Anger and Political Conflict Dynamics

Keith E. Schnakenberg* Carly N. Wayne†

SUPPLEMENTARY INFORMATION

Contents

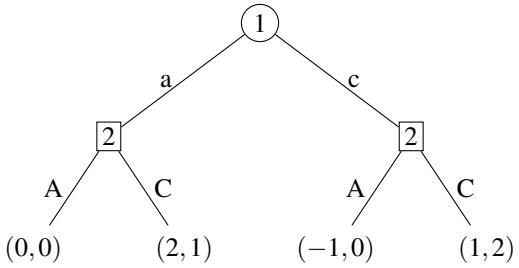
*Associate Professor of Political Science, Washington University in St. Louis. ORCID: 0000-0001-6997-5656. Contact: keith.schnakenberg@gmail.com.

†Assistant Professor of Political Science, Washington University in St. Louis. ORCID: 0000-0002-4183-3071. Contact: carlywayne@gmail.com.

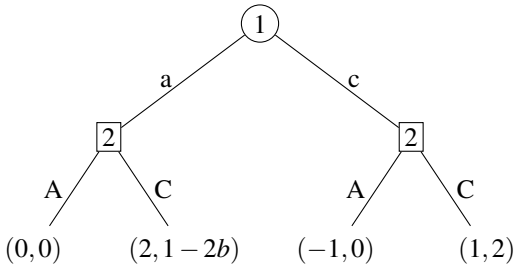
A A simple psychological game with anger

In this section we present a simple one-period sequential game modeled loosely off of the interaction described in the main model. Though we focus on a setting with similar features to the model in the paper, this section is intended to be self-contained. Our purpose here is not to reproduce the results from the paper but instead to explain the basics of how psychological preferences work in a more accessible setting. Consider an interaction between two players, 1 and 2. The sequence of play is that player 1 chooses an action in $\{a,c\}$ and then player 2 observes player 1's action and then chooses an action in $\{A,C\}$. As in the model in the main text, we interpret a or A as "aggressive" and c or C as "conciliatory."

The material payoffs of the players as well as a depiction of the sequence of the game are given in Figure A.1a. The key features of material payoffs are that Player 1 would most prefer to get away with being aggressive and receiving a conciliatory response from Player 2, but will play a conciliatory action if she thinks Player 2 would respond to an aggressive action with an aggressive response. However, Player 2 has a dominant strategy to play the conciliatory action at both information sets. Applying backward induction, the unique equilibrium is one in which player 1 chooses a and player 2 chooses C at both information sets, so the path of play is (a,C) and the equilibrium payoffs are $(2,1)$.



(a) Example game with only material payoffs



(b) Example game augmented with psychological payoffs

Figure A.1: Simple example game. Version (a) shows the game with only material payoffs and version (b) shows the game augmented with psychological payoffs.

However, we augment this game by providing psychological preferences for Player 2. To repre-

sent these preferences, let b denote player 2's conjecture about player 1's probability of playing c. That is, b denotes player 2's conjecture about player 1's mixed strategy. In line with the assumptions about anger in the paper, we assume that Player 2 becomes angry when Player 1 unexpectedly chooses a. Since the difference between Player 2's best payoff when Player 1 chooses c and Player 2's best payoff when Player 1 chooses a is 1, the decrease in Player 2's expected material payoff due to Player 1's action is equal to b. We assume that anger creates an aversion to increasing the payoff of the other player, so Player 2's decision utility is equal to her material payoff minus her anger level (equal here to the decrease in her expected material payoff) times the payoff of Player 1. This is captured by replacing Player 2's payoff of 1 in the game with purely material payoffs to $1-2b$ at the (a,C) terminal node.[1] Informally, a psychological sequential equilibrium requires that both players are sequentially rational in the usual sense of backward induction and also that Player 2's conjecture about Player 1's strategy be correct. That is, Player 1 plays c with probability p, where p is a best response to Player 2's strategy, and Player 2's strategy is a best response given the correct conjecture $b = p$. A look at the equilibrium concept reveals why the game cannot be analyzed using classical game theoretic tools: Player 2's preferences depend on her conjecture b, but this conjecture is itself endogenous.

We can analyze this game as follows. First, we analyze Player 2's best response at each information set given a particular conjecture b. Clearly, regardless of Player 2's conjecture, her best response to c is always C. Furthermore, C is a best response to a if $b \leq \frac{1}{2}$ and A is a best response to a if $b \geq \frac{1}{2}$. This captures the idea from the main model in the text that anger can change behavior only if the observed behavior deviates significantly enough from the expected behavior.

Second, we consider different possible strategies for Player 1 to find equilibria. Notice that this must go beyond the usual backward induction exercise: we must start from a candidate solution and then check whether both players are playing best responses *given that Player 2's beliefs are computed at that strategy.* First, consider the possibility that $p = 0$ as in the unique equilibrium to the game with only material payoffs. In that case, Player 2's action off-the-path of play following a choice of c would be C, which would give Player 1 a payoff of 1. On the path of play, Player 2's best response remains C since $b = p = 0$ so anger does not come into play. Thus, the unique equilibrium in the game with only material payoffs is also an equilibrium to the psychological game.

Next, consider the other pure strategy for Player 1, $p = 1$. In this case, Player 1 still knows

[1] Anger does not appear in the (a,A) terminal node because Player 1's payoff is zero, though we can think of this as being implicitly factored in but just equal to zero. Similarly anger does not factor into either payoff when Player 1 chooses c since Player 2 achieves a maximal material payoff.

that she can guarantee herself a payoff of 1 by choosing c. If she chooses a, then Player 2's best response will require her to play A since $b = p = 1$. Thus, Player 1 expects a payoff of 0 from choosing a and 1 from choosing c. Thus, there is also a pure strategy equilibrium in which Player 1 chooses c and Player 2 chooses A in response to a and C in response to C.

For there to be any equilibrium with $p \in (0,1)$, Player 1 must be indifferent between choosing a or c. Since Player 2 always must play C in response to c, this implies that Player 2 must play C in response to a with probability one half. This also requires that Player 2 is indifferent between playing A and C in response to a, which is true when $1 - 2b = 0$, i.e. when $b = \frac{1}{2}$. Thus, there is also a mixed strategy equilibrium in which Player 1 chooses c with probability $\frac{1}{2}$ and a with probability $\frac{1}{2}$ and Player 2 chooses C in response to c and chooses A with probability $\frac{1}{2}$ and C with probability $\frac{1}{2}$ in response to a.

This simple example deliberately leaves out certain components of the game in text that are crucial to the main results, namely the presence of incomplete information and the possibility of mistakes. However, it provides an illustration of how anger might operate in a much simpler setting.

B Examples of computing psychological payoffs

We now characterize the psychological motives of the players. Anger depends on three factors. First, agents are angry when there is a deviation between their expected payoff and actual payoff. Agents can therefore only be angry when $\tilde{y}_{t-1} = a$ and anger from $\tilde{y}_{t-1} = a$ is greater the more the agent initially expected a conciliatory outcome. Second, agents are angry when a negative outcome is blamed on the actions of another. The agent's anger therefore depends on the probability that the aggressive outcome was caused by an aggressive *intended* action rather than being unintentional. Finally, anger, once triggered, causes agents to negatively weight the payoffs of players from the other group which motivates aggression.

Detailed calculations of anger motivations for general strategies are provided in Appendix C.1 but the following examples illustrate the mechanisms that drive anger. For illustration we consider three illustrative strategy profiles in order to show how anger is derived in the model and how it depends on the agents' conjectures about the strategies played by others. The purpose here is not to illustrate equilibria – in fact we will compute psychological preferences for strategy profiles that will turn out not to be equilibria – but instead to show the mechanics of how emotional responses to the same situation are driven by beliefs about the context.

As highlighted in Proposition 2, when Friendly types are always aggressive in response to aggressive outcomes, agents do not get angry because observing an aggressive outcome does not constitute a deviation between actual and expected payoffs.

Example 1 (Certain conflict)**.** Suppose an agent's conjecture is that Friendly types always take aggressive actions toward the next player in response to aggressive outcomes. As we have discussed, the long-run probability of an aggressive action in an arbitrary time period is one given these strategies. Therefore, the psychological component of payoffs is zero in this situation since $\Pr[\tilde{y}_{t-1} = c | b_t^A, b_t^B] = 0$.

As we show in Proposition 2, the strategy profile analyzed in Example 1 is not an equilibrium to the psychological game. The example shows why perpetual conflict cannot be an equilibrium outcome even if sensitivity to anger is very large, since an aggressive outcome never produces a deviation between expected and actual payoffs.

Next we consider a situation in which the agent believes that Friendly types never behave aggressively. When Friendly types are always conciliatory, observing an aggressive outcome constitutes a deviation between expected and actual payoffs. Since it is possible that the other group has Hostile types, an aggressive outcome was possibly caused by the deliberate actions of another agent.

Example 2 (Friendly types are always conciliatory)**.** Suppose an agent's conjecture is that Friendly types always take conciliatory actions toward the next player. Then in any period $\Pr[\tilde{y}_{t-1} = c | b_t^A, b_t^B] = (1-\mu_0)(1-\pi)$. Furthermore, agents are not angry if they believe aggressive outcomes occur purely unintentionally, so our calculation of the agent's anger must compute $\Pr[y_{t-1} = a | \tilde{y}_{t-1} = a]$, the probability that an aggressive outcome followed from an aggressive intended action. If Friendly types are always conciliatory, then the only time $y_{t-1} = a$ is when $\theta_{t-1} = H$, so $\Pr[y_{t-1} = a | \tilde{y}_{t-1} = a] = \mu_t$ where μ_t is the updated belief that the other groups is made up of Hostile types given the outcome $\tilde{y}_{t-1} = a$. Additionally, since only Hostile types could have intentionally caused an aggressive outcome given these strategies, only the payoffs of Hostile types are negatively weighted in the agent's psychological payoff. The expected gain to a Hostile player $t+1$ from choosing $y_t = c$ rather than $y_t = a$ is $(1-\pi)r(H)$. Therefore, the agent's psychological payoffs reduce the expected decision utility of taking the conciliatory action by $\alpha_j(1-\mu_0)(1-\pi)\mu_t((1-\pi)r(H))$.

Though the extremes of Example 1 and 2 are useful benchmarks, the typical equilibrium to the game occurs in mixed strategies. We illustrate how psychological payoffs are derived in this case using an example in which an agent believes that the groups play symmetric mixed strategies.

Example 3 (Friendly types retaliate with probability one half)**.** Suppose an agent's conjecture is that Friendly types in both groups retaliate against player $t+1$ in response to aggressive results from interacting with player $t-1$ with probability $\frac{1}{2}$. We want to compute each player's anger level when they observe $\tilde{y}_{t-1} = a$, which we will first do for a Friendly member of group A and then show how it extends to a Friendly member of group B. As before, the first component of this anger level

is the deviation between this outcome and the expected outcome, so we need to know the agent's prior probability that they would have observed $\tilde{y}_{t-1} = c$ instead. We can compute this probability as follows. First, if the other group is made up of Hostile types, then the agent will observe $\tilde{y}_{t-1} = a$ with probability one. Thus, the probability that $\tilde{y}_{t-1} = c$ is $(1-\mu_0)(1-\Pr[\tilde{y}_{t-1} = a|\rho_B = 0])$. To compute this last probability, let q_A denote the probability that a group A member observes $\tilde{y}_{t-1} = a$ when both groups are made up of Friendly types and let q_B denote the comparable probability for group B. Then we have $q_A = \pi + (1-\pi)q_B 1/2$. That is, when both groups are made up of Friendly types, a member of group A observes $\tilde{y}_{t-1} = a$ either when there was a mistake, which happens with probability π, or when there was not a mistake but the previous member of group B also observed an aggressive result and chose to retaliate. Given the conjectured strategy profile, this retaliation happens with probability $\frac{1}{2}$. Since the conjectured strategy profile is symmetric, we should have $q_A = q_B$ and this probability reduces to $\frac{2\pi}{1+\pi}$. Thus, $\Pr[\tilde{y}_{t-1} = c|b_t^A, b_t^B] = (1-\mu_0)\left(1-\frac{2\pi}{1+\pi}\right)$.

The next component of the agent's anger is the likelihood that the observed outcome of $\tilde{y}_{t-1} = a$ resulted from a truly aggressive intended action from the previous player. If player $t-1$ is Hostile, they always choose an aggressive action. If player $t-1$ was Friendly, the probability that they took an aggressive action conditional on observing an aggressive result is one half.[2] Finally, a conciliatory action gives player $t+1$ an expected payoff of $(1-\pi)r(H)$ if that player is Hostile and an expected payoff of $1-\pi$ if that player is Friendly. Putting these together, anger reduces the Friendly group A member's decision utility for choosing the conciliatory action when $\tilde{y}_{t-1} = a$ by $\alpha_A(1-\mu_0)\left(1-\frac{2\pi}{1+\pi}\right)(1-\pi)\left(\mu_t r(H) + (1-\mu_t)\frac{1}{2}\right)$.

These examples only derive preferences for a given conjectured strategy profile. The equilibrium concept adds to this that the conjectured strategy profile is correct. Thus, the full process for computing equilibria first computes preferences as in the examples above for all possible strategy profiles. Then, this process is nested in a larger fixed point problem: the equillibrium strategy profile must involve all agents playing a best response given preferences computed at that strategy profile.

C Proof of Proposition 1

We begin by proving parts (1) and (2) of Proposition 1 which do not depend on Assumption 1. The result for Hostile types is stated in Lemma 1.

Lemma 1. *Hostile types play $x_t = a$ and $y_t = a$ at any information set in any equilibrium.*

[2]This can be computed via Bayes rule as follows: $\Pr[y_{t-1} = a|\tilde{y}_{t-1} = a \text{ and } \theta_{t-1} = F] = \frac{q_B\frac{1}{2}}{\Pr[y_{t-1}=a \text{ and } \theta_{t-1}=F]} = \frac{q_A\frac{1}{2}}{q_B} = \frac{1}{2}$ where the last step follows from the fact that $q_A = q_B$.

Proof. In either interaction, if the other player chooses a then the payoff to the Hostile type from choosing a is 0 and the payoff from choosing c is $-s < 0$. Similarly, if the other player chooses c then the payoff for choosing a is $r(H) > 1$ and the payoff for choosing c is 1. Thus, Hostile types have a dominant strategy to play $x_t = y_t = a$.■

Lemma 2 concerns a Friendly type's choice of x_t.

Lemma 2. *Friendly types at all $t > 0$ play $x_t = \tilde{y}_{t-1}$ in any equilibrium.*

Proof. If $\tilde{y}_{t-1} = a$ then a Friendly type's payoff for choosing $x_t = a$ is 0 and her payoff for choosing $x_t = c$ is $-s < 0$. If $\tilde{y}_{t-1} = c$ then a Friendly type's payoff for choosing $x_t = a$ is $r(F) < 1$ and her payoff for choosing $x_t = c$ is 1. Thus, in both cases the Friendly type's best response is to choose $x_t = \tilde{y}_{t-1}$. ■

Remark 1. Notice that Lemma 1 and Lemma 2 do not depend on anger. In the case of Lemma 1, anger would only make aggressive responses more attractive, but we show that Hostile types already have a dominant strategy to play a. In the case of Lemma 2, anger only increases the Friendly types incentive to play a following a signal of a, but it is already true that their best response is to play $x_t = a$ in that situation.

The rest of the results concern the optimal choice of y_t for Friendly types. We start with the benchmark of no anger ($\alpha_A = \alpha_B = 0$).

Lemma 3. *Let $\alpha_A = \alpha_B = 0$ or let $B_t = 0$. Then a Friendly type of player t plays $y_t = c$ if*

$$\mu_t \leq \frac{1-(1+s)(\pi+(1-\pi)\underline{\rho})}{(1+s)(1-\pi)(\overline{\rho}-\underline{\rho})}.$$

and plays $y_t = a$ otherwise.

Proof. A Friendly type's expected payoff for playing $y_t = c$ given a belief μ_t is

$$\mu_t((1-\overline{\rho})(1-\pi)-(1-\overline{\rho})\pi s-\overline{\rho}s)+(1-\mu_t)((1-\underline{\rho})(1-\pi)-(1-\underline{\rho})\pi s-\underline{\rho}s). \qquad (1)$$

That is, for a Friendly type of player t from group j and denoting the other group as $-j$, the probability that a conciliatory action produces a conciliatory signal and that player $t+1$ is Friendly is $(1-\rho_{-j})(1-\pi)$ and this leads to a payoff of 1. If either player $t+1$ is Hostile (probability ρ_{-j}) or player $t+1$ is Friendly but a conciliatory action produces an aggressive signal (probability $(1-\rho_{-j})\pi$ then player $t+1$ will play $x_{t+1} = a$ and player t will get a payoff of $-s$. Finally, taking expectations over values of ρ_t where $\rho_{-j} = \overline{\rho}$ with probability μ_t and $\rho_{-j} = \underline{\rho}$ with probability $(1-\mu_t)$ yields (1).

A Friendly type's expected payoff for playing $y_t = a$ is zero since this action induced a response of $x_{t+1} = a$ with probability one. Thus, setting (1) to be weakly positive and solving the inequality for μ_t yields the following incentive compatibility condition for a Friendly type to play $y_t = c$:

$$\mu_t \leq \frac{1-(1+s)(\pi+(1-\pi)\underline{\rho})}{(1+s)(1-\pi)(\overline{\rho}-\underline{\rho})}. \tag{2}$$

This completes the proof. ■

Lemma 3 immediately implies that Friendly types at $t = 0$ choose the conciliatory action given the assumptions of the model.

Lemma 4. *Under Assumption 1, Friendly types at $t = 0$ always choose $y_0 = c$.*

Proof. Player 0 is not angry since there is no history at time 0. Thus, the incentive compatibility condition for choosing $y_0 = c$ is the same as in the case of $\alpha_A = \alpha_B = 0$. By Lemma 3, the Friendly type of player 0 chooses $y_0 = c$ if

$$\mu_t \leq \frac{1-(1+s)(\pi+(1-\pi)\underline{\rho})}{(1+s)(1-\pi)(\overline{\rho}-\underline{\rho})}.. \tag{3}$$

Under Assumption 1 this always holds at the prior, which proves that Friendly types always choose $y_0 = c$. ■

We next derive beliefs at each information set. By Lemmas 1, 2, and 4, we may limit conjectures to strategies that differ only in values of $b_t^A(F,a)$ and $b_t^B(F,a)$, the probability that Friendly types from each group play an aggressive intended action y_t given that $\tilde{y}_{t-1} = a$. To derive beliefs, we first derive, for members of each group, the probability of observing $\tilde{y}_{t=1} = a$ for given values of ρ_A and ρ_B. Since the the agents do not know the value of t they use the long-run probability of observing $\tilde{y}_{t-1} = a$ which is given by q_j for members of group j and is defined as follows:

$$q_A(\rho_A,\rho_B) = \rho_B + (1-\rho_B)\left[\pi + (1-\pi)q_B(\rho_B,\rho_A)b_t^B(F,a)\right] \tag{4}$$

$$q_B(\rho_B,\rho_A) = \rho_A + (1-\rho_A)\left[\pi + (1-\pi)q_A(\rho_A,\rho_B)b_t^A(F,a)\right]. \tag{5}$$

The explanation for these equations is as follows: With probability ρ_{-j} player $t-1$ was Hostile in which case $\tilde{y}_{t-1} = a$ is observed with probability one. With probability $(1-\rho_{-j})$ player $t-1$ was friendly. In this case, with probability π a conciliatory action $y_{t-1} = c$ would have still resulted in $\tilde{y}_{t-1} = a$. With probability $(1-\pi)$ the signal is equal to the action. In this case, with probability $(1-q_{-j})$ player $t-1$ observed $\tilde{y}_{t-2} = c$ in which case she chose $y_{t-1} = c$ leading to $\tilde{y}_{t-1} = c$. With probability q_{-j} player $t-1$ observed $\tilde{y}_{t-2} = a$ in which case with probability $b_t^{-j}(F,a)$ she

chose $y_{t-1} = a$ leading to $\tilde{y}_{t-1} = a$ and with probability $1 - b_t^{-j}(F,a)$ she chose $y_{t-1} = c$ leading to $\tilde{y}_{t-1} = c$. Solving these equations for q_A and q_B gives equations for these long-run probabilities:

$$q_A(\rho_A,\rho_B) = \frac{\rho_B + \pi(1-\rho_B) + (1-\pi)(1-\rho_B)b_t^B(F,a)(\pi + \rho_A(1-\pi))}{1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a)} \tag{6}$$

$$q_B(\rho_B,\rho_A) = \frac{\rho_A + \pi(1-\rho_A) + (1-\pi)(1-\rho_A)b_t^A(F,a)(\pi + \rho_B(1-\pi))}{1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a)}. \tag{7}$$

Two implications of (6) and (7) is that $q_j(\rho_j,\rho_{-j})$ converges to 1 as ρ_{-j} goes to one or as both $b_t^A(F,a)$ and $b_t^B(F,a)$ go toward one. That is, the long-run probability of observing aggressive actions goes to one as either the other group is sure to be Hostile or when Friendly types from both groups always respond to aggressive actions by choosing aggressive actions going forward.

Lemma 5. *For each $j \in \{A,B\}$, $q_j(\rho_j,\rho_{-j})$ has the following properties:*

1. *$q_j(\rho_j,\rho_{-j})$ is increasing in ρ_j and ρ_{-j}*
2. *$q_j(\rho_j,\rho_{-j})$ is increasing in $b_t^A(F,a)$ and $b_t^B(F,a)$*

Proof. Proof of part 1 We start by showing that q_j is increasing in ρ_j. Using 6 or 7 and differentiating with respect to ρ_j gives:

$$\frac{\partial q_j}{\partial \rho_j} = \frac{b_t^{-j}(F,a)(1-\pi)^2(1-\rho_{-j})\left[1 - b_t^j(F,a)\left[(1-\rho_{-j})(\pi + (1-\pi)b_t^{-j}(F,a)) + \rho_{-j}\right]\right]}{(1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a))^2} \tag{8}$$

The numerator is a product of probabilities and is therefore positive. The denominator is a squared probability and is therefore positive. Hence, $\frac{\partial q_j}{\partial \rho_j} > 0$.

Next, we show that q_j is increasing in ρ_{-j}. Using 6 or 7 and differentiating with respect to ρ_{-j} gives:

$$\frac{\partial q_j}{\partial \rho_{-j}} = \frac{(1-\pi)\left[1 - b_t^{-j}(F,a)\left((1-\rho_j)(\pi + (1-\pi)b_t^j(F,a)) + \rho_j\right)\right]}{(1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a))^2}. \tag{9}$$

The numerator is positive because it is the product of probabilities and the denominator is a squared probability so is also positive. Hence, $\frac{\partial q_j}{\partial \rho_{-j}} > 0$.

Proof of part 2 Next, we show that q_j is increasing in the value of $b_t^j(F,a)$. We have

$$\frac{\partial q_j}{\partial b_t^j(F,a)} = \frac{b_t^{-j}(F,a)(1-\pi)^2(1-\rho_j)(1-\rho_{-j})\left[b_t^{-j}(F,a)(1-\pi)(1-\rho_{-j})((1-\pi)\rho_j + \pi)_\pi(1-\rho_{-j}) + \rho_j\right]}{(1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a))^2}. \tag{10}$$

The numerator is positive since it is a product of sums of probabilities and the denominator is also positive since it is a squared probability, hence, $\frac{\partial q_j}{\partial b_t^j(F,a)} > 0$.

Finally, we show that q_j is increasing in the value of $b_t^{-j}(F,a)$. We have

$$\frac{\partial q_j}{\partial b_t^{-j}(F,a)} = \frac{(1-\pi)(1-\rho_{-j})\left[\rho_j + (1-\rho_j)\left[b_t^j(F,a)(1-\pi)(\rho_{-j}+(1-\rho_{-j})\pi)+\pi\right]\right]}{(1-(1-\pi)^2(1-\rho_A)(1-\rho_B)b_t^A(F,a)b_t^B(F,a))^2}. \quad (11)$$

The numerator is positive since each term is positive and the numerator is positive as in the previous parts of this argument. ■

The beliefs of the Friendly types are derived in two steps. First, each player updates on their own group j's value of ρ_j based on their own type realizations. For Friendly types this interim belief is

$$\hat{\mu}^j = \frac{\mu_0(1-\overline{\rho})}{\mu_0(1-\overline{\rho})+(1-\mu_0)(1-\underline{\rho})}. \quad (12)$$

Second, the player updates based on the realization of $\tilde{y}_{t-1}$ as follows:

$$\mu_t^a = \frac{\hat{\mu}^j\mu_0 q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)\mu_0 q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j\mu_0 q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)\mu_0 q(\underline{\rho},\overline{\rho})+\hat{\mu}^j(1-\mu_0)q(\overline{\rho},\underline{\rho})+(1-\hat{\mu}^j)(1-\mu_0)q(\underline{\rho},\underline{\rho})} \quad (13)$$

$$\mu_t^c = \frac{\hat{\mu}^j\mu_0(1-q(\overline{\rho},\overline{\rho}))+(1-\hat{\mu}^j)\mu_0(1-q(\underline{\rho},\overline{\rho}))}{\hat{\mu}^j\mu_0(1-q(\overline{\rho},\overline{\rho}))+(1-\hat{\mu}^j)\mu_0(1-q(\underline{\rho},\overline{\rho}))+\hat{\mu}^j(1-\mu_0)(1-q(\overline{\rho},\underline{\rho}))+(1-\hat{\mu}^j)(1-\mu_0)(1-q(\underline{\rho},\underline{\rho}))}. \quad (14)$$

Each of these beliefs is a marginal probability of $\rho_{-j}=\overline{\rho}$ given the outcome.

The next result establishes the order of the agents' beliefs.

Lemma 6. *For any friendly player t, we have* $\mu_t^a > \mu_0 > \mu_t^c$.

Proof. We prove this result by comparing the likelihood ratio $LR_\mu = \frac{\mu}{1-\mu}$ with respect to each beliefs, which has the same order as the beliefs. We have:

$$LR_{\mu_t^a} = \frac{\hat{\mu}^j\mu_0 q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)\mu_0 q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j(1-\mu_0)q(\overline{\rho},\underline{\rho})+(1-\hat{\mu}^j)(1-\mu_0)q(\underline{\rho},\underline{\rho})} \quad (15)$$

$$= \frac{\mu_0}{1-\mu_0}\frac{\hat{\mu}^j q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j q(\overline{\rho},\underline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\underline{\rho})} \quad (16)$$

$$= LR_{\mu_0}\frac{\hat{\mu}^j q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j q(\overline{\rho},\underline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\underline{\rho})}. \quad (17)$$

By Lemma 5 part 1, we have $q(\overline{\rho},\overline{\rho}) > q(\overline{\rho},\underline{\rho})$ and $q(\underline{\rho},\overline{\rho}) > q(\underline{\rho},\underline{\rho})$. Hence $\frac{\hat{\mu}^j q(\overline{\rho},\overline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j q(\overline{\rho},\underline{\rho})+(1-\hat{\mu}^j)q(\underline{\rho},\underline{\rho})} > 1$ which implies that $LR_{\mu_t^a} > LR_{\mu_0}$.

Likewise, we have:

$$LR_{\mu_t^c} = \frac{\hat{\mu}^j \mu_0 (1-q(\overline{\rho},\overline{\rho})) + (1-\hat{\mu}^j)\mu_0(1-q(\underline{\rho},\overline{\rho}))}{\hat{\mu}^j(1-\mu_0)(1-q(\overline{\rho},\underline{\rho})) + (1-\hat{\mu}^j)(1-\mu_0)(1-q(\underline{\rho},\underline{\rho}))} \tag{18}$$

$$= \frac{\mu_0}{1-\mu_0} \frac{\hat{\mu}^j(1-q(\overline{\rho},\overline{\rho})) + (1-\hat{\mu}^j)(1-q(\underline{\rho},\overline{\rho}))}{\hat{\mu}^j(1-q(\overline{\rho},\underline{\rho})) + (1-\hat{\mu}^j)(1-q(\underline{\rho},\underline{\rho}))} \tag{19}$$

$$= LR_{\mu_0} \frac{\hat{\mu}^j(1-q(\overline{\rho},\overline{\rho})) + (1-\hat{\mu}^j)(1-q(\underline{\rho},\overline{\rho}))}{\hat{\mu}^j(1-q(\overline{\rho},\underline{\rho})) + (1-\hat{\mu}^j)(1-q(\underline{\rho},\underline{\rho}))}. \tag{20}$$

Lemma 5 part 1 implies that $1-q(\overline{\rho},\overline{\rho}) < 1-q(\overline{\rho},\underline{\rho})$ and $1-q(\underline{\rho},\overline{\rho}) < 1-q(\underline{\rho},\underline{\rho})$. Thus, $\frac{\hat{\mu}^j(1-q(\overline{\rho},\overline{\rho}))+(1-\hat{\mu}^j)(1-q(\underline{\rho},\overline{\rho}))}{\hat{\mu}^j(1-q(\overline{\rho},\underline{\rho}))+(1-\hat{\mu}^j)(1-q(\underline{\rho},\underline{\rho}))} < 1$ which implies that $LR_{\mu_t^c} < LR_{\mu_0}$. ■

Lemma 7. *Under Assumption 1, Friendly types at times $t > 0$ always choose $y_t = c$ when $\tilde{y}_{t-1} = c$.*

Proof. Agents' anger is zero when $\tilde{y}_{t-1} = c$, so Lemma 3 shows that a Friendly type plays $y_t = c$ if

$$\mu_t \leq \frac{1-(1+s)(\pi+(1-\pi)\underline{\rho})}{(1+s)(1-\pi)(\overline{\rho}-\underline{\rho})}. \tag{21}$$

Lemma 4 shows that this condition holds when $\mu_t = \mu_0$ under Assumption 1. Lemma 6 shows that $\mu_t^c < \mu_0$, which implies that Friendly agents will always play $y_t = c$. ■

Finally, we must characterize Friendly type's choices of y_t following $\tilde{y}_{t-1} = a$. Two considerations arise. First, since we established in Lemma 6 that $\mu_t^a > \mu_0$, it may be the cause that beliefs change enough following an aggressive signal that Friendly types should choose $y_t = a$ for some realizations of s even in the absence of anger. Second, agents may become angry following $\tilde{y}_{t-1} = a$ which could make them more inclined to choose $y_t = a$ at a given belief.

C.1 Anger derivations

We will begin by characterizing anger motivations when $\tilde{y}_{t-1} = a$ given the strategies of the players. Informally, recall that anger motivations are a product of (a) the deviation between the expected and actual payoff, (b) the extent to which this deviation is blamed on the actions of player $t-1$, and (c) the expected payoff to player $t+1$ (as well as player $t-1$ but this is not affected by the choice of y_t so we will ignore it).

To compute (a) recall that player t's payoff from the interaction with $t-1$ is 0 when $\tilde{y}_{t-1} = a$ and 1 when $\tilde{y}_{t-1} = c$. Thus, the deviation from the between the expected and actual payoff when $\tilde{y}_{t-1} = a$ is simply $\Pr[\tilde{y}_{t-1} = c | b_t^A, b_t^B] = 1 - \Pr[\tilde{y}_{t-1} = a | b_t^A, b_t^B]$. Since $\Pr[\tilde{y}_{t-1} = a | b_t^A, b_t^B, \rho_j, \rho_{-j}] =$

$q_j(\rho_j, \rho_{-j})$, the interim probability of $\tilde{y}_{t-1} = a$ for Friendly type is found by taking expectations with respect to μ_0 and $\hat{\mu}^j$. We define this interim belief for a Friendly type below:

$$Q_j \coloneqq \Pr[\tilde{y}_{t-1} = a | b_t^A, b_t^B] = \sum_{\rho_j \in \{\underline{\rho}, \overline{\rho}\}} \sum_{\rho_{-j} \in \{\underline{\rho}, \overline{\rho}\}} \Pr[\rho_j, \rho_{-j} | \theta_t] q(\rho_j, \rho_{-j}), \tag{22}$$

where

$$\Pr[\rho_j, \rho_{-j} | \theta_t = F] = \begin{cases} \hat{\mu}^j \mu_0 & \text{if } \rho_j = \rho_{-j} = \overline{\rho} \\ \hat{\mu}^j (1 - \mu_0) & \text{if } \rho_j = \overline{\rho} \text{ and } \rho_{-j} = \underline{\rho} \\ (1 - \hat{\mu}^j) \mu_0 & \text{if } \rho_j = \underline{\rho} \text{ and } \rho_{-j} = \overline{\rho} \\ (1 - \hat{\mu}^j)(1 - \mu_0) & \text{if } \rho_j = \rho_{-j} = \underline{\rho}. \end{cases} \tag{23}$$

Then the deviation between expected and actual outcomes for a Friendly type from group j is $\Pr[\tilde{y}_{t-1} = c] = 1 - Q_j$.

Parts (b) and (c) of the anger calculation each depend on beliefs about ρ_{-j}.For part (b) we must compute the probability that $y_{t-1} = a$ given $\tilde{y}_{t-1} = a$. We compute the joint prior probability of $(y_{t-1} = a, \rho_j, \rho_{-j})$ for each value of ρ_j and ρ_{-j} below:

$$\Pr[\rho_j = \underline{\rho} \wedge \rho_{-j} = \underline{\rho} \wedge y_{t-1} = a] = (1 - \hat{\mu}^j)(1 - \mu_0) \left[\underline{\rho} + (1 - \underline{\rho}) q_{-j}(\underline{\rho}, \underline{\rho}) b_t^{-j}(F, a) \right] \tag{24}$$

$$\Pr[\rho_j = \overline{\rho} \wedge \rho_{-j} = \underline{\rho} \wedge y_{t-1} = a] = \hat{\mu}^j (1 - \mu_0) \left[\underline{\rho} + (1 - \underline{\rho}) q_{-j}(\underline{\rho}, \overline{\rho}) b_t^{-j}(F, a) \right] \tag{25}$$

$$\Pr[\rho_j = \underline{\rho} \wedge \rho_{-j} = \overline{\rho} \wedge y_{t-1} = a] = (1 - \hat{\mu}^j) \mu_0 \left[\overline{\rho} + (1 - \overline{\rho}) q_{-j}(\overline{\rho}, \underline{\rho}) b_t^{-j}(F, a) \right] \tag{26}$$

$$\Pr[\rho_j = \overline{\rho} \wedge \rho_{-j} = \overline{\rho} \wedge y_{t-1} = a] = \hat{\mu}^j \mu_0 \left[\overline{\rho} + (1 - \overline{\rho}) q_{-j}(\overline{\rho}, \overline{\rho}) b_t^{-j}(F, a) \right]. \tag{27}$$

To explain these probabilities first note that ρ_j and ρ_{-j} are independent so the prior probability of each combination of these proportions is simply the product of probabilities. Furthermore, given ρ_{-j} the probability of $y_{t-1} = a$ is computed in the following way: with probability ρ_{-j} a given member of group $-j$ is Hostile and then player $y_{t-1} = a$ with probability 1. With probability $1 - \rho_{-j}$ a given member of group $-j$ is Friendly. In this case player $t-1$ never plays $y_{t-1} = a$ when $\tilde{y}_{t-2} = c$. When $\tilde{y}_{t-2} = a$, which happens with probability $q_{-j}(\rho_{-j}, \rho_j)$, player $t-1$ behaves according to the mixed strategy $b_t^{-j}(F, a)$, which implies playing $y_{t-1} = a$ with probability $b_t^{-j}(F, a)$. Marginalizing over ρ_j by summing (24) and (25) as well as (26) and (27) gives the joint

probability of ρ_{-j} and $y_{t-1} = a$:

$$\Pr[\rho_{-j} = \underline{\rho} \wedge y_{t-1} = a] = (1-\mu_0)\left[\underline{\rho} + (1-\underline{\rho})b_t^{-j}(F,a)(\hat{\mu}^j q_{-j}(\underline{\rho},\overline{\rho}) + (1-\hat{\mu}^j)q_{-j}(\underline{\rho},\underline{\rho})\right] \tag{28}$$

$$\Pr[\rho_{-j} = \overline{\rho} \wedge y_{t-1} = a] = \mu_0\left[\overline{\rho} + (1-\overline{\rho})b_t^{-j}(F,a)(\hat{\mu}^j q_{-j}(\overline{\rho},\overline{\rho}) + (1-\hat{\mu}^j)q_{-j}(\overline{\rho},\underline{\rho})\right]. \tag{29}$$

The conditional probabilities are then:

$$\Pr[\rho_{-j} = \underline{\rho} \wedge y_{t-1} = a | \tilde{y}_{t-1} = a] = \frac{\Pr[\rho_{-j} = \underline{\rho} \wedge y_{t-1} = a \wedge \tilde{y}_{t-1} = a]}{\Pr[\tilde{y}_{t-1} = a]} = \frac{\Pr[\rho_{-j} = \underline{\rho} \wedge y_{t-1} = a]}{Q_j} \tag{30}$$

$$\Pr[\rho_{-j} = \overline{\rho} \wedge y_{t-1} = a | \tilde{y}_{t-1} = a] = \frac{\Pr[\rho_{-j} = \overline{\rho} \wedge y_{t-1} = a \wedge \tilde{y}_{t-1} = a]}{\Pr[\tilde{y}_{t-1} = a]} = \frac{\Pr[\rho_{-j} = \overline{\rho} \wedge y_{t-1} = a]}{Q_j}, \tag{31}$$

since $\Pr[\tilde{y}_{t-1} = a | y_{t-1} = a] = 1$ and $\Pr[\tilde{y}_{t-1} = a] = Q_j$.

For part (c), we must compute the expected utility to player $t+1$ if player t chooses $y_t = c$. There are three possibilities: with probability π we have $\tilde{y}_t = a$ which gives player $v_{t+1} = 0$. With probability $1-\pi$ we have $\tilde{y}_t = c$ in which case a Hostile type plays $x_{t+1} = a$ and gets a payoff of $v_{t+1} = r(H)$ and a Friendly type plays $x_{t+1} = c$ and gets a payoff of $v_{t+1} = 1$. Thus, the expected payoff to player $t+1$ when $y_t = c$ is

$$\mathbb{E}[v_{t+1} | y_t = c, \rho_{-j}] = (1-\pi)(\rho_{-j} r(H) + (1-\rho_{-j})). \tag{32}$$

Additionally, $\mathbb{E}[v_{t+1} | y_t = a, \rho_{-j}] = 0$.

Putting these elements together, a Friendly type of player t from group j experiences the following psychological cost to playing $y_t = c$ when $\tilde{y}_{t-1} = a$:

$$\alpha_j \mathbb{E}[\beta_t(\tilde{s}_j, \tilde{s}_{-j}, \tilde{y}_{t-1} = a, y_{t-1} = c) v_{t+1}] = \alpha_j \frac{1-Q_j}{Q_j} \sum_{\rho \in \{\underline{\rho},\overline{\rho}\}} \Pr[\rho_{-j} = \rho \wedge y_{t-1} = a] \mathbb{E}[v_{t+1} | y_t = c, \rho_{-j} = \rho], \tag{33}$$

where $\Pr[\rho_{-j} = \underline{\rho} \wedge y_{t-1} = a]$ is defined according to (28) and (29) and $\mathbb{E}[v_{t+1} | y_t = c, \rho_{-j}]$ is defined according to (32).

D Proof of Proposition 2

The perpetual conflict equilibrium amounts to a profile in which $p_A = p_B = 1$. We show that this equilibrium cannot exist under Assumption 1. We first establish that the long-run probability of being in a state where $\tilde{y}_{t-1} = a$ must be one if a player's conjecture is that both groups play such a strategy.

Lemma 8. *If $b_t^A(F,a) = b_t^B(F,a) = 1$ then $q_j(\rho_j, \rho_{-j}) = 1$ for both $j \in \{A,B\}$ and any $(\rho_j, \rho_{-j}) \in \{\underline{\rho}, \overline{\rho}\}^2$.*

Proof. Plugging in $b_t^A(F,a) = 1$ and $b_t^B(F,a) = 1$ to (6) and (7) yields

$$q_A(\rho_A, \rho_B) = \frac{\rho_B + \pi(1-\rho_B) + (1-\pi)(1-\rho_B)(\pi + \rho_A(1-\pi))}{1-(1-\pi)^2(1-\rho_A)(1-\rho_B)} \tag{34}$$

$$q_B(\rho_B, \rho_A) = \frac{\rho_A + \pi(1-\rho_A) + (1-\pi)(1-\rho_A))(\pi + \rho_B(1-\pi))}{1-(1-\pi)^2(1-\rho_A)(1-\rho_B)}. \tag{35}$$

We will show the result just for q_A since the functions are symmetric. We start by expanding the numerator:

$$q_A(\rho_A, \rho_B) = \frac{\rho_B - \rho_A\rho_B\pi^2 + 2\rho_A\rho_B\pi - \rho_A\rho_B + \rho_A\pi^2 - 2\rho_A\pi + \rho_A + \rho_B\pi^2 - 2\rho_B\pi - \pi^2 + 2\pi}{1-(1-\pi)^2(1-\rho_A)(1-\rho_B)}. \tag{36}$$

Next, we expand the denominator:

$$q_A(\rho_A, \rho_B) = \frac{\rho_B - \rho_A\rho_B\pi^2 + 2\rho_A\rho_B\pi - \rho_A\rho_B + \rho_A\pi^2 - 2\rho_A\pi + \rho_A + \rho_B\pi^2 - 2\rho_B\pi - \pi^2 + 2\pi}{\rho_B - \rho_A\rho_B\pi^2 + 2\rho_A\rho_B\pi - \rho_A\rho_B + \rho_A\pi^2 - 2\rho_A\pi + \rho_A + \rho_B\pi^2 - 2\rho_B\pi - \pi^2 + 2\pi} = 1. \tag{37}$$

The same argument holds for q_B. ■

Proof of Proposition 2. We will show that under Assumption 1 there cannot exist an equilibrium in which $p_A = p_B = 1$. Consider a strategy profile with $p_A = p_B = 1$. In equilibrium, players' conjectures must set $b_t^A(F,a) = b_t^B(F,a) = 1$. By Lemma 8 we have $q_j(\rho_j, \rho_{-j}) = 1$ for any pair of values for ρ_j and ρ_{-j} under this strategy profile. As a result, $Q_A = Q_B = 1$. Plugging this into (33) we have zero anger motivation following $\tilde{y}_{t-1} = a$ for all players. Thus, by Lemma 3, a Friendly type of player t plays $y_t = c$ if

$$\mu_t \leq \frac{1-(1+s)(\pi+(1-\pi)\underline{\rho})}{(1+s)(1-\pi)(\overline{\rho}-\underline{\rho})}..$$

Recall that a Friendly player's belief following $\tilde{y}_{t-1} = a$ is

$$\mu_t^a = \frac{\hat{\mu}^j \mu_0 q(\overline{\rho},\overline{\rho}) + (1-\hat{\mu}^j)\mu_0 q(\underline{\rho},\overline{\rho})}{\hat{\mu}^j \mu_0 q(\overline{\rho},\overline{\rho}) + (1-\hat{\mu}^j)\mu_0 q(\underline{\rho},\overline{\rho}) + \hat{\mu}^j(1-\mu_0)q(\overline{\rho},\underline{\rho}) + (1-\hat{\mu}^j)(1-\mu_0)q(\underline{\rho},\underline{\rho})} \tag{38}$$

$$= \frac{\hat{\mu}^j \mu_0 + (1-\hat{\mu}^j)\mu_0}{\hat{\mu}^j \mu_0 + (1-\hat{\mu}^j)\mu_0 + \hat{\mu}^j(1-\mu_0) + (1-\hat{\mu}^j)(1-\mu_0)} \tag{39}$$

$$= \frac{\mu_0}{\mu_0 + (1-\mu_0)} \tag{40}$$

$$= \mu_0. \tag{41}$$

By Assumption 1,

$$\mu_0 \leq \frac{1-(1+\overline{s})(\pi-(1-\pi)\underline{\rho})}{(1+\overline{s})(1-\pi)(\overline{\rho}-\underline{\rho})}.$$

Thus, a Friendly type's best response is to always play $y_t = c$, which shows that this strategy profile is not an equilibrium. ■

E Proof of Proposition 3 and Corollary 1

Proof of Proposition 3. The proof of Proposition 3 concentrates on the case in which $\underline{\rho} = 0$ and $\overline{\rho} = 1$. Consider an increase in α_A. We need to show that p_A is increasing in α_A and that p_B is decreasing in α_A. The argument for α_B works symmetrically.

We have:

$$q_A = \pi + (1-\pi)q_B p_B \tag{42}$$

$$q_B = \pi + (1-\pi)q_A p_A \tag{43}$$

$$\tag{44}$$

which is solved by

$$q_A = \frac{\pi^2(p_B) - \pi p_B - \pi}{\pi^2 p_A p_B - 2\pi p_A p_B + p_A p_B - 1} \tag{45}$$

$$q_B = \frac{\pi^2(p_A) - \pi p_A - \pi}{\pi^2 p_A p_B - 2\pi p_A p_B + p_A p_B - 1}. \tag{46}$$

A Friendly agent $j's$ decision utility for playing $y_t = c$ following $\tilde{y}_{t-1} = a$ is

$$U_j^c = \mu_t^a \left[-s - \alpha_j(1-\mu_0)(1-q_j)\frac{q_{-j}}{q_j} p_{-j}(1-\pi)r(H) \right] + (1-\mu_t^a)\left[\pi s + (1-\pi)\left(1 - \alpha_j(1-\mu_0)(1-q_j)\frac{q_{-j}}{q_j} p_{-j} \right) \right]. \tag{47}$$

We first consider non-degenerate mixed strategies. At a pair of equilibrium strategies we have, for both $j \in \{A,B\}$, that $U_j^c = 0$. Let $(q_A^*(s), q_B^*(s_t'))$ be critical values of q_A and q_B that satisfy this equation. By the implicit function theorem we have

$$\frac{\partial q_A^*}{\partial \alpha_A} = -\frac{\frac{\partial U_{A1}}{\partial \alpha_A}}{\frac{\partial U_{A1}}{\partial q_A}} \tag{48}$$

$$\frac{\partial q_B^*}{\partial \alpha_A} = 0. \tag{49}$$

Since $\frac{\partial U_{A1}}{\partial \alpha_A} < 0$ and $\frac{\partial U_{A1}}{\partial q_A} > 0$ this tells us that $\frac{\partial q_A^*}{\partial \alpha_A} > 0$. By (45) and (46) the strategies must satisfy

$$\frac{\pi^2(p_B) - \pi p_B - \pi}{\pi^2 p_A p_B - 2\pi p_A p_B + p_A p_B - 1} - q_A^* = 0 \tag{50}$$

$$\frac{\pi^2(p_A) - \pi p_A - \pi}{\pi^2 p_A p_B - 2\pi p_A p_B + p_A p_B - 1} - q_B^* = 0. \tag{51}$$

By the implicit function theorem:

$$\begin{pmatrix} \frac{\partial p_A}{\partial \alpha_A} \\ \frac{\partial p_B}{\partial \alpha_A} \end{pmatrix} = -\begin{pmatrix} \frac{\partial}{\partial p_A}\frac{\pi^2(p_B)-\pi p_B-\pi}{\pi^2 p_A p_B-2\pi p_A p_B+p_A p_B-1} & \frac{\partial}{\partial p_B}\frac{\pi^2(p_B)-\pi p_B-\pi}{\pi^2 p_A p_B-2\pi p_A p_B+p_A p_B-1} \\ \frac{\partial}{\partial p_A}\frac{\pi^2(p_A)-\pi p_A-\pi}{\pi^2 p_A p_B-2\pi p_A p_B+p_A p_B-1} & \frac{\partial}{\partial p_B}\frac{\pi^2(p_A)-\pi p_A-\pi}{\pi^2 p_A p_B-2\pi p_A p_B+p_A p_B-1} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial q_A^*}{\partial \alpha_A} \\ 0 \end{pmatrix} \tag{52}$$

$$= -\begin{pmatrix} -\frac{p_A(1-p_A p_B(1-\pi)^2}{\pi(1-p_A(1-\pi))} & \frac{1-p_A p_B(1-\pi)^2}{(1-p_B(1-\pi))(1-\pi)\pi} \\ \frac{1-p_A p_B(1-\pi)^2}{(1-p_A(1-\pi))(1-\pi)\pi} & -\frac{p_B(1-p_A p_B(1-\pi)^2}{\pi(1-p_B(1-\pi))} \end{pmatrix} \begin{pmatrix} \frac{\partial q_A^*}{\partial \alpha_A} \\ 0 \end{pmatrix} \tag{53}$$

$$= -\frac{\partial q_A^*}{\partial \alpha_A} \begin{pmatrix} -\frac{p_A(1-p_A p_B(1-\pi)^2)}{\pi(1-p_A(1-\pi))} \\ \frac{1-p_A p_B(1-\pi)^2}{(1-p_A(1-\pi))(1-\pi)\pi} \end{pmatrix}. \tag{54}$$

This shows that $\frac{\partial p_A}{\partial \alpha_A} > 0$ and $\frac{\partial p_B}{\partial \alpha_A} < 0$.

Our reliance on the implicit function theorem assumes interior cutpoints, so to complete the proof we must consider the possibility that $p_A = 0$ or $p_B = 0$ (that is, at least one group is always conciliatory). In these cases increasing α_A *weakly* increases (decreases) p_A (p_B). In this case, the analysis above shows that increasing α_A either increases p_A or leaves it unchanged. This in turn weakly decreases p_B, also by the analysis above. Similarly, if $p_A = 1$ then increasing α_A has no effect. ■

Corollary 1. This corollary follows from the calculations for the equilibrium values of q_A and q_B in the previous proof. As (48) shows, the long-run probability of $\tilde{y}_{t-1}$ when player t is a member of group A and both groups are Friendly is increasing in α_A. Furthermore, the prrobability that $\tilde{y}_{t-1}$ when player t is a member of group A, group A is Friendly, and group B is Hostile is equal to

one regardless of the value of α_A. Thus, the overall probability that a Friendly group A observes an aggressive outcome in an arbitrary time period is increasing in α_A. The same argument holds for group B. ■

F Proof of Proposition 4

Proof. By Lemma 8, the long-run probability of conflict is one if $p_A = p_B = 1$. We will show that, for some values of α_j, a Friendly type from group j will choose $y_t = a$ following $\tilde{y}_{t-1}$ which implies that $p_A = p_B = 1$. Since we only need to find some α^* that guarantees perpetual conflict we consider the hardest case when $s = 0$ which implies the result for any $s > 0$ since the expected utility for choosing a conciliatory action is necessarily decreasing in s. When $s = 0$ a Friendly type of player t will choose $y_t = a$ following $\tilde{y}_{t-1} = a$ when

$$\begin{aligned} &\mu_t^a((1-\overline{\rho})(1-\pi) - \alpha_j(1-\pi)(\overline{\rho}r(H)) + (1-\overline{\rho}))) \\ &+(1-\mu_t^a)((1-\underline{\rho})(1-\pi) - \alpha_j(1-\pi)(\underline{\rho}r(H)) + (1-\underline{\rho}))) < 0. \end{aligned} \tag{55}$$

The explanation of this condition is as follows: player t beliefs that $\rho_{-j} = \overline{\rho}$ with probability μ_t^a and $\rho_{-j} = \underline{\rho}$ with probability $1 - \mu_t^a$. Given a value of ρ_{-j}, a Friendly type of player t's expected material payoff for playing $y_t = c$ is equal to the probability that $x_{t+1} = c$, which is $(1-\rho_{-j})(1-\pi)$. This is because $s = 0$ so the payoff when $x_{t+1} = a$ is equal to zero. Finally, player t subtracts from this payoff α_j times the expected benefit to player $t+1$ from playing $y_t = c$. This benefit is zero if palyer $t+1$ mistakenly gets an aggressive signal, which happens with probability π. With probability $1-\pi$ player $t+1$ gets a conciliatory signal, in which case the Hostile type plays $x_t = a$ and bets a benefit of $r(H)$, and the Friendly type plays $x_t = c$ and gets a benefit of 1.

This condition always holds if

$$\alpha_j > \frac{1 - \mu_t^a\bar{\rho} - (1-\mu_t^a)\underline{\rho}}{1 + (r(H)-1)\mu_t^a\bar{\rho} + (r(H)-1)(1-\mu_t^a)\underline{\rho}}. \tag{56}$$

Let

$$\alpha^* = \frac{1 - \mu_0\bar{\rho} - (1-\mu_0)\underline{\rho}}{1 + (r(H)-1)\mu_0\bar{\rho} + (r(H)-1)(1-\mu_0)\underline{\rho}}. \tag{57}$$

by Lemma 6, $\mu_t^a > \mu_0$. Thus, if $\alpha_j > \alpha^*$ then (56) holds for any posterior belief. Thus, $p_A = p_B = 1$ which means the long-run probability of conflict is one. ■

G Extension: More information about history

We consider the case in which an agent receives a signal $(\tilde{y}_{t-3}, \tilde{y}_{t-1})$ if $t > 3$. This resembles a situation with limited intergenerational transfer of information in which agents learn about experiences of the previous generation of her own group, or a diplomatic context in which knowledge transfer across, e.g. presidential administrations, does occur, but may become less reliable over longer time periods.

We assume each agent learns their own type and $\tilde{y}_{t-3}$ first, then updates their beliefs, and then receive $\tilde{y}_{t-1}$. This fits our intergenerational transfer interpretation: agents receive information about history before personally engaging in interactions with the other group. The consequence is that agents become angry when their expected payoff after learning $\tilde{y}_{t-1}$ is low relative to the reference point of their interim expected payoff *after learning about the history of the game*. Proposition G.1 states our result.

Proposition G.1. *In the model in which agents learn about the history of the game:*

1. *For small enough values of α_j, Friendly types from group $j \in \{A,B\}$ play the conciliatory action $y_t = c$ with probability one when either or $\tilde{y}_{t-3} = c$ or $\tilde{y}_{t-1} = c$.*

2. *For large enough values of α_j, Friendly types from group $j \in \{A,B\}$ play the aggressive action $y_t = a$ with probability one when $\tilde{y}_{t-3} = c$ and $\tilde{y}_{t-1} = a$.*

Proof. Agents believe that the probability of a Hostile type is zero when either $\tilde{y}_{t-3} = c$ or $\tilde{y}_{t-1} = c$ since no Hostile type ever plays a conciliatory action. The long-run probability of $\tilde{y}_{t-3} = \tilde{y}_{t-1} = a$ for an agent with conjectures b_t^A and b_t^B when both groups only consist of Friendly types is

$$\begin{aligned} q_j^{aa} =& \pi^2 + \pi(1-\pi)\left(\pi + (1-\pi)b_t^j(F,a)\right) b_t^{-j}(F,a) \\ &+ (1-\pi)\pi q_{-j} b_t^{-j}(F,a) + (1-\pi)^2 q_{-j} b_t^{-j}(F,a)^2 \left(\pi + (1-\pi)b_t^j(F,a)\right). \end{aligned} \tag{58}$$

The explanation is as follows: with probability π^2 player t would observe $\tilde{y}_{t-3} = \tilde{y}_{t-1} = a$ regardless of the actions of the players. With probability $\pi(1-\pi)$ player t will observe $\tilde{y}_{t-3} = a$ regardless of the action of that player and will observe $\tilde{y}_{t-1} = a$ only if $y_{t-1} = a$. Since player $t-2$ always observes an aggressive signal in this circumstance, the probability of this occurring is $\left(\pi + (1-\pi)b_t^j(F,a)\right) b_t^{-j}(F,a)$, since $\pi + (1-\pi)b_t^j(F,a)$ is the probability of $\tilde{y}_{t-2} = a$ given $\tilde{y}_{t-3} = a$ and $b_t^{-j}(F,a)$ is the probability of $y_{t-1} = a$ given $\tilde{y}_{t-2} = a$. Next, with probability $(1-\pi)\pi$ player t observes $\tilde{y}_{t-1} = a$ regardless of the action chosen and observes $\tilde{y}_{t-3} = a$ only if $y_{t-3} = a$. The probability that $y_{t-3} = a$ is $q_{-j} b_t^{-j}(F,a)$. Finally, with probability $(1-\pi)^2$ player t observes $\tilde{y}_{t-3} = \tilde{y}_{t-1} = a$ only if $y_{t-3} = y_{t-1} = a$. In this case, probability that $y_{t-3} = a$ is $q_{-j} b_t^{-j}(F,a)$,

the probability that $\tilde{y}_{t-2} = a$ conditional on $\tilde{y}_{t-3} = a$ is $\pi + (1-\pi)b_t^j(F,a)$, and the probability of $y_{t-1} = a$ given $\tilde{y}_{t-2} = a$ is $b_t^{-j}(F,a)$. All of this yields the equation above. Clearly $q_j^{aa} > q_j$, also implies that beliefs after seeing $\tilde{y}_{t-3} = \tilde{y}_{t-1} = a$ are higher than μ_t^a in the baseline model. Part 1 of the proposition follows from the fact that, for $\alpha_j \approx 0$, playing $y_t = c$ is always a best response to the belief that all members of the other groups are Friendly types with probability one. To prove Part 2, we must compute the agent's anger levels following the sequential revelation of information. First, the probability of a Friendly type of t observing $\tilde{y}_{t-1} = a$ after observing $\tilde{y}_{t-3} = c$ is $\pi(1 + (1-\pi)b_t^{-j}(F,a))$. That is, since player $t-2$ observes a conciliatory signal as well and plays a conciliatory action, and since both groups must be all Friendly types at this information set, there are only two ways to observe $\tilde{y}_{t-1} = a$: (1) player t observes an aggressive signal regardless of the action, which happens with probability π, or (2) player t's signal depends on the action (with probability $1-\pi$), player $t-1$ mistakenly sees an aggressive signal (which happens with probability π), and player $t-1$ then takes an aggressive action (which happens with probability $b_t^{-j}(F,a)$).[3] The probability that $y_{t-1} = a$ given $\tilde{y}_{t-3} = c, \tilde{y}_{t-1} = a$ is then $\frac{\pi b_t^{-j}(F,a)}{\pi(1+(1-\pi)b_t^{-j}(F,a)}$.

Given this information, agent t's decision utility for choosing $y_t = c$ after observing $\tilde{y}_{t-3} = c, \tilde{y}_{t-1} = a$ is

$$U_t^{ca} = 1 - \alpha_j(1 - \pi(1 + (1-\pi)b_t^{-j}(F,a)))\frac{\pi b_t^{-j}(F,a)}{\pi(1 + (1-\pi)b_t^{-j}(F,a))}. \tag{59}$$

Thus, if $\alpha_j > \frac{1}{(1-\pi(1+(1-\pi)b_t^{-j}(F,a)))\frac{\pi b_t^{-j}(F,a)}{\pi(1+(1-\pi)b_t^{-j}(F,a))}}$ then $U_t^{ca} < 0$ and therefore the agent chooses $y_t = a$ for all values of s. ■

Proposition G.1 shows that, when agents have more information about history, strategies depend on the history of the game in ways that are very distinct from a standard model without endogenous reference points. When either previous signal was conciliatory, the agent is certain that the other group is Friendly. With no anger – or with low enough sensitivity to anger – a Friendly type is therefore always conciliatory. However, consider how anger interacts with this information. When $\tilde{y}_{t-3} = c$, agent t thinks that it is very likely that $\tilde{y}_{t-1} = c$. However, there is some chance that $\tilde{y}_{t-1} = a$ and, when that happens, the deviation between agent t's expected and actual payoff is even larger than it would have been had she not learned about the history of the game. This triggers an anger response which, for large enough values of α_j, causes an aggressive response. Thus, anger changes the effect of history on current behavior. Furthermore, this behavior cannot be rationalized by beliefs about material payoffs: agent t is certain that agent $t+1$ is a Friendly type but chooses the aggressive action anyway.

[3]Recall that playing c is a best response to c at all anger levels.

This result highlights how the psychological game produces outcomes that are impossible when agents care about only the material payoffs. An agent may *know for sure* that the other group is Friendly and still act aggressively.

The extension here expands the baseline model only by giving each player to one additional signal about recent actions taken by members of the other group. An alternative approach would be to expand the player's information set to include all recent history including actions taken by members of one's own group (for instance by showing player t outcomes from $t-3$, $t-2$, and $t-1$ rather than only $t-1$ and $t-3$). However, we took this more limited approach in order to maintain an important and realistic feature from the original model, which is that players observe outcomes of interactions with the other group without fully understanding the context of the other group members' actions. This is important here because joint knowledge of θ_{t-1} and $\tilde{y}_{t-2}$ would allow player t to perfectly learn the intended action of player $t-1$ regardless of the outcome. That said, Proposition G.1 is robust to letting players learn some information about the actions of members of their own group as long as this feature is retained. For instance, the proof of Proposition G.1 accounts for the fact that player t makes inferences (sometimes perfect ones) about the actions of player $t-2$ given the information she receives. Thus, giving the agent information about y_{t-2} (as opposed to $\tilde{y}_{t-2}$) does not significantly alter the argument. Furthermore, giving player t a noisy signal of $\tilde{y}_{t-2}$ would complicate the proof but not significantly change the result.

H Extension: Concern for others' beliefs

We modify psychological preferences in the baseline model so that

$$\beta_i t(\tilde{y}_{t-1}, y_{t-1}; b_t^j, b_t^{-j}) := \begin{cases} \Pr[\tilde{y}_{t-1} = c | b_t^j, b_t^{-j}] I_a(y_{t-1}) \left[\mu_t^a + (1-\mu_t^a)(1-\mu_{t-1}^a)\right] & \text{if } \tilde{y}_{t-1} = a \\ 0 & \text{if } \tilde{y}_{t-1} = c. \end{cases} \tag{60}$$

Thus, the model with concern for beliefs is identical to the baseline model except that anger motivations depend on second order beliefs through the inclusion of the term $\mu_t^a + (1-\mu_t^a)(1-\mu_{t-1}^a)$. The idea of this component is that, when the other player is a Friendly type, the agent's anger is muted to the extent that the action could have been justified by the belief that player t was a Hostile type. Since a Friendly type who chooses y_{t-1} must have had $\tilde{y}_{t-2} = a$, μ_{t-1}^a is also player t's expectation of player $t-1$'s belief in the event that they were Friendly and chose y_{t-1}, so in this way concern for intentions involves second-order beliefs. The reasoning is that player t is forgiving of aggressive actions by player $t-1$ to the extent that she believes those actions to have been motivated by a belief that she was a Hostile type. Note that this understanding differs from that in Battigalli, Dufwenberg and Smith (2019), since agent $t-1$ would still expect to reduce a Hostile

type's payoff by choosing $y_{t-1} = a$ rather than $y_{t-1} = c$. One way of reconciling these views is that we imagine player t views the Hostile type as a different player from herself and views reducing a Hostile type's payoff as morally defensible and therefore not something that triggers anger.

Proposition H.1. *Under Assumption 1, in the model where agents have concern for others' beliefs, in addition to actions:*

1. *Hostile types play $x_t = a$ and $y_t = a$ at any information set, Friendly types at any time $t > 0$ play $x_t = \tilde{y}_{t-1}$, Friendly types play $y_t = c$ when $\tilde{y}_{t-1} = c$ or when $t = 0$, and Friendly types from group j play $y_t = a$ with probability $p_j \in [0,1]$ when $\tilde{y}_{t-1} = a$.*

2. *Certain conflict cannot occur in equilibrium*

3. *Increasing a group's sensitivity to anger (i.e. increasing α_j for some group j) increases the probability that Friendly members of that group choose an aggressive action in response to an aggressive outcome and decreases the probability that Friendly members of the other group choose an aggressive action in response to an aggressive outcome.*

4. *Holding the rest of the parameters constant, the effect of anger on the probability that any Friendly type chooses an aggressive action is lower than in the baseline model.*

Proof. Part 1 follows from the same analysis as above, since incentive compatibility calculations are unchanged for Hostile types, for choices of x_t, and for Friendly types when $\tilde{y}_{t-1} = c$. Furthermore, Part 2 follows from the proof of Proposition 2 since anger still vanishes if the long-run probability that $\tilde{y}_{t-1} = a$ goes to one. To prove Part 3, note that a Friendly type from group j's decision utility for playing $y_t = c$ when $\tilde{y}_{t-1} = a$ is

$$U_j^c = \mu_t^a \left[-s - \alpha_j (1-\mu_0)(1-q_j) \frac{q_{-j}}{q_j} p_{-j} (1-\pi) r(H) \right] + (1-\mu_t^a) \left[\pi s + (1-\pi) \left(1 - \alpha_j (1-\mu_0)(1-q_j) \frac{q_{-j}}{q_j} p_{-j} (1-\mu_{t-1}^a) \right) \right]. \tag{61}$$

As in the proof of Proposition 3, we let $(q_A^*(s), q_B^*(s_t'))$ be critical values of q_A and q_B that make $U_A^c = 0$ and $U_B^c = 0$ for particular values of s and s_t'. By the implicit function theorem we have

$$\frac{\partial q_A^*}{\partial \alpha_A} = -\frac{\frac{\partial U_{A1}}{\partial \alpha_A}}{\frac{\partial U_{A1}}{\partial q_A}} \tag{62}$$

$$\frac{\partial q_B^*}{\partial \alpha_A} = 0, \tag{63}$$

as before. The rest of the proof follows from the same reasoning as the proof of Proposition 3. Finally, Part 4 follows from the fact that anger is strictly lower at any information set when agent's consider intentions. ■

I Extension: Informational effects of anger

Instead of forming beliefs using Bayes's rule, agents' beliefs minimize a separable function of accuracy motives and "directional" motives. Our directional motives reflect the idea that an agent who is made angry will be more apt to conclude that members of the other group are Hostile types.

Letting μ_t^a and μ_t^c be the proper Bayesian beliefs following $\tilde{y}_{t-1} = a$ and $\tilde{y}_{t-1} = c$ respectively, agent t's beliefs are equal to

$$\hat{\mu}_t = \begin{cases} \arg\max_{\hat{\mu}} \left[\alpha_j \Pr[\tilde{y}_{t-1} = c | b_t^j, b_t^{-j}] \mathbb{E}_{\hat{\mu}}[\rho_{-j}] - KL(\hat{\mu} || \mu_t^a) \right] & \text{if } \tilde{y}_{t-1} = a \\ \mu_t^c & \text{if } \tilde{y}_{t-1} = c, \end{cases} \tag{64}$$

where $KL(\hat{\mu}||\mu_t^a)$ is the Kullback-Liebler (KL) divergence between $\hat{\mu}$ and μ_t^a and ρ_{-j} is the proportion of Hostile types in the other group. The use of KL divergence is due to Little (2021) and is a common way to measure the difference between probability distributions. Furthermore, agents maximize only their material payoffs so that anger in this model enters only through beliefs and not preferences.

In this language of motivated reasoning, $\alpha_j \Pr[\tilde{y}_{t-1} = c | b_t^j, b_t^{-j}] \mathbb{E}_{\hat{\mu}}[\rho_j]$ is the agent's directional motive. We make four observations about directional motives. First, α_j is a weight on directional motives relative to accuracy motives. In this sense, α_j serves as a measure of the agents' sensitivity to anger just as in the original model, though referring to sensitivity of beliefs rather than preferences. Second, we retain the context-dependence of anger by including $\Pr[\tilde{y}_{t-1} = c | b_t^j, b_t^{-j}]$ in the directional motives. As in the original model, anger is relevant when outcomes diverge from expected outcomes. Third, we omitted the term $I_a(y_{t-1})$ which appeared in the baseline model from directional motives here, so the agent does not explicitly consider the likelihood that harm was caused by the actions of the other player. This is done purely for convenience: since the likelihood that $y_{t-1} = a$ depends on posterior beliefs about types, the directional motive would depend on beliefs and solving for optimal beliefs would require an additional fixed point argument. Motivated reasoning with fully endogenous beliefs is an interesting area of potential future research. Finally, the term $\mathbb{E}_{\hat{\mu}}[\rho_j]$ implies that angrier agents are more motivated to adopt beliefs that lead to higher expected proportions of Hostile types in the other group.

By Theorem 1 in Little (2021) and the fact that $\Pr[\tilde{y}_{t-1} = c | b_t^j, b_t^{-j}] = 1 - q_j$, the optimal motivated belief when $\tilde{y}_{t-1} = a$ is

$$\hat{\mu}_t^a = \frac{\mu_t^a e^{\alpha(1-q_j)}}{\mu_t^a e^{\alpha(1-q_j)} + (1 - \mu_t^a)}. \tag{65}$$

We establish in Proposition I.1 that the main results from the baseline model can be recovered by the model in which anger operates through information processing, though we highlight some

qualitative differences below the proposition.

Proposition I.1. *In the motivated reasoning model, under Assumption 1, the following are true:*

1. *Hostile types play $x_t = a$ and $y_t = a$ at any information set, Friendly types at any time $t > 0$ play $x_t = \tilde{y}_{t-1}$, Friendly types play $y_t = c$ when $\tilde{y}_{t-1} = c$ or when $t = 0$, and Friendly types from group j play $y_t = a$ with probability $p_j \in [0,1]$ when $\tilde{y}_{t-1} = a$.*

2. *Certain conflict cannot occur in equilibrium*

3. *Increasing a group's sensitivity to anger (i.e. increasing α_j for some group j) increases the probability that members of that group choose an aggressive action and decreases the probability that members of the other group choose an aggressive action.*

Proof. Since

$$LR(\hat{\mu}_t^a) = \frac{\mu_t^a}{1-\mu_t^a} e^{\alpha_j(1-q_j)} > \frac{\mu_t^a}{1-\mu_t^a}, \tag{66}$$

observing $\tilde{y}_{t-1} = a$ weakly increases the likelihood that t plays $y_t = a$. The calculations for Hostile types, for choices of x_t, and for choices following $\tilde{y}_{t-1}$ do not depend on anger calculations so the rest of the conclusions of Part 1 follow from previous analysis.

Furthermore, under perpetual conflict, we have $q_j = 1$ by Lemma 8 and the optimal belief is

$$\frac{\mu_t^a e^0}{\mu_t^a e^0 + (1-\mu_t^a)} = \mu_t^a = \mu_0 \tag{67}$$

where the last equality follows from the fact that $\mu_t^a = \frac{\mu_0}{\mu_0 + (1-\mu_0)q_j} = \mu_0$ when $q_j = 1$. Thus, under Assumption 1 agent t's best response is to play $y_t = c$ which implies that perpetual conflict is not an equilibrium.

To prove Part 3, consider again a change in α_A, since the argument works symmetrically for α_B. The direct effect of increasing α_A is clearly to increase $\hat{\mu}_t^a$, which increases the likelihood that a member of group A plays $y_t = a$, through the informational effects already outlined. Furthermore, α_A does not appear in the decision utility of B members so the only effect of increase α_A is to increase q_B, which reduces $\hat{\mu}_t^a$ for group B members and therefore makes them more likely to choose $y_t = c$. ■

The insight of Proposition I.1 is that we can recover the main predictions of our model using a formulation of anger that works through beliefs rather than preferences. However, the two models are not identical. First, in the preference-based model, Friendly types are more likely to take the aggressive action when $r(H)$ is larger because they get disutility when another agent benefits from their conciliatory actions. This preference for punitiveness does not drive behavior in the motivated

reasoning model so we do not make this prediction. Secondly, we informally compare what would happen with additional signals as in Section 6.2. In that model, if anger is high enough, the Friendly agents choose $y_t = a$ when they get a conciliatory signal followed by an aggressive signal. They do this even though they are fully convinced that the other group is made up of Friendly types. That cannot happen here. Why? Note the optimal motivated belief. When $\mu_t = 0$, the optimal belief also is zero. Since beliefs rather than preferences drive behavior in the motivated reasoning model, and since motivated reasoning converges to Bayesian reasoning when all uncertainty is eliminated, this prediction would fail under these circumstances. If players believe the other side is certainly Friendly, they will never take an aggressive action.

References

Battigalli, Pierpaolo, Martin Dufwenberg and Alec Smith. 2019. "Frustration, aggression, and anger in leader-follower games." *Games and Economic Behavior* 117:15–39.

Little, Andrew T. 2021. "Detecting Motivated Reasoning.".
URL: *osf.io/b8tvk*