Cycles of Distrust: An Economic Model*

Daron Acemoglu
MIT

Alexander Wolitzky
Stanford and Microsoft Research

July 17, 2012

Abstract

We propose a model of cycles of distrust and conflict. Overlapping generations of agents from two groups sequentially play coordination games under incomplete information about whether the other side consists of "extremists" who will never take the good/trusting action. Good actions may be mistakenly perceived as bad/distrusting actions. We also assume that there is limited information about the history of past actions, so that an agent is unable to ascertain exactly when and how a sequence of bad actions originated. Assuming that both sides are not extremists, spirals of distrust and conflict get started as a result of a misperception, and continue because the other side interprets the bad action as evidence that it is facing extremists. However, such spirals contain the seeds of their own dissolution: after a while, Bayesian agents correctly conclude that the probability of a spiral having started by mistake is sufficiently high, and bad actions are no longer interpreted as evidence of extremism. At this point, one party experiments with a good action, and the cycle restarts. We show how this mechanism can be useful in interpreting cycles of ethnic conflict and international war, and how it also emerges in models of political participation, dynamic inter-group trade, and communication—leading to cycles of political polarization, breakdown of trade, and breakdown of communication.

Keywords: communication, cooperation, coordination, ethnic conflict, distrust, polarization, trust, overlapping generations.
JEL Classification: D74, D72.

*We thank Sandeep Baliga, Sylvain Chassang, Edward Glaeser, James Fearon, Jon Levin, Qingmin Liu, and seminar participants at the Canadian Institute for Advanced Research, Microsoft Research, the NBER Political Economy Conference, the Northwestern Conflict and Cooperation Conference, and Stanford for useful comments. Acemoglu gratefully acknowledges financial support from the Canadian Institute for Advanced Research and the ARO.

1 Introduction

Mutual benefits from trust, cooperation, and communication not withstanding, inter-group conflict is pervasive. In his study of the Peloponnesian War, Thucydides (2000) traces the origins of conflict as much to fear and distrust as to other factors such as greed and honor. He argues that the Peloponnesian War became inevitable precisely because each side saw war as inevitable and did not want to relinquish the first mover advantage to the other (see also Kagan, 2004).[1] This view of conflict, sometimes referred as the Hobbesian view or spiral model, has a clear dynamic implication: if Group A's actions look aggressive, Group B infers that Group A is likely to be aggressive and acts aggressively itself (e.g., Jervis, 1976, Kydd, 1997). Moreover, unless Group A can fully understand that Group B is acting aggressively in response to its own actions, it will take this response as evidence that Group B is aggressive. As a result, conflict spirals.

The ubiquity of "conflict spirals" throughout history provides *prima facie* support for this view. A leading example is ethnic conflict: Donald L. Horowitz argues that "The fear of ethnic domination and suppression is a motivating force for the acquisition of power as an end" (Horowitz, 2000, p. 187), and suggests that such fear of ethnic domination was the primary cause of the rise in ethnic violence following the withdrawal of colonial powers.[2] Horowitz also suggests (p. 189, italics in the original): "The imminence of independence in Uganda aroused 'fears of future ill-treatment' along ethnic lines. In Kenya, it was 'Kikuyu domination' that was feared; in Zambia, 'Bemba domination'; and in Mauritius, ... ['Hindu domination']... Everywhere the word *domination* was heard. Everywhere it was equated with political control."

More recent examples of such spirals are provided by conflicts in Northern Ireland, the Balkans, Lebanon, Iraq, Gaza and the West Bank, and Turkey. For instance, many accounts of the dynamics of the Serbian-Croatian war emphasize Croatian fears from the aggressive posturing of Milosevic, which were instrumental in triggering more aggressive Croatian actions, including the adoption as the national symbol of the sahovnica, associated with the fascist pre-Yugoslavia Ustasha regime, and a variety of discriminatory policies towards the Serbian minority (e.g., Posen, 1993). These actions then spiraled into all-out war.[3] Spiral effects might account not only for violent conflict between nations and

[1] The fear motive for conflict is also referred to as the "Hobbesian trap" or the "security dilemma" (following Schelling, 1960). It is modeled by, among others, Baliga and Sjostrom (2004) and Chassang and Padro i Miquel (2010).

[2] Horowitz also writes: "In this atmosphere of uncertainty, the greatest group anxiety was to avoid creating an old colonialism for a new one..." (p. 188), and quotes from James S. Coleman's (1958) study of ethnic conflict in Nigeria that "in a self-governing Nigeria the north would in effect be a backward protectorate governed by Southerners... The threat of sudden domination, fancied or real, was the major stimulant in the northern awakening." Robert N. Kearney (1967) describes the intensification of ethnic conflict in Sri Lanka (Ceylon): "The gradual transfer of power from foreign to Ceylonese hands quickly created concern for the relative political strength of the various communities. The basic assumption upon which this concern rested was that the share of political power held by members of one community would be used for the exclusive benefit of that community or to the detriment of other communities."

[3] DellaVigna et al. (2011) provide further evidence highly suggestive of a conflict spiral in this context. They show that Croatians who received nationalistic radio broadcasts from the Serbian side were more nationalistic and more supportive of anti-Serbian actions. Kaplan et al. (2005) provide evidence consistent with a "cycle of violence" from the Israeli-Palestinian conflict (but see also Jaeger and Paserman, 2008).

ethnic groups, but also for lack of trust and communication between groups and within organizations. Guiso, Sapienza and Zingales (2009) document deep-rooted distrust among some nations, and show that it is associated with lower international trade and foreign direct investment, and Bottazzi, Da Rin, and Hellmann (2011) show a similar pattern of international business ventures. Kramer (1999) surveys a large social psychology literature documenting the emergence and persistence of distrust within organizations. In most cases, distrust is triggered by suspicion that others are untrustworthy or are pursuing ulterior motives. Some see the root cause of the increasing polarization in US politics also in this type of spiral effects which may have made each side segregate into their own "echo-chambers" where they only listen to like-minded communication (e.g., Sunstein, 2006).[4]

This classical view of conflict and distrust is incomplete, however, because it only explains how conflict *starts* and not how it *stops*—even though most conflict spirals come to an end sooner or later. For example, sectarian conflict in Northern Ireland has ended starting with a cease-fire in 1994 ultimately leading to the Good Friday agreement in 1998; the widespread distrust and conflict between blacks and whites in South Africa has largely subsided after the fall of the apartheid regime in 1994; war and conflict between different ethnic and national groups in the Balkans have mostly ended; and the historical Franco-German distrust and animosity has made way to vibrant trade and economic and diplomatic cooperation. The bloody ethnic wars that seemed intractable after the end of World War II have dramatically abated over the past two decades. Even if political polarization in the US seems incorrigible today, a similar era of polarization in the first third of the 20th century was followed by a long period of non-partisan politics (McCarty, Poole and Rosenthal, 2008). So rather than *infinite* conflict spirals—where conflict once initiated never subsides—history for the most part looks like a series of *conflict cycles*, where even long periods of conflict eventually end, often quite suddenly and unexpectedly.

This paper proposes a simple model of conflict spirals, and then shows that such spirals contain the seeds of their own dissolution—thus accounting not only for the onset but also the end of conflict. The basic idea of our approach is simple: once Groups A and B get into a phase in which they are both acting aggressively, the likelihood that a conflict spiral has been triggered by mistake at some point increases over time. This implies that aggressive actions—which are typically informative about how aggressive the other side is—eventually become uninformative. Once this happens, one group will find it beneficial to experiment with cooperation and, unless the other group is truly aggressive, cooperation will resume—until the next conflict spiral begins.

Formally, our model features a coordination game between overlapping generations of (representatives of) two groups. The "bad" action in the coordination game may correspond to fighting or initiating other types of conflicts, and is a best response to bad actions from the other party, while the "good" action is optimal when good actions are expected. Each side is uncertain about the type of their opponents, who may be—with small probability—committed to the bad action. The two distinguishing features of our approach are: (1) noisy Bayesian updating, so that individuals (groups)

[4] See McCarty, Poole and Rosenthal (2008) and Abramowitz (2011) on polarization of US politics.

understand that conflict may be initiated because of a misperception or unintended action; and (2) "finite memory," so that there is limited information about the exact sequence of events in the past, and when and for what reason conflict has started will be unknown. These features together generate a distinctive pattern where, in the unique sequential equilibrium of this dynamic game, a spiral of distrust and conflict is sometimes initiated and persists, but must also endogenously come to an end. The first contribution of our model is to show that because of limited information about the past, when the current generation sees conflict (but not how it came about), it often responds by choosing a bad action, perpetuating the spiral.[5] The main contribution of our model is to show that such spirals of conflict will eventually terminate: when an individual or group reasons that there have been "enough" chances for a conflict spiral to have gotten started (call this number T), they will conclude that the likelihood that it started by mistake—rather than being started intentionally by a truly aggressive adversary—is sufficiently high, and they will experiment with the good action. In our baseline model, these two forces lead to a unique equilibrium which features a mixture of deterministic and stochastic cycles. In particular, a single misperceived action stochastically initiates a conflict spiral, which then concludes deterministically at the next time t that is a multiple of T.[6]

The rest of the paper examines the robustness of the forces we isolate in this simple dynamic game and argues that they are relevant for thinking about cycles of distrust in a variety of situations. First, in subsection 2.3 we argue that our baseline game is a natural model of ethnic or international war. In addition to capturing the essence of the Hobbesian view of conflict, ours is a direct overlapping generations analog of the models used by Baliga and Sjostrom (2004) and Chassang and Padro i Miquel (2010) to study related issues in a static setting. Our model implies that the unique equilibrium in the context of ethnic and international conflict may involve cycles of distrust fueling cycles of conflict—so that neither war nor peace is an absorbing state.

We then turn to extensions of the baseline model in Section 3, which show that our basic insights are robust to relaxing a variety of assumptions. These include versions of the model without deterministic dependence on calendar time and versions with more information about past actions.

We believe that forces similar to those in our baseline model are important for understanding other instances of the cycle of distrust. However, these applications typically necessitate somewhat different and more detailed assumptions than in our baseline model. Section 4 develops one such application. There, we study the dynamics of political partisanship and participation using a model

[5] While informal accounts of conflict often invoke spirals (e.g., Posen 1993), they do not clarify the conditions under which such spirals will emerge. For example, full observation of the history of signals and actions would preclude spirals for the following reason: if Group A *knows* that Group B perceived her initial action as aggressive and responded aggressively, then Group A should *not* respond aggressively in turn (as she knows that Group B would have behaved this way even if he were not inherently aggressive). In our model, limited information about past signals, as well as about when conflict started, prevents this type of perfect inference and makes spirals possible.

[6] We certainly do not claim that every possible model of conflict spirals leads to cycles. For example, Rohner, Thoenig and Zilibotti (2011) develop a dynamic "Hobbesian" model of conflict where information about a group's type accumulates over time, leading asymptotically to either permanent war or permanent peace. The key difference is that, because of finite memory, information does not accumulate in our model—as decision makers have limited information about the past—so there is no "asymptotic learning."

where extremist politicians care primarily about a partisan political issue, while moderate politicians care primarily about a common-interest issue. A cycle of partisanship emerges, where politicians act on the partisan issue when they think the other party is extremist, and act on the major issue when they think the other party is moderate. Augmenting this model with a model of voter turnout, we predict a cycle of voter turnout that mirrors the cycle of partisanship: turnout and propaganda are both high in the partisan phase of the cycle, while turnout is low and communication between the parties is informative in the cooperative phase. This model may capture aspects of the cycle of political partisanship documented by McCarty, Poole and Rosenthal (2008), Abramowitz (2011), and others.[7]

Two further applications are examined in Appendix A. The first considers cycles of trade breakdown: We posit that some groups are unable to produce high-quality goods, while others can produce them at a cost. This yields a coordination game because this cost is only worth paying when the other group will also produce high-quality goods for trade, and we show that equilibrium is essentially unique and cyclical as in the baseline model. The second additional application is to cycles of "miscommunication" between two groups with different political views. Each group may be extremist and thus stubbornly repeat their own views, while normal groups are willing to moderate their communication in order to influence the other party. This model yields a cycle where phases of informative communication alternate with phases of uninformative, "ideological," communication.[8]

Our paper also relates to several strands of the literature on dynamic games, including anti-folk theorems in overlapping generations games (e.g., Lagunoff and Matsui, 1997) and reputation with limited records (e.g., Liu and Skrzypacz, 2011). We discuss the relationship of our paper to these and other literatures in Section 5. Section 6 concludes, while Appendix B contains proofs omitted from the text.

2 Baseline Model: Trust Game

In this section, we present our baseline model, which formalizes in the simplest possible way how conflict spirals can form but cannot last forever when individuals are Bayesian and have limited information about the history of conflict. At the end of this section, we present our first application,

[7] Notably, our model generates such cycles even though the electorate itself does not become more polarized, which is consistent with the evidence presented in Fiorina (2011).

Other related work includes Bernhardt, Krasa, and Polborn (2008), Chan and Suen (2008), and Gul and Pesendorfer (2011), who present models linking polarization to competition among media outlets. One interpretation of "cooperation" in our model is that the cooperative action is engaging in an informative political dialogue that gives the other party the information it needs to set policy, while the uncooperative action is broadcasting propaganda to one's base to try to maximize turnout. On this interpretation, our model complements these papers by studying the incentives for political parties themselves to provide useful information rather than propaganda.

[8] This may be viewed as a dynamic model of social learning, where ideological extremism—or merely the fear that the other side is extreme—can prevent information aggregation. In this respect, it generalizes Morris (2001) as well as Prendergast (1993), Canes-Wrone, Herron, and Shotts (2001), Maskin and Tirole (2004), and Acemoglu, Egorov and Sonin (2011). Relatedly, Banerjee and Somanathan (2001) and Sethi and Yildiz (2009) present models where social learning is precluded because of heterogeneous and unknown biases among members of society.

ethnic conflict and international war. We relax several of the simplifying assumptions adopted here in Section 3.

2.1 Model and Equilibrium Characterization

Two groups, Group A and Group B, interact over time $t = 0, 1, 2, \ldots$. At every time t, there is one active player ("player t") who takes a pair of actions $(x_t, y_t) \in \{0,1\} \times \{0,1\}$, where $x_t = 1$ and $y_t = 1$ are "good" (or "honest") actions and $x_t = 0$ and $y_t = 0$ are "bad" (or "cheating") actions; as will become clear, x_t is player t's action toward player $t-1$, and y_t is player t's action toward player $t+1$. In even periods the active player is a member of Group A, and in odd periods the active player is a member of Group B. A different player is active every period. A key assumption is that players observe very little about what happened in past periods: in particular, before player t takes her action, she observes only a signal $\tilde{y}_{t-1} \in \{0,1\}$ of player $t-1$'s action toward her. This assumption captures the feature that agents may know that there is currently conflict or cheating without fully knowing when and how this was initiated in the past. We assume that $\tilde{y}_{t-1}$ is determined as:

$$\begin{aligned} \Pr\left(\tilde{y}_{t-1} = 1 | y_{t-1} = 1\right) &= 1 - \pi \\ \Pr\left(\tilde{y}_{t-1} = 1 | y_{t-1} = 0\right) &= 0, \end{aligned}$$

where $\pi \in (0,1)$.[9] Thus, a good action sometimes leads to a bad signal, but a bad action never leads to a good signal (both this in the assumption that nothing from the past history beyond last period's conflict is observed is relaxed Section 3).

Each group consists either entirely of normal types or entirely of bad types. The probability that a group is bad (i.e., consists of bad types) is $\mu_0 > 0$. Playing $(x_t = 0, y_t = 0)$ is a dominant strategy for the bad type of player t. For $t > 0$, the normal type of player t has utility function

$$u\left(x_t, \tilde{y}_{t-1}\right) + u\left(\tilde{y}_t, x_{t+1}\right),$$

so her payoff is the sum of her payoff against player $t-1$ and her payoff against player $t+1$.[10] By writing payoffs as a function of the realized signal of player $t-1$'s action, we are following the literature on dynamic games of incomplete information in ensuring that no additional information is obtained from realized payoffs. The normal type of player 0 has utility function $u\left(\tilde{y}_0, x_1\right)$.[11] We assume that each "subgame" between neighboring players is a coordination game, and that $(1,1)$ is the Pareto-dominant equilibrium:

[9] There are several ways of interpreting the misperception probability π. One group may literally misperceive the other's action, or a group's leaders may try to do one thing but mistakenly do another.

An alternative assumption which is entirely identical is that a fraction π of each group's membership may be "bad types" (even when the group's type is normal) who always play 0, perhaps because they are "provacateurs" who benefit from sending the groups into conflict (cf. Rabushka and Shepsle, 1972, Glaeser, 2005, Baliga and Sjostrom, 2011).

[10] Changing this utility function to $(1-\lambda)\, u\left(x_t, \tilde{y}_{t-1}\right) + \lambda u\left(\tilde{y}_t, x_{t+1}\right)$ for $\lambda \in (0,1)$ would have no effect on the results or in fact on the expressions that follow.

[11] Note that this makes action x_0 irrelevant, so we ignore it (equivalently, assume that player 0 only chooses $y_0 \in \{0,1\}$).

Assumption 1 1. $u(1,1) > u(0,1)$.

2. $u(0,0) > u(1,0)$.

3. $u(1,1) > u(0,0)$.

We also assume that the probability that a group is bad is below a certain threshold $\mu^* \in (0,1)$:

Assumption 2

$$\mu_0 < \mu^* \equiv \frac{u(1,1) - u(0,0)}{u(1,1) - u(1,0)}.$$

Assumption 2 is equivalent to assuming that normal player 0, with belief μ_0, plays $y_0 = 1$ when she believes that player 1 plays $x_1 = 1$ if and only if he is normal and sees signal $\tilde{y}_0 = 1$.

We can now explain the logic of the model. Assumption 1 ensures that in any sequential equilibrium player t does indeed play $x_t = 1$ if and only if he is normal and sees signal $\tilde{y}_0 = 1$. In view of this, Assumption 2 implies that normal player 0's prior about the other group is sufficiently favorable that she starts out with $y_0 = 1$.

Next, consider the problem of normal player 1. If he sees signal $\tilde{y}_0 = 1$, then he knows the other group is normal—since bad types take the bad action which never generates the good signal. In this case, his belief about the other group is even better than player 0's, so he plays $y_1 = 1$ (in addition to playing $x_1 = 1$). But what if he sees signal $\tilde{y}_0 = 0$? He clearly plays $x_1 = 0$, but moreover, by Bayes rule, his posterior belief that the other group is bad rises to

$$\mu_1 = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\pi} > \mu_0,$$

which follows in view of the fact that $\tilde{y}_0 = 0$ may have resulted from the other side being extremist, probability μ_0, or due to a bad signal following from the good action when the other side is normal, probability $(1 - \mu_0)\pi$. Now if μ_1 is sufficiently high—in particular, if it is above the cutoff belief μ^*—then player 1 plays $y_1 = 0$ after she sees signal 0.[12]

Now suppose that up until time t normal players play $y_t = 0$ after seeing signal 0, and consider the problem of normal player t. Again, if she sees signal 1, she knows the other group is normal and plays $(x_t = 1, y_1 = 1)$. But if she sees signal 0, she knows that this could be due to a bad signal arriving at *any* time before t, because a single bad signal starts a spiral of bad actions. Thus, her posterior is

$$\mu_t = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\left(1 - (1 - \pi)^t\right)},$$

which follows since the probability of no bad signal arriving at any time before t, conditional on the other side being normal, is $(1 - \pi)^t$, and thus the total probability of player t seeing $\tilde{y}_{t-1} = 0$ is $\mu_0 + (1 - \mu_0)\left(1 - (1 - \pi)^t\right)$.

[12] We do not assume that $\mu_1 > \mu^*$. But if $\mu_1 < \mu^*$, then the conflict "cycle" that emerges is the trivial cycle where cooperation always restarts immediately after a misperception.

If μ_t is above the cutoff belief μ^*, then player t again plays $y_t = 0$ after seeing signal 0. Crucially, note that μ_t is decreasing in t, and that furthermore $\mu_t \to \mu_0$ as $t \to \infty$. Recall that $\mu_0 < \mu^*$. Thus, there is some first time T—given by (2) in Appendix B—at which $\mu_T \leq \mu^*$. And at this time, player T plays $y_T = 1$ *even if she sees signal* 0. Thus, any spiral of bad actions that started before time T ends at T.

Finally, consider the problem of normal player $T+1$. He knows that player T plays $y_T = 1$ if and only if she is normal. Thus, player $T+1$ is in exactly the same situation as player 1, and play from period $T+1$ on is exactly like play from period 1 on. Hence, play is characterized by cycles of length T, in which a single bad signal starts at some time t starts a spiral of bad actions that lasts until the next multiple of T.

A central feature of the above argument is that it holds regardless of beliefs about future play. Consequently, equilibrium is unique up to one technicality: if μ_T exactly equals μ^*, then cycles can be of length either T or $T+1$, and this can eventually lead to "restarts" of cooperation occurring at a wide range of times. To avoid this possibility, we make the following genericity assumption on the parameters:

Assumption 3 $\mu_t \neq \mu^*$ for all $t \in \mathbb{N}$.

We now state our main result for the baseline model, establishing that there is a unique equilibrium which is cyclic. The same cyclic equilibrium structure will arise in all of the more detailed applications studied later in the paper.

Proposition 1 *Under Assumptions 1-3, the baseline model has a unique sequential equilibrium. It has the following properties:*

1. *At every time $t \neq 0 \bmod T$, normal player t plays good actions $(x_t = 1, y_t = 1)$ if she gets the good signal $\tilde{y}_{t-1} = 1$, and plays bad actions $(x_t = 0, y_t = 0)$ if she gets the bad signal $\tilde{y}_{t-1} = 0$.*

2. *At every time $t = 0 \bmod T$, normal player t plays the good action $x_t = 1$ toward player $t-1$ if and only if she gets the good signal $\tilde{y}_{t-1} = 1$, but plays the good action $y_t = 1$ toward player $t+1$* regardless of her signal.

3. *Bad players always play bad actions $(x_t = 0, y_t = 0)$.*

It is straightforward to turn the above discussion into a proof of Proposition 1, and we omit this formal proof.[13]

Figures 1 and 2 graph the probability of observing conflict (i.e., the bad signal) and a normal player's posterior assessment of the probability that the other group is bad after observing conflict in the last period (as a function of time t) when both groups are normal (these are respectively

[13] The proof is similar to—but simpler than—the proof of Propostion 10, which is in Appendix B.

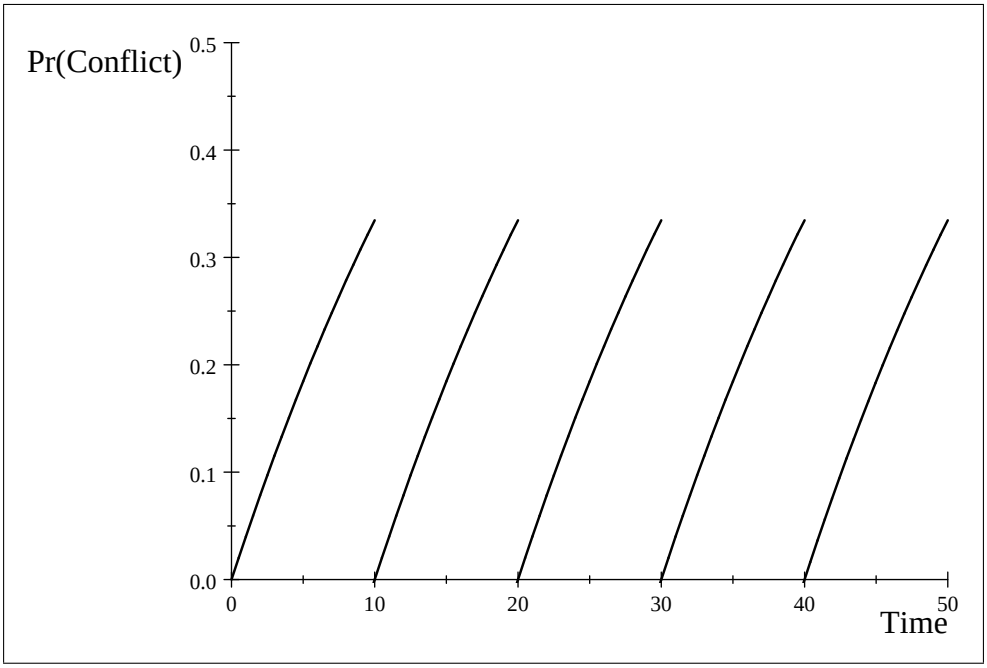

Figure 1: A Cycle of Conflict

given by $1-(1-\pi)^{t \bmod T}$ and $\mu_{t \bmod T}$). The parameter values used for constructing this figure are: $\mu_0 = 1/10$, $\pi = 1/25$, and $(u(1,1) - u(0,0)) / (u(1,1) - u(1,0)) = 1/4$.[14] These figures illustrate the distinguishing features of cycles in this model, which have both stochastic and deterministic elements. With these parameters, the period of the conflict cycle T equals 10, and consequently the figures show that cooperation restarts every 10 periods. That is, the onset of conflict is stochastic and its end is deterministic: the probability that a conflict spiral will have started by time t increases until $t = 10$. At this point this probability is so high that conflict is no longer sufficiently informative about the other group's type (as indicated by the posterior belief in Figure 2 hitting the dotted line), and if conflict had already started, it stops and cooperation restarts.

2.2 Additional Results

To help build intuition about the mechanics of the baseline model, we next present simple comparative statics on the cycle length T and results on social welfare when the probability of a misperception π is small.

Our comparative statics result is as follows: First, cycles are longer when $u(0,0)$ is higher, $u(1,0)$ is lower, or $u(1,1)$ is lower, as all of these changes make experimenting with the good action less appealing (i.e., they decrease μ^*). Second, cycles are longer when the prior probability of the bad type is higher, as this makes players more pessimistic about the other group (i.e., increases μ_t for all

[14] The figures graph $\mu_{t \bmod T}$ and $1-(1-\pi)^{t \bmod T}$ as continuous functions of t, even though time is discrete in the model.

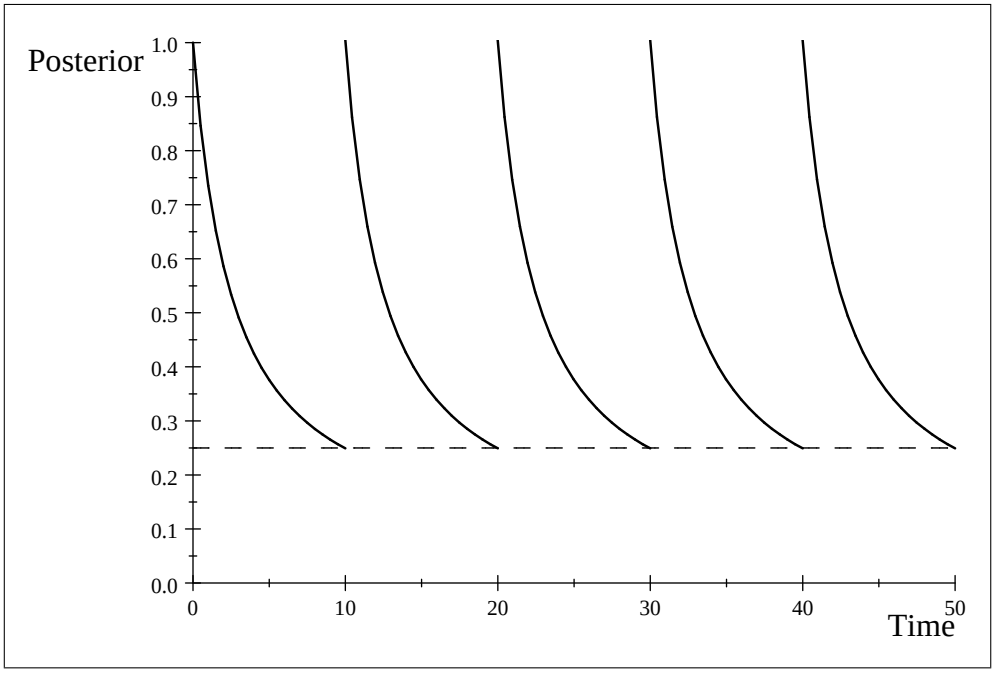

Figure 2: The Corresponding Cycle of Beliefs

t). Finally, cycles are shorter when the misperception probability π is higher, as this makes the bad signal less informative of the opposing group's type, which makes players more optimistic (decreases μ_t). Summarizing, we have the following result:

Proposition 2 *The length (period) of the cycle in the baseline model T is increasing in $u(0,0)$, decreasing in $u(1,0)$, decreasing in $u(1,1)$, increasing in the prior probability of the bad type μ_0, and decreasing in the misperception probability π.*

Another interesting property of the baseline model is that expected social welfare when both groups are normal—averaged across all players—is bounded away from the efficient level $2u(1,1)$, even as the probability of a misperception π goes to 0. Thus, not only do some players receive payoff less than $2u(1,1)$ for all $\pi > 0$ (which is immediate), but the fraction of players who get less than this does not vanish as $\pi \to 0$. The intuition is that, while the probability of a conflict spiral starting each period goes to 0 as $\pi \to 0$, the expected length of a conflict spiral conditional on its starting goes to ∞ as $\pi \to 0$, because when π is small conflict is very informative and it therefore takes a long time for cooperation to restart after a misperception. This result is in stark contrast to what would happen in a static setting, where, as $\pi \to 0$, the players could coordinate on the good outcome with probability approaching 1.[15]

In contrast, expected social welfare when both groups are normal does converge to the efficient level $2u(1,1)$ when *both* the probability of a misperception π and the prior probability that a group

[15] More precisely, in the "static" (i.e., two-period version) of our model, when both groups are normal, the probability that both players play 1 would converge to 1 and payoffs would converge to full information payoffs as $\pi \to 0$.

is bad μ_0 go to 0 (regardless of the relative sizes of these probabilities). Thus, both the probability of accidental conflict and the fear of the other group's true intentions must be small for efficiency to prevail. The intuition here can be seen from examining the formula for μ_t: if μ_0 is vanishingly small, then any positive probability of conflict $1-(1-\pi)^t$ is large enough that a player who observes conflict will restart cooperation. Hence, the probability that conflict actually occurs at any point in a given T-period cycle goes to 0 when both π and μ_0 go to 0.

Formally, we have the following result, where social welfare is evaluated according to the limit-of-means criterion (proof in Appendix B).[16]

Proposition 3 *Suppose that both groups are normal. Then the following hold:*

1. *The limit of expected social welfare as $\pi \to 0$ is less than the efficient level $2u(1,1)$.*
2. *For any sequence $(\pi_n, \mu_{0,n})$ converging to $(0,0)$ as $n \to \infty$ (such that Assumptions 1-3 hold for all n), the limit of expected social welfare as $n \to \infty$ equals the efficient level $2u(1,1)$.*

2.3 Application to Ethnic Conflict and International War

An immediate application of our baseline model is to ethnic conflict and international war. Consider two ethnic groups (or two countries) who repeatedly face the potential for conflict. For each potential conflict, the groups choose between two actions sequentially, and one of these corresponds to aggression or war. We assume that representatives of the two ethnic groups alternate in taking the first action as in the baseline model. The "security dilemma," or the "Hobbesian trap," suggests a game form in which a group or country dislikes taking the peaceful action when the other side is aggressive. In our overlapping-generations setup, this exactly corresponds to parts 1-2 of Assumption 1, implying that aggression is a best response to the belief that the other side has been aggressive so far or is expected to be aggressive in the future. Part 3 of Assumption 1 then implies that both sides are better off without such aggression.[17]

It is plausible in the context of such conflict that non-aggressive acts are sometimes viewed as aggressive by the other party, justifying our assumption concerning the relationship between actions and signals (i.e., between y and $\tilde{y}$).[18] Finally, we believe that, though extreme, the assumption that the past history of signals is not fully observed is also reasonable in this context. Even though we all have access to history books, it is difficult to ascertain and agree on how and exactly when a

[16] That is, if player t's payoff is u_t, social welfare is defined to be $\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N} u_t$.

[17] One might argue that our baseline model would better capture the "first-mover advantage" aspect of ethnic conflict if we allowed player t's payoff from choosing war after getting the peaceful signal from player $t-1$ to differ from her payoff from choosing war prior to player $t+1$'s playing peace; that is, if we allowed a player's payoff to depend on whether she moves first or second in a given conflict. Our results would not be affected by this generalization so long as each potential conflict remains a coordination game (i.e., a player wants to match her opponent's action or signal, regardless of whether she moves first or second).

[18] An alternative that leads to identical results is to assume that even when a group mostly consists of normal, non-aggressive types, a few of its members may be aggressive (e.g., military commanders bent on initiating conflict even when most politicians prefer peace).

given conflict started (see also Section 3). Consequently, the structure of the baseline model together with Assumptions 1-3 can be applied to the analysis of ethnic conflict or international war, and this leads to the equivalent of Proposition 1, accounting for potential conflict spirals and their endogenous cessation.

3 Extensions

This section presents four extensions of the baseline model, each relaxing a simplifying assumption made in the main analysis. The goal is to show that our main explanation of cycles of distrust—and thus our analysis of the applications—is not overly sensitive to these assumptions. Section 3.1 assumes that players do not observe calendar time, ruling out the strict dependence on calendar time found in the baseline model. Section 3.2 shows that behavior can be independent of calendar time even when time is observed, if additional information about the previous player's type is available. Section 3.3 allows the bad action to generate the good signal as well as the other way around. Section 3.4 allows players to observe the signals of actions taken more than one period ago.

3.1 Unobserved Calendar Time

One highly stylized aspect of the baseline model is the strict dependence of behavior on calendar time. The most direct way of eliminating this feature is to assume that players do not know calendar time: each player observes a signal $\tilde{y}_{t-1}$ of her predecessor's action and then chooses actions (x_t, y_t), without knowing t. This approach is intuitively appealing, but it introduces the somewhat delicate issue of what players believe about calendar time when they enter the game. Here, we simply assume that players have an "improper uniform prior" about calendar time, in that they take the probability of observing signal $\tilde{y}_{t-1}$ to equal the long-run fraction of periods in which the signal equals $\tilde{y}_{t-1}$.[19] Player 0, however, is assumed to know calendar time (e.g., she can infer this from the fact that she does not observe a signal).

In such a model, normal players play $y_t = 1$ in response to $\tilde{y}_{t-1} = 1$, as they know the other group is normal after observing $\tilde{y}_{t-1} = 1$. There can be no equilibrium in which normal players play $y_t = 0$ in response to $\tilde{y}_{t-1} = 0$ with probability 1: if there were, then $\tilde{y}_{t-1} = 0$ would be observed almost surely in the long run, so a normal player would believe that the opposing group is good with probability $\mu_0 < \mu^*$ after observing $\tilde{y}_{t-1} = 0$ and would therefore play $y_t = 1$ (by Assumption 2). So suppose that in response to $\tilde{y}_{t-1} = 0$ normal players play $y_t = 0$ with some probability $p \in [0, 1]$ and play $y_t = 1$ with probability $1 - p$. Then when both groups are good the long-run fraction of periods in which $\tilde{y}_{t-1} = 0$—denoted by q—is given by $q = \pi + (1 - \pi) qp$, or

$$q = \frac{\pi}{1 - (1 - \pi) p}.$$

[19] See Appendix C of Liu and Skrzypacz (2011) for a more rigorous foundation of this approach.

This expression follows because if there is a misperception then $\tilde{y}_{t-1} = 0$ with probability 1 while if there is not a misperception then $\tilde{y}_{t-1} = 0$ with probability qp.

Now a normal player's assessment of the probability that the other group is bad after observing $\tilde{y}_{t-1} = 0$ is given by

$$\mu = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\, q}.$$

For her to be indifferent between playing $y_t = 0$ and $y_t = 1$, it must be that $\mu = \mu^*$, or

$$q = q^* \equiv \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*}.$$

This holds if and only if

$$p = p^* \equiv \frac{q^* - \pi}{q^* (1 - \pi)}.$$

Summarizing, we have the following result:[20]

Proposition 4 *Under Assumptions 1 and 2, the model with an improper uniform prior over calendar time has a unique symmetric sequential equilibrium. It has the following properties:*

1. *Normal player* 0 *plays* $y_0 = 1$.
2. *At every time* $t > 0$, *normal player* t *plays good actions* $(x_t = 1, y_t = 1)$ *if she gets the good signal* $\tilde{y}_{t-1} = 1$. *If she gets the bad signal* $\tilde{y}_{t-1} = 0$, *she plays the bad action* $x_t = 0$ *toward player* $t-1$ *and plays the good action* $y_t = 1$ *toward player* $t+1$ *with probability* $1 - p^*$.
3. *Bad players always play bad actions* $(x_t = 0, y_t = 0)$.

This result implies that even when calendar time is not observed, our model generates (irregular) cycles: when both groups are good, a conflict spiral starts in each period with probability π and ends in each period with probability $1-p^*$. Therefore, this important aspect of our model does not depend on players' observing calendar time.

3.2 Independence from Observable Calendar Time

In the baseline model, the Tth player plays the good action (toward the next player) regardless of her signal, while the $T - 1^{st}$ player plays the good action only if she gets the good signal. This "discontinuous" behavior is driven by the endogenously changing informativeness of the bad signal (or conflict) about the other group's type. In particular, the Tth player, correctly, thinks observing conflict is sufficiently likely that it is uninformative, while it is slightly more informative for the $T-1^{st}$ player; and this difference is enough to cause them to behave differently in equilibrium. Thus, the different behavior of the Tth and $T - 1^{st}$ players is not driven by calendar time per se—after all, nobody directly cares about calendar time—but rather by the informativeness of conflict.

[20] Here, a *symmetric* equilibrium is one in which all normal players (except for player 0) use the same strategy.

This subsection clarifies this idea by assuming that with probability $\kappa \in (0,1)$, independently of all other random variables, player t is "crazy." Thus, each group consists of either $1-\kappa$ normal players and κ crazy players, or $1-\kappa$ bad players and κ crazy players. Crazy players are more inclined to play the bad actions $x_t = 0$ and $y_t = 0$ than normal players are; for simplicity, we just assume that they always play $(x_t = 0, y_t = 0)$ though this is not necessary. Crazy players also differ from both normal and bad players in that their type is observable to the next player. That is, every player t observes whether or not player $t-1$ was crazy, in addition to observing $\tilde{x}_{t-1}$.

The point of adding crazy players to the baseline model is that now if player t sees that player $t-1$ was crazy, then she becomes like player T in the baseline model in that for her conflict is uninformative of the other group's type. Therefore, normal players always cooperate when the previous player was crazy. In addition, if there are enough crazy players, there are no longer deterministic restarts of trust, because getting the bad signal from a non-crazy player is always a strong signal that the other group is bad. In this case, whether or not players observe calendar time in the model becomes irrelevant, but cycles of distrust emerge again for reasons similar to those in the baseline model.[21]

To see this formally, first note that the analog of Assumption 2, which guarantees that normal player 0 plays $y_0 = 1$, is now:

Assumption 2′

$$\mu_0 < \mu^*_{CRAZY} \equiv 1 - \frac{1}{1-\kappa}\left(\frac{u(0,0) - u(1,0)}{u(1,1) - u(1,0)}\right).$$

As in the baseline model, let μ_t be the probability that normal player t assigns to the other group being normal after getting a bad signal from a non-crazy player, under the hypothesis that normal players play $y_t = 0$ if and only if they get a bad signal from a non-crazy player. The following lemma characterizes μ_t:

Lemma 1 *We have*

$$\mu_t = \frac{\mu_0}{\mu_0 + (1-\mu_0)\left(\frac{(1-\kappa)\pi}{\kappa+(1-\kappa)\pi}\right)\left(1-(1-\kappa)^t(1-\pi)^t\right)}.$$

This is decreasing in t, and satisfies $\lim_{t\to\infty}\mu_t \equiv \mu_\infty < \mu^*_{CRAZY}$ if and only if

$$\kappa < \kappa^* \equiv \frac{\pi\left(\mu^*_{CRAZY} - \mu_0\right)}{\pi\left(\mu^*_{CRAZY} - \mu_0\right) + \mu_0\left(1 - \mu^*_{CRAZY}\right)}, \tag{1}$$

which is a positive number under Assumption 2′.

The genericity assumption is now modified to:

Assumption 3′ $\mu_t \neq \mu^*_{CRAZY}$ for all $t \in \mathbb{N}$.

[21] The presence of crazy players may also be realistic: for example, it may be that in any ethnic group there is a probability β that an aggressive military leader, who always (regardless of the preferences of the group) wants to wage war, comes to power, and this is observed by the next generation.

We now describe the unique equilibrium. When both groups are normal, conflict phases always begin with a misperception and end with the arrival of a crazy player. Whether there is also a deterministic component of conflict cycles (as there is in the baseline model) depends on whether κ is greater or less than κ^*. If $\kappa < \kappa^*$, then eventually being in a conflict phase is weak enough evidence that the other group is bad that normal players will eventually switch to the good action—because $\mu_\infty < \mu^*_{CRAZY}$—and so there are deterministic restarts. But if $\kappa > \kappa^*$, then being in a conflict cycle is always strong enough evidence that the other group is bad that normal players do not restart (because $\mu_\infty > \mu^*_{CRAZY}$), so restarts are purely stochastic.

Proposition 5 *Under Assumptions 1, 2′, and 3′, the model with crazy types has a unique sequential equilibrium. Let κ^* be given by (1) and $T_{CRAZY} \equiv \min\{t : \mu_t < \mu^*_{CRAZY}\}$. Then:*

If $\kappa < \kappa^*$, then the equilibrium has the following properties:

1. *At every time $t \neq 0 \bmod T_{CRAZY}$, normal player t plays the good actions $(x_t = 1, y_t = 1)$ if she gets the good signal $\tilde{y}_{t-1} = 1$, and plays the bad actions $(x_t = 0, y_t = 0)$ if she gets the bad signal $\tilde{y}_{t-1} = 0$.*

2. *At every time $t = 0 \bmod T_{CRAZY}$, normal player t plays the good action $x_t = 1$ toward player $t-1$ if and only if she gets the good signal $\tilde{y}_{t-1} = 1$, but plays the good action $y_t = 1$ toward player $t+1$ regardless of her signal.*

3. *Bad players and crazy players always play bad actions $(x_t = 0, y_t = 0)$.*

If $\kappa > \kappa^*$, then the equilibrium has the following properties:

1. *At every time $t > 0$, normal player t plays good actions $(x_t = 1, y_t = 1)$ if she gets the good signal $\tilde{y}_{t-1} = 1$, and plays the bad actions $(x_t = 0, y_t = 0)$ if she gets the bad signal $\tilde{y}_{t-1} = 0$.*

2. *Normal player 0 plays the good action $y_0 = 1$ toward player 1.*

3. *Bad players and crazy players always play bad actions $(x_t = 0, y_t = 0)$.*

The proof is straightforward given Proposition 1 and Lemma 1: If $\kappa < \kappa^*$ then $\mu_\infty < \mu^*_{CRAZY}$, and hence $\mu_t < \mu^*_{CRAZY}$ for sufficiently large t, so T_{CRAZY} is well-defined. The result in this case is analogous to Proposition 1. In contrast, if $\kappa > \kappa^*$ then $\mu_\infty > \mu^*_{CRAZY}$, and hence $\mu_t > \mu^*_{CRAZY}$ for all t. So normal player t plays $y_t = 1$ if and only if $\tilde{y}_{t-1} = 1$.

3.3 Two-Sided Errors

The analysis of the baseline model was simplified by the assumption that only the good action can generate the good signal. This section shows that our main conclusions still apply when either action can generate either signal.

In particular, assume now that the signal $\tilde{y}_{t-1}$ is distributed as follows:

$$\begin{aligned} \Pr\left(\tilde{y}_{t-1}=1|y_{t-1}=1\right) &= 1-\pi \\ \Pr\left(\tilde{y}_{t-1}=1|y_{t-1}=0\right) &= \pi', \end{aligned}$$

where $\pi, \pi' \in (0,1)$ and $\pi + \pi' < 1$. The assumption that $\pi + \pi' < 1$ means that the good action is more likely to generate the good signal than is the bad action, and is thus essentially a normalization.

As in the baseline model, Assumption 1 guarantees that normal player t plays $x_t = 1$ if and only if $\tilde{y}_{t-1} = 1$, and Assumption 2 guarantees that she plays $y_t = 1$ if and only if her assessment of the probability that the other group is bad after observing $\tilde{y}_{t-1}$ is less than μ^*. Denote her assessment of this probability after observing $\tilde{y}_{t-1} = 0$ by μ_t (as usual), and denote her assessment of this probability after observing $\tilde{y}_{t-1} = 1$ (which equals 0 in the baseline model, due to one-sided errors) by μ'_t. To compute these probabilities, let

$$M = \begin{pmatrix} 1-\pi & \pi' \\ \pi & 1-\pi' \end{pmatrix}$$

be the Markov transition matrix governing the evolution of $\tilde{y}_t$ in the event that both groups are normal, under the hypothesis that normal players play $y_t = 1$ if and only if $\tilde{y}_{t-1} = 1$. That is, if both groups are normal and $\tilde{y}_t = 1$, then $\tilde{y}_{t+1} = 1$ with probability $1 - \pi$; if, on the other hand, $\tilde{y}_t = 0$, then $\tilde{y}_{t+1} = 1$ with probability π'. Then, by Bayes rule,

$$\mu_t = \frac{\mu_0\left(1-\pi'\right)}{\mu_0\left(1-\pi'\right) + \left(1-\mu_0\right)\left(1 - M^t_{(1,1)}\right)},$$

where $M^t_{(1,1)}$ is the $(1,1)$ coordinate of the tth power of M. This is simply because the probability of observing $\tilde{x}_{t-1} = 0$ conditional on the other group being bad equals $1 - \pi'$, while the probability of observing $\tilde{x}_{t-1} = 0$ conditional on the other group being good is $1 - M^t_{(1,1)}$. Similarly,

$$\mu'_t = \frac{\mu_0 \pi'}{\mu_0 \pi' + \left(1-\mu_0\right) M^t_{(1,1)}}.$$

In the baseline model, it was the case that $\mu_t \to \mu_0$ as $t \to \infty$, so Assumption 2 guaranteed the existence of a time T such that $\mu_T < \mu^*$. With two-sided errors, $M^t_{(1,1)} \to \frac{\pi'}{\pi+\pi'}$ as $t \to \infty$, so $\mu_t \to \mu_\infty$ as $t \to \infty$, where

$$\mu_\infty = \frac{\mu_0\left(1-\pi'\right)}{\mu_0\left(1-\pi'\right) + \left(1-\mu_0\right)\frac{\pi}{\pi+\pi'}}.$$

If $\mu_\infty < \mu^*$, then Assumption 2 guarantees the existence of a smallest time $T_{2-SIDED}$ such that $\mu_{T_{2-SIDED}} < \mu^*$, and there is a deterministic cycle with period $T_{2-SIDED}$, as in the baseline model. If on the other hand $\mu_\infty \geq \mu^*$, then there is no deterministic cycle, and in particular a bad signal always leads to a spiral of bad actions that lasts until the next accidental good signal.

Summarizing, we have the following result (proof in Appendix B):

Proposition 6 *Under Assumptions 1-3, the model with two-sided errors has a unique sequential equilibrium. If $\mu_\infty < \mu^*$, then the equilibrium has the following properties:*

1. *At every time $t \neq 0 \bmod T_{2-SIDED}$, normal player t plays the good actions $(x_t = 1, y_t = 1)$ if she gets the good signal $\tilde{y}_{t-1} = 1$, and plays the bad actions $(x_t = 0, y_t = 0)$ if she gets the bad signal $\tilde{y}_{t-1} = 0$.*

2. *At every time $t = 0 \bmod T_{2-SIDED}$, normal player t plays the good action $x_t = 1$ toward player $t-1$ if and only if she gets the good signal $\tilde{y}_{t-1} = 1$, but plays the good action $y_t = 1$ toward player $t+1$ regardless of her signal.*

3. *Bad players always play bad actions $(x_t = 0, y_t = 0)$.*

If instead $\mu_\infty \geq \mu^*$, then the equilibrium has the following properties:

1. *At every time $t > 0$, normal player t plays the good actions $(x_t = 1, y_t = 1)$ if she gets the good signal $\tilde{y}_{t-1} = 1$, and plays the bad actions $(x_t = 0, y_t = 0)$ if she gets the bad signal $\tilde{y}_{t-1} = 0$.*

2. *Normal player 0 plays the good action $y_0 = 1$ toward player 1.*

3. *Bad players always play the bad actions $(x_t = 0, y_t = 0)$.*

3.4 More Information About the Past

Another stylized aspect of the baseline model is the assumption that players observe a signal of only the most recent action y_{t-1}, and get no information about any earlier actions. Though this simple information structure allowed us to explicitly characterize equilibrium and show that it features "restarts" of trust every T periods, it is not necessary for our main intuition for cycling. This section shows that when players observe the previous K signals, for any integer K, there are still deterministic restarts of trust (though not necessarily at regular intervals); in particular, we show that there are still infinitely many times t at which a normal player plays the good action $x_t = 1$ even if she observes K bad signals, and that this occurs for essentially the same reason as in the baseline model.[22]

Formally, let us modify the baseline model by supposing that players observe the previous K signals, for some fixed integer K. That is, before choosing her action, player t observes $\left(\tilde{y}_{t-K}, \tilde{y}_{t-(K-1)}, \ldots, \tilde{y}_{t-1}\right)$, where this vector is truncated at 0 if $t < K$. Player t's utility function is still given by $u\left(x_t, \tilde{y}_{t-1}\right) + u\left(\tilde{y}_t, x_{t+1}\right)$, exactly as in the baseline model.

Proposition 7 *Under Assumptions 1 and 2, in any sequential equilibrium of the model where players observe the last K signals, there are infinitely many times t at which normal player t plays the good action $y_t = 1$ toward player $t+1$ with positive probability when she observes all bad signals (i.e., when $\tilde{y}_{t-k} = 0$ for all $k \in \{1, \ldots, K\}$).*

[22] In addition, the event that player t observes K bad signals always occurs with positive probability, in particular with probability at least π^K.

Proposition 7 and its proof, which is presented in Appendix B, show that our main intuition for cycling goes through when players observe any number of past signals, not just one. However, when $K > 1$ cycling is no longer regular (i.e., there is no longer a restart of trust every T periods), and explicitly characterizing equilibrium seems very challenging.[23]

A final remark on more information about the past: In the context of war or ethnic conflict, it is sometimes argued that grievances from the distant past can be a salient source of conflict and distrust (e.g., massacres or desecration of holy sites). Proposition 7 can be modified to allow for this possibility. Suppose that every instance of conflict (i.e., every time that $\tilde{y}_t = 0$ or $x_t = 0$) leads to a lasting grievance for the opposing group with some probability. If players always remember the exact timing of their group's last $\tilde{K}$ grievances for some (potentially arbitrarily large but finite) $\tilde{K}$, no matter how long ago they originated (but forget earlier grievances), then the same argument leading to Proposition 7 (see Apendix B) implies that there are again infinitely many times at which normal players restart cooperation even if the last K signals are all bad and they remember $\tilde{K}$ grievances. In consequence, conflict cycles emerge even when players can remember unboundedly distant grievances with positive probability.

4 Application to Political Partisanship and Participation Cycles

This section develops a detailed application of our model to cycles of political partisanship and political participation. A standard—though not uncontroversial—narrative is that polarization in American politics was high in the first third of the 20th century, low in the middle third, and high in the last third. For example, McCarty, Poole, and Rosenthal (2008) document this pattern of polarization between Democrats and Republicans in congressional roll call votes. Other prominent views are that underlying political preferences have not polarized to nearly the same extent as have the major parties' political positions (Fiorina, 2011) and that political partisanship and participation are closely linked (Abramowitz, 2011).

The model we present in this section provides an approach for thinking about polarization between political parties in the absence of polarization of the electorate. In particular, we consider political competition between two parties that can each be either moderate or extremist. Every period, the representative of the party in power can take action on either a major, common interest issue (e.g., the economy) or a partisan issue (e.g., school prayer). Taking action on the major issue requires some form of cooperation from the previous political leader; we model this by assuming that the previous leader may learn some information about the economy that must be passed on for the current leader to set economic policy effectively.[24] Whether or not this cooperation is granted is observed by the

[23] The reason for this is as follows: the good action $y_t = 1$ is a best response to a single good signal (i.e., $\tilde{y}_{t-k} = 1$ for some $k \in \{1, \ldots, K\}$). In turn, when a player observes all bad signals, she has to update her beliefs about the last time a player restarted trust, which can be an intractable updating problem.

[24] One could alternatively assume, with identical results, that the previous leader may need to give the current leader a vote of confidence in order for her to be able to make major policy changes.

political supporters of the incumbent party—referred to as its "base"—and can influence their decision on whether or not to turn out to vote (in an extension of the model, we allow both parties' bases to turn out). The difference between moderate and extremist politicians is that moderates care more about the major issue, while extremists care more about the partisan issue; thus, unlike in our earlier models, we do not assume that extremists are committed to some course of action, and the behavior of both moderates and extremists is determined by strategic considerations.

We show that this model leads to a distinctive pattern of partisanship and turnout that resembles the cycles of distrust in our baseline model. Extremists never cooperate or communicate with the other party, because in equilibrium cooperating would signal to the base that the other party is moderate and the base would therefore not bother turning out. Moderates cooperate as long as they believe that the other party is sufficiently likely to be moderate. In equilibrium, a single accidental failure of cooperation leads to a partisan phase in which no parties cooperate, the bases turn out, and politicians act only on the partisan issue. However, eventually a politician realizes that a partisan phase was quite likely to have started by accident, and hence cooperates with the other party, which leads to a trusting phase where both parties cooperate, the bases do not turn out, and politicians act only on the major issue. As in the baseline model, these phases alternate, and the dynamics are driven by how much the parties trust each other.

We split this section into three subsections. Subsection 4.1 presents the model, subsection 4.2 characterizes the unique equilibrium, and subsection 4.3 extends the model by allowing both parties' bases to turn out.

4.1 Model

Group A is a left-wing party and Group B is a right-wing party. In every period $t = 0, 1, 2, \ldots$, a representative of the group in power is the leader ("player t"). A different representative is chosen every period, and representatives do not know what happened in past periods (including which group was in power—the dynamics of power holding are described below). Each group can either consist entirely of moderates or entirely of extremists. The probability that a group consists entirely of extremists is $\mu_0 > 0$. Both moderates and extremists are policy-motivated but they differ in what political issues they find more important.

Every period, player t can take action on one of two issues: the "major issue" w_t or the "partisan issue" z_t. Players' ideal actions on the partisan issue are fixed over time and commonly known. In contrast, the *period t state* $\theta_t \in \{-1, 1\}$ is drawn independently every period, with probability $\frac{1}{2}$ on either state. Player t observes the period t state θ_t and has the option of attempting to communicate this to player $t+1$. If player t attempts communication (denoted $s_t = 1$), communication succeeds (denoted $\tilde{s}_t = 1$) with probability $1 - \pi \in (0, 1)$, in which case player $t+1$ observes θ_t (and can infer that communication was attempted), and communication fails (denoted $\tilde{s}_t = 0$) with probability π, in which case player $t+1$ does not observe θ_t (and also does not observe whether communication was attempted).

The timing is as follows: At the beginning of period t, the identity of the group in power and its representative, player t, is determined. Player t then observes θ_t, and also observes θ_{t-1} if player $t-1$ successfully communicated it. We also assume that player 0 directly observes or has received communication about the initial state θ_{-1}. Player t then chooses which issue to "take action" on. Taking action on the major issue means choosing a policy $w_t \in [-1, 1]$ for the major issue, while taking action on the partisan issue means choosing a policy $z_t \in [-1, 1]$ for the partisan issue, and in either case it is assumed that the period t policy on the issue on which action is not taken is set equal to the default policy 0. Finally, player t decides whether to attempt to communicate θ_t to player $t+1$. Note that player t makes these decisions without knowing which group will be in power in period $t+1$.

Player t cares about policy on both issues in periods t and $t+1$. The difference between left-wing players and right-wing players is that they have opposite preferences about the partisan issue: left-wing players have bliss point $z_t = -1$, and right-wing players have bliss point $z_t = 1$. The difference between moderates and extremists is that moderates think the major issue is more important than the partisan issue, while extremists think the partisan issue is more important.[25] Formally, the relative importance of the two issues is measured by a constant $\alpha \in (0, 1)$. Moderate player t of type $\eta \in \{-1, +1\}$, where -1 corresponds to left-wing and $+1$ corresponds to right wing, has payoff

$$u_\eta^m (w_t, w_{t+1}, z_t, z_{t+1}, \theta_t, \theta_{t+1}) = - |w_t - \theta_{t-1}| - |w_{t+1} - \theta_t| - \alpha |z_t - \eta| - \alpha |z_{t+1} - \eta| .$$

Extremist player t of type η has payoff

$$u_\eta^e (w_t, w_{t+1}, z_t, z_{t+1}, \theta_t, \theta_{t+1}) = -\alpha |w_t - \theta_{t-1}| - \alpha |w_{t+1} - \theta_t| - |z_t - \eta| - |z_{t+1} - \eta| .$$

These payoff functions make it clear that the only difference between moderates and extremists is in the weights that they put on the major and the partisan issues, and the only difference between left and right-wing players is in their bliss point for the partisan issue.

We now describe the dynamics of power holding. Which group holds power in the initial period $t = 0$ is determined by the flip of a fair coin, and in every subsequent period $t = 1, 2, \ldots$ which group holds power is determined by an election at the end of period $t-1$. The potential voters are either "independents" or are members of a group of organized supporters of the left or the right, referred to henceforth as the left and right *bases*. A base has to pay some cost $c > 0$ in order to organize and motivate their supporters to turn out for the election, and we assume (for now) that a base can only do this if its own party is in power. Thus, the incumbent party's base will turn out for the election if the potential gain compensates for the cost of organization. We assume that the number of net votes from independent voters for the incumbent party in each period is given by an independent draw from some distribution G symmetric about 0; hence, without turnout from the bases, the incumbent party wins the election with probability $G(0) = \frac{1}{2}$. Let k denote the size of each base, so when the supporters of

[25] In particular, all differences in the propensities of moderates and extremists to *communicate* are derived rather than assumed. In other words, in contrast to the baseline model, extremists are not behavioral types committed to a particular action, but choose a different action because of the different weights they put on the major and the partisan issue.

the incumbent party turn out, that party wins with probability $G(k) > \frac{1}{2}$, and to simplify notation we let $1-\rho \equiv G(k)$. Note that turning out increases the probability that the incumbent party wins by $(1-\rho) - \frac{1}{2} = \frac{1}{2} - \rho$. We make the following assumption, which serves only to focus attention on the most interesting region of parameter space:

Assumption POL 1. $\rho < \frac{1-\alpha}{4}$.

2. $c \in \left(\left(\frac{1}{2}-\rho\right)(1+\alpha)\mu_0, \left(\frac{1}{2}-\rho\right)(2\alpha)\right)$.

3. $\mu_0 < \mu^*_{POL} \equiv \frac{4\alpha\rho+2(1-\alpha)}{1+\alpha}$.

We also assume that the incumbent's base at time t consists of individuals who are in their youth at time t, and have the same payoff functions as moderate politicians. This in particular implies that the base will not turn out in response the propaganda because it has extreme preferences or it is fooled by this, but because even with moderate preferences it will find it beneficial to turn out given its expectations about the type of the other side.

The last element of the model is the information of the base. We assume that the period t base knows only whether its group is in power (but not whether it is moderate or extremist) and the realization $\tilde{s}_t$. The interpretation is that $\tilde{s}_t = 1$ represents an "informative" message to the next leader, while $\tilde{s}_t = 0$ represents a "propaganda" message to the base (and the probability law for $\tilde{s}_t$ reflects the fact that a message that is meant to be informative may be misinterpreted as propaganda). We also assume that, as in the baseline model, all players observe calendar time.

4.2 Equilibrium

We now show that this model has a sequential equilibrium in which politicians behave much as in the baseline model—where attempting communication is like the good action, and sending propaganda is like the bad action—and where the base turns out if and only if it receives propaganda. Furthermore, equilibrium is unique for generic parameter values (proof in Appendix B).[26]

Proposition 8 *Under Assumption POL, the political partisanship model has a generically unique sequential equilibrium. There exists an integer $T_{POL} > 0$ such that in this equilibrium:*

1. *At every time $t \neq 0 \bmod T_{POL}$, moderate player t plays $(w_t = \theta_{t-1}, s_t = 1)$ if $\tilde{s}_{t-1} = 1$, and otherwise plays $(z_t = -1, s_t = 0)$ (if left-wing) or $(z_t = 1, s_t = 0)$ (if right-wing).*

2. *At every time $t = 0 \bmod T_{POL}$, moderate player t plays $(w_t = \theta_{t-1}, s_t = 1)$ if $\tilde{s}_{t-1} = 1$, and otherwise plays $(z_t = -1, s_t = 1)$ (if left-wing) or $(z_t = 1, s_t = 1)$ (if right-wing).*

3. *Extremist player t always plays $(z_t = -1, s_t = 0)$ if left-wing or $(z_t = 1, s_t = 0)$ if right-wing.*

[26] In general, sequential equilibrium is not well-defined in games with a continuum of actions. This is not a problem here, because the probability distribution over player $t+1$'s information sets depends on player t's strategy only through the binary variable s_t. So one can define sequential equilibrium by supposing that players "tremble" only on s_t.

4. The base turns out if and only if it receives propaganda.

There are several differences between Proposition 8 and Proposition 1. First, the analogy between sending informative messages and the good action in the baseline model must be established. A preliminary observation is that extremists always choose to act on the partisan issue, because for them the partisan issue is more important than the major issue. Moderates, on the other hand, act on the major issue if and only if they receive information about the current state, since they cannot improve on the default policy on the major issue unless they are informed of the current state. Now, given that the base turns out if and only if it receives propaganda, the benefit of attempting communication is greater when the opposing group is more likely to be moderate, while the cost of attempting communication (i.e., the opportunity cost of not sending propaganda) does not depend on whether the opposing group is moderate (as after propaganda the next leader always acts on the partisan issue, as she is not informed of the state). Thus, sending informative messages is like the good action in the baseline model.

Second, the incentives of the base must be accounted for. When the base hears propaganda it knows that it is in a partisan phase, so its belief about the opposing group is irrelevant (as in a partisan phase moderates and extremists play the same way), and the assumption that $c < \left(\frac{1}{2} - \rho\right)(2\alpha)$ implies that the base turns out. When the base hears communication, it infers that the opposing group is extremist with at most the prior probability μ_0 (i.e., hearing communications is "good news" about the other group's type), in which case the assumption that $c > \left(\frac{1}{2} - \rho\right)(1 + \alpha)\mu_0$ implies that the base does not turn out.

Third, the probability of power changes is now endogenous and depends on the two groups' types, so that now the fact that a particular group finds itself in power is informative of the other group's type. This makes Bayesian updating somewhat subtle. Nonetheless, Lemma 2 in Appendix B shows that moderate player t's assessment of the probability that the opposing group is extremist, μ_t (conditional on player $t-1$ failing to communicate and player t's group holding power in period t), behaves as in the baseline model in the long run, and also that $\mu_1 \to 1$ as $\pi \to 0$ (which implies that $T_{POL} > 1$ for small π, so that partisan cycles are not trivial).

4.3 Allowing Both Bases to Turn Out

We now dispense with the simplifying assumption that only the base of the incumbent party can turn out. The rest of the model is as above, and in particular the out-of-power period t base cannot receive propaganda from its party, so it cannot infer whether or not the period t leader successfully communicates (unlike the in-power period t base). Therefore, the out-of-power period t base can condition its turnout decision only on calendar time.

The following result shows that there are two possible consequences of letting the out-of-power base turn out: either it never turns out in equilibrium, and in this case, the equilibrium is exactly as in Proposition 8, or it turns out deterministically in the later periods of each cycle, in which case the

cycle length can also be different from T_{POL}.

Proposition 9 *Under Assumption POL, the political partisanship model where the out-of-power base can also turn out has a generically unique sequential equilibrium. In this equilibrium, one of the following two statements hold, and each holds for some parameter values:*

1. *The out-of-power base never turns out in sequential equilibrium, and the sequential equilibrium strategies for the politicians and the in-power base are exactly as in the main partisanship model.*
2. *There exist integers $\hat{T} < \tilde{T}$ such that the sequential equilibrium strategies for the politicians and the in-power base are as in the main partisanship model with cycle length $\tilde{T}$, and the out-of-power base turns out at time t if and only if $t \geq \hat{T} \bmod \tilde{T}$.*

Proposition 9 shows that partisan cycles behave much as in the main partisanship model when both bases can turn out. However, the corresponding turnout cycle now has an additional deterministic component: in each cycle, the out-of-power base turns out deterministically after a certain cutoff time $\hat{T}$. Thus, the overall turnout cycle is given by stochastic turnout of the in-power base coinciding with the onset of the partisan phase, deterministic turnout by the out-of-power base in the latter part of each cycle, and a deterministic drop in turnout at the end of each cycle.

The intuition for Proposition 9 is that the out-of-power base would like to turn out if and only if the partisan cycle has begun. Since this is more likely later in each cycle, the out-of-power base turns out deterministically after a certain cutoff time $\hat{T}$ in each cycle. If the implied cutoff time is greater than the cycle length $\tilde{T}$, then the out-of-power base never turns out (in which case $\tilde{T}$ must equal the cycle length in the main partisanship model, T_{POL}).

5 Related Theoretical Literature

Before concluding the paper, we take a moment to relate it, from a more theoretical perspective, to several existing classes of models that provide explanations for cyclic behavior in dynamic games: repeated games with imperfect public monitoring, stochastic games, reputation models of credibility, reputation models with limited records, and dynamic games with overlapping generations of players. In abstract terms, our baseline model is best described as a reputation model with limited records and overlapping generations, and it is to our knowledge the first such model in the literature. A central and distinguishing feature of our model is the uniqueness of equilibrium and the associated cycles which have both stochastic and deterministic elements.[27] More specifically, the fact that our key mechanism can lead to unique equilibrium cycles where the end date of the cycle is a deterministic function of time highlights that it is in essence very different from the existing literature (though we

[27] A notable partial exception here is Pesendorfer's (1995) model of fashion cycles. In his model, a durable-goods monopolist faces a population of consumers who signal their types by purchasing the good, and the logic of cycles is entirely different.

have shown in Section 3 that variants of our model can also generate patterns of behavior that are more complicated and realistic than regular deterministic cycles).

Repeated games with imperfect public monitoring date back to Green and Porter (1984). In their model, cyclic equilibrium behavior is due to moral hazard: bad signals lead to phases of bad actions, and vice versa. Abreu, Pearce, and Stachetti (1988, 1990) show this behavior can in fact emerge in *optimal* equilibria. Yared (2010) applies these insights to cycles of war and peace. These models differ from ours in that they have no incomplete information about types, they do not feature deterministic cycles, and they do not have equilibrium uniqueness.

Two branches of the stochastic games literature are particularly related. First, there are stochastic games of perfect information in which behavior cycles with the state. A leading example is Dixit, Grossman, and Gul's (2000) model of political compromise, which extends Alesina (1988) to the case of more than two states. See also Baron (1996), Battaglini and Coate (2008), and Acemoglu, Golosov, and Tsyvinski (2010). Second, there are reputation games where players' types follow a Markov process (Mailath and Samuelson, 2001; Phelan, 2006; Wiseman, 2009; Ekmecki, Gossner, and Wilson, 2011). Letting players' types change in our model would only produce a second reason for cycling. Typically, these models do not have deterministic cycles or equilibrium uniqueness.

Among reputation models, the literature on "credibility," starting with Sobel (1985) and Benabou and Laroque (1992), is particularly related. In these models, there is a deterministic or stochastic cycle in which a long-run player builds her reputation by being trustworthy against a series of short-run players, before cheating them and thus burning her reputation. As such, cycling is a short-run phenomenon that ends when players' types are learned.

The recent literature on reputation with limited records (Liu and Skrzypacz, 2011; Liu, 2011; Monte, 2011) is closely related to our paper. Most closely related is Liu and Skrzypacz (2011), where a long-run player facing a series of short-run player with limited records repeatedly builds her reputation up to a point and then exploits it. These models do not have deterministic cycles and the mechanism for cycles is completely different.

Finally, our paper is related to the literature on dynamic games with overlapping generations of players. The folk theorem often fails in these models—and indeed the overlapping generations aspect of our model is important for equilibrium uniqueness. The first anti-folk theorem in this literature is derived by Lagunoff and Matsui (1997), who show uniqueness in an overlapping generations coordination game. Bhaskar (1998) presents additional anti-folk theorems in a consumption-loan model. Lagunoff and Matsui (2004) develop a model where two groups interact over time, with a new representative of each group drawn every period and communication between generations, as in our model. In their model, the groups play a prisoner's dilemma (while they play a coordination game in our model), and there is no incomplete information in their model (while incomplete information and learning about the other group's type is key in our model). They show that "all defect" is the only equilibrium, but that a folk theorem holds if communication costs or altruism within groups is introduced. Anderlini, Gerardi, and Lagunoff (2010) present a model of war related to Lagunoff and

Matsui (2004), and show that there can be equilibrium cycles of war that hold each group below what its minmax payoff would be if it were a single decision-maker. Finally, Acemoglu and Jackson (2011) study a coordination game with overlapping generations and imperfect monitoring, where social norms change over time and "prominent" players can try to shift the social norm to a good static equilibrium. Their model does not have incomplete information about player types or deterministic cycles.

6 Conclusion

This paper has proposed a model of cycles of distrust and inter-group conflict based on the classical idea that conflict often results from distrust of the other side. In a dynamic context, a real or perceived aggression from one group makes it appear as innately aggressive to the other side, which in response acts more aggressively itself. When the first group cannot be sure whether this new aggression is a response to its own action or is due to the other side's actually being aggressive, a spiral of aggression and conflict forms. But—as our model shows—such a spiral cannot last forever, because it eventually becomes almost certain that a conflict spiral will have gotten started accidentally, at which point aggressive actions become completely uninformative of the other group's type. At such a time, a group experiments with the trusting action, and cooperation is restored.

We have also argued that this mechanism is robust and can be useful in understanding a range of situations in which there are (endogenous) cycles of distrust. First, the presence of a first-mover advantage in violent conflict makes our approach relevant to cycles of Hobbesian ethnic conflict or international war. Second, similar forces can lead to cycles of political partisanship in a dynamic model of policy choice and voter turnout. In addition, we show in Appendix A that these forces also emerge in a simple model of production and trade as well as in a model of communication between groups that may be moderate or extremist, and show that they can account for cyclic breakdowns of trade and communication.

Though our basic mechanism is simple, it is both different from existing explanations for cyclic behavior in dynamic games and, we believe, potentially relevant for understanding why seemingly unending conflicts end, and why cooperation and communication often follow periods of distrust. Our model points to several possible areas for future research. On the theoretical side, it would be interesting to study the more complex reputational incentives that would emerge if players lived for more than one period, and also to consider different ways in which players might learn about the history of conflict and cooperation between the groups (though we make some progress in this direction in Section 3). Finally, empirical analysis is needed to determine whether the mechanism we highlight—agents concluding that long-lasting conflicts are no longer informative about the true intentions of the other party—can indeed account for cycles of distrust, conflict, polarization, and communication breakdown in practice.

Appendix A: Additional Applications

Application to Trade

We now present a simple model of inter-group trade where members of one group will produce high-quality goods only if they expect members of the other group to produce high-quality goods for which they can trade. If everyone is afraid that the other group is unable or unwilling to produce high-quality goods (i.e., be a "bad" trading partner), then the equilibrium involves a cycle in which phases of trust and trade alternate with phases of distrust and the breakdown of production and trade.

Group A produces apples and Group B produces bananas. Group A members can produce rotten ("bad") apples for free regardless of whether Group A is normal or bad, but if it is normal (which occurs with probability $1 - \mu_0$), then they can also produce good apples at cost $c > 0$. Similarly for Group B and bananas.

All players live for 2 periods, and get utility from consuming one piece of fruit in each period. Members of Group A have a taste for bananas, and get utility $b > c$ from consuming a good banana, but only get utility $d \in (0, c)$ from consuming a good apple. Members of Group B get b from consuming a good apple and get d from consuming a good banana. No one gets utility from consuming rotten fruit. Assume also that a player gets utility $-\varepsilon$ if she trades for her opponent's rotten fruit, where $\varepsilon > 0$ is interpreted as a (small) transaction cost; with this interpretation, the payoff b of consuming the other group's fruit is the payoff net of the transaction cost.[28]

At time $t = 1, 2, \ldots$, a market opens in which players t and $t-1$ can exchange goods. Production by players t and $t-1$ for the time t market is staggered as we describe next. Each period is subdivided into three subperiods, which we denote as times $t - \frac{2}{3}$, $t - \frac{1}{3}$, and t for convenience.

At time $t - \frac{2}{3}$, normal player $t-1$ chooses whether to produce a good fruit or a rotten fruit for the time t market (bad players always produce rotten fruit), and her quality choice is denoted $y_{t-1} \in \{0, 1\}$ ($\{rotten, good\}$). If she produces a good fruit, it immediately rots with probability π. Both players then observe the final quality of her fruit, denoted $\tilde{y}_{t-1} \in \{0, 1\}$.

At time $t - \frac{1}{3}$, normal player t chooses whether to produce a good fruit or a rotten fruit (having observed $\tilde{y}_{t-1}$), and her quality choice is denoted by $x_t \in \{0, 1\}$ (thus, x_t is player t's quality choice at time $t - \frac{1}{3}$ for the time t market, and y_t is her quality choice at time $t + \frac{1}{3}$ for the time $t + 1$ market). Again, if she produces a good fruit, it rots with probability π, and both players observe the final quality of her fruit, denoted $\tilde{x}_t \in \{0, 1\}$.

Finally, players t and $t - 1$ arrive at the time t market. They then simultaneously decide on whether they would like to exchange goods: the exchange occurs if and only if both decide to do so. Each player then consumes the fruit she is left with (and pays the transaction cost if trade occurred). Player $t - 1$ then dies, player $t + 1$ is born, and the game continues with player t making her quality choice at time $t + \frac{1}{3}$.

The trade model differs from the baseline model in that both players' good actions (or production

[28] The transaction cost plays only a minor technical role in the analysis as discussed below.

choices) can turn bad in each period, and that players makes decisions about trade as well as production. Nonetheless, as we now show, equilibrium behavior in the trade model is closely related to that in the baseline model.

Consider first the production decision of player 0. Let us conjecture that if player 0's fruit does not rot ($\tilde{y}_0 = 1$), then normal player 1 will produce a good fruit and trade will occur provided that his fruit does not rot. Player 0's expected payoff from producing a good fruit is:[29]

$$(1-\mu_0)\left[(1-\pi)^2 b + \pi(1-\pi) d\right] + \mu_0 (1-\pi) d - c.$$

To see this, note that player 0 gets payoff b if player 1 is normal and neither fruit rots, while he gets payoff d if her fruit doesn't rot and either player 1 is normal but his fruit rots or player 1 is bad. On the other hand, player 0's expected payoff from producing a rotten fruit is 0, because rotten fruit never generates a good signal (and is worth zero in consumption). Therefore, player 0 produces a good fruit if and only if

$$\mu_0 < \mu^*_{TRADE} \equiv 1 - \frac{1}{(1-\pi)^2}\left(\frac{c-(1-\pi)d}{b-d}\right).$$

Suppose that this is the case, and consider the production decisions of normal player 1 at time $\frac{2}{3}$. Clearly, she will choose to produce a bad fruit if player 0's fruit is bad, as in this case trade never occurs. In contrast, if player 0's fruit is good, the assumption that $\mu_0 < \mu^*_{TRADE}$ is sufficient to guarantee that she will choose to produce a good fruit. Intuitively, a good fruit is more attractive when one's partner's fruit is good, and the assumption that $\mu_0 < \mu^*_{TRADE}$ implies that producing a good fruit is optimal even before knowing the quality of one's partner's fruit.

Next, consider the production decisions of normal player 1 at time $1 + \frac{1}{3}$. His position is now similar to that of player 0 at time $\frac{1}{3}$, except that he has a different assessment of the probability that the other group is bad. In particular, if player 0's fruit was good, then he is certain that the other group is good. If player 0's fruit was bad, then he believes that the other group is bad with probability

$$\mu_1 = \frac{\mu_0}{\mu_0 + (1-\mu_0)\pi}$$

as in the baseline model. Thus, if μ_1 is above the cutoff belief μ^*_{TRADE}, then player 1 produces a bad fruit if player 0's fruit was bad.

The analogy with the baseline model should now be clear. Assuming that $\mu_{t'} > \mu^*_{TRADE}$ for all $t' < t$, then when player $t-1$'s fruit is bad player t believes that the other group is bad with probability

$$\mu_t = \frac{\mu_0}{\mu_0 + (1-\mu_0)\left(1-(1-\pi)^t\right)}.$$

Once μ_t drops below μ^*, player t will produce a good fruit for the time $t+1$ market even if player $t-1$'s fruit was bad, and trade will resume until the next time a good fruit rots.

Formally, impose the following versions of Assumptions 2 and 3 from the baseline model.

[29] Here and in what follows, we take the limit $\varepsilon \to 0$.

Assumption 2″ $\mu_0 < \mu^*_{TRADE}$.

Assumption 3″ $\mu_t \neq \mu^*_{TRADE}$ for all $t \in \mathbb{N}$.

We obtain the following result (proof in Appendix B):

Proposition 10 *Under Assumptions 2″ and 3″, the trade model has a unique trembling-hand perfect equilibrium. It has the following properties (where* $T_{TRADE} \equiv \min\{t : \mu_t < \mu^*_{TRADE}\}$*):*

1. *For every time* $t \neq 0 \bmod T_{TRADE}$*, normal player* t *plays* $x_t = 1$ *if and only if* $\tilde{y}_{t-1} = 1$*; approves trade in market* t *if and only if* $\tilde{y}_{t-1} = 1$*; plays* $y_t = 1$ *if and only if* $\tilde{y}_{t-1} = 1$*; and approves trade in market* $t+1$ *if and only if* $\tilde{x}_{t+1} = 1$*.*

2. *For every time* $t = 0 \bmod T_{TRADE}$*, normal player* t *plays* $x_t = 1$ *if and only if* $\tilde{y}_{t-1} = 1$*; approves trade in market* t *if and only if* $\tilde{y}_{t-1} = 1$*;* always *plays* $y_t = 1$*; and approves trade in market* $t+1$ *if and only if* $\tilde{x}_{t+1} = 1$*.*

3. *Bad players always play* $x_t = 0$*; approve trade in market* t *if and only if* $\tilde{y}_{t-1} = 1$*; always play* $y_t = 0$*; and approve trade in market* $t+1$ *if and only if* $\tilde{x}_{t+1} = 1$*.*

Proposition 10 is the analog of Proposition 1 for the model of trade, and has a similar intuition. When an agent receives a rotten fruit, she reckons there is a sufficiently high probability that the other side is not a good trading partner, and decides not to incur the cost for producing high-quality fruit herself. This then creates a spiral effect where distrust in the ability of the other sides to be a good trading partner perpetrates over time. But, for the same reason as in the baseline model, this spiral also comes to an end—after a while, one of the sides concludes that the observed lack of trade is not very informative, and thus experiments with producing a high-quality fruit, which restarts trade.

It is also worth noting that the role of the staggered nature of production for each market, the ε transaction cost, and the strengthening of the solution concept from sequential equilibrium to trembling-hand perfect equilibrium is to ensure equilibrium uniqueness. Without these features, the strategy profile described in Proposition 10 would still be an equilibrium, but there would also be other, more "artificial" equilibria. For example, if production for each market were simultaneous, there would be an equilibrium in which players always produce low-quality fruit because they are sure that their trading partners do so as well. Without the transaction cost and trembling-hand perfection, there would be equilibria where, when both fruits are rotten, players use their trade approval decisions to send cheap talk messages (which are payoff irrelevant for them but matter for future generations).

Application to Communication

Another example of spiral effects, and hence of potential cycles, is communication between two groups on opposite sides of an issue. Even though informative communication may be in the interest of both parties, the fear that the other side is extremist can prevent communication between "moderates"

(normal players). We present a simple model illustrating this possibility. We assume that the interests of moderates on the two sides are sufficiently aligned to permit equilibrium communication in the absence of extremists—indeed, we take the extreme case where moderates on the two sides prefer the same policy, conditional on any state of the world. However, moderates will tune out the other side when they believe that they are listening to extremists. This implies that moderate communicators will not send extreme messages (so as to separate themselves from extremists), provided they believe the other group is likely to be moderate, but when they receive an extreme message, they will themselves also send an extreme message in return. This leads to cycles of moderation and extremism in communication, with phases of no communication (reminiscent to the "echo-chambers" of Sunstein (2006)) being followed by phases of informative and moderated communication.

Suppose that Group A is left-wing, while Group B is right-wing. There are overlapping generations, and each player lives for two periods, staggered across the groups. Informally, the model is as follows:

1. Every period, a state of the world, θ_t, is randomly drawn and is observed by the old player.
2. The old player sends a message, s_t, to the young player (who is from the other group), who then takes an action a_t (as in a standard cheap talk game).
3. The young player then ages, and the stage game repeats.

In particular, if a representative of Group A is the "sender" (resp., "receiver") in period t, then a representative of Group B is the "sender" (resp., "receiver") in period $t+1$.

More precisely, the model is as follows: The first representative of Group A is active in period 0 only, while subsequent representative of Group A are active in periods $\{t, t+1\}$ for t odd; and representatives of Group B are active in periods $\{t, t+1\}$ for t even. In every period $t = 0, 1, 2, \ldots$, the *period t state* $\theta_t \in \{-1, 0, 1\}$ is drawn independently, with probability $\frac{1}{3}$ on each state, and is observed by the active Group A representative if t is even, and by the active Group B representative if t is odd. The player who observes the state (the "sender") then sends a message $s_t \in \{-1, 0, 1\}$ to the other player (the "receiver"), who receives message $\tilde{s}_t \in \{-1, 0, 1\}$, where $\tilde{s}_t$ equals s_t with probability $1 - \frac{2}{3}\pi$, and $\tilde{s}_t$ equals each of the other two possible messages with probability $\frac{\pi}{3}$ each (thus, the message is replaced with "white noise" with probability π). Finally, the receiver takes an action $a_t \in \{-1, 0, 1\}$, and the game moves on to period $t+1$.

We now assume that there are three types of agents: moderates, extremists, and naifs (i.e., "naive" players). Each group consists entirely of one of these types. The probability that a group consists entirely of extremists is $\mu_0 > 0$, and the probability that a group consists entirely of naifs is $\nu_0 > 0$. Extremists and naifs are modelled as "behavioral types." It is assumed that left-wing extremists always send message $s_t = -1$ and take action $a_t = -1$, and right-wing extremists always send message $s_t = 1$ and take action $a_t = 1$. Naifs, on the other hand, always send message $s_t = \theta_t$ and take action $a_t = \tilde{s}_t$; this implies that naifs can be influenced by even extreme messages. Finally, a left-wing moderate who is active in period t gets utility $u_L(a_t, \theta_t)$ when action a_t is taken in state θ_t, and a right-wing

moderate gets utility $u_R(a_t, \theta_t)$; thus, for example, a left-wing moderate who is active in periods t and $t+1$ gets utility

$$u_L(a_t, \theta_t) + u_L(a_{t+1}, \theta_{t+1}).$$

These payoffs are realized at the end of period $t+1$, so that there is no information revealed from payoff realizations.

In addition, we assume that both left and right-wing moderates have preferences that satisfy single-crossing in (a, θ) and are single peaked with bliss point equal to the state θ_t at time t. This implies that left-wing and right-wing moderates always agree on the best action when the state of the world is known. However, we assume that a left-wing moderate prefers action -1 when she believes that the state is distributed "fairly evenly" on $\{-1, 0, 1\}$, while a right-wing moderate prefers action 1 when he believes this, in a sense formalized below. In addition, we assume that $u_L(a, \theta) = u_R(-a, -\theta)$ for all (a, θ). This simply allows us to avoid stating separate assumptions for u_L and u_R; none of our analysis relies on this symmetry.

We again impose a range of assumptions on u_L in order to focus on the most interesting region of parameter space (where the corresponding assumptions on u_R are implied by symmetry). To facilitate exposition, we state these assumptions "parametrically"; we view them as being fairly weak in the leading case where π is small relative to the other parameters.[30] In particular, using the notation "belief (x, y, z)" to stand for the belief that the state is -1 with probability x, 0 with probability y, and 1 with probability z, we impose:

Assumption COM 1. Action -1 is optimal for a left-wing moderate given belief

$$\left(\frac{(1-\pi)\mu_0 + \pi/3}{(1-\pi)(1+2\mu_0)+\pi}, \frac{(1-\pi)\mu_0 + \pi/3}{(1-\pi)(1+2\mu_0)+\pi}, \frac{1-\pi+\pi/3}{(1-\pi)(1+2\mu_0)+\pi}\right)$$

(i.e., when the state is equally likely to be -1 and 0, and somewhat more likely to be 1).

2. Action 0 is optimal for a left-wing moderate given belief

$$\left(\frac{\pi/3}{(1-\pi)(1-\mu_0)+\pi}, \frac{(1-\pi)(1-\mu_0)+\pi/3}{(1-\pi)(1-\mu_0)+\pi}, \frac{\pi/3}{(1-\pi)(1-\mu_0)+\pi}\right)$$

(i.e., when the state is equally likely to be -1 and 1, and much more likely to be 0).

Action 0 is optimal for left-wing moderate given belief

$$\left(\frac{\pi/3}{(1-\pi)(2-2\mu_0-\nu_0)+\pi}, \frac{(1-\pi)(1-\mu_0)+\pi/3}{(1-\pi)(2-2\mu_0-\nu_0)+\pi}, \frac{(1-\pi)(1-\mu_0-\nu_0)+\pi/3}{(1-\pi)(2-2\mu_0-\nu_0)+\pi}\right)$$

[30] For concreteness, a "typical" example satisfying the next assumption is given by $\pi = .01$, $\mu_0 = .5$, $\nu_0 = .1$, and $u_L(a, \theta)$ given by the following table:

		state		
		-1	0	1
action	-1	15	6	1
	0	5	7	3
	1	0	3	4

(i.e., when state is most likely to be 0, somewhat less likely to be 1, and much less likely to be -1).

3. $$\mu_0 + \frac{u_L(-1,-1) - u_L(1,-1)}{u_L(0,-1) - u_L(1,-1)}\nu_0 \in \left(1 - \frac{6(1-\pi)}{1+6\pi}\left(\frac{u_L(-1,-1) - u_L(0,-1)}{u_L(0,-1) - u_L(1,-1)}\right), 1\right).$$

4. $u_L(-1,0) > u_L(1,0)$.

Finally, for reasons familiar from cheap talk games, sequential equilibrium is not unique in this model, even under appropriate payoff restrictions (unlike in our other models). For example, one can always "switch" messages 0 and 1, say, so that message 0 corresponds to state 1 and vice versa.[31] We therefore impose the mild restriction of message-monotonicity: a sequential equilibrium is *message monotone* if for all histories, if a normal sender sends signal s with positive probability when the state is θ, then she never sends a signal strictly lower than s when the state is strictly higher than θ.[32]

The result is the following (proof in Appendix B):

Proposition 11 *Under Assumption COM, the communication model has a unique message-monotone sequential equilibrium. It has the following properties:*

1. *For all times t, left-wing moderate senders send $s_t = 0$ if $\theta_t = -1$ and $\tilde{s}_{t-1} \in \{-1, 0\}$ (or $t = 0$), and otherwise send $s_t = \theta_t$. Right-wing moderate senders send $s_t = 0$ if $\theta_t = 1$ and $\tilde{s}_{t-1} \in \{0, 1\}$, and otherwise send $s_t = \theta_t$.*

2. *For all times t, left-wing moderate receivers play $a_t = -1$ if $\tilde{s}_t \in \{-1, 1\}$, and play $a_t = 0$ if $\tilde{s}_t = 0$. Right-wing moderate receivers play $a_t = 1$ if $\tilde{s}_t \in \{-1, 1\}$, and play $a_t = 0$ if $\tilde{s}_t = 0$.*

The existence part of Proposition 11 is straightforward: one can compute a player's posterior belief about both the state and the other group's type following every possible signal $\tilde{s} \in \{-1, 0, 1\}$, and show that Assumption COM implies that the prescribed behavior is a best response. The intuition for uniqueness is as follows. For concreteness, focus on a left-wing sender and right-wing receiver. By Assumption COM, a moderate receiver plays $a = 1$ after signal $\tilde{s} = -1$, as this signal is always fairly likely to have come from an extremist sender, in which case it contains no information about θ. Also,

[31] This is true even though, because of the presence of naifs, ours is not quite a cheap talk game. If there are enough naifs (ν_0 large), then one could recover uniqueness, but it is natural to think of ν_0 as small.

[32] This is adapted from Chen's (2011) definition of message monotonicity to allow for mixed strategies. Chen assumes concave loss functions and a continuous action space, ensuring that all sequential equilibria are in pure strategies. Here, we allow for mixed strategies, even though we will show that all message-monotone sequential equilibria will still be in pure strategies.

It is also worth noting that while messages are assumed to be monotone in θ_t, receiver actions are not monotone in $\tilde{s}_t$ in the unique equilibrium of our model. This is fundamentally for the same reason as in Chen (2011): extreme messages may be more likely to come from biased senders, and therefore are discounted. In fact, our model is quite similar to a (special case of a) dynamic and two-sided version of Chen's model with three kinds of senders (moderate, extreme, and naive) rather than two (normal and naive). See also Chen, Kartik, and Sobel (2008) and Kartik, Ottaviani, and Squintani (2007) for static and one-sided communication models with a mix of normal and naive players.

a moderate receiver plays $a = 1$ after signal $\tilde{s} = 1$, as this signal is informative of $\theta = 1$ by message monotonicity, and Assumption COM implies that a moderate receiver would play $a = 1$ in the absence of an informative signal. Finally, a moderate receiver plays $a = 0$ after signal $\tilde{s} = 0$. This follows because if he played $a = -1$ or 1 after $\tilde{s} = 0$, then moderate senders would play $s = 0$ when $\theta = 0$, targeting naive receivers; and this, by Assumption COM, implies that moderate receivers should play $a = 0$ after $\tilde{s} = 0$ after all. Given this characterization of moderate receivers' strategies, Bayes rule and Assumption COM imply that moderate senders must play as specified in the proposition.

Intuitively, Proposition 11 shows that, in the unique message-monotone equilibrium, moderates start by moderating their "own-extreme" signals. In particular, left-wing senders will misreport a signal of -1 (and right-wingers will misreport 1) as 0 so as to separate themselves from the extremists on their side. Moderating one's own-extreme signal in this way is analogous to playing the good action in the baseline model in that it leads to Pareto-improving communication when the other group is moderate, but is not a best response when the other group is thought to consist of extremists. Then signals other than the sender's own-extreme signal are interpreted as good signals, and a moderate player who receives a good signal plays action equal to the signal, and then moderates her own signal next period. But this also implies that an extreme signal is evidence that the sender is an extremist, and a moderate from the other side who receives such a signal takes the opposite action and does not moderate her own signal in the next period (i.e., she sends her own-extreme signal if she thinks this is the state, in an effort to persuade naive receivers). This then starts a spiral analogous to the spirals we have seen in the baseline and trade models. Nevertheless, there are also important differences from the baseline model. First, here the reason why moderates refrain from sending extreme messages is precisely because they would like to distinguish themselves from extremists—thus the presence of extremists affects the form of communication between moderates even outside the cycle of distrust. Second, and perhaps more importantly, there are no longer any deterministic restarts of trust here. This is because the stochastic restarts of trust are frequent enough, ensuring that an own-extreme signal is always a strong indication that the sender is extremist.[33]

[33] A similar effect arises in Sections 3.2 and 3.3.

Appendix B: Omitted Proofs

Proof of Proposition 3. Rearranging the definition of T, one can check that T is the least integer greater than $\log\left(\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}\right)/\log(1-\pi)$, i.e.,

$$T = \left\lceil \frac{\log\left(\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}\right)}{\log(1-\pi)} \right\rceil. \tag{2}$$

This of course implies that $(1-\pi)^T \leq \frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)} \leq (1-\pi)^{T-1}$. Now, by Proposition 1, expected (limit-of-means) social welfare equals expected average social welfare within each T-period block. Consider for example the first block, consisting of periods 1 to $T-1$. Continuing to let u_t be player t's payoff, and assuming that both groups are normal, this equals

$$\frac{1}{T}\left[\begin{array}{c} E\left[u_0+u_{T-1}\right] \\ +\sum_{t=1}^{T-2}\left[\left(1-(1-\pi)^t\right)2u(0,0)+(1-\pi)^t\pi\left(u(1,1)+u(0,0)\right)+(1-\pi)^{t+1}2u(1,1)\right] \end{array}\right].$$

We are interested in evaluating this expression as $\pi \to 0$, which also implies (from (2)) $T \to \infty$. Thus, the expression of interest is

$$\begin{aligned}
&\lim_{\pi\to 0}\frac{1}{T}\left[\begin{array}{c} E\left[u_0+u_{T-1}\right]+\left(T-1-\frac{1-(1-\pi)^{T-1}}{\pi}\right)2u(0,0) \\ +\left(1-(1-\pi)^{T-1}\right)\left(u(1,1)+u(0,0)\right)+\left(\frac{1-(1-\pi)^T}{\pi}\right)2u(1,1) \end{array}\right] \\
= &\lim_{\pi\to 0}\frac{1}{T}\left[\begin{array}{c} E\left[u_0+u_{T-1}\right]+(T-1)2u(0,0) \\ +\left(1-(1-\pi)^{T-1}\right)\left(u(1,1)+u(0,0)\right)+2\left(\frac{1-(1-\pi)^T}{\pi}\right)\left(u(1,1)-u(0,0)\right) \\ +2\left(\left(\frac{1-(1-\pi)^{T-1}}{\pi}\right)-\left(\frac{1-(1-\pi)^T}{\pi}\right)\right)u(0,0). \end{array}\right] \\
= &\ 2u(0,0)+2\lim_{\pi\to 0}\frac{1}{T}\left(\frac{1-(1-\pi)^T}{\pi}\right)\left(u(1,1)-u(0,0)\right),
\end{aligned}$$

where the second equality follows by state rearrangement, and the third one simply from canceling the terms that go to zero and nothing that $(T-1)/T \to 1$ as $T \to \infty$. The first part of the proposition then follows by observing that

$$\begin{aligned}
\lim_{\pi\to 0}\frac{1}{T}\left(\frac{1-(1-\pi)^T}{\pi}\right) &= \lim_{\pi\to 0}\left(\frac{1-\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}}{T\pi}\right) = \lim_{\pi\to 0}\left(\frac{1-\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}}{T\log(1-\pi)}\right) \\
&= \lim_{\pi\to 0}\left(\frac{1-\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}}{-\log\left(\frac{\mu^*-\mu_0}{\mu^*(1-\mu_0)}\right)}\right) < 1,
\end{aligned}$$

where the inequality holds for all $\mu_0 > 0$. Finally, proof of the second part of the proposition is completed by observing that the limit of expected social welfare as $n \to \infty$ equals

$$2u(0,0)+2\lim_{n\to\infty}\frac{1}{T}\left(\frac{1-(1-\pi_n)^T}{\pi_n}\right)\left(u(1,1)-u(0,0)\right),$$

and that

$$\lim_{n\to\infty}\frac{1}{T}\left(\frac{1-(1-\pi_n)^T}{\pi_n}\right) = \lim_{n\to\infty}\left(\frac{1-\frac{\mu^*-\mu_{0,n}}{\mu^*(1-\mu_{0,n})}}{T\pi_n}\right) = \lim_{n\to\infty}\left(\frac{1-\frac{\mu^*-\mu_{0,n}}{\mu^*(1-\mu_{0,n})}}{T\log(1-\pi_n)}\right)$$

$$= \lim_{n\to\infty}\left(\frac{1-\frac{\mu^*-\mu_{0,n}}{\mu^*(1-\mu_{0,n})}}{-\log\left(\frac{\mu^*-\mu_{0,n}}{\mu^*(1-\mu_{0,n})}\right)}\right) = 1,$$

where the final equality uses $\mu_{0,n}\to 0$. ■

Proof of Lemma 1. Suppose that normal players choose $y_t = 1$ if and only if $\tilde{y}_{t-1} = 1$ or player $t-1$ is crazy. Now if both groups are normal, then non-crazy player t plays $y_t = 0$ if and only if the most recent "misperception" (i.e., time t' such that $y_{t'} = 1$ but $\tilde{y}_{t'} = 0$) came after the most recent arrival of a crazy player. Then denote the probability of this happening in period t by γ_t. Then one can check that

$$\gamma_{t_0} = \left(\sum_{t=0}^{t_0-1}(1-\kappa)^t\kappa\left(1-(1-\pi)^t\right)\right) + (1-\kappa)^{t_0}\left(1-(1-\pi)^{t_0}\right),$$

where the firm term sums over times at which the most recent arrival of a crazy player could have occurred, and the second term accounts for the possibility that there might not have been any crazy players by time t_0. Summing the geometric series yields

$$\gamma_t = \left(\frac{(1-\kappa)\pi}{\kappa+(1-\kappa)\pi}\right)\left(1-(1-\kappa)^t(1-\pi)^t\right). \tag{3}$$

Now, by Bayes rule,

$$\mu_t = \frac{\mu_0}{\mu_0+(1-\mu_0)\gamma_t}.$$

Note that μ_t is decreasing in t because γ_t is increasing in t. Moreover, from (3),

$$\mu_\infty = \frac{\mu_0(\kappa+(1-\kappa)\pi)}{\mu_0\kappa+(1-\kappa)\pi}.$$

Therefore, $\mu_\infty < \mu^*_{CRAZY}$ if and only if

$$\kappa < \kappa^* \equiv \frac{\pi(\mu^*_{CRAZY}-\mu_0)}{\pi(\mu^*_{CRAZY}-\mu_0)+\mu_0(1-\mu^*_{CRAZY})},$$

which is a positive number under Assumption 2′. ■

Proof of Proposition 6. Since player $t+1$ plays $x_{t+1} = 1$ if and only if he is normal and $\tilde{y}_t = 1$, it follows that (normal) player t plays $y_t = 1$ if and only if his belief that the other group is bad is below the cutoff μ^*. Now one can compute that $M^t_{(1,1)} = \frac{\pi'+\pi(1-\pi-\pi')^t}{\pi+\pi'}$ (for example, by adopting the proof of Lemma 1). In particular, $M^t_{(1,1)} > \pi'$ for all t, and hence $\mu'_t < \mu_0$ for all t. Therefore, Assumption 2 implies that player t always plays $y_t = 1$ after seeing signal $\tilde{y}_{t-1} = 1$. Finally, $\mu_t > \mu^*$ for all

$t < T_{2-SIDED}$ (with the convention that $T_{2-SIDED} = \infty$ if $\mu_\infty \geq \mu^*$), by definition of $T_{2-SIDED}$, so player t plays $y_t = 0$ after seeing $\tilde{y}_{t-1} = 1$, for all $t < T_{2-SIDED}$. That player t plays $y_t = 0$ after seeing $\tilde{y}_{t-1} = 1$ for all $t \neq 0 \bmod T_{2-SIDED}$ can now be established by induction, as in the baseline model. ■

Proof of Proposition 7. Suppose not. Then there exists a time $\bar{T}$ such that at all times $t \geq \bar{T}$ normal player t plays $y_t = 0$ after observing all bad signals. Suppose that both groups are normal, and observe that the probability that player $\bar{T} + K$ observes all bad signals is at least π^K. In this event, all subsequent players play $y_t = 0$ and thus observe all bad signals. In the alternative event that player $\bar{T} + K$ observes at least one good signal, the probability that player $\bar{T} + 2K$ observes all bad signals is still at least π^K. Hence, the overall probability that player $\bar{T} + 2K$ observes all bad signals is at least $1 - \left(1 - \pi^K\right)^2$. Now it is easy to see by induction on m that player $\bar{T} + mK$ observes all bad signals with probability at least $1 - \left(1 - \pi^K\right)^m$. Hence, normal player $\bar{T} + mK$'s belief that the other group is bad when she observes all bad signals is at most

$$\tilde{\mu}_{\bar{T}+mK} = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\left(1 - (1 - \pi^K)^m\right)}.$$

This belief converges to μ_0 as $m \to \infty$, so it follows from Assumption 2 that $\tilde{\mu}_{\bar{T}+MK} < \mu^*$ for some integer M. Therefore, normal player $\bar{T} + MK$ would deviate to playing $y_{\bar{T}+MK} = 1$ after observing all bad signals, which yields a contradiction and establishes the desired result. ■

Before presenting the proof of Proposition 8, we state and prove the following lemma:

Lemma 2 *In the political partisanship model, let μ_t be moderate player t's assessment of the probability that the opposing group is extremist, conditional on player $t-1$ failing to communicate and player t's group holding power in period t. Then $\lim_{t\to\infty} \mu_t = \mu_0$, and $\lim_{\pi\to 0} \mu_1 = 1$.*

Proof. We work up to computing μ_t by first computing three other probabilities, which will be denoted by χ_t, ϕ_t, and ψ_t. Let χ_t be the probability that whichever group is in power today is also in power t periods from now, when the probability that power switches each period is ρ. Note that this event is equivalent to power switching an even number of times out of t chances. Thus, it is straightforward to check that

$$\begin{aligned}
\chi_t &= \sum_{k=0}^{\frac{t}{2}} \binom{t}{2k} \rho^{2k} (1-\rho)^{t-2k} \text{ if } t \text{ is even,} \\
\chi_t &= \sum_{k=0}^{\frac{t-1}{2}} \binom{t}{2k} \rho^{2k} (1-\rho)^{t-2k} \text{ if } t \text{ is odd.}
\end{aligned}$$

In addition, it can easily be shown that if $\rho = \frac{1}{2}$ (contrary to our assumptions) then $\chi_t = \frac{1}{2}$.

Next, let ϕ_t be the probability that the left-wing group (say) is in power in period t *and* does not observe communication, conditional on the event that the left-wing group is moderate and the

right-wing group is extremist. By Bayes rule,

$$\phi_t = \frac{1}{2}\left[\left(\frac{1}{2}\right)^t (1-\pi)^{t-1}\pi + \sum_{m=1}^{t-1}\left(\frac{1}{2}\right)^{m-1}(1-\pi)^{m-1}\left(\frac{1}{2}\left(1-\chi_{t-m}\right)+\frac{1}{2}\pi\chi_{t-m}\right)\right] + \frac{1}{2}\left(1-\chi_t\right).$$

This expression requires some explanation. The first term corresponds to the event that the left-wing group is in power in period 0. This event is then divided into the events that the first time m at which a player sets $s_m = 0$ is $m = 1, \ldots, t$. Conditional on the left-wing group holding power in period 0, the probability that $m = t$, the left-wing group holds power in period t, and player $t-1$ does not successfully communicate is $\left(\frac{1}{2}\right)^t (1-\pi)^{t-1}\pi$. And for any $m < t$, conditional on the left-wing group holding power in period 0, the probability that the first time at which a player sets $s = 0$ is m, the left-wing group holds power in period t, and player $t-1$ does not successfully communicate is $\left(\frac{1}{2}\right)^{m-1}(1-\pi)^{m-1}\left(\frac{1}{2}\left(1-\chi_{t-m}\right)+\frac{1}{2}\pi\chi_{t-m}\right)$, as after $m-1$ periods of the left-wing group holding power and setting $s = 1$, there is probability $\frac{1}{2}$ that the (extremist) right-wing group takes power, and probability $\frac{1}{2}\pi$ that the left-wing group retains power but starts setting $s = 0$ (and in either event all subsequent power switches occur with probability ρ). Finally, the $\frac{1}{2}\left(1-\chi_t\right)$ term corresponds to the possibility that the right-wing group is in power in period 0.

Finally, let ψ_t be the probability that the left-wing group is in power in period t and does not observe communication, conditional on the event that the left-wing group is moderate and the right-wing group is *moderate*. By Bayes rule,

$$\psi_t = \frac{1}{2}(1-\pi)^{t-1}\pi + \sum_{m=1}^{t-1}(1-\pi)^{m-1}\pi\left(\frac{1}{2}\left(1-\chi_{t-m}\right)+\frac{1}{2}\chi_{t-m}\right).$$

This formula is similar but somewhat simpler than the formula for ϕ_t. The first term corresponds to the event that there are $t-1$ successful communications. For each term in the sum, the probability that the first time at which a player sets $s = 0$ is m, the left-wing group holds power in period t, and player $t-1$ does not successfully communicate is $(1-\pi)^{m-1}\pi\left(\frac{1}{2}\left(1-\chi_{t-m}\right)+\frac{1}{2}\chi_{t-m}\right)$, as after $m-1$ periods of successful communication and a single failed communication, the posterior probability that either group holds power is $\frac{1}{2}$, and both groups subsequently set $s = 0$ (so subsequent power switches occur with probability ρ).

The assessment of moderate player t of the probability that the opposing group is extremist, conditional on player $t-1$ not successfully communicating (and conditional on player t's group holding power in period t) is now simply

$$\mu_t = \frac{\mu_0\phi_t}{\mu_0\phi_t + (1-\mu_0)\psi_t} = \frac{\mu_0}{\mu_0 + (1-\mu_0)\frac{\psi_t}{\phi_t}}.$$

Now note that $\lim_{t\to\infty}\chi_t = \frac{1}{2}$, so the sum in the formula for ϕ_t converges to $\sum_{m=0}^{\infty}\left(\frac{1}{2}\right)^m(1-\pi)^m\left(\frac{1}{2}-\frac{1}{2}\pi\right) = 1$, and hence $\phi_t \to \frac{1}{2}\left[0+\frac{1}{2}\right]+\frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{2}$. Similarly, the sum in the formula for ψ_t converges to $\sum_{m=0}^{\infty}(1-\pi)^m\pi\left(\frac{1}{2}\right) = \frac{1}{2}$, and therefore $\psi_t \to 0+\frac{1}{2} = \frac{1}{2}$. Therefore,

$\lim_{t\to\infty} \mu_t = \mu_0$. In addition, it is easy to verify that

$$\frac{\psi_1}{\phi_1} = \frac{\pi}{\frac{1}{2}\pi + \rho},$$

which goes to 0 as $\pi \to 0$, so $\mu_1 \to 1$ as $\pi \to 0$. ■

Proof of Proposition 8. Fix a time $t < T_{POL}$, where T_{POL} is to be computed. We first derive the turnout decisions of the supporters of the incumbent party. When they see communication about θ, they infer that their own party is moderate, so their net benefit from turning out is

$$\left(\frac{1}{2} - \rho\right) \mu \left(1 + \alpha\right),$$

where the first term is the change in probability of the same party keeping power because of the greater turnout, the second term, μ, is the belief after receiving informative communication that the other side is an extremist, and the last term is the net gain from having their own party rather than an extremist opposing party in power, taking into account that moderates on both sides will choose the right policy on the major issue. Under the hypothesized behavior of politicians, μ is determined by Bayes rule and satisfies $\mu \leq \mu_0$. Intuitively, this is because both the information that the period t base's party is in power and the information that the period t leader successfully communicates suggest that the opposing party is more likely to be moderate. To see this formally, note that if the opposing party is extremist, then the only way that the period t base's party holds power in period t and the period t leader successfully communicates is if the period t base's party holds power and successfully communicates in periods $0, \ldots, t$, which occurs with probability $\left(\frac{1}{2}\right)^{t+1} (1-\pi)^{t+1}$. In addition, if the opposing party is moderate, then the only way that the period t base's party holds power in period t and the period t leader successfully communicates is if there is successfully communication in periods $0, \ldots, t$ (which occurs with probability $(1-\pi)^{t+1}$), and the period t base's party holds power in period t (which, conditional on the opposing party being moderate and there being successful communication in periods $0, \ldots, t$, occurs with probability $\frac{1}{2}$). Thus,

$$\mu = \frac{\mu_0 \left(\frac{1}{2}\right)^{t+1} (1-\pi)^{t+1}}{\mu_0 \left(\frac{1}{2}\right)^{t+1} (1-\pi)^{t+1} + (1-\mu_0) \left(\frac{1}{2}\right) (1-\pi)^{t+1}} \leq \mu_0.$$

Finally, the fact that $\mu \leq \mu_0$ combined with the assumption that $c > \left(\frac{1}{2} - \rho\right)(1+\alpha)\mu_0$ implies that $c > \left(\frac{1}{2} - \rho\right)(1+\alpha)\mu$, so the incumbent's base never turns out following successful communication.

Next consider the behavior of the incumbent's base following propaganda. Since every politician acts on the partisan issue following propaganda, the net benefit from turning out in this case can be written as

$$2\alpha \left(\frac{1}{2} - \rho\right).$$

The assumption that $c < 2\alpha\left(\frac{1}{2} - \rho\right)$ now implies that the incumbent's base always turns out following propaganda.

We now turn to moderate politicians' incentives. First, consider the payoff from playing $s_t = 0$. Given the above assumptions, this ensures turnout but will also cause the next leader to act on that partisan issue. When the belief that the other side is extremist is μ, this has payoff

$$-(1-\rho)(1) - \rho(1+2\alpha).$$

On the other hand, playing $s_t = 1$ yields payoff

$$\pi\left[-(1-\rho)(1) - \rho(1+2\alpha)\right] + (1-\pi)\left[-\frac{1}{2}(\alpha) - \frac{1}{2}\left(\mu(1+2\alpha) + (1-\mu)(\alpha)\right)\right].$$

The difference between these two payoffs is given by

$$(1-\pi)\left[2\alpha\rho + (1-\alpha) - \frac{1}{2}(1+\alpha)\mu\right],$$

which is positive for

$$\mu < \mu^*_{POL} = \frac{4\alpha\rho + 2(1-\alpha)}{1+\alpha},$$

and negative for μ above this threshold. Thus, moderate politicians attempt communication if and only if $\mu < \mu^*_{POL}$ (which is always a positive number).

In contrast, the payoff to an extremist incumbent from setting $s_t = 0$ equals

$$-(1-\rho)(\alpha) - \rho(\alpha+2),$$

while the payoff to an extremist incumbent from setting $s_t = 1$ equals

$$\pi\left[-(1-\rho)(\alpha) - \rho(\alpha+2)\right] + (1-\pi)\left[-\frac{1}{2}(\alpha) - \frac{1}{2}\left(\mu(\alpha+2) + (1-\mu)(1)\right)\right].$$

The difference between these payoffs equals

$$(1-\pi)\left[2\rho - \frac{1}{2}(1-\alpha) - \frac{1}{2}(1+\alpha)\mu\right],$$

which is positive for

$$\mu < \frac{4\rho - (1-\alpha)}{1+\alpha}.$$

However, the assumption that $\rho < \frac{1-\alpha}{4}$ implies that $\frac{4\rho-(1-\alpha)}{1+\alpha} < 0$. Thus, extremist politicians never attempt communication.

Now to compute T_{POL}, note that Lemma 2 and the assumption that $\mu_0 < \mu^*_{POL}$ imply that there exists a smallest integer t such that $\mu_t \leq \mu^*_{POL}$. Denoting this integer by T_{POL} and following the same argument as in the baseline model now implies that the conjectured strategy profile (and beliefs) is a sequential equilibrium, with cycle length T_{POL}.[34]

[34] In addition, the fact that $\lim_{\pi\to 0}\mu_1 = 1$ implies that $T > 1$ when π is sufficiently small.

Finally, this sequential equilibrium is unique—up to indifference at non-generic parameter values—by the same argument as in the baseline model, with the addition that the beliefs and strategies of the base are also uniquely determined by induction on t. In particular, the $t = 0$ base does not turn out after observing communication (as its assessment that the opposing party is extreme must be μ_0, regardless of whether or not it observes communication, in any sequential equilibrium), and the $t = 0$ base does turn out after observing propaganda (regardless of its assessment that the opposing party is extreme). The rest of the induction argument proceeds exactly as in the baseline model (or as in the proof of Proposition 10, where more details are provided). ■

Proof of Proposition 9. Suppose that $t < \tilde{T}$, where $\tilde{T}$ is the cycle length when both bases can turn out, which remains to be computed. When the out-of-power base turns out, it increases the probability that its party holds power in period $t+1$ by $\frac{1}{2} - \rho$ (note that this is true even if the in-power base also turns out, as in that case turning out increases the probability of a power switch from $G(-k) = \rho$ to $\frac{1}{2}$). If player t successfully communicated and both parties are moderate, then the benefit to the out-of-power base of having its own party in power in period $t+1$ (compared to having the other party in power) is 0. If player t successfully communicated and only the opposing party is moderate, then this benefit is $-1+\alpha$ (so that in this case the out-of-power base would rather its party lose the election). If player t did not successfully communicate, then the benefit is 2α. Finally, the event that player t successfully communicated and only the out-of-power party is moderate occurs with probability 0. We now claim that the net benefit to the out-of-power period t base of turning out (for $t < \tilde{T}$) is

$$NB_t = \left(\frac{1}{2} - \rho\right) \left[\begin{array}{c} (1-\mu_0)^2 (1-\pi)^t (0) + \mu_0 (1-\mu_0) \left(\frac{1}{2}\right)^t (1-\pi)^t (-1+\alpha) \\ + \left(1 - (1-\mu_0)^2 (1-\pi)^t - \mu_0 (1-\mu_0) \left(\frac{1}{2}\right)^t (1-\pi)^t\right) (2\alpha) \end{array} \right].$$

To see this, note that the probability that player t successfully communicated and both parties are moderate, conditional on the out-of-power base's party being out of power, is $(1-\mu_0)^2 (1-\pi)^t$; the probability that player t successfully communicated and only the opposing party is moderate, conditional on the out-of-power base's party being out of power, is $\mu_0 (1-\mu_0) \left(\frac{1}{2}\right)^t (1-\pi)^t$; and, since the probability that player t successfully communicated and only the out-of-power base's party is moderate is 0, the probability that player t did not successfully communicate is $1 - (1-\mu_0)^2 (1-\pi)^t - \mu_0 (1-\mu_0) \left(\frac{1}{2}\right)^t (1-\pi)^t$.

Note that NB_t is increasing in t, and that $\lim_{t\to\infty} NB_t = \left(\frac{1}{2} - \rho\right)(2\alpha)$. Hence, the assumption that $c < \left(\frac{1}{2} - \rho\right)(2\alpha)$ implies that there is some time $\hat{T}$ such that the out-of-power base turns out if and only if $\hat{T} < \tilde{T}$ and $t \geq \hat{T} \bmod \tilde{T}$. Now observe that if $\hat{T} \geq T_{POL}$ (defined in the proof of Proposition 8), then μ_t is as in Proposition 8, which implies that $\tilde{T} = T_{POL}$ and the out-of-power base never turns out. If instead $\hat{T} < T_{POL}$ then the computation of μ_t in the proof of Proposition 8 must be modified to take into account that the probability of a power switch changes deterministically at time $\hat{T}$ (from $\frac{1}{2}$ to $1-\rho$ if the in-power base does not turn out, and from ρ to $\frac{1}{2}$ if the in-power base does turn out). Without explicitly calculating $\tilde{T}$ in this case, it is easy to see that if $\hat{T} < T_{POL}$ then

also $\hat{T} < \tilde{T}$, as if $\hat{T} \geq \tilde{T}$ then the formula for $\tilde{T}$ would be the same as for T, contradicting $\hat{T} < T_{POL}$. Thus, if $\hat{T} < T_{POL}$, then there is a deterministic time $\tilde{T} > \hat{T}$ such that partisan phases always begin with probability π and end deterministically at times $t = 0 \bmod \tilde{T}$, the in-power base turns out only during partisan phases, and the out-of-power base turns out for the last $\tilde{T} - \hat{T}$ periods of every $\tilde{T}$ period cycle (regardless of whether the partisan phase has begun or not). ■

Proof of Proposition 10. First, note that in any trembling-hand perfect equilibrium (henceforth, "equilibrium") a player (good or bad) approves trade if and only if her opponent's fruit is good. This follows because trade strictly increases a player's payoff if her opponent's fruit is good and strictly decreases a player's payoff if her opponent's fruit is bad—regardless of the quality of her own fruit—and perfection requires that a player's approval decision is a best response to a completely mixed strategy of her opponent's.[35]

Second, note that in any equilibrium normal player t plays $x_t = 1$ if and only if $\tilde{y}_{t-1} = 1$. To see this, note that her expected payoff in market t when she produces a good fruit and her opponent's fruit is good equals

$$(1 - \pi)\, b - c.$$

Her expected payoff in market t when she produces a bad fruit is 0. In addition, her expected payoff in market $t+1$ does not depend on her choice of x_t, (because player $t+1$'s strategy cannot depend on the outcome of market t, and player $t-1$'s trade approval decision cannot give player t any information about the other group's type since a player's trade approval decision is independent of her type). Finally, Assumption 2″ implies that $\mu^*_{TRADE} > 0$, which in turn implies that $(1 - \pi)\, b - c > 0$. So Assumption 2″ implies that normal player t plays $x_t = 1$ if and only if $\tilde{y}_{t-1} = 1$.

It remains only to determine when normal player t plays $y_t = 1$. We proceed by induction on t. Suppose that play in every equilibrium is as specified by the proposition for all $t' \leq t < T$, $t \in \mathbb{N}$. Then at time $t + \frac{1}{3}$, normal player t's assessment of the probability that the other group is good is 1 if $\tilde{y}_{t-1} = 1$ and is μ_t if $\tilde{y}_{t-1} = 0$ (and is μ_0 if $t = 0$). Then normal player t's expected payoff in market $t + 1$ when she produces a good fruit and her assessment is μ equals

$$(1 - \mu) \left[(1 - \pi)^2\, b + \pi\, (1 - \pi)\, d \right] + \mu\, (1 - \pi)\, d - c,$$

while her expected payoff in market $t+1$ when she produces a bad fruit equals 0. Hence, she produces a good fruit if $\mu < \mu^*_{TRADE}$ and produces a bad fruit if $\mu > \mu^*_{TRADE}$. Since $\mu_t > \mu^*_{TRADE}$ by definition of T, it follows that she plays $y_t = 1$ if and only if $\tilde{y}_{t-1} = 1$. This proves that play in every equilibrium is as specified by the proposition for all $t < T$.

The same argument now implies that, as specified in the proposition, normal player T's assessment is below μ^*_{TRADE} regardless of $\tilde{y}_{T-1}$ (strictly so, by Assumption 3″). This implies that she always plays $y_T = 1$. Repeating the argument from the previous paragraph now implies that play in every equilibrium is as in the proposition for all $t \in \{T + 1, \ldots, 2T\}$, and inducting on $k \in \mathbb{N}$ then implies

[35] Note that this argument relies on the presence of the transaction cost. It is the only place where the transaction cost is used.

that play is as in the proposition for all $t \in \{kT+1, \ldots, (k+1)T\}$ for any $k \in \mathbb{N}$, completing the proof. ■

Proof of Proposition 11. We first establish the existence of an equilibrium of this form, and then show that it is unique.

Let γ_t denote the probability that player t does not moderate her signal (i.e., sends $s_t = \theta_t$ for all θ_t) when both players are normal and follow the strategies described in the proposition; this is the same as the probability that $\tilde{s}_{t-1}$ equals player $t-1$'s own-extreme signal. We calculate player t's posterior belief about (1) the state, and (2) her opponent's type, in terms of γ_t, conditional on the received signal. For these calculations, suppose that the receiver is right-wing—the other case is symmetric.

Posteriors about the state are as follows:

After signal $\tilde{s} = -1$:

$$\begin{aligned}
\Pr(\theta = -1) &= \frac{(1-\pi)(\mu_0 + \nu_0 + (1-\mu_0-\nu_0)\gamma_t) + \frac{\pi}{3}}{(1-\pi)(3\mu_0 + \nu_0 + (1-\mu_0-\nu_0)\gamma_t) + \pi} \\
\Pr(\theta = 0) &= \frac{(1-\pi)\mu_0 + \frac{\pi}{3}}{(1-\pi)(3\mu_0 + \nu_0 + (1-\mu_0-\nu_0)\gamma_t) + \pi} \\
\Pr(\theta = 1) &= \frac{(1-\pi)\mu_0 + \frac{\pi}{3}}{(1-\pi)(3\mu_0 + \nu_0 + (1-\mu_0-\nu_0)\gamma_t) + \pi}.
\end{aligned}$$

This follows from Bayes rule by noting that the probabilities of the following joint events are:

$$\begin{aligned}
\Pr(\theta = -1, \tilde{s} = -1, extreme) &= \Pr(\theta = 0, \tilde{s} = -1, extreme) \\
&= \Pr(\theta = 1, \tilde{s} = -1, extreme) = \frac{(1-\pi)\mu_0}{3} + \frac{\pi}{9} \\
\Pr(\theta = -1, \tilde{s} = -1, naif) &= \frac{(1-\pi)\nu_0}{3} + \frac{\pi}{9} \\
\Pr(\theta = -1, \tilde{s} = -1, normal) &= \frac{(1-\pi)(1-\nu_0-\mu_0)\gamma_t}{3} + \frac{\pi}{9} \\
\Pr(\theta = 0, \tilde{s} = -1, extreme) &= \frac{(1-\pi)\mu_0}{3} + \frac{\pi}{9} \\
\Pr(\theta = 0, \tilde{s} = -1, naif) &= \Pr(\theta = 1, \tilde{s} = -1, naif) = \Pr(\theta = 0, \tilde{s} = -1, normal) \\
&= \Pr(\theta = 1, \tilde{s} = -1, normal) = \frac{\pi}{9}.
\end{aligned}$$

After signal $\tilde{s} = 0$:

$$\begin{aligned}
\Pr(\theta = -1) &= \frac{(1-\pi)(1-\mu_0-\nu_0)(1-\gamma_t) + \frac{\pi}{3}}{(1-\pi)(\nu_0 + (1-\mu_0-\nu_0)(2-\gamma_t)) + \pi} \\
\Pr(\theta = 0) &= \frac{(1-\pi)(1-\mu_0) + \frac{\pi}{3}}{(1-\pi)(\nu_0 + (1-\mu_0-\nu_0)(2-\gamma_t)) + \pi} \\
\Pr(\theta = 1) &= \frac{\frac{\pi}{3}}{(1-\pi)(\nu_0 + (1-\mu_0-\nu_0)(2-\gamma_t)) + \pi}.
\end{aligned}$$

After signal $\tilde{s} = 1$:

$$\begin{aligned}
\Pr(\theta = -1) &= \frac{\frac{\pi}{3}}{(1-\pi)(1-\mu_0)+\pi} \\
\Pr(\theta = 0) &= \frac{\frac{\pi}{3}}{(1-\pi)(1-\mu_0)+\pi} \\
\Pr(\theta = 1) &= \frac{(1-\pi)(1-\mu_0)+\frac{\pi}{3}}{(1-\pi)(1-\mu_0)+\pi}.
\end{aligned}$$

For the prescribed strategies to be an equilibrium it is required that, for all t, the receiver plays $a = 1$ after $\tilde{s} = -1$, plays $a = 0$ after $\tilde{s} = 0$, and plays $a = -1$ after $\tilde{s} = 1$. The first and third of these conditions are implied by Part 1 of Assumption COM (noting that after $\tilde{s} = -1$, playing $a = 1$ is less appealing when γ_t is higher, and Part 1 of Assumption COM implies that playing $a = 1$ is optimal for $\gamma_t = 1$). The second of these conditions is implied by Part 2 of Assumption COM (as this implies that the condition holds for $\gamma_t = 0$ and $\gamma_t = 1$, and preferences are single-crossing in (a, θ)).

Posteriors about the opponent's type (again assuming a right-wing receiver and again using the above expressions) are as follows:

After signal $\tilde{s} = -1$:

$$\begin{aligned}
\Pr(extreme) &= \mu_t = \frac{\mu_0(3-2\pi)}{(1-\pi)(3\mu_0+\nu_0+(1-\mu_0-\nu_0)\gamma_t)+\pi} \\
\Pr(naive) &= \nu_t = \frac{\nu_0}{(1-\pi)(3\mu_0+\nu_0+(1-\mu_0-\nu_0)\gamma_t)+\pi} \\
\Pr(normal) &= 1-\mu_t-\nu_t = \frac{(1-\mu_0-\nu_0)((1-\pi)\gamma_t+\pi)}{(1-\pi)(3\mu_0+\nu_0+(1-\mu_0-\nu_0)\gamma_t)+\pi}.
\end{aligned}$$

After signal $\tilde{s} = 0$:

$$\begin{aligned}
\Pr(extreme) &= \mu_t = \frac{\mu_0\pi}{(1-\pi)(\nu_0+(1-\mu_0-\nu_0)(2-\gamma_t))+\pi} \\
\Pr(naive) &= \nu_t = \frac{\nu_0}{(1-\pi)(\nu_0+(1-\mu_0-\nu_0)(2-\gamma_t))+\pi} \\
\Pr(normal) &= 1-\mu_t-\nu_t = \frac{(1-\mu_0-\nu_0)((1-\pi)(2-\gamma_t)+\pi)}{(1-\pi)(\nu_0+(1-\mu_0-\nu_0)(2-\gamma_t))+\pi}.
\end{aligned}$$

After signal $\tilde{s} = 1$:

$$\begin{aligned}
\Pr(extreme) &= \mu_t = \frac{\mu_0\pi}{(1-\pi)(1-\mu_0)+\pi} \\
\Pr(naive) &= \nu_t = \frac{\nu_0}{(1-\pi)(1-\mu_0)+\pi} \\
\Pr(normal) &= 1-\mu_t-\nu_t = \frac{1-\mu_0-\nu_0}{(1-\pi)(1-\mu_0)+\pi}.
\end{aligned}$$

Here, we need that it is always optimal for the receiver to moderate his signal in the next period (i.e., send 0 when $\theta = 1$) after $\tilde{s} = -1$ and $\tilde{s} = 1$, but not after $\tilde{s} = 0$. It is easy to see that when she does

so following $\tilde{s} = -1$, she will also do it a fortiori following $\tilde{s} = 1$. Now it is optimal for a sender with beliefs μ_t and ν_t to moderate her signal if and only if

$$(1 - \nu_t) u_L (1, -1) + \nu_t u_L (-1, -1) < (1 - \mu_t) u_L (0, -1) + \mu_t u_L (1, -1),$$

or

$$\mu_t + \left(\frac{u_L (-1, -1) - u_L (1, -1)}{u_L (0, -1) - u_L (1, -1)} \right) \nu_t < 1.$$

So we need to show that, for any $\gamma_t \in [0, 1]$, this inequality holds whenever μ_t and ν_t result from signal $\tilde{s} = 0$, while the opposite inequality holds whenever μ_t and ν_t result from signal $\tilde{s} = -1$.[36] A bit of algebra shows that Part 3 of Assumption COM is a sufficient condition for this to be the case.

The existence of an equilibrium of the specified form follows by noting that $s_t = \theta_t$ is optimal when θ_t is not one's own-extreme state (and this is in fact true regardless of the receiver's type).

The proof of uniqueness of the message-monotone sequential equilibrium (henceforth "equilibrium") proceeds in five steps. Let us continue to assume a right-wing receiver.

Step 1: In any equilibrium, moderate receivers play $a = 1$ after signal $\tilde{s} = -1$.

Since preferences satisfy single-crossing in (a, θ), it suffices to check that a moderate receiver plays $a = 1$ after signal $\tilde{s} = -1$ when moderate senders send $s = -1$ after $\theta = -1$ and $\theta = 0$, but not after $\theta = 1$. This condition is exactly Part 1 of Assumption COM.[37]

Step 2: In any equilibrium, moderate receivers play $a = 1$ after signal $\tilde{s} = 1$.

By message-monotonicity, the posterior distribution after signal $\tilde{s} = 1$ is weakly higher (in the sense of first-order stochastic dominance) than it would be if all senders sent signal $s = 1$ for all θ. Part 1 of Assumption COM then implies that a moderate receiver plays $a = 1$ in that case, so the claim follows from single-crossing.

Step 3: In any equilibrium, moderate senders send $s = 1$ when $\theta = 1$.

This follows from Step 2, as sending $s = 1$ is then optimal against every type of receiver, and strictly optimal against naive and moderate receivers.

Step 4: In any equilibrium, moderate receivers play $a = 0$ after signal $\tilde{s} = 0$, and moderate senders send $s = 0$ when $\theta = 0$.

First, suppose moderate receivers play $a = 1$ with probability 1 after $\tilde{s} = 0$. Then moderate senders would send $s = 0$ when $\theta = 0$ (as the moderate and extreme receivers would be playing $a = 1$ always, so moderate senders would target the naive receivers). But then Part 2 of Assumption COM would imply that moderate receivers should play $a = 0$ after $\tilde{s} = 0$, yielding a contraction.

Therefore, a moderate receiver must play either $a = -1$ or $a = 0$ with positive probability after $\tilde{s} = 0$. This implies that moderate senders send $s = 0$ when $\theta = 0$, as this is clearly optimal against naive and extreme receivers, and it is also strictly optimal against moderate receivers by Part 4 of Assumption COM. Hence, Part 2 of Assumption COM ensures that regardless of what moderate

[36] This is actually stronger than necessary, because γ_t does not take on all values in $[0, 1]$.

[37] This is in fact the reason why Part 1 of Assumption COM is a stronger condition than is needed for existence.

senders do when $\theta = -1$, it is strictly optimal for moderate receivers to play $a = 0$ after $\tilde{s} = 0$. Thus moderate receivers must play $a = 0$ with probability 1 after $\tilde{s} = 0$.

Step 5: In any equilibrium, moderate senders send $s = 0$ *when* $\theta = -1$ *if* $\mu_t + \left(\frac{u_L(-1,-1)-u_L(1,-1)}{u_L(0,-1)-u_L(1,-1)}\right)\nu_t < 1$*, and send* $s = -1$ *when* $\theta = -1$ *if the opposite inequality holds.*

This follows from the characterization of the moderate receiver's strategy, single-crossing (which implies that sending $s = 0$ is always preferable to sending $s = 1$), and the derivation of this inequality in the existence part of the proof. ■

References

[1] Abramowitz, A.I. (2010), *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*, New Haven: Yale University Press.

[2] Abreu, D., D. Pearce, and E. Stacchetti (1986), "Optimal Cartel Equilibria with Imperfect Monitoring," *Journal of Economic Theory*, 39, 251-269.

[3] Abreu, D., D. Pearce, and E. Stacchetti (1990), "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, 58, 1041-1063.

[4] Acemoglu, D., M. Golosov, and A. Tsyvinski (2010), "Power Fluctuations and Political Economy," *Journal of Economic Theory*, 146, 1009-10041.

[5] Acemoglu, D., G. Egorov, and K. Sonin (2011), "A Political Theory of Populism," mimeo.

[6] Acemoglu, D. and M.O. Jackson (2011), "History, Expectations, and Leadership in the Evolution of Social Norms," mimeo.

[7] Alesina, A. (1988), "Credibility and Policy Convergence in a Two-Party System with Rational Voters," *American Economic Review*, 78, 796-805.

[8] Anderlini, L. Gerardi, D., and R. Lagunoff (2010), "Social Memory, Evidence, and Conflict," *Review of Economic Dynamics*, 13, 559–574.

[9] Baliga, S. and T. Sjostrom (2004), "Arms Races and Negotiations," *Review of Economic Studies*, 17, 129-183.

[10] Baliga, S. and T. Sjostrom (2010), "The Hobbesian Trap," *Oxford Handbook of the Economics of Peace and Conflict*, eds. Michelle Garfinkel and Stergios Skaperdas, Oxford: Oxford University Press.

[11] Baliga, S. and T. Sjostrom (2011), "The Strategy of Manipulating Conflict," *American Economic Review*, forthcoming.

[12] Banerjee, A.V. and R. Somanathan (2001), "A Simple Model of Voice," *Quarterly Journal of Economics*, 116, 189-227.

[13] Baron, D.P. (1996), "A Dynamic Theory of Collective Goods Programs," *American Political Science Review*, 90, 316-330.

[14] Battaglini, M. and S. Coate (2008), "A Dynamic Theory of Public Spending, Taxation and Debt," *American Economic Review*, 98, 201-236.

[15] Benabou, R. and G. Laroque (1992), "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility," *Quarterly Journal of Economics*, 107, 921-958.

[16] Bernhardt, D., S. Krasa, and M. Polborn (2008), "Political Polarization and the Electoral Effects of Media Bias," *Journal of Public Economics*, 92, 1092-1104.

[17] Bhaskar, V. (1998), "Informational Constraints and the Overlapping Generations Model: Folk and Anti-Folk Theorems," *Review of Economic Studies*, 65, 135-149.

[18] Bottazzi, L., M. Da Rin, and T. Hellmann (2011), "The Importance of Trust for Investment: Evidence from Venture Capital," mimeo.

[19] Canes-Wrone, B., M. Herron, and K. Shotts (2001), "Leadership and Pandering: A Theory of Executive Policymaking," *American Journal of Political Science*, 45, 532-550.

[20] Chan, J. and W. Suen (2008), "Spatial Theory of News Consumption and Electoral Competition," *Review of Economic Studies*, 75, 699-728.

[21] Chassang, S. and G. Padro-i-Miquel (2010), "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, 125, 1821-1858.

[22] Chen, Y. (2011), "Perturbed Communication Games with Honest Senders and Naive Receivers," *Journal of Economic Theory*, 401-424.

[23] Chen, Y., N. Kartik, and J. Sobel (2008), "Selecting Cheap-Talk Equilibria," *Econometrica*, 76, 117-136.

[24] Coleman, J.S. (1958), *Nigeria: Background to Nationalism,* University of California Press, Berkeley and Los Angeles.

[25] DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2011), "Unintended Media Effects in a Conflict Environment: Serbian Radio and Croatian Nationalism," mimeo.

[26] Dixit, A., G.M. Grossman, and F. Gul (2001), "The Dynamics of Political Compromise," *Journal of Political Economy*, 108, 531-568.

[27] Ekmekci, M., O. Gossner, and A. Wilson (2011), "Impermanent Types and Permanent Reputations," *Journal of Economic Theory*, forthcoming.

[28] Fiorina, M.P. (2011), *Culture War? The Myth of a Polarized America*, Third Edition, Boston: Longman.

[29] Glaeser, E. (2005) "The Political Economy of Hatred," *Quarterly Journal of Economics*, 120, 45-86.

[30] Green, E.J. and R.H. Porter (1984), "Noncooperative Collusion under Imperfect Price Formation," *Econometrica*, 52, 87-100.

[31] Guiso, L., P. Sapienza, and L. Zingales (2009), "Cultural Biases in Economic Exchange?" *Quarterly Journal of Economics*, 124, 1095-1131.

[32] Gul, F. and W. Pesendorfer (2011), "Media and Policy," mimeo.

[33] Horwitz, D.L. (1985), *Ethnic Groups in Conflict,* University of California Press, Berkeley and Los Angeles.

[34] Jaeger, D.A. and M.D. Paserman (2008), "The Cycle of Violence? An Empirical Analysis of Fatalities in the Palestinian-Israeli Conflict," *American Economic Review*, 98, 1591-1604.

[35] Jervis, R. (1976), *Perception and Misperception in International Politics*, Princeton: Princeton University Press.

[36] Kagan, D. (2004), *The Peloponnesian War*, Penguin, New York.

[37] Kaplan, E.H., A. Mintz, S. Mishal, and S. Claudio (2005), "What Happened to Suicide Bombings in Israel? Insights from a Terror Stock Model," *Studies in Conflict and Terrorism*, 28, 225-235.

[38] Kartik, N., M. Ottaviani, and F. Squintani (2007), "Credulity, Lies, and Costly Talk," *Journal of Economic Theory*, 134, 117-136.

[39] Kearney, R.N. (1967), *Communalism and Language in the Politics of Ceylon*, Duke University Press, Durham.

[40] Kramer, R.M. (1999), "Trust and Distrust in Organizations: Emerging Perspectives, Enduring Questions," *Annual Review of Psychology*, 50, phi 69-598.

[41] Kydd, A. (1997), "Game Theory and the Spiral Model," *World Politics*, 49, 371-400.

[42] Lagunoff, R. and A. Matsui (1997), "Asynchronous Choice in Repeated Coordination Games," *Econometrica*, 65, 1467-1477.

[43] Lagunoff, R. and A. Matsui (2004), "Organizations and Overlapping Generations Games: Memory, Communication, and Altruism," *Review of Economic Design*, 8, 383-411.

[44] Liu, Q. (2011), "Information Acquisition and Reputation Dynamics," 78, 1400-1425.

[45] Liu, Q. and A. Skrzypacz (2011), "Limited Records and Reputation," mimeo.

[46] Mailath, G.J. and L. Samuelson (2001), "Who Wants a Good Reputation?" *Review of Economic Studies*, 68, 415-441.

[47] Maskin, E. and J. Tirole (2004), "The Politician and the Judge: Accountability in Government," *American Economic Review*, 94, 1034-1054.

[48] McCarty, N., K.T. Poole, and H. Rosenthal (2008), *Polarized America: The Dance of Ideology and Unequal Riches*, Cambridge: MIT Press.

[49] Monte, D. (2010), "Bounded Memory and Permanent Reputations," mimeo.

[50] Morris, S. (2001), "Political Correctness," *Journal of Political Economy*, 109, 231-265.

[51] Pesendorfer, W. (1995), "Design Innovation and Fashion Cycles," *American Economic Review*, 85, 771-792.

[52] Phelan, C. (2006), "Public Trust and Government Betrayal," *Journal of Economic Theory*, 130, 27-43.

[53] Posen, B.R. (1993), "The Security Dilemma and Ethnic Conflict," *Survival*, 35, 27-47.

[54] Prendergast, C. (1993), "A Theory of 'Yes Men'," *American Economic Review*, 83, 757-770.

[55] Rabushka, A. and K. Shepsle (1972), *Politics in Plural Societies*, Columbus: Merrill.

[56] Rohner, D., M. Thoenig and F. Zilibotti (2011) "War Signals: A Theory of Trade, Trust and Conflict" CEPR Discussion Paper 8352.

[57] Schelling, T. (1960), *The Strategy of Conflict*, Cambridge: Harvard University Press.

[58] Sethi, R. and M. Yildiz (2011), "Public Disagreement," *American Economic Journal: Microeconomics*, forthcoming.

[59] Sobel, J. (1985), "A Theory of Credibility," *Review of Economic Studies*, 52, 557-573.

[60] Sunstein, C. (2006), *Going to Extremes: How Like Minds Unite and Divide*, Oxford: Oxford University Press.

[61] Thucydides (2000), *The History of the Peloponnesian War*, HardPress Publishing.

[62] Wiseman, T. (2008), "Reputation and Impermanent Types," *Games and Economic Behavior*, 62, 190-210.

[63] Yared, P. (2010), "A Dynamic Theory of War and Peace," *Journal of Economic Theory*, 145, 1921–1950.