance of fighting. To ensure that both types are at local maxima, it must first be the case that being slightly tougher than β_l leads to conflict with the tough type, or $\beta_l + \beta_h = 2k$. This condition

^{||}As the labels “h” and “l” are chosen for mnemonic ease, the $\beta_l < \beta_h$ assumption is without loss of generality.

Figure 4: Illustration of Proposed SP-SPE. In the top two panels, β_l and β_h are not global maximizers of the objective fitness function, and hence they do not represent equilibria. In the bottom panels, β_l and β_h and both global maxima, indicating an equilibrium



also implies that becoming marginally tougher than the tough type leads to conflict with the weak type, and hence ensures that β_h is a local maximum as well.

The top right panel of figure 4 shows an example where this condition holds, but is still not an SP-SPE because the tougher (β_h) types – which constitute half of the population – get a higher objective payoff. By the intuitions developed in section ??, we might then expect the tough types to reproduce faster, leading to more tough types.

The bottom left panel shows that this intuition holds. By increasing the proportion of tough types from $1/2$ to $2/3$, the weak and tough types now get the same objective payoff, and hence this distribution of types and the strategies in lemma ?? constitutes an SP-SPE. This is not the only SP-SPE: as shown in the bottom right panel, there is another equilibrium where the types are “closer” to each other and the proportion of weak types is higher.

There are two important points to take from this graph. First, there are multiple SP-SPE; in fact, as shown below for any $\beta_l \in [0, k]$ there will be a corresponding $\beta_h > \beta_l$ and p_h that constitute an equilibrium. Second, the probabilities of conflict are different in these equilibria. In particular, the probability of conflict is p_h^2 , which is $4/9$ in the bottom left panel and $1/9$ in the bottom right panel.

This is indicative of a general pattern: when the types are closer together there is less conflict. Analogous to the model in section ??, let $\delta = \beta_h - k = k - \beta_l$ measure the distance (divided

by two) between the types, which also represents the level of *exploitation* in the bargaining game. That is, the larger δ , the larger a proportion of the prize taken by the tough type when matched with a weak type.

In order to balance the higher level of exploitation when δ is large, the high types must engage in inefficient conflict more often in order to get the same objective payoff as the weak types. Algebraically, the expected objective payoff for the tough type is $v + (1 - p_h)\delta - p_h k$, and the objective payoff for the weak type is $v - p_h \delta$ (see the appendix for more detail). To ensure both types get the same objective payoff it must be the case that these are equal, which simplifies to:

$$p_h = \frac{\delta}{k}$$

That is, when the level of exploitation δ is high, the proportion of tough types must be high, meaning there is a high level of conflict.** More generally:

Proposition 3. *For any $\delta \in [0, k]$, the model has a SP-SPE where $\beta_l = k - \delta$, $\beta_h = k + \delta$, $p_h = \frac{\delta}{k}$, and the probability of conflict is $p_c = \left(\frac{\delta}{k}\right)^2$.*

Proof See the appendix.

Proof of Proposition 3

For a two-type class of preference distributions, the expected fitness payoff to being type β given σ^* is:

$$\Pi(\beta; \beta_l, \beta_h, p_h, \sigma^*) \equiv p_h \pi(\beta; \beta_h, \sigma^*) + (1 - p_h) \pi(\beta; \beta_l, \sigma^*)$$

For the triple (β_l, β_h, p_h) to be a part of a SP-SPE, it must be the case that actors with toughness parameters β_l and β_h get same payoff as each other, and that no “invader” with a different toughness parameter would get a higher payoff. Formally, for any $\beta' \in \mathbb{R}$:

$$\Pi(\beta_l; \beta_l, \beta_h, p_h, \sigma^*) = \Pi(\beta_h; \beta_l, \beta_h, p_h, \sigma^*) \geq \Pi(\beta'; \beta_l, \beta_h, p_h, \sigma^*) \quad (4)$$

Given the equilibrium strategies derived in lemma ??, the fitness for type β' given a two-type distribution is:

$$\Pi(\beta'; \beta_l, \beta_h, p_h, \sigma^*) = \begin{cases} (1 - p_h) \left(v + \frac{\beta' - \beta_l}{2} \right) + p_h \left(v + \frac{\beta' - \beta_h}{2} \right) & \beta' \leq 2k - \beta_h \\ (1 - p_h) \left(v + \frac{\beta' - \beta_l}{2} \right) + p_h (v - k) & \beta' \in (2k - \beta_h, 2k - \beta_l] \\ v - k & \beta' > 2k - \beta_l \end{cases}$$

This is a piecewise linear function with discontinuities at $\beta' = 2k - \beta_h$ and $\beta' = 2k - \beta_l$. Other than at the discontinuities, for $\beta_j < \beta_h$ the payoff is increasing – i.e., tougher types get a higher payoff due to getting better deals when in the responder role without leading to more conflict. So, the condition for both types to be at local maxima is $\beta_l + \beta_h = 2k$.

**To connect to the analysis in section ??, note that the deadweight loss from conflict there was equal to $1/2$, and this equation becomes $p_h = 2\delta$ when $k = 1/2$.

For both types to get the same payoff, it must be the case that:

$$\begin{aligned} (1 - p_h) \left(v + \frac{2k - \beta_h - \beta_l}{2} \right) + p_h \left(v + \frac{2k - \beta_h - \beta_h}{2} \right) \\ = (1 - p_h) \left(v + \frac{2k - \beta_l - \beta_l}{2} \right) + p_h (v - k) \end{aligned}$$

which simplifies to

$$(1 - p_h)v + p_h(v - \delta) = (1 - p_h)(v + \delta) + p_h(v - k) \implies p_h = \delta/k$$

where $\delta = k - \beta_l$.

The probability of conflict follows from the fact that conflict occurs if and only if two high types are matched, which happens with probability p_h^2 .

Other Classes of Equilibria

The definition of an SP-SPE allows for any distribution for the β 's. The analysis above characterizes all SP-SPE with two types. To demonstrate that our central results are not sensitive to this restriction, we demonstrate that (1) No type can have a $\beta < 0$ in an SP-SPE, (i.e., be irrationally conciliatory) (2) for any finite n , there is a class of SP-SPE with properties similar to the two-type equilibrium, and (3) there is no SP-SPE that admits a density.

Proposition 4. *In any SP-SPNE, $Pr(\beta_j \leq 0) = 0$*

Proof The intuition behind the result is to divide the pool of types into those with a toughness less than or equal to $2k$ and those with a toughness strictly greater than $2k$. There is no incentive to have a toughness less than 0 against the first group because one can always become tougher and get better deals without fighting. There is also no incentive to have toughness less than zero against the latter group because any deal than can be struck with them is worse than fighting. So, it is always strictly better to be type $\beta' = 0$ than any $\beta' < 0$.

Formally, write the objective payoff for being type β' given type distribution F as:

$$\begin{aligned} \Pi(\beta', F) = \int_{-\infty}^{2k} \left(\mathbf{1}_{\beta' + \beta_{-j} \leq 2k} \left(v + \frac{\beta' - \beta_{-j}}{2} \right) + \mathbf{1}_{\beta' + \beta_{-j} > 2k} (v - k) \right) dF(\beta_{-j}) \\ + \int_{2k}^{\infty} \left(\mathbf{1}_{\beta' + \beta_{-j} \leq 2k} \left(v + \frac{\beta' - \beta_{-j}}{2} \right) + \mathbf{1}_{\beta' + \beta_{-j} > 2k} (v - k) \right) dF(\beta_{-j}). \end{aligned}$$

I.e., the first integral captures the payoff from being matched with a $\beta_{-j} \leq 2k$ type and the second being matched with a $\beta_{-j} > 2k$ type. There are two cases to consider:

- i. If $F(2k) = 1$ (i.e., all types are less than $2k$), then the second integral drops out and $\beta' + \beta_{-j} \leq 2k$ for $\beta' < 0$, hence the objective payoff is strictly increasing for $\beta' < 0$.
- ii. If $F(2k) < 1$, then $Pr(\beta_{-j} > 2k) > 0$. For the range $\beta' < 0$, $\beta' + \beta_{-j} \leq 2k$ for all β_{-j} corresponding to the first integral, so the payoff is strictly increasing in β' in this range (i.e., one

gets better deals without fighting this group by getting tougher). If $\beta' < 0$ and $\beta_{-j} > 2k$, then $v + \frac{\beta' - \beta_{-j}}{2} < v - k$, so increasing β' can only lead to more fighting among this group, but fighting gives a higher payoff than striking a deal, so the objective payoff when matched against a $\beta_{-j} > 2k$ is weakly increasing for $\beta' < 0$.

So, in either case the objective payoff is strictly increasing for $\beta' < 0$ violating the condition for a SP-SPE which places positive probability on $\beta_j < 0$. ■

Proposition 5. *For any finite integer m , i. there exists a class of SP-SPE with m distinct types, and*

ii. in this class of equilibria conflict can occur with any probability greater than 0.

As in the two type case, there is always an equilibrium where all types have toughness greater than $2k$ and always fight. So, what remains is to show that there are equilibria with an interior probability of conflict.

First consider distributions with an even number of types, so the number of types can be written $m = 2n$ for some integer n . Order the types such that $\beta_1 < \beta_2 < \dots < \beta_{2n}$, and let $Pr(\beta_j = \beta_i) = p_i$. The payoff for being type β' in such an equilibrium is:

$$Pr(\beta_{-j} \leq 2k - \beta') \left(v + \frac{\beta' - \mathbb{E}[\beta_{-j} | \beta_{-j} \leq 2k] - \beta'}{2} \right) + (1 - Pr(\beta_{-j} \leq 2k - \beta'))(v - k)$$

As long as $Pr(\beta_j < 2k) > 1$, the objective payoff for being $\beta' = 2k - \beta_1$ is strictly higher than the payoff to being $\beta' > 2k$, so $\beta_{2n} < 2k$. Further $\beta_j \geq 0$, so this is a piecewise linear function that is weakly increasing on each segment (and strictly increasing for $\beta' < 2k$, with discontinuities at $2k - \beta_j$ for all $j \in \{1, \dots, 2n\}$). So for each type to be at a local maximum, a more general symmetry condition must hold:

Lemma 1. *In any finite even type distribution with $2n$ types, it must be the case that $0 < \beta_1 < \dots < \beta_n < k$ and $\beta_i = 2k - \beta_{2n+1-i}$.*

Proof If not, the objective payoff must be strictly increasing at some $\beta' = \beta_j$, violating the condition for a SP-SPE. If this condition holds, all β_j are at a local maximum.

Given this restriction, type j strikes a deal with types $1, \dots, 2n - j + 1$ and fights the rest, giving objective payoff:

$$\begin{aligned} \Pi(\beta_j) &= P_{2n-j+1} \left(v + \frac{\beta_j - \bar{\beta}_j}{2} \right) + (1 - P_{2n-j+1})(v - k) \\ &= v - k + P_{2n-j+1} \left(k + \frac{\beta_j - \bar{\beta}_j}{2} \right) \end{aligned}$$

where

$$\begin{aligned} P_j &= \sum_{i=1}^j p_i \\ \bar{\beta}_j &= \frac{\sum_{i=1}^{2n-j+1} p_i \beta_i}{\sum_{i=1}^{2n-j+1} p_i} \end{aligned}$$

Next we show that for any β_1, \dots, β_n meeting this condition, there exists a probability distribution where $\beta_i = 2k - \beta_{2n+1-i}$ and $Pr(\beta_i = \beta) = p_i$ is a SP-SPE. Adjacent types getting the same payoff requires:

$$\begin{aligned}
P_{2n-j+1} \left(k + \frac{\beta_j - \bar{\beta}_j}{2} \right) &= P_{2n-j} \left(k + \frac{\beta_{j+1} - \bar{\beta}_{j+1}}{2} \right) \\
p_{2n-j+1} 2k &= P_{2n-j} \beta_{j+1} - P_{2n-j+1} \beta_j + \left(\sum_{i=1}^{2n-j+1} p_i \beta_i - \sum_{i=1}^{2n-j} p_i \beta_i \right) \\
p_{2n-j+1} 2k &= (P_{2n-j+1} - p_{2n-j+1}) \beta_{j+1} - P_{2n-j+1} \beta_j + (\beta_{2n-j+1} p_{2n-j+1}) \\
p_{2n-j+1} 2k &= P_{2n-j+1} (\beta_{j+1} - \beta_j) + p_{2n-j+1} (\beta_{2n-j+1} - \beta_{j+1}) \\
p_{2n-j+1} (2k - \beta_{2n-j+1} + \beta_{j+1}) &= P_{2n-j+1} (\beta_{j+1} - \beta_j) \\
p_{2n-j+1} &= P_{2n-j+1} \frac{\beta_{j+1} - \beta_j}{\beta_{j+1} + \beta_j}
\end{aligned}$$

for $j = 1, \dots, 2n - 1$. Setting $j = 1$ gives:

$$p_{2n} = \frac{\beta_2 - \beta_1}{\beta_1 + \beta_2}$$

since $P_{2n} = 1$. Given this, setting $j = 2$ gives:

$$p_{2n-1} = P_{2n-1} \frac{\beta_{j+1} - \beta_j}{\beta_{j+1} + \beta_j} = (1 - p_{2n}) \frac{\beta_{j+1} - \beta_j}{\beta_{j+1} + \beta_j}$$

where p_{2n} is given above. More generally:

$$p_{2n-j+1} = \left(1 - \sum_{i=2n-j}^{2n} p_i \right) \frac{\beta_{j+1} - \beta_j}{\beta_{j+1} + \beta_j} \quad (5)$$

This gives a recursive definition for the p_2, \dots, p_n 's, which are all strictly positive. So, as long as $\sum_{j=2}^{2n} p_j \leq 1$, setting $p_1 = 1 - \sum_{j=2}^{2n} p_j$ makes the p_i 's a proper probability distribution. The $j = 2n - 1$ adjacency condition gives:

$$\begin{aligned}
p_2 &= \left(1 - \sum_{j=3}^{2n} p_j \right) \frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}} \\
\frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}} p_2 &\leq \left(1 - \sum_{j=3}^{2n} p_j \right) \frac{\beta_{2n} - \beta_{2n-1}}{\beta_{2n-1} + \beta_{2n}} \\
\sum_{i=2}^{2n} p_i &\leq 1
\end{aligned}$$

So, there is a unique p_i which is a probability distribution that meets the adjacency conditions. This completes part i.

For part ii, the probability of conflict is:

$$p_c = \sum_{i=2}^{2n} \sum_{j=2n+i-j} p_i p_j$$

Since $\beta_{n+1}, \dots, \beta_{2n}$ are uniquely determined by β_1, \dots, β_n by the symmetry condition and the p_i 's are recursively defined by the β_j 's and continuous in each β_j , can be written as $p_c(\beta_1, \dots, \beta_n)$. Further, as $\beta_i \rightarrow k$ from below for $i = 1, \dots, n$, which implies $\beta_i \rightarrow k$ from above for $i = n+1, \dots, 2n$, then $p_i \rightarrow 0$ for all $i > 1$, hence $p_1 \rightarrow 1$ and $p_c \rightarrow 0$. Similarly, if $\beta_i \rightarrow 0$ for $i = 1, \dots, n$ and hence $\beta_i \rightarrow 2k$, for $i = n+1, \dots, 2n$, then $p_{n+1} \rightarrow 1$ and $p_c \rightarrow 1$. So, by the continuity of p_c in the β 's, for any $p > 0$ there exists a β_1, \dots, β_n and hence distribution of preferences such that $p_c = p$ by the Intermediate Value Theorem, proving the result for even m .

An analogous result holds for odd m , with the restriction that $\beta_{(m+1)/2} = k$ and the other $\beta_i = \beta_{2n+1-i}$. ■

Proof that there is no SP-SPE that admits a density

Suppose the type distribution admits a density f . Let $\underline{\beta} \in -\infty \cup \mathbb{R}$ be the lowest β in the support of f and $\bar{\beta} \in \infty \cup \mathbb{R}$ the highest. There is a class of SP-SPE where $\underline{\beta} > 2k$, so suppose this is not the case. The expected payoff for being type $\beta_j \in [\underline{\beta}, \bar{\beta}]$ is then:

$$\Pi(\beta_j; \underline{\beta}, \bar{\beta}, \sigma^*) = \begin{cases} \int_{\underline{\beta}}^{2k-\beta_j} \left(\frac{\beta_j + \beta_{-j}}{2} \right) f(\beta_{-j}) d\beta_{-j} + Pr(\beta_j + \beta_{-j} > 2k) (v - k) & \beta_j + \underline{\beta} < 2k \\ (v - k) & \text{otherwise} \end{cases}$$

Proposition 6. *There is no SP-SPE such that $\underline{\beta} < 2k$ where the type distribution admits a density.*

Proof Suppose not, let the density be f . This implies that $\Pi(\beta_j; f, \sigma^*)$ is continuous in β_j . We derive two contradictory inequalities:

i. It must be the case that $\bar{\beta} < 2k - \underline{\beta}$. If not, $\Pi(\bar{\beta}; f, \sigma^*) = v - k$. Consider a type $\beta_j = 2k - \underline{\beta} - \epsilon$. This type will strike a bargain with types $[\underline{\beta}, \underline{\beta} + \epsilon]$, and since $\underline{\beta} < 2k$ the expected payoff from these bargains will be greater than $v - k$. So for some $\epsilon > 0$, $\Pi(2k - \underline{\beta} - \epsilon; f, \sigma^*) > \Pi(\bar{\beta}; f, \sigma^*)$.

ii It must be the case that $\underline{\beta} > 2k - \bar{\beta}$. If not, types $\underline{\beta}$ never fight, and for small $\epsilon > 0$ a type $\beta_j = \underline{\beta} + \epsilon$ would never fight, and get a strictly higher payoff than $\underline{\beta}$.

So $\underline{\beta} + \bar{\beta} < 2k$ and $\underline{\beta} + \bar{\beta} > 2k$, a contradiction ■

In words, if the aggregate toughness of the lowest and highest types is too high, a type slightly less tough than the toughest type would get a strictly higher payoff than the toughest type. If the aggregate toughness of the lowest and highest types is too low, a type slightly tougher than the least tough could always extract a better bargain than the lowest type.

Technical Details for Uniform Case

This function is continuous and decreasing in β_m on the defined domain, with range $[0, \infty)$, which ensures a unique β^* such that $\beta_{max}(\beta^*) = \beta^*$. This intersection will be on the first segment if and only if:

$$\begin{aligned}\beta^{\max}(\beta_m)(2(k - \epsilon)) &< 2(k - \epsilon) \\ 2k - (2(k - \epsilon) + \epsilon) &< 2(k - \epsilon) \\ k &< \frac{3}{2}\epsilon\end{aligned}$$

and when the intersection happens at the second segment, β^* is given by:

$$\begin{aligned}\beta^* &= \frac{2k - (\beta^* - \epsilon)}{3} \\ \beta^* &= \frac{k}{2} + \frac{\epsilon}{4}.\end{aligned}$$

When $k > 32\epsilon$ the intersection happens at the second segment, and β^* solves:

$$\begin{aligned}2k - (\beta^* + \epsilon) &= \beta^* \\ \beta^* &= k - \frac{\epsilon}{2}\end{aligned}$$

For the probability of conflict, the cumulative density function of $\beta_j + \beta_{-j}$ follows a triangle distribution:

$$Pr(\beta_j + \beta_{-j} < x) = \begin{cases} \frac{(x - 2(\beta^* - \epsilon))^2}{8\epsilon^2} & x \in [2(\beta^* - \epsilon), 2\beta^*] \\ 1 - \frac{(2(\beta^* + \epsilon) - x)^2}{8\epsilon^2} & x \in [2\beta^*, 2(\beta^* + \epsilon)] \end{cases}$$

Since $2k > 2\beta^*$, the probability of conflict comes from the second segment (and plugging in the values of β^* from above):

$$\begin{aligned}Pr(\beta_j + \beta_{-j} > 2k) &= \frac{(2(\beta^* + \epsilon) - 2k)^2}{8\epsilon^2} \\ &= \begin{cases} \frac{(2(k - \epsilon/2 + \epsilon) - 2k)^2}{8\epsilon^2} & \epsilon \leq 2/3k \\ \frac{(2(k/2 + \epsilon/4 + \epsilon) - 2k)^2}{8\epsilon^2} & \epsilon \geq 2/3k \end{cases} \\ &= \begin{cases} 1/8 & \epsilon \leq 2/3k \\ \frac{5\epsilon - k}{8\epsilon^2} & \epsilon \geq 2/3k \end{cases}\end{aligned}$$

Proof of Theorem 1

If G has an upper bound, then $\beta^* < k - \epsilon/2$. Write the toughness of two randomly selected players (in the stable distribution) as $\beta_i = \beta^* + \nu_i$ So:

$$\begin{aligned}Pr(\beta_1 + \beta_2 < 2k) &= Pr(2\beta^* + \nu_1 + \nu_2 < 2k) \\ &= Pr(2k - \epsilon + \nu_1 + \nu_2 < 2k) \\ &= Pr(\nu_1 + \nu_2 < \epsilon) > 0\end{aligned}$$

where the final inequality follows from the fact that $G(\underline{\epsilon}) - G(0) > 0$. Writing out this final inequality in integral form gives the desired result.

Proof of Theorem 2

Suppose not, and let the highest type in the support of f^* be $\bar{\beta}$ (this must be finite, as we have assumed no conflict occurs). For there to be no conflict, it must be the case that $\bar{\beta} < k$, and hence $\Pi(\beta_j, f^*, \sigma^*)$ is strictly increasing in the support of f^* . So, the $\bar{\beta}$ type gets the highest payoff, and hence $w(\Pi(\bar{\beta}, f^*, \sigma^*)) > 0$. (This follows from w increasing and the fact that $f^*(\beta)w(\Pi(\beta; f^*, \sigma^*))$ must be a density.) Finally, since $G(0) \in (0, 1)$, the left-hand side of equation 3 must place positive density on $\bar{\beta} + \epsilon$ for some $\epsilon > 0$, contradicting the equilibrium condition. ■

Model With Imperfect Observability

Suppose the type of the responder is observed with probability $q \in [0, 1]$. When observed, the bargain game plays out as in the baseline. When unobserved, the proposer's belief about the responder's type is equal to the equilibrium distribution. Since the responder's payoffs are unaffected by the proposer's type and hence does not affect the acceptance strategy, equilibrium behavior is not dependent on whether player 1's type is observed or not.

So, a strategy in the extension with partial observability is: a mapping from the type to (1) an offer to make when not observing the type of the responder, (2) the offer to make as a function of the responder type when observed, and (3) a set of offers to accept in the responder role. Given the addition of incomplete information, our solution concept for fixed preferences is now Perfect Bayesian Equilibrium (PBE).

Motivated by the complete information analysis, we first search for an equilibrium of the following form:

- There are two types, β_l and β_h , and proportion p_h of the players are β_h
- $0 < \beta_l < k$ and $\beta_h = 2k - \beta_l$

As in the baseline, let $\delta = \beta_h - k = k - \beta_l$. When the responder or player 2's type is observed, the equilibrium strategies given the preferences are as derived above: the low type offers $v - k + \beta_l$ which is accepted, and the high type offers $v - k + \beta_l$ to the l type (accepted) and makes a lowball offer to the high type (rejected).

When player 2's type is unobserved, by a standard argument there are three candidate offers: a lowball offer which is always rejected x_0 , offering enough to buy the low type with $x_l = v - k + \beta_l = v - \delta$ and offering enough to buy off the high type with $x_h = v - k + \beta_h = v + \delta$. The subjective expected payoff to a high type for making these offers is:

$$\begin{aligned} u_h(x_0) &= v - k + \beta_h = v + \delta \\ u_h(x_l) &= p_h(v - k + \beta_h) + (1 - p_h)(v + k - \beta_l) = v + \delta \\ u_h(x_h) &= v + k - \beta_h = v - \delta \end{aligned}$$

So, the high type is indifferent between the lowball offer and buying off the low type, which are both strictly preferred to buying off both types. However, the objective payoff for making the offer x_l is strictly higher than the payoff from making offer x_0 , so if there is any heterogeneity within the tough types those more apt to make an offer of x_l will reproduce faster. To rule out offering x_0 , make the following assumption:

Assumption 1. *Whenever a type is indifferent between two actions based on their subjective payoffs, they always select the action that gives a higher objective payoff.*

The payoff to the low type for making these offers is:

$$\begin{aligned} u_l(x_0) &= v - k + \beta_l = v - \delta \\ u_l(x_l) &= p_h(v - k + \beta_l) + (1 - p_h)(v + k - \beta_l) = v + (1 - 2p_h)\delta \\ u_l(x_h) &= v + k - \beta_h = v - \delta \end{aligned}$$

So the low type always offers x_l for any $p_h < 1$.

For most configurations, the equilibrium outcome is the same with partial information as with complete information. Low types strike a bargain determined by the low type's reservation point, and high types always fight. When the high type and low type are matched with the high type in the proposer role, the high type offers x_l and it is accepted. The difference between the two cases arises when a low type is in the proposer role and a high type is in the responder role. With perfect observability, the low type would offer enough to buy off the high type, but with imperfect visibility it is not worth it for the low type to offer enough to buy him off (though she would under perfect information). So, in this case, *both* players are worse off in objective terms, as they get payoffs $(v - k, v - k)$ rather than $(v - \delta, v + \delta)$, respectively.

In sum, a high type gets an objected expected payoff of

$$\Pi_h = p_h(v - k) + (1 - p_h) [1/2(1 - q)(v - k) + (q + 1/2(1 - q))(v + \delta)]$$

Which is linearly decreasing in p_h , approaches $v - k$ as $p_h \rightarrow 1$, and approaches $1/2(1 - q)(v - k) + (q + 1/2(1 - q))(v + \delta)$ as $p_h \rightarrow 0$.

The average objective payoff for being a low type is:

$$\Pi_l = p_h[(1 - q)(1/2)(v - k) + (q/2 + 1/2)(v - \delta)] + (1 - p_h)v$$

Which approaches $v - k(1 - q)/2 - \delta(1/2 + q/2)$ as $p_h \rightarrow 0$ and v as $p_h \rightarrow 1$. The first equilibrium condition is that both the high and low type get the same payoff. Setting $\Pi_l = \Pi_h$ and solving for δ gives

$$\delta = k \frac{1 - q(1 - 2p_h)}{1 + q} \quad (6)$$

Recall that q and k are exogenous parameters of the model, so the first condition for an equilibrium is that there exist a pair (δ, p_h) that meets equation 6. The right-hand side of equation 6 is linearly increasing in p_h and always on $[0, k]$. So for any $p_h \in [0, 1]$ and $q > 0$, there exists a $\delta \in [0, k]$ solving this equation.

However, a pair (δ, p_h) meeting equation 6 ensures that the high and low types get the same payoff, but not that no invader type β' would get a higher payoff.

Given the information structure, the type of the invader is observed with probability q , and with probability $1 - q$ the opponent believes that their partner is drawn from the proposed equilibrium distribution.

When the types are observed, the analysis is the same as the complete information case, giving objective payoffs:

$$\Pi^{\text{obs}}(\beta'; p_h, \delta) = \begin{cases} v - k & \beta' > k + \delta \\ p_h(v - k) + (1 - p_h) \left(v + \frac{\beta' - \beta_l}{2} \right) & k - \delta < \beta' \leq k + \delta \\ p_h \left(v + \frac{\beta' - \beta_h}{2} \right) + (1 - p_h) \left(v + \frac{\beta' - \beta_l}{2} \right) & \beta' \leq k - \delta \end{cases}$$

When the type is unobserved and the invader is in the proposer role, his subjective expected payoffs to offering x_0 , x_l and x_h are:

$$\begin{aligned} u_l(x_0) &= v - k + \beta' \\ u_l(x_l) &= p_h(v - k + \beta') + (1 - p_h)(v + k - \beta_l) = v + p_h(\beta' - k) + (1 - p_h)\delta \\ u_l(x_h) &= v + k - \beta_h = v - \delta \end{aligned}$$

Comparing the first two payoffs, offering x_0 is preferred if $\beta' > \beta_h$, and if this holds x_0 is also preferred to x_h . Offering x_h is preferred to offer x_l if $\beta' < \beta_h - \frac{2\delta}{p_h}$. So, by assumption 1 the invader offers x_h if $\beta' \leq \beta_h - \frac{2\delta}{p_h}$, x_l if $\beta' \in (\beta_h - \frac{2\delta}{p_h}, \delta + k]$, and x_0 if $\beta' > \delta + k$.

When the type is unobserved and the invader is in the player 2 role, they accept any offer such that $x \geq 1/2 - k + \beta'$, and both types offer x_l , so the objective payoff is $v - \delta$ if $\beta' \leq k - \delta$ and $1/2 - k$ if $\beta' > k - \delta$.

As before, the total payoff to being a β' type is always increasing in β' except at points where a marginal increase leads to more fighting. So, the only invader type that can possible generate a higher objective payoff than β_l or β_h is $\beta_h - \frac{2\delta}{p_h} \equiv \beta_w$. Such a “mega-wimp” ($\beta_w < \beta_l$) gets a very low payoff if their type is observed and they are in the responder role, but if their type is rarely observed this is outweighed by the fact that they are willing to make a more conciliatory offer which is superior to conflict when they are in the proposer role.

Formally, when the type is observed, such an invader gets a payoff of $v - k + \beta_w$ when in the responder role and $v - \beta_{-j} + k$ when in the proposer role. When the type is unobserved, the invader gets a payoff of $v - \delta$ when in the proposer role (as he offers x_h and it is accepted) and also $v - \delta$ when in the responder role (as the proposer always offers x_l) and this is accepted. So, the expected payoff for the invader is:

$$\begin{aligned} \Pi(-\beta_h; p_h, \delta) &= q((1/2)(v + k + \beta_w) + (1/2)(v - p_h\beta_h - (1 - p_h)\beta_l + k) \\ &\quad + (1 - q)(v - \delta)) \end{aligned}$$

Plugging in the δ meeting equation and subtracting this from the proposed equilibrium payoff for the β_h and β_l types gives:

$$\frac{k(1 - p_h)[(1 - q)q(1 + p_h^2) + p_h(q - 2q + 3q^2)]}{(1 + q)p_h} > 0$$

Which implies the β_w type and hence no invader can get a strictly higher payoff than the proposed equilibrium types as long as equation 6 is met.

Finally, the equilibrium probability of conflict is:

$$p_c = (p_h^*)^2 + (1 - q)p_h^*(1 - p_h^*)/2,$$

as conflict occurs when either two tough types meet when the type is not observed with a weak proposer and a tough responder. This probability approaches 0 when $p_h^* \rightarrow 0$ and 1 when $p_h^* \rightarrow 1$, so, as in the complete information model, any probability of conflict can be sustained in equilibrium.

Model with Asymmetry of Toughness based on Role

In the baseline model the actors vary on a single toughness parameter that changes their conflict payoff whether in the proposer or in responder role. Suppose instead that the type is now a double $\beta = (\beta^r, \beta^p) \in \mathbb{R}^2$,

Parameterization with Noisy Evolution