

# Mistrust, Misperception, and Misunderstanding: Imperfect Information and Conflict Dynamics\*

Daron Acemoglu and Alexander Wolitzky

MIT

April 28, 2023

## Abstract

Building on theories of international relations, we analyze how *mistrust* (uncertainty about an adversary’s preferences or capabilities), *misperception* (imperfect observation of an adversary’s actions), and *misunderstanding* (non-degenerate higher-order beliefs) can lead to conflict and drive its dynamics. We develop our analysis in the context of three classic models: a one-shot *security dilemma* or *spiral model*; a repeated version of the security dilemma that allows for gradual learning about the opponent’s type, as well as the possibility of *conflict spirals*, *traps*, and *cycles*; and a *deterrence model*. We relate these models to the empirical literature and to current and historical episodes of conflict.

## 1 Introduction

A central question of international relations and political economy is what determines the onset and dynamics of conflict between actors such as sovereign nations, alliances, political parties, and ethnic or religious groups. Many conflicts are hugely costly for all parties involved; indeed, conflict within and between countries is often viewed as a major cause of

---

\*We thank Jim Fearon and Carlos Seiglie for helpful comments.

economic underdevelopment. This basic fact motivates a large body of research attempting to understand conflict as resulting from “imperfect information”—meaning that conflict arises and persists due to parties’ differing assessments of their own and their opponents’ preferences and capabilities, as well their opponents’ past actions and future intentions.

“Fear” has been viewed as a key factor explaining the onset and duration of conflict since antiquity. In the *History of the Peloponnesian War*, Thucydides (2003) identified the causes of war between Athens and Sparta as “three of the strongest motives: fear, honour, and interest.” Thucydides seems to have viewed fear as the most important of these motivations, writing that “the nature of the case first compelled us [Athens] to advance our empire to its present height; fear being the primary motive, though honour and interest afterwards came in,” and “The growth of the power of Athens, and the alarm which this inspired in Lacedaemon [Sparta], made war inevitable.” In Richard Crawley’s classic translation of Thucydides, the word “fear” appears 187 times, against 75 for “interest” or “interests” and 91 for “honour.” Fear and misperception have likewise been viewed as major causes of many historical conflicts—including World War I (Tuchman, 1962), the Korean War (Kydd, 2005), and the Iraq War (Debs and Montero, 2014; Coe and Vaynman, 2020)—as well as current geopolitical tensions, such as those surrounding Taiwan and the South China Sea (Kaplan, 2014).

For example, several prominent accounts of the onset and intensification of World War I emphasize mistrust, misperceptions, and misunderstandings. A well-known account by Edward Grey, the British Foreign Secretary at the time, argued that arms procurements by both sides before 1914 were interpreted as hostile actions, which led to an intensifying arms race, the formation of a web of military alliances, and aggressive international posturing (Grey, 1925). Many scholars believe that no European leaders truly desired a major war, but rather “sleepwalked” into a conflict that they either did not expect to occur, or expected to win easily (e.g., Clark, 2012). Further, once the war began, allegations of German atrocities early in the war contributed to its escalation, but this cause of escalation was not well-understood by the German side (Horne and Kramer, 2001).

Contemporary tensions between the United States and China over Taiwan illustrate similar dynamics. Both sides argue that relations have worsened because the status quo—wherein

the US “acknowledges” China’s claim to Taiwan but also tacitly guarantees Taiwan’s security if it is attacked—has come under stress. But each side blames the other for this development. For instance, in the 2022 Shangri-La Dialogue, US Defense Secretary Lloyd Austin maintained, “Our policy [over Taiwan] is unchanged and unwavering. It has been consistent across administrations,” and admonished China for changing the status quo, stating “unfortunately, that doesn’t seem to be true for the PRC.” In a starkly different interpretation, Chinese Defense Minister Wei Fenghe asserted, “It is not the mainland that is changing the status quo. It is Taiwan independence forces... and outside forces that are trying to change the status quo.” In this context, numerous high-profile actions have been understood completely differently by each side. For example, US House Speaker Nancy Pelosi’s visit to Taiwan was interpreted by most Americans as consistent with the status quo, while China viewed it as a major escalation; similarly, when Defense Minister Wei asserted in the Shangri-La Dialogue that China will “fight to the very end” in case of Taiwanese secession, this was interpreted by the American side as an escalation, while the Chinese maintain that this statement simply reaffirmed the status quo.

This chapter synthesizes and contributes to the literature on the role of imperfect information in driving conflict and its dynamics. We distinguish three types of imperfect information, which we can informally define as follows:

**Mistrust:** Uncertainty about an adversary’s preferences or capabilities.

**Misperception:** Imperfect observation of an adversary’s actions.

**Misunderstanding:** Uncertainty about an adversary’s past perceptions/observations or current level of trust.

For example, Sparta’s mistrust of Athens corresponds to fundamental uncertainty about whether Athens’ leaders truly wanted to dominate the Peloponnese, or about the true strength of the Athenian navy. Sparta could have misperceived a move by Athens if, for example, a low-level Athenian commander or an ally of Athens took an action that was more aggressive than what Athens’ leaders desired. And a misunderstanding could have arisen between Athens and Sparta if, similar to the Taiwan example above, Sparta perceived

a naval buildup by Athens as an aggressive move, but Athens could view the buildup as an affirmation of the status quo and believe that Sparta interpreted it similarly.

In terms of game theory, mistrust corresponds to *incomplete information* about an opponent's type, as in models of adverse selection or screening; misperception corresponds to *imperfect monitoring*, as in moral hazard or repeated-game models; and misunderstanding corresponds to *imperfect private monitoring*, where players have imperfect information about their opponent's observations of their own actions. Misunderstanding entails that the players' *higher-order beliefs*—their beliefs about their opponent's beliefs about their own actions and/or type—are non-degenerate. This feature plays an important role in many of the settings we consider.

We explore how these mistrust, misperception, and misunderstanding drive conflict in the context of three canonical models. We first cover the *one-shot security dilemma*—also called the *spiral model*—which is a classic model of how mistrust (or fear) and misperception can cause conflict between parties that would both prefer to avoid it.<sup>1</sup> The security dilemma is a coordination game where each party is afraid that her opponent is a “bad type” who always takes a “bad action” such as conflict. It thus incorporates mistrust, but not misperceptions or misunderstandings. If the level of mistrust is sufficiently high—meaning that each party assigns a sufficiently high probability to her adversary being a bad type—then both parties always take the bad action in the unique Bayesian Nash equilibrium of the game, even though both might rather coordinate on a peaceful “good action.”

If the security dilemma unfolds sequentially rather than simultaneously, then the possibility that a good action might be misperceived as bad can also contribute to conflict. The presence of a “first-mover” (or “first-strike”) advantage makes this misperception more consequential, as it makes the second-mover more worried about being caught off guard by an aggressive first move. While misperception—imperfect observation of actions—has long been a critical ingredient of moral hazard and repeated-game models in economics, it has received less attention in theories of conflict. The importance of misperceptions and their interaction with mistrust and misunderstanding is a key theme of the current chapter.

---

<sup>1</sup>The seminal reference on mistrust and misperception in international relations is Jervis (1976). Despite having “misperception” in its title, Jervis focuses more on what we call mistrust, as well as on psychological factors. Our terminology is closer to Kydd (1997, 2005).

We then analyze a *repeated security dilemma*, which additionally introduces misunderstanding (as the parties may not observe each other's past signals) as well as richer conflict dynamics. Here a key idea is that if (mis)perceptions are *private*—so that a player does not know when her opponent misperceives her action—they lead to divergent beliefs between the two players, and thus to misunderstanding. Misunderstanding has received even less attention in the literature on conflict, but we will see that it can profoundly influence the onset and dynamics of conflict. These dynamics include the possibility of *conflict spirals* (extended periods of conflict resulting from a single misperception), *traps* (permanent conflict spirals), and *cycles* (recurrent spirals interspersed with extended periods of peace).

We finally consider a *deterrence model*, where one party is tempted to initiate conflict, which the other party tries to deter through the threat of retaliation. In this model, misperception and misunderstanding drive a range of issues, including attribution problems in the presence of multiple adversaries, salami tactics and *faits accomplis*, and the potential utility of tripwires and other trigger strategies in dynamic settings.

Overall, we argue that the analysis of mistrust, misperception, and misunderstanding in the context of security dilemma and deterrence games can unify a large portion of theoretical literature on conflict in international relations and economics; provide new theoretical insights; inform the empirical literature; and contribute to our understanding of several historical and current episodes of conflict.

This chapter is organized as follows. We begin in Section 2 by briefly reviewing classic work on mistrust and misperception in international relations theory, and by relating this work to the more modern models we focus on. In Section 3, we lay out both simultaneous and sequential versions of the classic one-shot security dilemma, and analyze the roles of mistrust and misperception in this game. The core of the chapter is Section 4, which covers the repeated the security dilemma, where misunderstanding and rich conflict dynamics enter the analysis. Finally, Section 5 analyzes deterrence games, with applications to imperfect attribution of attacks and deterrence in long-run relationships. While this chapter is primarily theoretical, we draw connections to the empirical and historical literatures throughout. We pay particular attention to relating the parameters of the models we cover to empirical variables of interest.

## 2 Fear and Misperception in International Relations

Since Thucydides, many leading thinkers have explored how mutual mistrust can cause conflict or other inefficiencies. In *Leviathan* (1651), Thomas Hobbes argued that without external enforcement—as in typically the case in international relations—people would live in a state of “continuall feare and danger of violent death,” leading to war “of every man, against every man”. Hobbes echoed Thucydides’ motives for war, writing that “in the nature of man, we find three principal causes of quarrel. First, competition; secondly, diffidence; thirdly, glory. The first maketh men invade for gain; the second, for safety; and the third, for reputation.” A century later, in his *Discourse on Inequality* (1755), Jean-Jacques Rousseau considered the problem of two hunters who each decide whether to hunt stag or hare, where hunting stag is successful only if both hunters hunt stag, while each hunter can catch a less valuable hare on his own. Rousseau recognized that the inefficient outcome—hunting hare—may result, just as conflict can result in the security dilemma, and proceeded to provide philosophical arguments in favor of the outcome where both hunters hunt stag.<sup>2</sup> Following these classical thinkers, the general position that fear or mistrust is a primary cause of conflict—even between parties that would prefer to avoid it—is alternately known as *Thucydides’ trap*, the *Hobbesian trap*, the *security dilemma*, or the *spiral model*.<sup>3</sup> The core model in this chapter is a simple, game-theoretic version of the security dilemma.

The modern theory of international conflict, especially its branch termed “structural realism,” builds on related ideas. Works such as Kenneth Waltz’s (1979) *Theory of International Politics* and John Mearsheimer’s (2001) *The Tragedy of Great Power Politics* emphasize how conflict can emerge from states’ efforts to gain territorial or other advantages over their adversaries, their attempts to preempt others’ attacks, or even their own defensive actions. Waltz, for example, writes, “In an anarchic domain, a state of war exists if all parties lust for power. But so too will a state of war exist if all states seek only to ensure their own safety” (1979, p. 44). Waltz’s view that states are usually “security-seeking” rather than

---

<sup>2</sup>David Hume’s *A Treatise of Human Nature* (1739) also contains similar examples.

<sup>3</sup>As well as *Schelling’s dilemma*, after Schelling’s parable of an armed homeowner encountering a burglar in his essay, “The Reciprocal Fear of Surprise Attack” (in Schelling, 1960). There is some inconsistency in terminology in the literature, with some authors using “security dilemma” for prisoner’s dilemma-type games rather than coordination games. We stick to the latter usage.

outright “expansionist” makes him a “defensive realist,” in contrast to “offensive realists” like Mearsheimer who emphasize expansionist and competitive motives as well as fear. However, our reading is that the offensive and defensive realists’ views of the security dilemma have much in common. Indeed, for Mearsheimer, the “tragedy” of great power politics is precisely that the logic of strategic competition can force conflict even on states that would prefer to remain at peace. This type conflict is primarily attributed to mistrust: Mearsheimer’s key “bedrock assumptions” include, “states can never be certain about other states’ intentions,” “survival is the primary goal of great powers,” and “great powers are rational actors,” and these assumptions (supposedly) lead to “three general patterns of behavior[...]: fear, self-help, and power maximization” (2001, pp.100–101). Similar factors are also central to many ethnic conflicts, as emphasized for example by Donald Horowitz, who in his seminal study of ethnic conflicts in Africa writes, “The fear of ethnic domination and suppression is a motivating force for the acquisition of power as an end,” (Horowitz, 2000, p. 187).

Mistrust and misperception have been studied more systematically in the literature on the security dilemma, pioneered by Herz (1950) and Butterfield (1951). The seminal contribution in this area is Jervis (1976, 1978), who provided a detailed discussion of how the balance between the offensive and defensive capabilities of rival states, as well as their uncertainty and (mis)perceptions regarding the other side’s intentions, determine the likelihood of conflict. For Jervis, misperceptions were related to psychological factors, errors of judgment, and other systematic mistakes. He also argued that misperception (in our terminology, mistrust) could lead to conflict “spirals,” where each side escalates in response to the other. The run-up to World War I discussed above is a classic example of such a spiral.<sup>4</sup>

The game-theoretic approach to mistrust was developed by Glaser (1992, 2010) and, especially, Kydd (1997, 2005), who provided a formal analysis of the spiral model, which we reformulate and extend in Section 3.<sup>5</sup> Our view of mistrust is the same as Kydd’s: in a

---

<sup>4</sup>See also Glaser (2010, Chapter 9) for a discussion of this example, as well as other examples of arms races interpreted through the lens of the security dilemma, in the context of the lead-ups to World War I and II and the Cold War.

<sup>5</sup>Other models of conflict based on uncertainty and asymmetric information are discussed by, e.g., Fearon (1995), Powell (1987, 1999), and Ramsay (2017). Early contributions to asymmetric information models of conflict include Brito and Intriligator (1985), Powell (1987), Morrow (1989), and Wagner (1994). Other explanations for conflict include excessive optimism on both sides of a potential conflict (e.g., Landes, 1971, Yildiz, 2004, Fey and Ramsay, 2007, and Slantchev and Tarar, 2011); dynamic commitment problems (e.g.,

coordination-type game, a state that suspects that its rival is a type that always takes a bad or aggressive action will respond by taking this action itself, even if each side would prefer a peaceful outcome with a very high probability. However, by additionally considering conflict dynamics, misperception, and misunderstanding, this chapter provides several new insights and further refines some of Glaser and Kydd’s results. To give one important example, Kydd shows that in his one-shot model, “tragic spirals between security seekers [normal, rational states] are likely to be a small proportion of observed conflicts, especially as information improves” (p. 75). However, we will see that this is not necessarily true in a dynamic model, where “tragic conflict” can occur with substantial probability even as the probability of misperception vanishes.

Finally, we also mention the literature on “cooperation theory” or “neo-liberal institutionalism” (e.g., Axelrod, 1984; Keohane, 1984; Oye, 1986). This literature conceptualizes conflict as defection in a repeated prisoner’s dilemma, rather than coordination on an inefficient equilibrium in a coordination game. A basic lesson of the repeated prisoner’s dilemma is that the precision of monitoring (along with the discount factor) determines the viability of equilibrium cooperation, so that “misperception” in our sense of imperfect monitoring plays a key role in this setting (e.g., Fearon, 1998). This literature has focused on the case where signals are public, so that, while misperceptions can arise, misunderstanding—which require uncertainty about an adversary’s past observations—cannot. Neglecting misunderstanding is a major omission: as we will see, the distinction between public and private signals has critical implications in both repeated security dilemma games and repeated deterrence games.

## **3 Misperceptions and Mistrust in the Security Dilemma**

### **3.1 The Security Dilemma**

This section lays out a simple version of the classic security dilemma or spiral model, which is a one-shot coordination game with incomplete information. We consider both simultaneous and sequential versions of the game. Both versions capture the importance of mistrust

---

Fearon, 1995, Acemoglu and Robinson, 2000, and Powell, 2004, 2006), and domestic audience concerns (Fearon 1994, Debs and Weiss, 2016, and Fergusson, Robinson, Torvik and Vargas, 2016).



(uncertainty about the opponent’s preferences), while the latter version also introduces misperception (imperfect observation of the first-mover’s action). These one-shot games are the building blocks of the models of conflict dynamics that we introduce in Section 4.

The basic security dilemma has two players, each with two possible actions, which we call *dovish* ( $D$ ) and *hawkish* ( $H$ ), and two possible types, which we call *normal* and *bad*.<sup>6</sup> The payoff of a normal player, as a function of her own action and her opponent’s action, is given by the matrix

	opponent’s action		
	$D$	$H$	
own action	$D$	$1 - l$	
	$H$	$g$	$0$

where  $l > 0$  and  $g < 1$ .<sup>7</sup> These assumptions ensure that for two normal players, the game is a coordination game (also called a *stag hunt* or *assurance game*). It has a Pareto-dominant equilibrium ( $D, D$ ) or *peace*, and a Pareto-dominated equilibrium ( $H, H$ ) or *conflict*. In contrast, a bad player is assumed to always take  $H$ , which is essentially equivalent to assuming that  $H$  is a strictly dominant action for a bad player, so that bad types have the same preferences as in a prisoner’s dilemma.<sup>8</sup> The prior probability that each player is bad is denoted by  $\mu_0$ , and the players’ types are assumed to be independent. The parameter  $\mu_0$  thus measures each player’s *fear* of facing a bad opponent. This parameter is our measure of *mistrust*.

In the international relations literature (e.g., Kydd, 2015; Fearon, 2020), it is standard to call the parameters  $g$  and  $l$  the *first-strike advantage* and the *second-strike disadvantage*, respectively, as these parameters measure the gain from taking  $H$  “first” and the loss when the opponent takes  $H$  “first” (relative to the payoff of 0). We follow this terminology, even when we consider the version of the game where the players move simultaneously.

The above general setup, where (i) each player can be normal or bad, (ii) normal types have coordination game preferences, (iii) bad types have prisoner’s dilemma preferences,

---

<sup>6</sup>See Fearon (2011) for a constructive critique of this type of simplifying assumptions.

<sup>7</sup>The game is assumed to be symmetric for simplicity.

<sup>8</sup>A small difference is that we do not allow bad players to take  $D$  even off the equilibrium path. This assumption simplifies our analysis, because it implies that a player who sees her opponent take  $D$  can be sure that he is normal.

and (iv) a player’s trust in her opponent is measured by her belief that he is normal, is canonical. In broad outline, the security dilemma dates to Jervis (1976, 1978), and it was formalized by Kydd (1997; 2005, Chapter 2) and Baliga and Sjöström (2004). This setup can be extended in several directions. One important direction is generalizing the assumption of independent binary types to investigate issues such as correlated types, higher-order beliefs, and communication. Papers on these topics include Baliga and Sjöström (2004, 2008, 2012), Chassang and Padro i Miquel (2010), and Acharya and Ramsay (2013). We do not cover this branch of the literature here, as it is addressed in the chapter by Baliga and Sjöström. Another direction is introducing dynamics, along with the concomitant issues of learning about the opponent’s type, as well the possibility of misperception and misunderstanding. This is the topic of the current chapter.

As discussed previously, we analyze two versions of the basic security dilemma. In the *simultaneous security dilemma*, the players take actions simultaneously. This is the canonical version of the security dilemma, which captures mistrust-induced conflict.

In the *sequential security dilemma*, player 1 takes her action  $a_1 \in \{D, H\}$  first, and then player 2 observes a *signal*  $s$  of player 1’s action before taking his own action  $a_2 \in \{D, H\}$ . For simplicity, we assume throughout that  $s \in \{D, H\}$  with probability distribution

		player 2’s signal	
		$D$	$H$
player 1’s action	$D$	$1 - \pi$	$\pi$
	$H$	$0$	$1$

Thus, if player 1 plays  $D$ , this action is *misperceived* as  $H$  with probability  $\pi$ ; while for simplicity if player 1 plays  $H$ , this action is always perceived correctly. We assume (again, largely for simplicity) that the signal is payoff-irrelevant, and that payoffs depend on the underlying chosen actions even if these actions are misperceived. We also note that, since player 1 does not take any further actions following player 2’s observation of the signal  $s$ , whether or not player 1 observes  $s$  as well—that is, whether the signal  $s$  is *public* or *private*—is irrelevant in this sequential case as well. In contrast, this feature makes a big difference in the dynamic models considered in Sections 4 and 5. In sum, whereas the

simultaneous security dilemma represents conflict induced by mistrust alone, the sequential security dilemma captures conflict induced by the interaction of mistrust and misperception.<sup>9</sup>

### 3.2 Empirical Determinants of Mistrust, Misperception, and Offense-Defense Balance

The basic security dilemma model has three parameters: the level of mistrust  $\mu_0$ , the first-strike advantage  $g$ , and the second-strike disadvantage  $l$ . The sequential security dilemma adds to this the misperception probability  $\pi$ . We now discuss what these parameters correspond to in reality and how they are determined.

The level of mistrust  $\mu_0$  is the probability that one party believes that her opponent is “truly bad,” in that he will take aggressive actions regardless of her own behavior. In principle, such negative beliefs about another party can have many sources, but one that is often crucial in practice is a history of conflict, violence, or ideological or ethnic division between two groups. The path from past conflict to current mistrust and conflict is the subject of several prominent empirical studies. For example, Besley and Reynal-Querol (2014) find that exposure to conflict in Africa in the precolonial period (1400 to 1700) predicts greater postcolonial conflict, as well as lower levels of generalized trust and increased levels of ethnic identity, as measured by Afrobarometer. Similarly, Rohner, Thoenig, and Zilibotti (2013) find that exposure to ethnic conflict in Uganda in the early 2000s decreased generalized trust and increased ethnic identity. It also appears that other sources of historical traumas can have similar effects. Key studies here include Michalopoulos and Papaioannou (2016), who show that ethnic groups that were split during the “Scramble for Africa” in the 19th century experienced more political violence in the early 21st century, including more ethnic conflict; and Nunn and Wantchekon (2011), who show that individuals who belong to ethnic groups that were heavily raided during the slave trade exhibit lower levels of generalized trust today.

---

<sup>9</sup>Kydd (1997; 2005, Chapter 3) considers a related model where the players receive exogenous signals of the opponent’s type before taking actions simultaneously. “Incorrect” signals in Kydd’s model are somewhat akin to misperceptions in our sequential model. Nevertheless, there are important differences between the models. For example, our sequential model typically has a unique sequential equilibrium, and it is easy to generalize it to richer models of conflict dynamics, as we do in Section 4.

The empirical content of the misperception probability  $\pi$ —the probability that, when one party takes an action that is intended to be peaceful or conciliatory, or at least consistent with the status quo, the other party perceives this action as aggressive—is less well-understood. We can speculate that misperceptions may be more likely between individuals or groups with less shared history or culture; when military actions are more difficult to observe (e.g., in the covert or cyber domains); or when media or other aspects of the prevailing information space are less open and free.<sup>10</sup> In Section 4.4, we will give some concrete examples of misperceptions in the context of ethnic conflicts in Colombia and Northern Ireland.

Finally, there is a large literature on the meaning and significance of the first-strike advantage  $g$  and the second-strike disadvantage  $l$ . This is the literature on “offense-defense” balance in international relations (e.g., Schelling, 1966; Jervis, 1978; Van Evera, 1998; see also Kydd, 2005, and Chassang and Padro i Miquel, 2010). This literature focuses on how changes in military technology, as well as other factors like geography or military or civilian tactics, make either aggressive actions or defensive and retaliatory actions more likely to succeed.

### 3.3 Analyzing the Simultaneous Security Dilemma

We now turn to the basic analysis of the security dilemma.

In the simultaneous security dilemma, it is always a (Bayesian Nash) equilibrium for normal types to take  $H$ . This follows because normal types have coordination game preferences (as  $l > 0$  and  $g < 1$ ), so if the normal opponent type as well as the bad opponent type take  $H$ , a normal player faces  $H$  for sure, and thus takes  $H$  in response. We call such an equilibrium *conflictual* or *inefficient*.

The key question is whether it is also an equilibrium for normal types to take  $D$ , in which case we say the equilibrium is *peaceful* or *efficient*.<sup>11</sup> A peaceful equilibrium exists if and only if it is optimal for a normal player to take  $D$  when the normal opponents take  $D$ . This

---

<sup>10</sup>However, greater media freedom can also lead to the spread of conspiracy theories and misinformation, obfuscating important elements of past events and in the extreme case, creating less agreement on history.

<sup>11</sup>When a peaceful equilibrium exists, there is also a mixed equilibrium, which we ignore.

is the case if and only if

$$\underbrace{(1 - \mu_0)(1) + \mu_0(-l)}_{\text{expected payoff from } D} \geq \underbrace{(1 - \mu_0)(g) + \mu_0(0)}_{\text{expected payoff from } H} \iff \mu_0 \leq \mu^{\text{sim}} := \left(1 + \frac{l}{1 - g}\right)^{-1}.$$

So, a peaceful equilibrium exists if and only if there is not too much mistrust—specifically, if and only if  $\mu_0$  is below a cutoff  $\mu^{\text{sim}}$  (for “simultaneous”). This is the basic insight of the security dilemma.

We summarize this discussion with a proposition. (In all of our propositions, we describe only the play of the normal type of each player, recalling that the bad type of each player always takes  $H$ .)

**Proposition 1** *In the simultaneous security dilemma,*

1. *If  $\mu_0 \leq \mu^{\text{sim}}$ , there are multiple equilibria, including a peaceful equilibrium where both players take  $D$ .*
2. *If  $\mu_0 > \mu^{\text{sim}}$ , there is a unique equilibrium, which is conflictual: both players take  $H$ .*

The cutoff level of mistrust  $\mu^{\text{sim}}$  displays intuitive comparative statics. Observe that  $\mu^{\text{sim}}$  is decreasing in the first-strike advantage  $g$ , because increasing  $g$  raises the expected payoff from playing  $H$ , which makes playing  $H$  more attractive; similarly,  $\mu^{\text{sim}}$  is decreasing in the second-strike disadvantage  $l$ , because increasing  $l$  decreases the expected payoff from  $D$ . Note also that our normalization of the payoffs from  $(D, D)$  and  $(H, H)$  implies that  $l$  and  $g$  are decreasing in the payoff difference between  $(D, D)$  and  $(H, H)$ . Altogether, we see that a conflictual equilibrium always exists, while a peaceful equilibrium exists if mistrust is low, if peace is much more efficient than conflict, and if the first-strike advantage and the second-strike disadvantage are both small. These are all standard results.<sup>12</sup>

### 3.4 Analyzing the Sequential Security Dilemma

The logic of the sequential security dilemma is slightly more complicated, as there are now two equilibrium conditions to consider.

---

<sup>12</sup>See, for example, Kydd (2005, Chapter 2).

First, if player 1 (the first-mover) anticipates that the normal type of player 2 will match his action to his perception of player 1's action (i.e., will take  $a_2 = D$  when  $s = D$ , and will take  $a_2 = H$  when  $s = H$ ), then player 1 will take  $D$  if and only if

$$\underbrace{(1 - \pi) ((1 - \mu_0)(1) + \mu_0(-l)) + \pi(-l)}_{\text{expected payoff from } D} \geq \underbrace{0}_{\text{payoff from } H} \iff \mu_0 \leq \mu^{\text{seq}} := \left(1 + \frac{l}{1 - \pi(1 + l)}\right)^{-1}.$$

(Here  $\mu^{\text{seq}}$  stands for “sequential.”)

Second, by Bayes' rule, if player 2 believes that the normal type of player 1 always takes  $D$  (while, as always, the bad type of player 1 always takes  $H$ ), his posterior belief that player 1 is bad when he observes signal  $s = H$  is given by

$$\mu_1 := \left(1 + \frac{1 - \mu_0}{\mu_0} \pi\right)^{-1}.$$

Note that  $\mu_1 > \mu_0$ : that is, observing signal  $s = H$  makes player 2 trust player 1 less. Now, just as in our analysis of the simultaneous security dilemma, it is optimal for player 2 to take  $D$  if and only if he believes that player 1 took  $H$  with probability less than  $\mu^{\text{sim}}$ . We thus see that, provided that the normal type of player 1 takes  $D$  in equilibrium, player 2 takes  $D$  when he observes signal  $s = H$  iff  $\mu_1 \leq \mu^{\text{sim}}$ . We also note that player 2 always takes  $D$  when he observes signal  $s = D$ , because this signal can arise only if player 1 took  $D$ .<sup>13</sup> Finally, since  $\mu_0 < \mu_1$ , we see that if  $\mu_1 \leq \mu^{\text{sim}}$ , then it is an equilibrium for both players to always take  $D$  (regardless of the signal).

Putting these observations together, we can now describe the (sequential) equilibria of the sequential security dilemma. We say that an equilibrium is *peaceful* if player 1 takes  $D$  (when normal) and player 2 takes  $D$  after either signal; an equilibrium is *conflictual* if player 1 takes  $H$  and player 2 takes  $H$  after signal  $H$  (which, recall, arises with probability 1 when player 1 takes  $H$ ); and an equilibrium is *escalatory* if player 1 takes  $D$  and player 2 matches his action to his perception of player 1's action, so that  $a_2 = s$  for  $s \in \{D, H\}$ . We say that both peaceful and escalatory equilibria are *non-conflictual*, since in both of these types of equilibria player 1 takes  $D$  and player 2 responds with  $D$  if he correctly perceives player 1's

<sup>13</sup>This fact relies on our assumption that action  $H$  is never misperceived.

action. However, crucially, in an escalatory equilibrium player 2 takes  $H$  if he misperceives player 1's action, so a non-conflictual equilibrium may not be entirely peaceful.

**Proposition 2** *In the sequential security dilemma,*

1. *If  $\mu_0 < \mu^{\text{seq}}$ , there is a (generically) unique sequential equilibrium, which is non-conflictual. If  $\mu_1 < \mu^{\text{sim}}$ , the equilibrium is peaceful: player 1 takes  $D$ , and player 2 takes  $D$  after either signal. If  $\mu_1 > \mu^{\text{sim}}$ , the equilibrium is escalatory: player 1 takes  $D$ , and player 2 takes  $a_2 = s$  for  $s \in \{D, H\}$ . If  $\mu_1 = \mu^{\text{sim}}$ , both peaceful and escalatory equilibria exist.*
2. *If  $\mu_0 \geq \mu^{\text{seq}}$  and  $\mu_1 \leq \mu^{\text{sim}}$ , there are multiple sequential equilibria, including a peaceful equilibrium and a conflictual equilibrium.*
3. *If  $\mu_0 > \mu^{\text{seq}}$  and  $\mu_1 > \mu^{\text{sim}}$ , there is a unique sequential equilibrium, which is conflictual.*

Several points are worth noting here. First, when  $\mu_0 < \mu^{\text{seq}}$ , not only does there exist a non-conflictual equilibrium, but this is the unique sequential equilibrium. In particular, there is no a conflictual equilibrium.<sup>14</sup> Uniqueness of a non-conflictual equilibrium in the sequential security dilemma with low initial mistrust contrasts with the situation in the simultaneously security dilemma, where a conflictual equilibrium always exists.<sup>15</sup>

Second, note that we have  $\mu^{\text{seq}} > \mu^{\text{sim}}$ —which, by Propositions 1 and 2, implies that it is easier to sustain a non-conflictual equilibrium in the sequential game than in the simultaneous game—if and only if

$$\pi < \frac{g}{1+l}. \tag{1}$$

Thus, if misperceptions are sufficiently unlikely and  $g > 0$ , it is easier to sustain a non-conflictual equilibrium in the sequential game. The intuition is that if  $\pi$  is small and  $g > 0$  then when player 2 switches from always taking  $D$  (his strategy in a simultaneous-move

---

<sup>14</sup>To see this, suppose that there is a conflictual equilibrium. But player 1 deviate to  $D$ , and with high probability this action will be correctly perceived, in which case player 2 infers that player 1 took  $D$  and will respond by taking  $D$ . This implies that the deviation to  $D$  is a profitable deviation, implying that there cannot be a conflictual equilibrium.

<sup>15</sup>This effect is similar to that observed by Lagunoff and Matsui (1997), who show that asynchronous moves force coordination on the good equilibrium in repeated coordination games.

peaceful equilibrium) to taking the action equal to his signal (his strategy in a sequential-move escalatory equilibrium), this has a small effect on player 1’s expected payoff from  $D$  but significantly reduces player 1’s expected payoff from  $H$ ; this makes  $D$  more attractive for player 1. Thus, far from a first-mover advantage sparking conflict, when misperceptions are unlikely sequential moves favor coordination on the efficient equilibrium. It is only when the fear of one’s action being misperceived is sufficiently large that sequential moves undermine the efficient equilibrium.

Third, when condition (1) holds, so that  $\mu^{\text{seq}} > \mu^{\text{sim}}$ , the second case considered in Proposition 2 cannot arise (recalling that  $\mu_0 < \mu_1$ ), so there is a (generically) unique sequential equilibrium, which is non-conflictual if  $\mu_0 < \mu^{\text{seq}}$  and conflictual if  $\mu_0 > \mu^{\text{seq}}$ . This equilibrium uniqueness property makes the sequential security dilemma a convenient building block for modeling conflict dynamics, which we turn to in the next section.

Before doing so, we close the current section by considering the comparative statics for  $\mu^{\text{seq}}$ , which as we have seen determines whether the unique equilibrium is conflictual or non-conflictual (when condition (1) holds). These comparative statics are intuitive, although they are a bit different from those in the simultaneous game. First, observe that  $\mu^{\text{seq}}$  is decreasing in both  $l$  and  $\pi$ , because increasing either  $l$  or  $\pi$  decreases the expected payoff from  $D$ . However,  $\mu^{\text{seq}}$  does not depend on  $g$ , because if player 1 takes  $H$  in the sequential game this action is always met by  $H$ . Thus, a non-conflictual equilibrium exists if mistrust is low; if misperceptions are unlikely; if peace is much more efficient than conflict; and if the second-strike disadvantage is small. In particular, a key lesson is that greater mistrust and more frequent misperceptions both threaten the existence of a non-conflictual equilibrium in the sequential security dilemma.<sup>16</sup>

---

<sup>16</sup>However, since  $\mu_1$  is decreasing in  $\pi$ , more frequent misperceptions make a non-conflictual equilibrium more likely to be peaceful rather than escalatory. In addition, if (1) fails, more frequent misperceptions can shift the game from Case 3 of Proposition 2 to Case 2, which permits a peaceful equilibrium. Intuitively, when misperceptions are frequent and normal player 1 takes  $D$ , player 2 is inclined to take  $D$  even after an  $H$  signal. For this reason, for some parameters (and possibly for some equilibrium selection) the probability that player 2 takes  $H$  is maximized at an intermediate value for  $\pi$ . This “so-so signals are worst” effect is also emphasized by Fearon (2020).



## 4 Misperceptions and Conflict Dynamics

We now move beyond the simple one-shot games of Section 3 to study the impact of misperceptions on conflict dynamics. We will find that these dynamics can be quite rich—encompassing eventual successful learning and coordination on peace, permanent *conflict traps* where learning ceases forever, and recurring *conflict cycles*. These dynamics nevertheless depend on the form of uncertainty and information facing the players in an intuitive manner. As compared to the classical literature on the security dilemma, the models we explore thus highlight a wider range of possible outcomes of international (or ethnic or political) conflict, as well as locating the causes of these dynamics in misperception and misunderstanding.

Formally, we consider a repeated version of the sequential security dilemma, where the identity of the first-mover alternates between periods. That is, the first-mover will be player 1 in odd periods and player 2 in even periods. This first mover takes an action  $a_t^1 \in \{D, H\}$ , and then the second-mover observes a signal  $s_t \in \{D, H\}$  of the first-mover's action before taking his own action  $a_t^2 \in \{D, H\}$ . To rule out dynamic incentives stemming from strategic experimentation, we assume that both players are infinitely impatient, so that each player maximizes her short-run payoff in each play of the sequential game.<sup>17</sup> However, the plays of the game are still tied together because each player's type is assumed to be perfectly persistent across periods. We also assume that misperceptions occur independently across periods.

We consider three variants of this repeated security dilemma game, and we now discuss these variance and the resulting dynamics informally. The rest of the section is devoted to a more formal analysis.

In the first variant, the players' perceptions are common knowledge, meaning that the signal  $s_t \in \{D, H\}$  is publicly observed by both players (in particular, by the first-mover herself as well as the second-mover). For example, if the first-mover plays  $D$  but her opponent misperceives this action as  $H$ , the first-mover knows that her action was misperceived. We

---

<sup>17</sup>Because of this assumption, our results will not depend on whether or not the second-mover's action is observed. For simplicity, we assume that that the second-mover's action is never observed. Note that this assumption implies that the players' payoffs are not measurable with respect to their information; however, we again emphasize that this assumption is inessential.

show that in this version of the game, when both players are normal they eventually manage to learn each other's types and successfully coordinate on a peaceful equilibrium, where they play  $(D, D)$  in every period. The key intuition is that when misperceptions are public, the scope for misunderstanding is limited. Repeating the security dilemma then gives each player as many chances as she needs to convince her opponent that she is normal, which eventually leads to learning and coordination on a peaceful equilibrium. This result is simple but it seems to be novel; we provide a proof in the Appendix to this chapter.

We next consider the case where the players' perceptions are private, so that the signal  $s_t$  is observed only by the second-mover in period- $t$ . In this case, a player does not know whether her own action is perceived correctly or not. This feature introduces a wider scope for misunderstanding. Now, if the period- $t$  first-mover takes  $D$  but perceives her opponent's first move in period  $t + 1$  to be  $H$ , she does not know if her own period- $t$  action was itself misperceived. If she had recognized that there was such a misperception, she could "excuse" her opponent's play of  $H$  in period  $t + 1$  and try to initiate a more peaceful outcome. In this setting, however, she cannot recognize such misperceptions, paving the way to *misunderstanding* between the two players. In these circumstances, we show that whenever misperceptions are sufficiently rare, normal players become trapped in permanent conflict with positive probability. Namely, with positive probability,  $(H, H)$  starts being played in every period. Intuitively, misperceptions at the beginning of the players' relationship can trap them in a state of conflict, where neither player trusts the other enough to try to initiate a switch to the peaceful equilibrium. As we explain, this result can be viewed as a variant of the "war trap" model of Rohner, Thoenig, and Zilibotti (2013).

Finally, we consider a variant of the model with limited memory. For simplicity, following Acemoglu and Wolitzky (2014) we focus on the extreme case where the players have "one-period memory," so that they are aware only of the action that the opponent took the last time the opponent was the first mover. As we elaborate below, a natural interpretation of this version of the model is that each player represents an infinite dynasty and the two dynasties/groups interact across overlapping generations. We show that this model leads to *conflict cycles*: there exists an integer  $T$  such that the groups fall into persistent conflict whenever a dovish action is misperceived, but revert to the peaceful equilibrium every  $T$

periods regardless of their history of actions. Intuitively, with one-period memory a player knows the current state of the groups' relationship but not how the relationship got into the current state. Consequently, once the groups have had many chances to fall into accidental conflict, the players realize that being in conflict is not very informative about the other group's type, and they rationally attempt to switch to the peaceful equilibrium.

The private misperceptions—and the resulting misunderstanding—that feature in the second and third versions of the model capture a salient aspect of many dynamic conflicts. It is a typical feature of conflict that the two sides disagree about what action initiated the conflict. As the historian of the Irish Republican Army Richard English puts it, “We all have different narratives of what happened. So, I’ll start it when you plant the bomb. You’ll start it when I invaded your country, which is why you planted the bomb. We all have different starting points for evil. Everyone claims to be reacting appropriately to someone else’s violence,” (quoted in Blattman, 2022, p. 152). A similar mechanism seems to be involved in the US-China conflict over Taiwan described in the Introduction. Related forces can also be seen in non-violent conflicts: for example, everyone agrees that polarization and distrust between Democrats and Republicans in the United States are at their highest points in a century or more, but the parties vehemently disagree about the sources of this polarization, each blaming it on the extremism of the other side. The models considered in this section are simple tool for analyzing this kind of effect.

Before beginning our analysis, we note that Kydd (2000; 2005, Chapter 7) considers a twice-repeated version of the simultaneous security dilemma, which he calls the *reassurance game*. Unlike the sequential-move games with impatient players we are considering here, Kydd’s reassurance game has a range of equilibria due to signaling considerations. However, Kydd nonetheless derives several interesting implications, for example providing conditions for the existence of a separating equilibrium in which normal players successfully coordinate on  $(D, D)$  in the second period, even when the level of mistrust is initially high. We refer the reader to his work for the specifics of these results.

## 4.1 Best Responses as a Function of Beliefs

To simplify the analysis, we assume throughout this section that condition (1) holds, so that (as we have seen) the one-shot sequential security dilemma has a unique sequential equilibrium. This assumption lets us make use of the following observation, which follows directly from Proposition 2 above.

**Lemma 1** *Consider any of the three variants of the repeated sequential security dilemma described above, and assume that condition (1) holds. In any sequential equilibrium, if the first-mover believes that the opponent is bad with probability less than  $\mu^{\text{seq}}$  in any period, then she takes  $H$ ; and if she believes that the opponent is bad with probability greater than  $\mu^{\text{seq}}$ , she takes  $H$ .*

*In addition, if the second-mover believes that the opponent took  $H$  with probability less than  $\mu^{\text{sim}}$ , she takes  $D$ ; and if she believes that the opponent took  $H$  with probability greater than  $\mu^{\text{sim}}$ , she takes  $H$  (where these beliefs are conditional on the history of actions and signals including the current-period signal).*

We will also impose genericity assumptions that imply that the first-mover's equilibrium belief is never exactly equal to  $\mu^{\text{seq}}$ , so that Lemma 1 fully determines the first-mover's action. Finally, we also assume throughout this section that  $\mu_0 < \mu^{\text{seq}}$ , so that the first-mover in the first period of the game takes  $D$ . If instead  $\mu_0 > \mu^{\text{seq}}$ , we are in the trivial situation, where in the unique equilibrium of all the model variants, all players take  $H$  at every history.

## 4.2 Public Misperceptions Lead to Eventual Cooperation

We first assume that the history of period- $t$  signals  $s_t$  is observed by both parties. For a given equilibrium, we let  $p_t$  denote the *ex ante* probability that the peaceful outcome ( $D, D$ ) prevails in period  $t$ , conditional on the event that both players are normal. We also define the key variable

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

This is the posterior probability that the opponent is bad when the history contains a total of  $n$  signals of  $H$  from the opponent's action (and zero  $D$  signals), under the hypothesis that

the opponent always plays  $D$  with probability one when he is normal. Note that  $\mu_0$  is the prior, and  $\mu_1$  matches the notation in Section 3.4. Moreover,  $\mu_n$  is increasing in  $n$  (as  $H$  signals are bad news about the opponent's type), and satisfies  $\lim_{n \rightarrow \infty} \mu_n = 1$ . We impose the genericity assumption that  $\mu_n \neq \mu^{\text{seq}}$  for all  $n$ . We also denote the smallest integer  $n$  such that  $\mu_n > \mu^{\text{seq}}$  by  $n^*$ . The main result of this subsection is then as follows.

**Proposition 3** *The repeated sequential security dilemma with public misperceptions has a unique sequential equilibrium. In the equilibrium,  $p_t \rightarrow 1$ : that is, normal players eventually coordinate on the peaceful outcome with probability 1.*

We defer the proof to the Appendix. To see the intuition, suppose the players are in fact both normal, but they are unlucky and get  $H$  signals for many consecutive periods at the start of their relationship, despite both playing  $D$ . In this event, at some point (specifically, in period  $T = 2n^*$ ) player 2's belief that player 1 is normal drops below  $\mu^{\text{seq}}$ , at which point he switches to taking  $H$ . However, because the signals are public, player 1 perfectly understands that player 2 will switch to taking  $H$  at this point even when he is normal, and therefore player 1 does not become more pessimistic herself, even as she keeps observing  $H$  signals after this point in the game. Instead, she understands that it is reasonably likely that player 2 is normal and is taking  $H$  because he is pessimistic about her own type, so she continues to play  $D$  until she sees that player 2 perceives her action correctly. At this point, if player 2 is in fact normal, he learns that player 1 is also normal and switches to  $D$ , thus ensuring eventual convergence to the peaceful equilibrium.

The key lesson of this model is that misperceptions need not cause conflict if they are commonly understood and recognized by both parties. This observation refines the classical accounts of the role of misperceptions discussed above: conflict is caused not merely by noisy observations per se, but by misperceptions that lead the players' interpretations of the history of their relationship to diverge. Jervis (1976) himself recognized that such occurrences can arise from psychological biases or irrationally misaligned beliefs, though the role of misunderstandings—as opposed to misperceptions—is not discussed in his work or other works that we are aware of in this literature. In the next subsection, we will see that these misunderstandings can also arise when players are rational but their signals are private

information, so that a player cannot easily recognize her opponent’s misperceptions.

### 4.3 Private Misperceptions Lead to Conflict Traps

We now consider the implications of private misperceptions. The previous subsection showed that public misperceptions do not generate persistent conflict. However, it is usually unrealistic to assume that a player can observe the other side’s misperceptions, especially when misperceptions depend on subjective interpretations or context-specific observations, which are probably the most relevant types of misperceptions in international or ethnic conflicts. For example, Country A is unlikely to accurately perceive whether Country B perceives Country A’s military spending to be at a normal level or a provocatively aggressive one, or whether Country B attributes a border skirmish to a mistake by a hotheaded lieutenant or to a strategic shift by Country A.

Formally, we now assume that signals are observed only by the second-mover. Let us define

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

This is the posterior probability that the opponent is bad given that *at least one* out of  $n$  signals of the opponent’s action takes value  $H$ , under the hypothesis that the opponent always plays  $D$  when he is normal. Note that  $\rho_1 = \mu_1$ ;  $\rho_n$  is decreasing in  $n$  (for  $n \geq 1$ ); and  $\lim_{n \rightarrow \infty} \rho_n = \mu_0$ . (Intuitively, the more signals are observed, the “less bad” is the news that at least one of them was  $H$ .) We impose the genericity assumption that  $\rho_n \neq \mu^{\text{seq}}$  for all  $n$ . For simplicity, we also focus on the case where  $\rho_2 > \max \mu^{\text{seq}}$ , which suffices to contrast the following result with Proposition 3.

**Proposition 4** *Assume that  $\rho_2 > \mu^{\text{seq}}$ , and consider any sequential equilibrium of the repeated sequential security dilemma with private misperceptions. Whenever a misperception occurs in the first period, the first-mover takes  $H$  in every subsequent period, even when both players are in fact normal.*

*If in addition  $\rho_2 > \mu^{\text{sim}}$ , whenever a misperception occurs in the first period, the second-mover also takes  $H$  in every subsequent period.*

We again defer the proof to the Appendix. To see the intuition, suppose that player 1 takes  $D$  in period 1, but player 2 misperceives this action as  $H$ . Player 2 then updates negatively about player 1’s type, and responds with  $H$  in period 2.<sup>18</sup> The key difference from the previous subsection is that player 2’s misperception itself is not observed by player 1, so player 1 does not know that player 2 updated negatively about her own type. Player 1 only observes signal  $H$  in period 2, which makes her update her belief to  $\rho_2$ : this follows by Bayes rule’, since player 1 understands that in equilibrium  $s_2 = H$  if and only if *either* player 2 misperceived her action in period 1 (and hence took  $H$  in period 2) *or* player 2 perceived her action correctly in period 1 (and hence took  $D$  in period 2) but she misperceived his action in period 2. When  $\rho_2 > \mu^{\text{seq}}$ , this belief update causes player 1 to take  $H$  in period 3. Player 2 then observes signal  $H$  in period 3, leaving his belief unchanged (since he understands that both types of player 1 to take  $H$  in period 3 in response to  $H$  in period 2), so he again takes  $H$  in period 4, and so on, leading both players to take  $H$  in every subsequent period. Thus, a private misperception in the first period of the game causes the players to get stuck in a permanent “conflict trap.”

As this argument shows, the conflict trap is not caused by player 2’s misperception alone, but by the combination of this misperception and *misunderstanding*—player 1’s imperfect understanding of why player 2 is subsequently taking the action  $H$ . Hence, our theoretical analysis is clarifying that misunderstanding is the key for permanent conflict, and is itself caused by mistrust and misperceptions.

The analysis and interpretation of this subsection are related to that of Rohner, Thoenig and Zilibotti (2013). They analyze a two-player dynamic game with one-sided incomplete information and two-sided noise (so both type-I and type-II errors are possible, in contrast to our model with two-sided incomplete information but only type-I errors). The reasoning of their main result is related to ours and can be explained as follows. Since signals are uninformative when both the normal type and the bad type take the hawkish action (in their model, starting a war), the uninformed player can never learn that the opponent is bad for sure. Because the uninformed player’s beliefs eventually converge (by the martin-

---

<sup>18</sup>That player 2’s negative inference is sufficiently strong to induce this response is a consequence of the assumption that  $\rho_2 > \mu^{\text{seq}}$ . In particular, player 2’s posterior belief equals  $\rho_1$ , which exceeds  $\mu^{\text{seq}}$  because  $\rho_1 > \rho_2 > \mu^{\text{seq}}$ .

gale convergence theorem), this implies that with positive probability, the normal and bad opponent types take the same actions in the limit, leading to a permanent “war trap.”<sup>19</sup>

#### 4.4 Limited Memory Leads to Conflict Cycles

In our final variant of the repeated sequential security dilemma, we continue to assume that signals are private, but replace the assumption that players observe the full history of signals with a one-period memory assumption. This setup is intended to capture situations where players have much more precise information about the current state of their relationship—e.g., whether the opponent’s most recent action towards them was aggressive or conciliatory—than about the distant past. As we will explain, it also models settings where each player represents an infinitely lived dynasty, and members of each group/dynasty interact with members of the opposite group in overlapping generations. Most importantly, this structure captures uncertainty and disagreement about how and why historical conflicts were initiated and thus brings another dimension to possible misunderstandings between groups. Our main result in this model will be that the equilibrium involves infinitely recurring *conflict cycles*, where misperceptions trigger conflicts that persist for a while but ultimately come to an end.

For a concrete example of the type of conflict cycle that we hope to capture, consider the recurrent civil wars in Colombia, fought between the Liberal and Conservative parties and their supporters starting in the early 1850s. The first civil war began in 1851, and it led to neither a long-last conflict trap nor an enduring peace. Instead, hostilities quickly came to an end, but flared up again in 1854. The two parties then continued to alternate between periods of war and peace, with violent episodes occurring in 1859–63, 1876, 1884–85, 1895, and 1899–1902. After four decades of more peaceful coexistence, conflict resumed in the 1940s, in part because the Conservatives feared that the Liberals were growing more popular and would soon be able to permanently exclude them from power.<sup>20</sup> The growing hostilities culminated in the murder of the Liberal leader Jorge Eliécer Gaitán, which triggered the

---

<sup>19</sup>This result is in turn related to “learning traps” in dynamic experimentation models such as Easley and Kiefer (1988) and Aghion et al. (1991), since switching from  $D$  to  $H$  corresponds to stopping experimentation.

<sup>20</sup>Classic studies of power shifts of this type as causes of war include Fearon (1996) and Powell (2004, 2006).



most notorious episode of civil conflict, *La Violencia*, in 1948. The subsequent widespread agitation led by street mobs in Bogotá was in turn interpreted by the Conservatives as a move against them by the Liberals. The Conservatives’ reaction then led to a further significant escalation, culminating in a long-running civil war. However, even this large-scale civil war did not lead to a permanent trap, as a power-sharing agreement was eventually signed in 1957. Moreover, and consistent with our framework, this agreement did not result because the two sides softened their preferences or resolve—in fact, it was engineered by hard-line leaders, such as the Conservative politician Laureano Gómez.<sup>21</sup> Other examples of stop-and-start conflicts driven by rising and falling levels of fear and distrust include centuries of conflict between France and Germany, the Troubles in Northern Ireland (as described, e.g., in the quote by Richard English above), and many ethnic conflicts in post-colonial Africa. Finally, more direct evidence of distrust as a driver of conflict spirals comes from DellaVigna et al. (2014), who show that Croatians who live in villages that receive Serbian radio signals—and who are thus exposed to Serbian nationalistic content that was created for a Serbian audience—vote more for extreme Croatian nationalist political parties and also plaster their villages with anti-Serbian graffiti.

Formally, in this subsection we continue to assume that the period- $t$  signal  $s_t$  is observed only by the second-mover in period  $t$ , but we now also assume that when a player is the first-mover in period  $t$ , she conditions her decision only on the signal she observed in period  $t - 1$ ; and when a player is the second-mover in period  $t$ , she conditions her decision only on the signal she observes in the current period. A natural interpretation is that the two “players” each represent an infinite dynasty consisting of players who each live for one period and interact with players in the other group according to an overlapping generations structure. Specifically, the player who is born at time  $t$  observes the period- $t$  signal  $s_t$  and acts as the second-mover in period  $t$ , and then also acts as the first-mover in period  $t + 1$ ; the players who are born in odd periods all come from dynasty 1, while the players who are born in even periods all come from dynasty 2; and all players from each dynasty have the same type.<sup>22</sup>

Let us define  $\rho_n$  as in the previous subsection and denote the smallest integer  $n$  such that

---

<sup>21</sup>For the details of this history, see Hartlyn (1988) or Safford and Palacios (2002).

<sup>22</sup>We call the first-mover in period 1 “player 0.” This player does not observe a signal  $s_t$ , and always take  $a_1^1 = D$  by the assumption that  $\mu^0 < \mu^{\text{seq}}$ .

$\rho_n < \mu^{\text{seq}}$  by  $T^*$ .<sup>23</sup> Our main result in this subsection is as follows.

**Proposition 5** *The repeated sequential security dilemma with one-period memory has a unique sequential equilibrium. In this equilibrium:*

1. *In every period  $t = 1 \bmod T^*$ , the first-mover (when normal) plays  $a_t^1 = D$  regardless of  $s_t$ , and the second-mover (when normal) plays  $a_t^2 = s_t$  for each  $s_t \in \{D, H\}$ .*
2. *In every period  $t \neq 1 \bmod T^*$ , the first-mover (when normal) plays  $a_t^1 = s_{t-1}$  for each  $s_t \in \{D, H\}$ , and the second-mover (when normal) plays  $a_t^2 = s_t$  for each  $s_t \in \{D, H\}$ .*

This result is a variation of Proposition 1 in Acemoglu and Wolitzky (2014). To see the intuition, suppose that  $\rho_1 > \mu^{\text{seq}}$ , so  $T^* \geq 2$ .<sup>24</sup> Since  $\mu_0 < \mu^{\text{seq}}$ , player 0 takes  $a_1^1 = D$ . If player 1 misperceives this action, she updates her belief to  $\rho_1 > \mu^{\text{seq}}$ , and takes  $a_2^1 = H$ . This causes player 2 to observe  $s_2 = H$ , but he only updates his belief to  $\rho_2$  rather than  $\rho_1$ , because he understands that  $s_2 = H$  may be observed even if the other group is normal because of a misperception in period 1 *or* period 2. That is, the players understand that the more chances the two groups have had to get into conflict, the less informative the fact that they are currently in conflict is about the other group’s underlying type. Indeed, so long as a single misperception triggers a prolonged conflict, conflict eventually becomes completely uninformative. Formally, since  $\lim_{t \rightarrow \infty} \rho_t = \mu_0$ , a player’s posterior when observing signal  $H$  following  $t$  periods in which any misperception sparks persistent conflict converges to the prior. But this observation implies that misperceptions cannot trigger permanent conflict: once the groups have had “enough” chances to get into conflict, the occurrence of a conflict becomes completely uninformative. This then induces a normal first-mover to take  $D$  after a while, regardless of the signal she observed in the previous period. By construction, this “restart” occurs for the first time in period  $T^*$ . As a result, a player who observes signal  $H$  in period  $T^* + 1$  against becomes very pessimistic about the other group’s type, and conflict resumes.<sup>25</sup>

---

<sup>23</sup>This is well-defined by the assumption that  $\mu_0 < \mu^{\text{seq}}$ .

<sup>24</sup>If  $T^* = 1$ , we are left with the trivial equilibrium where the first-mover always takes  $D$ .

<sup>25</sup>Banerjee (1993) notes a somewhat similar mechanism in a model of rumors about an investment opportunity. He studies how a rumor becomes less informative over time, as there are more and more chances for the rumor to start. However, his model does not involve “restarts” or cycling.

The equilibrium of the one-period memory model displays some interesting comparative statics. Perhaps the most striking of these relies on the endogeneity of the “restart time”  $T^*$ . For example,  $T^*$  is decreasing in  $\pi$ : when misperceptions are rarer, observing an  $H$  signal is worse news about the opponent’s type, so the groups must have had more chances to get into conflict before a player is willing to risk playing  $D$ . In fact, it is not hard to show that  $T^*$  must converge to infinity as  $\pi \rightarrow 0$ , and that moreover the long-run fraction of the periods in which the groups are in conflict remains bounded away from 0 as  $\pi \rightarrow 0$ . Intuitively, as  $\pi \rightarrow 0$  misperceptions becomes extremely rare, but they also become extremely damaging to trust when they do occur, so that on average the groups take longer to return to cooperation following a misperception.

This result that the long-run probability of conflict remains bounded away from 0 even as misperceptions become extremely rare contrasts sharply with lessons from classical static models such as Kydd’s. One of Kydd’s key predictions is that “tragic spirals between [normal types] are likely to be a small proportion of observed conflicts, especially as information improves.” This prediction is valid in the context of one-shot games, but it is not valid in an infinite-horizon setting with limited memory, because while making misperceptions rarer does make the onset of “tragic spirals” less frequent, it also makes spirals last longer when they do occur (precisely because the onset of a tragic spiral is “worse news” about the opponent’s type). In this regard, the predictions of our infinite-horizon, limited-memory model are actually closer to Waltz and Mearsheimer’s earlier claims that tragic conflict is likely to be quite prevalent than to Kydd’s game-theoretic result that tragic conflict is rare in a one-shot model.<sup>26</sup>

A final remark is that the result that cooperation restarts exactly every  $T^*$  periods is of course very stylized. This result stems from the assumption that players perfectly observe calendar time  $t$ , and thus can count how many opportunities there have been for the groups to fall into conflict. An alternative, possibly more realistic model would assume that the

---

<sup>26</sup>Another difference from Kydd’s result is that Kydd emphasizes “convergence on correct beliefs is more likely than convergence on incorrect beliefs,” (2005, p.18). Instead, with limited memory, and arguably often in reality, the players’ beliefs about each other’s type do not converge at all, and beliefs can spend a lot of time far from the truth. This feature would also arise in models where memory is unbounded but players’ types change over time according to a Markov process, as in, e.g., Mailath and Samuelson (2001) or Phelan (2006).

players cannot observe calendar time, and instead update their beliefs based on the long-run frequencies of the different signals.<sup>27</sup> In Acemoglu and Wolitzky (2014), we show that such a model also generates cycles, but now cycles are irregular and result from players mixing between  $D$  and  $H$  after observing  $s = H$  in the previous period.

## 4.5 Additional Applications

We now briefly discuss a number of additional applications of the framework developed in this section.

The first application concerns the role of third-party mediators—including international organizations such as the United Nations—in supporting peace. Although mediators are often thought to reduce the likelihood of conflict among nations, why this is so is theoretically unclear, since mediators generally lack enforcement power, and hence any agreement they broker is non-binding. A literature in international relations and economics studies mediators as communication intermediaries that can reduce conflict by reducing mistrust between the parties. One strand of this literature—which is closely related to standard mechanism design (e.g., Myerson, 1982)—considers mediators who can commit to the messages that they send to each party (e.g., recommendations to take aggressive or peaceful actions) as a function of the messages they receive from both of them (e.g., their reporting military strength). The literature often finds such mediators to be effective, and in some cases just as effective as if they could compel the parties to implement a recommended agreement (Goltsman et al., 2009; Horner, Morelli, and Squintani, 2015; Meierowitz et al., 2019).<sup>28</sup> A second strand of the literature considers mediators without commitment power. The effectiveness of such mediators depends on their preferences. For example, mediators who only want to achieve peace may be ineffective—intuitively, because they always tell the parties not to fight, and thus cannot credibly convey any useful information—but mediators who are somewhat biased towards one party and are willing to tolerate occasional conflict can still be useful (Kydd,

---

<sup>27</sup>This is equivalent to assuming that each player has an improper uniform prior over the time at which she enters the game.

<sup>28</sup>However, in a setting closer to the models considered in this chapter—where the parties' private information concerns only their own payoffs—Fey and Ramsay (2010) show that mediation is ineffective even with commitment power.

2003, 2006; Smith and Stam, 2003; Rauchhaus, 2006).

However, both of these strands of literature focus on mediation as reducing mistrust ( $\mu_0$ ) rather than misperceptions ( $\pi$ ). In our framework, the consequences of reducing mistrust and misperceptions are distinct, and they interact in some interesting ways. To see this, consider the conflict cycle model of the previous subsection. If international organizations reduce  $\pi$ , this makes conflict less likely to be initiated. However, as already noted, this also increases  $T^*$ , and thus any conflict lasts longer. As a result, the overall impact on conflict intensity and duration from such communication may be small. In contrast, if interactions within the auspices of international organizations also reduce mistrust (lowering  $\mu_0$ ), then the gains in terms of reduced conflict are much larger. It can be verified that as both  $\pi$  and  $\mu_0$  go to zero, conflict disappears—but this does not happen when only  $\pi$  goes to zero (Acemoglu and Wolitzky, 2014). Hence, in this instance, our analysis highlights the importance of reducing mistrust and misperceptions simultaneously.

Another potentially important consequence of communication and cooperation under the umbrella of international organizations may be to allow countries to recognize each other's misperceptions, thus reducing misunderstanding. As we have seen, reducing misunderstanding can move the equilibrium from one with conflict traps or cycles (as in Sections 4.3 and 4.4) to one where, while there can still be some short-lived conflicts, normal parties always manage to secure peace in the long run (as in Section 4.2). Reducing misunderstandings is thus potentially one of the most important roles—and, perhaps, one of the greatest historical successes—of international organizations. At the same time, the general discussion in this chapter also highlights that achieving the degree of communication necessary for removing all misunderstanding may be quite difficult.

The second application we discuss briefly is the idea of Kantian peace—the notion that democracies are less likely to go to war with each other. Kantian peace is one of the major regularities in international relations, though there exists no consensus explanation for this pattern (see Baliga, Lucca, and Sjöström, 2011, for a discussion of this literature, a model of Kantian peace, and new evidence on this pattern). Our framework emphasizes the same types of forces we discussed in the context of the role of international organizations. Democracies may be better able to communicate, reducing misperceptions ( $\pi$ ), but once again,

unless this also reduces mistrust, it will not fully eliminate conflict. In this instance, there may be additional reasons why mistrust can be reduced as well. In particular, democratic leaders may have less to gain from aggressive actions than autocratic leaders, and if this is broadly recognized, it will translate into lower  $\mu_0$ , and can thus significantly limit conflict between democracies. Additionally, better communication between democracies may once again reduce misunderstandings. However, free media and open political competition in democracies can also create room for conspiracy theories, rumors, misinformation, and extremist positions, which can increase misperceptions or mistrust between nations. Hence, our framework highlights some theoretical limits to the possibility that the Kantian peace works through better communication and comprehension between polities.

The third application concerns the question of whether and how conflict traps and conflict cycles can be broken. One possibility is that this could be achieved by some generation of one of the parties exhibiting “leadership”, for example, acting in a more forward-looking manner or taking actions that are more distinctive, which can become more informative signals of peaceful intentions. A prominent example of such leadership is Nelson Mandela’s actions during the 1995 Rugby World Cup. These took place in the context of the end of white Afrikaner rule in South Africa, which created a period of uncertainty and fear, especially among the white community. This era was thus rife with mistrust and potential misperceptions, which could easily have led to heightened racial conflict. However, Mandela not only advocated peaceful reconciliation and the rights of the white minority, but also took various symbolic actions to demonstrate his commitment to the peaceful resolution of outstanding problems. Famously, during the Rugby World Cup, Mandela wore the jersey of the South African national team, the Springboks, which was until then associated with the apartheid regime. Mandela also personally presented the trophy to team captain Francois Pienaar, an Afrikaner. These symbolic gestures were interpreted as signals of conciliatory intent and helped a smoother transition to democratic rule and broadly peaceful relations between the white minority and the black majority.

Within our model, the possibility for such leadership can be introduced in two complementary ways. The first is to assume that some generations could have more forward-looking preferences and thus internalize the benefits of experimenting with peaceful actions, even in

the middle of conflict traps or cycles. The second is to assume that they may have actions that are less likely to be misperceived.<sup>29</sup> Via either or both channels, such agents can try to shift the equilibrium to a more cooperative one. In the first case, this will be by taking the peaceful action  $D$ , while regular players would have continued to play  $H$ . This will be costly for these “leaders” in the short run but if it can induce more peaceful behavior in the future, their forward-looking utility could increase. In the second case, they may choose to take the peaceful action  $D$  precisely because they recognize that this is less likely to be misperceived, and this can also shift the equilibrium in a more cooperative direction. Formally combining leadership with the kind of models analyzed in this section is an interesting direction for future research.

## 5 Misperceptions and Deterrence

Our analysis so far has built on the classic security dilemma or spiral model. The basic insight of this model is that two parties that would both rather coordinate on a peaceful equilibrium may nonetheless be drawn into conflict due to mistrust or misperception. The underlying model is thus a coordination game, and the question of interest is whether the parties can successfully coordinate on the peaceful equilibrium.

We now turn to a related but distinct model, which we term *the deterrence model*. In the deterrence model, the two parties are asymmetric: for the first-mover (or *attacker*), the hawkish action is dominant, while the second mover (or *defender*) wants to match the first-mover’s action. The basic game can thus be viewed as a prisoner’s dilemma on one side and

---

<sup>29</sup>This second mechanism is studied in Acemoglu and Jackson (2015), which is a related overlapping-generations model with incomplete information. In their setup, each generation also cares about the actions of the next generation, which thus introduces forward-looking behavior in this context. Leadership emerges when there are “prominent” agents whose actions are observed with less noise, and not just by neighboring agents but also by all future generations. This enables these prominent agents to leverage the expectation-anchoring role of their action, potentially shifting the equilibrium from a conflictual one to a more peaceful/cooperative one.

a coordination game on the other. We represent this game with the following payoff matrix:

		defender	
		$D$	$H$
attacker	$D$	$1, 1$	$-z, g$
	$H$	$x, -l$	$0, 0$

where  $l > 0$  and  $g < 1$  (so the defender's preferences are the same as in the security dilemma/spiral model), but also  $z > 0$  and  $x > 1$  (so  $H$  is dominant for the attacker). As in the sequential version of the security dilemma game, player 1 (here the attacker) moves first, and then player 2 (here the defender) observes a signal  $s$  of player 1's action before choosing his own action. While this is not essential, for simplicity we start with the case where the signal distribution is the same as in the sequential security dilemma game: if  $a_1 = D$ , then the action is misperceived as  $H$  with probability  $\pi$ , while if  $a_1 = H$ , the action is always perceived correctly. In contrast to the security dilemma, the key insights of the deterrence game can already be seen in the case where there is perfect information about the players' preferences, and we will therefore begin with this case. (In other words, we assume that it is common knowledge that the players' preferences are given by the above payoff matrix.)

We can begin with the immediate observation that if  $\pi = 0$  (so the attacker's action is perfectly observed), the unique subgame perfect equilibrium of the deterrence game has both players taking  $D$  along the equilibrium path. This follows because even though  $H$  is dominant for the attacker in the above payoff matrix, she understands that a play of  $D$  will be met with  $D$ , while a play of  $H$  will be met with  $H$ , so she prefers to take  $D$ . This observation captures the basic logic of deterrence. In contrast, if  $\pi = 1$  then the defender observes  $H$  regardless of the attacker's action, and it is easy to see that the unique subgame perfect equilibrium then has both players taking  $H$ : the attacker takes  $H$  because this is dominant for him for any fixed action of the defender's, and the defender therefore anticipates that the attacker takes  $H$  and hence takes  $H$  in response. As this simple discussion indicates, the scope for deterrence depends on the precision of the defender's signal.

The basic deterrence model we have laid out is a simple example of an *inspection game* (Avenhaus, von Stengel, and Zamir, 2002). In an inspection game, an *inspectee* (the attacker)



chooses whether to *violate* or *not violate* (here corresponding to actions  $H$  and  $D$ ); an *inspector* (the defender) then observes a signal of the inspectee’s action and chooses whether to *sound* an alarm or *not sound* it (again corresponding to  $H$  and  $D$ , respectively); and the players’ preferences over the four pure outcome  $(D, D)$ ,  $(D, H)$ ,  $(H, D)$ , and  $(H, H)$  are the same as in the above payoff matrix. Inspection games emerged in the 1960s in the context of military and operations research applications, especially arms control and disarmament, and were later used to analyze pollution control. Inspection-type models entered economics in the guise of *auditing games*, where, for instance, taxpayers decide whether to cheat on their taxes, and the government decides which taxpayers to audit.<sup>30</sup> Similar models also appear in the conflict literature in the context of unobserved decisions to acquire arms (e.g., Baliga and Sjöström, 2008; Meiorowitz and Sartori, 2008; Jackson and Morelli, 2009; Debs and Monteiro, 2014; Meiorowitz et al., 2019).<sup>31</sup> As we will see, equilibria in deterrence or inspection games are typically mixed, so that, in particular, the conflict outcome  $(H, H)$  is played with positive probability, even though it is common knowledge that both players prefer the peaceful outcome  $(D, D)$ . This result echoes a key theme of the papers on unobserved arming just mentioned—inefficient conflict occurs with positive probability.

An important simplifying assumption in inspection games (as well as the deterrence game above) is that while  $H$  is dominant for the inspectee/attacker, the inspector/defender truly has coordination game preferences. This assumption sidesteps another important theoretical issue in deterrence theory, which is that the defender’s threat to play  $a_2 = H$  after a signal that indicates  $a_1 = H$  may not be credible. This issue is known as the *search for credibility* (e.g., Powell, 1990). It played a critical role in nuclear deterrence during the Cold War, where the key question was, if the US or the USSR detected a massive wave of incoming missiles, would they actually retaliate in kind? While this is a fascinating issue, in most situations it seems reasonable to suppose that an actor truly prefers to act dovishly towards a potential attacker who refrains for attacking while preferring to act hawkishly against the

---

<sup>30</sup>For references, see Avenhuas, von Stengel and Zamir (2002) and Baliga, Bueno de Mesquita, and Wolitzky (2020).

<sup>31</sup>In Baliga and Sjöström (2008), what is stochastic is whether an attempt to acquire arms is successful or not, rather than whether it is undertaken. In Jackson and Morelli (2009), arming is observed but is undertaken simultaneously by both parties, which again generates a mixed equilibrium. In Meiorowitz and Sartori (2008), Debs and Monteiro (2014), and Meiorowitz et al. (2019), arming is unobserved, and the basic logic is similar to that of an inspection game.

same attacker when she does attack. In this chapter we proceed under this assumption.

## 5.1 Analysis of the Basic Deterrence Game

The key insight of the basic deterrence game is that there is typically a unique sequential equilibrium, which is in mixed strategies: the attacker attacks with a probability  $p$  that makes the defender indifferent between  $H$  and  $D$ , and conditional on observing signal  $s = H$  the defender takes  $H$  with a probability  $r$  that makes the attacker indifferent between  $H$  and  $D$ . The logic is that if the attacker always attacked, then conditional on observing  $s = H$  the defender would be very confident that  $a_1 = H$  (if we assume that the misperception probability  $\pi$  is sufficiently small, which seems realistic) and would thus always take  $a_2 = H$ ; but then the attacker would be better off not attacking. Similarly, if the attacker never attacked, then the defender would always take  $a_2 = D$ , even after observing signal  $s = H$  (as in this case the defender would be confident that she misperceived the signal). But then the attacker would prefer to attack.

Let us work out this argument more formally, which will also lead us to some interesting comparative static results. First, let  $\beta(p)$  denote the defender's belief that  $a_1 = H$  conditional on observing  $s = H$ , when the attacker's equilibrium probability of attacking equals  $p$ . We will refer to this belief  $\beta(p)$  as measuring how *suspect* the attacker is following signal  $H$ . By Bayes' rule, this belief is given by

$$\beta(p) = \left(1 + \frac{1-p}{p}\pi\right)^{-1}.$$

It is easy to check that the defender's best response is to take  $a_2 = H$  following signal  $s = H$  if and only if:<sup>32</sup>

$$\beta(p) \geq \beta^* := \left(1 + \frac{l}{1-g}\right)^{-1}.$$

Note that  $\beta(0) = 0 < \beta^*$  and  $\beta(1) = 1 > \beta^*$ , so by continuity there exists a threshold attack probability  $p^* \in (0, 1)$  such that  $\beta(p) = \beta^*$  if and only if  $p = p^*$ . That is,  $p^*$  is the probability such that, if the attacker attacks with probability  $p^*$  and the defender observes

---

<sup>32</sup>Notice that this is the same threshold as  $\mu^{sim}$  in the security dilemma: that is,  $\beta^* = \mu^{sim}$ .

signal  $H$ , she is indifferent between responding with  $a_2 = D$  or  $a_2 = H$ .

Next, note that when the defender takes  $a_2 = H$  (which we can refer to as *retaliating*, i.e., taking a hawkish response to the perceived action of the attacker) with probability  $r$  following signal  $s = H$ , and never retaliates following signal  $s = D$ , then taking  $a_1 = D$  is optimal for the attacker if and only if

$$\underbrace{(1 - \pi r)(1) - \pi r(z)}_{\text{expected payoff from } D} \geq \underbrace{(1 - r)(x) + r(0)}_{\text{expected payoff from } H} \iff r \geq r^* := \frac{x - 1}{x - \pi(1 + z)}.$$

Assuming that  $\pi(1 + z) < 1$  (so that misperceptions are sufficiently unlikely), we have  $r^* \in (0, 1)$ . In this case, in the unique sequential equilibrium of the deterrence game, the attacker attacks with probability  $p^* \in (0, 1)$ , the defender never retaliates following signal  $s = D$ , and the defender retaliates with probability  $r^* \in (0, 1)$  following signal  $s = H$ . We can summarize this discussion with a proposition.

**Proposition 6** *In the basic deterrence game where misperceptions are sufficiently unlikely (i.e.,  $\pi(1 + z) < 1$ ), there is a unique sequential equilibrium, which is in mixed strategies: the attacker attacks with probability  $p^* \in (0, 1)$ , the defender never retaliates following signal  $s = D$ , and the defender retaliates with probability  $r^* \in (0, 1)$  following signal  $s = H$ .*

The comparative statics of the unique equilibrium follow a standard mixed strategy logic, but some of them are somewhat counterintuitive at first glance.

For example, both  $p^*$  and  $r^*$  are increasing in the misperception probability  $\pi$ : that is, a greater misperception probability leads to both more attacks and more retaliation in equilibrium.<sup>33</sup> This is a key lesson of the basic deterrence game. To see the intuition, first note that as  $\pi$  increases for fixed  $p$ , the attacker becomes less suspect following signal  $s = H$  (as this signal becomes more likely to have resulted from a misperception); this makes the defender less inclined to retaliate. Hence, to keep the defender willing to retaliate, the attack probability  $p$  must increase. Second, as  $\pi$  increases for fixed  $r$ , the attacker's payoff from  $D$

---

<sup>33</sup>Note that an increase in  $\pi$  actually leads to greater retaliation for three distinct reasons: First, holding  $p$  and  $r$  fixed, an increase in  $\pi$  increases the frequency of  $H$  signals. Second, an increase in  $\pi$  increases  $p$ , which also increases the frequency of  $H$  signals. Third, an increase in  $\pi$  increases  $r$ , the probability of retaliation following an  $H$  signal. (Recall that there is never retaliation following a  $D$  signal.)

decreases while his payoff from  $H$  remains constant. Hence, to keep the attacker willing to not attack, the retaliation probability  $r$  must increase.

Another standard, but still notable, observation is that each party's mixing probability is determined only by the other party's preferences. As attacking becomes more attractive for the attacker ( $x$  and/or  $z$  increases), the attack probability  $p$  stays constant, while the retaliation probability  $r$  increases to keep the attacker indifferent. As retaliating becomes more attractive for the defender ( $g$  and/or  $l$  increases), the retaliation probability  $r$  stays constant, while the attack probability  $p$  decreases, as this makes the attacker less suspect following  $s = H$ , which keeps the defender indifferent.

These basic observations concerning the logic of the deterrence or inspection game have some interesting implications. For example, Tsebelis (1989) used this logic to challenge the canonical idea in the law and economics literature that increasing the fine for criminal activity reduces crime. As Tsebelis pointed out, criminal and law enforcement activities take the form of a deterrence or inspection game, where the police will bother exerting effort to catch criminals only if the crime rate reaches a certain level. From this perspective, the equilibrium crime rate is determined by the preferences of the *police*, not those of the criminals. Hence, greater fines for criminals will reduce the equilibrium effort level of the police (as is required to keep the criminals indifferent), while the equilibrium crime rate will remain constant (as the crime rate required to keep the police indifferent has not changed).<sup>34</sup>

Another interesting feature of inspection games is that making the defender's signal more precise (in the sense of Blackwell, 1951) does not necessarily reduce the equilibrium attack probability or increase the defender's equilibrium expected payoff.<sup>35</sup> This effect can arise when there are three or more possible realizations of the defender's signal. To see the

---

<sup>34</sup>Of course, this extreme result requires that the equilibrium elasticity of law enforcement effort with respect to the crime rate is infinite. In a more realistic model where this elasticity is finite (as in the inspection game variant considered in the next subsection), an increase in the fine for criminal activity would decrease both law enforcement effort and the crime rate.

The literature on deterrence/inspection games in law and economics and related areas has remained active since Tsebelis's paper. A key theme of the recent literature is the importance of random inspections. See, e.g., Lazear (2006), Eeckhout, Persico, and Todd (2010), Ortner and Chassang (2018), Dilme and Garrett (2019), Varas, Marinovic, and Skrzypacz (2020), Pei and Strulovici (2021), Kapon (2022), and Ball and Knoepfle (2023).

<sup>35</sup>This observation is due to Baliga, Bueno de Mesquita, and Wolitzky (2020), although a related point was made by Cremér (1995) in the context of a principal-agent model with renegotiation.

intuition, first suppose that the defender can only get signal  $s = D$  or  $s = H$ , where  $s = H$  is more indicative of  $a_1 = H$ , and in equilibrium signal  $s = H$  is sufficiently informative that the defender retaliates following signal  $s = H$  given the equilibrium attack probability  $p$ . Now suppose that the defender’s information structure becomes more refined in the Blackwell sense, so that when signal  $s = H$  arose under the original information structure, now one of two signals arise: either  $s = H_L$ , indicating a “probable attack,” or  $s = H_H$ , indicating a “certain attack.” If the attack probability remains fixed at  $p$ , the defender may now stop retaliating following signal  $s = H_L$ , because his posterior belief that  $a_1 = H$  following signal  $s = H_L$  under the new information structure is lower than his posterior following signal  $s = H$  under the original information structure.<sup>36</sup> But this change in the defender’s strategy would make  $a_1 = H$  uniquely optimal for the attacker, which is inconsistent with equilibrium, so the attacker’s equilibrium attack probability must increase to the point where the defender is willing to retaliate following signal  $s = H_L$  as well as  $s = H_H$ .

## 5.2 Multiple Possible Attackers and Imperfect Attribution

Classical deterrence theory along the lines just described was developed in the context of bilateral interactions, with the US and the USSR in the Cold War being the canonical example. In today’s more multi-polar world—as well as in more mundane problems arising in areas like pollution control or law and economics—another important aspect of deterrence is the presence of multiple possible aggressors and *imperfect attribution of attacks*, so that the defender’s information may not definitely determine *which* attacker is responsible for a given attack, as well as whether or not an attack actually occurred. In the context of international relations, imperfect attribution is especially important in cyberwarfare, where examples of false alarms, detection failure, and misidentification of the perpetrator of an attack abound. Thus, whereas so far this chapter has focused on misperceptions of whether or not a single adversary took an aggressive action, we now consider a second important type of misperception: *which* adversary is more likely to have acted aggressively

Baliga, Bueno de Mesquita, and Wolitzky (2020) develop a simple model of deterrence

---

<sup>36</sup>The follows because the defender’s posterior following  $s = H$  “splits” into a lower posterior following  $s = H_L$  and a higher posterior following  $s = H_H$ .

with imperfect attribution, which builds on the standard deterrence/inspection game described above. Consider a situation with  $n$  possible attackers and one defender. One of the attackers may randomly get a chance to launch an attack (for example, by identifying a weakness in the defender’s computer network), and the defender then receives a signal that probabilistically indicates whether an attack occurred, and if so which attacker is responsible. To make the model more tractable and to generate more realistic comparative statics, assume that the benefit  $x$  from attacking without facing retaliation is stochastic and varies across attackers, with the realization of this variable for each attacker being her private information. Similarly, let us assume that the loss  $l$  from failing to retaliate against each attacker is stochastic and heterogeneous across attackers, and is the private information of the defender.<sup>37</sup> These random payoffs “purify” the mixed equilibrium in the standard deterrence game, leading to an equilibrium where each attacker  $i$  attacks if and only if his attack benefit  $x_i$  exceeds an (attacker-specific) threshold  $x_i^*$ , while the defender retaliates against attacker  $i$  after a given signal realizations if and only if her loss from failing to retaliate  $l_i$  is exceeds a threshold  $l_i^*$ . We also observe that an attacker with a sufficiently high value of  $x_i$  will always attack in equilibrium, so the probability that an attacker has a high value of  $x_i$  plays a similar role to the level of mistrust of this attacker, in the terminology of the previous sections.

A key feature of this model is that although there are no direct payoff externalities among the attackers (who each care only about whether they strike the defender, and whether they face retaliation), the model features a kind of endogenous strategic complementarity among the attackers, which works through the defender’s Bayesian updating problem of attributing responsibility for a given attack. To see the idea, suppose that the distribution of attack benefits  $x_i$  for one attacker  $i$ —say,  $i$  =Russia—shifts up, so that Russia attacks with higher probability in equilibrium. This implies that, after each signal, the defender believes that Russia is *guilty* (responsible for the attack) with higher probability, and correspondingly believes that each other potential attacker is guilty with lower probability.<sup>38</sup> This in turn implies that the defender’s retaliation threshold increases for each potential attacker other

---

<sup>37</sup>Baliga, Bueno de Mesquita, and Wolitzky use a slightly different, but equivalent, payoff parameterization.

<sup>38</sup>Except after a “business as usual” or *null* signal, which indicates that no attack is likely to have occurred. The strategic complementarity logic holds as long as the defender never retaliates following the null signal.

than Russia. Finally, this sequence of reasoning implies that the other attackers increase their equilibrium attack probabilities as well. In sum, due to the defender’s finite supply of “suspicion” following each attack, when one attacker becomes more aggressive, this uses up some of the defender’s suspicion, and causes the other attackers to also become less suspect and therefore more aggressive.<sup>39</sup>

The model can be used to understand the impact of changes on the defender’s signal technology on the profile of equilibrium attack probabilities. There are some surprising effects here, as well as some expected ones. As noted above, improving the defender’s information in the Blackwell sense need not reduce the equilibrium frequency of attacks or increase the defender’s equilibrium payoff.<sup>40</sup> Another finding is that, in a certain sense, it is better for the defender to fail to detect an attack at all, rather than detecting the attack but attributing it to the wrong attacker. The intuition is that failing to detect attacks by Attacker 1 or misattributing to Attacker 2 both reduce retaliation against Attacker 1, and hence make him more aggressive; but the latter type of mistake also makes the defender more hesitant to retaliate against Attacker 2 when the signal points to her (because these signals may result from misattributed attacks by Attacker 1), and hence makes Attacker 2 more aggressive as well. This result therefore highlights the subtle effects of misperceptions on the likelihood of the onset of conflict in a Bayesian framework, especially in settings with multiple potential aggressors.

### 5.3 Continuous Claims and Salami Tactics

In the basic deterrence model, the attacker makes a binary choice between attacking ( $H$ ) and not attacking ( $D$ ). In many scenarios, the first-mover in a conflict can also choose from among a range of possible *claims*, of different levels of aggressiveness, while the second mover

---

<sup>39</sup>This logic is somewhat akin to the law and economics literature on “crime waves,” where crime is modeled as a game of strategic complements among potential criminals, because high crime rates overwhelm the police (e.g., Glaeser, Sacerdote, and Scheinkman, 1996). However, a key theme of this literature is that strategic complements can lead to multiple equilibria, which may explain the high variance of crime rates across time and space. In contrast, in the imperfect attribution model, there is always a unique equilibrium. Intuitively, this is because, by Bayes’ rule, the defender’s posterior beliefs (and hence the retaliation probabilities) are determined by the *ratio* of the equilibrium attack probabilities rather than their levels, so there cannot exist multiple equilibria with different frequencies of attacks.

<sup>40</sup>This can happen even with a single potential attacker.

then observes a signal of the claim and then has to decide whether to acquiesce to the claim (that is, *accept*, which corresponds to the dovish action  $D$  in the deterrence game), or to initiate a costly conflict (that is, *reject* the claim, corresponding to the hawkish action  $H$  in the deterrence game). For example, the first-mover might choose how much territory to occupy or how many arms to acquire, and the second-mover might misperceive exactly how much territory was occupied or how many arms were acquired before deciding whether to acquiesce or contest the occupation or arming. In addition to being more realistic, allowing a range of possible claims can capture the possibility of *salami tactics* (Schelling, 1966), where deterrence is hindered by the second-mover’s inability to clearly and credibly indicate what signals will trigger conflict; this allows the first-mover to gradually claim more and more resources (“slicing the salami”).<sup>41</sup> One prominent recent example comes from Chinese naval activities in the South China Sea, which are viewed by several experts as instances of using the country’s first-mover status in their neighborhood and uncertainty/ambiguity about their exact actions in order to gain a strategic advantage (e.g., Kaplan, 2014; Coy, 2021).

As this discussion indicates, a deterrence game with a range of possible claims may alternatively be viewed as a game of *ultimatum bargaining with imperfectly observed offers*. In this game, the first-mover (who we will call the *claimant*) chooses a claim  $a \in [-M, M]$  (where  $M$  can be any number greater than 1), and the second-mover (who we will call the *responder*) observes a noisy signal  $s$  of the claim  $a$  before accepting or rejecting the claim. The size of the total available economic surplus (“the pie”) is normalized to 1, so that if the responder accepts a claim of  $a$ , payoffs are  $a$  for the claimant and  $1 - a$  for the responder; if instead the claimant rejects, each player’s payoff is 0. This game is studied by Wolitzky (2022).<sup>42</sup> The assumption that  $M > 1$  implies that the claimant can demand “more than the entire pie”: that is, he can claim so many resources that the responder is better-off contesting the claim rather than accepting it.

A couple remarks on the model are in order. First, if the responder’s signal  $s$  is perfectly

---

<sup>41</sup>A different perspective on salami tactics is provided by Powell (1996), who models such situations as wars of attrition with incomplete information.

<sup>42</sup>Ravid (2020) and Denti, Marinacci, and Rustichini (2021) develop related models of unobserved-offers bargaining, where the responder endogenously acquires a signal of the claim at some cost.



informative, the game reduces to the standard ultimatum bargaining game, where in the unique subgame-perfect equilibrium the claimant demands  $a = 1$  and the responder accepts. In this situation, the assumption that the claimant can demand more than the entire pie is irrelevant, since the responder would reject these demands. However, once the claim is observed with noise, this assumption will preclude the possibility of an equilibrium where conflict is entirely absent (i.e., where the responder accepts the claim with probability 1). Intuitively, if the responder accepts with probability 1, then she must accept with probability 1 after every signal realization, in which case the claimant will make the greatest possible claim (since the claim is always accepted). But if this claim is greater than 1, the responder would be better-off rejecting.

More precisely, under some standard assumptions on the distribution of  $F$ —including, crucially, the assumption that  $F$  has a non-moving support—in every Nash equilibrium of the imperfectly-observed offers bargaining game, either the claim is rejected with probability 1, or the claimant demands the entire pie ( $a = 1$ ) with probability 1, and the responder accepts after some signal realizations but rejects after others.<sup>43</sup> Costly conflict thus occurs with positive probability, even though the pie is perfectly divisible and there is no uncertainty about the parties’ preferences.<sup>44</sup> The key friction in the model is that the claimant cannot demand slightly less than the entire pie (by taking action  $a = .99$ , say) while credibly communicating this reduced claim to the responder. If the claim were perfectly observable then the responder would always accept when  $a = .99$ —leaving both parties better off—but since the responder’s signal is noisy, conflict risk is only slightly lower when  $a = .99$  rather than 1, so that the claimant is better off claiming  $a = 1$  and incurring a slightly higher conflict risk.

The friction that generates conflict in bargaining with imperfectly-observed claims is somewhat akin to the mixed-strategy nature of the equilibrium in the basic deterrence game. In the deterrence game, if the attacker could take the dovish action  $D$  and could credibly re-

---

<sup>43</sup>In addition to full support, the required assumptions are that the conditional signal distribution is smooth, satisfies the strict monotone likelihood ratio property in  $s$  and  $a$ , and is log-concave in  $a$ . These assumptions imply that there is no equilibrium where the claimant mixes. Note that there is also always a trivial equilibrium where the claim is always rejected, by the same logic as in Bagwell (1995).

<sup>44</sup>Imperfect observation of claims thus seems distinct from the other “rationalist explanations of war” proposed by Fearon (1995).

veal this action to the defender, the resulting outcome of  $(D, D)$  would be preferred by both players to the mixed equilibrium outcome. Thus, misperception of the first-mover's action is the root cause of conflict in both the basic deterrence game and the richer imperfectly-observed-offers bargaining game. The bargaining game also displays some natural comparative statics that can be useful for interpreting various conflict episodes. For instance, noisier observations of claims is associated with greater conflict risk, and (in an extension of the model where rejecting a claim leads to temporary rather than permanent conflict) conflict risk persists even as the time between successive claims goes to zero.

Given that the attacker in a deterrence game or the claimant in an imperfectly-observed-claims bargaining game is better-off if she can reveal her action to the other party, a natural critique of these models is that in practice parties should be able to avoid conflict by revealing this information. However, in reality adversaries often do choose to conceal the extent of their aggressive actions, such as the size of the weapons stockpiles they have acquired. There are likely several reasons for this. One obvious reason is that revealing one's weapons makes one vulnerable if conflict does break out. This mechanism is investigated by Coe and Vaynman (2020), who argue that it is a key reason why arms control failed to prevent the Iraq War. Another reason is that, just as the level of arming itself is imperfectly observed, so may the extent to which a country is complying with an arms control agreement. In other words, enriching a deterrence or bargaining model by letting the aggressor reveal his action might just push the scope for misperception from the action itself to the extent to which the aggressor has actually fully revealed it.

## 5.4 Dynamic Deterrence with Misperceptions

Given that one-shot sequential bargaining with imperfectly-perceived claims results in a positive probability of inefficient conflict, it is natural to ask whether repeated bargaining—which is a more realistic model of a long-run relationship between two adversaries—can lead to a more efficient outcome. In a similar context, Schelling (1966) argued informally that “tripwire” or “plate glass window” strategies could play an important role in deterring aggressive claims and averting conflict. Intuitively, if there is some critical threshold  $s^*$  such that the responder is expected to accept if and only if the perceived claim  $s$  is below  $s^*$ , the

aggressor can be incentivized to moderate his claim, and the responder can be incentivized to reject claims above  $s^*$ , for fear of triggering a shift to a more conflictual equilibrium regime. In Schelling’s words (1966, p. 56),

“Our deterrence rests on Soviet expectations. This, I suppose, is the ultimate reason why we have to defend California—aside from whether or not Easterners want to. There is no way to let California go to the Soviets and make them believe nevertheless that Oregon and Washington, Florida and Maine, and eventually Chevy Chase and Cambridge cannot be had under the same principle.”

In an extension of the bargaining model described in the previous subsection, Wolitzky (2022) shows that such a “tripwire” strategy can succeed, provided that three conditions are satisfied. First, it must be possible to choose the cutoff signal  $s^*$  so that the signal  $s$  is unlikely to exceed  $s^*$  along the equilibrium path, but so that  $s$  becomes much more likely to exceed  $s^*$  if the aggressor marginally increases his claim. This can be done provided that extreme signal realizations are sufficiently informative about the claim.<sup>45</sup> Second, the signal  $s$  must be *publicly observed*, so that both players’ play can shift to a bad equilibrium if the responder accepts following a signal above  $s^*$ . For example, in Schelling’s parable, it is essential that the Soviets understand the Americans’ signals of their actions, so that they can see whether or not the Americans are upholding their deterrence strategy. Third, the players must be sufficiently patient, so that the threat of shifting to a bad equilibrium if  $s$  exceeds  $s^*$  is strong enough to deter aggressive claims even when the equilibrium probability that  $s > s^*$  is small. Under these conditions, any division of the pie can be sustained in equilibrium with a minimal risk of conflict: formally, for any numbers  $x \in (0, 1)$  and  $\eta > 0$ , there exists a threshold discount factor  $\bar{\delta} < 1$  such that, for every  $\delta > \bar{\delta}$ , there exists a sequential equilibrium of the repeated imperfectly-observed-claims bargaining game with discount  $\delta$  where the players’ expected per-period payoffs are within a distance  $\eta$  of the pair

---

<sup>45</sup>For example, this is the case if  $x = s + \varepsilon$ , where  $\varepsilon$  is an independent normal random variable of fixed variance. In general, what is required is that if the density of the signal  $s$  conditional on the claim  $a$  is given by  $f(s|a)$ , then  $\lim_{s \rightarrow +\infty} f_a(s|a)/f(s|a) = \infty$ . The relevant limit is the one as  $s \rightarrow +\infty$  rather than  $-\infty$  because higher signals are assumed to be more informative of aggressive claims, and the relevant signals are the ones that most strongly indicate an aggressive claim.

$(x, 1 - x)$ . This result is a variant of the folk theorem for repeated games with imperfect public monitoring (Fudenberg, Levine, and Maskin, 1994).

The situation is very different if the signal is observed only by the responder and there is room for misunderstanding. This would be the case, for example, if the Soviets do not understand the Americans' signals and the Americans do not recognize this. In this case, the claimant cannot tell when the responder's signal exceeds  $s^*$ , so the responder is tempted to accept the claim even after unfavorable signal realizations (as she understands that, in equilibrium, these signal realizations are due to "bad luck" rather than truly aggressive claims). However, such behavior by the responder undermines deterrence. More precisely, it can be shown that when the signal is observed only by the responder, there is no equilibrium within a large class of strategies where the responder obtains any positive surplus. This finding is in the spirit of earlier impossibility results for repeated games with imperfect private monitoring (e.g., Matsushima, 1991). The contrast between the prospects of dynamic deterrence with public and private signals is another example where misperceptions need not lead to conflict as long as their occurrence is commonly understood by both parties, while private misperceptions that cause misunderstandings inevitably do lead to conflict.

## 6 Conclusion

This chapter has reviewed, extended, and applied some canonical models of imperfect information and conflict dynamics. We argue that "3Ms"—mistrust, misperception, and misunderstanding—are critical to the study of conflict dynamics. *Mistrust*—incomplete information regarding an adversary's preferences or capabilities—can lead to conflict in settings with a first-mover advantage, such as the classic *security dilemma*, where each party takes an aggressive action out of fear that the other party is doing the same. *Misperceptions*—imperfect observation of an adversary's action—can amplify mistrust and thus increase conflict risk in dynamic interactions. Finally, *misunderstanding*—uncertainty about an adversary's observations or perceptions—is essential for understanding more complex and realistic conflict dynamics, including conflict spirals, traps, and cycles.

Of the "3Ms," mistrust has been the main emphasis of the conflict literature. Misper-

ceptions have played an important role in discussions of the causes of several major wars, but have been less often incorporated into formal models of conflict. Misunderstanding as a major determinant of the dynamics of conflict has been largely ignored in this literature. This chapter has stressed the central roles of misperception and misunderstanding in the onset, continuation, and cessation of conflict. We have also shown how the same forces determine conflict risk and dynamics in *deterrence* (or *inspection*) games.

Our main aim in this chapter has been theoretical—clarifying the distinct and complementary roles of mistrust, misperception, and misunderstanding in driving conflict onset and dynamics. Nevertheless, we hope that a better appreciation of these forces will also be of use for more historical and empirical research in this area. Although there are already some empirical studies suggesting that mistrust, misperception, and misunderstanding can be important factors in conflict, much work remains to be done. From a theoretical viewpoint, we believe that further work on conflict dynamics in the presence of misperceptions and misunderstandings can generate sharper predictions about the form, severity, and dynamics of war and other conflicts. One important avenue here is to extend these insights from the simple models considered in this chapter to richer settings within the broader political economy area, such as conflict and cooperation between ethnic groups, political parties, or interest groups.

Another important direction for future research is theoretical, historical, or empirical work elucidating different kinds of misperception and misunderstanding and their implications. For simplicity, we have focused on the simplest kind of misunderstanding, where each party is completely unaware of any misperception by the other party. Alternative, richer forms of misunderstanding (i.e., divergent higher-order beliefs) can result from, for example, miscommunication or differing interpretations of history. Another important area is to incorporate within-group heterogeneity into the theory of conflict (e.g., different attitudes towards war directly resulting from preference heterogeneity or because of the consequences of conflict for other outcomes, such as redistribution). Finally, another set of important questions concern a more in-depth study of how international organizations and other third-party mediators can help ameliorate the effects of misperception and misunderstanding.

# A Appendix: Omitted Proofs

## A.1 Proof of Proposition 3

Since bad players always take  $H$ , we need only consider normal players' strategies. We first note that if the second-mover observes  $s_t = D$  in any period  $t$ , she learns with certainty that her opponent is normal (since  $s_t = D$  can result only if  $a_t^1 = D$ , and  $a_t^1 = D$  can be played only if the period- $t$  first-mover is normal). This player's belief  $\mu$  that her opponent is bad is then 0 in every future period, which by Lemma 1 implies that she plays  $D$  in every future period. Given this, if the first-mover observes  $s_t = D$  in some period  $t$ , he knows that his opponent will take  $D$  in every future period if she is normal, and will take  $H$  in every future period if she is bad, so that his own optimal continuation play is determined by Bayes' rule as in Lemma 1. Moreover, conditional on the event  $s_t = D$ , it is clear that the players eventually coordinate on the dovish outcome with probability 1 in the event that both of them are normal, because the period- $t$  second-mover takes  $D$  forever (in periods where she is the first-mover), and this eventually generates another  $D$  signal, which causes her opponent to start taking  $D$  forever as well. It thus remains only to determine the normal players' strategies at histories where the signal has equalled  $H$  in every period, and to show that these strategies eventually lead to a  $D$  signal arising with probability 1. Let  $E_T$  denote this event:  $E_T = \{s_t = H \forall t \leq T - 1\}$ .

We argue by induction that for every period  $T < 2n^*$ , conditional on the event  $E_T$  the first-mover believes that her opponent is bad with probability  $\mu_T$ , and moreover  $\mu_{\lfloor T/2 \rfloor} < \mu^{\text{seq}}$  so the first-mover takes  $a_T^1 = D$  by Lemma 1.<sup>46</sup> This holds trivially for the first period ( $T = 1$ ), since the first-mover's belief is equal to the prior,  $\mu_0$ . Subsequently, if  $T < 2n^*$  is even, the first-mover (player 2) has observed  $T/2 = \lfloor T/2 \rfloor$  periods  $t$  where  $s_t = H$ , yet the strategy of the normal type of player 1 prescribed  $a_t^1 = D$ , so the first-mover's belief is  $\mu_{\lfloor T/2 \rfloor}$ . If instead  $T < 2n^*$  is odd, the first-mover (player 1) has observed  $(T - 1)/2 = \lfloor T/2 \rfloor$  periods  $t$  where  $s_t = H$ , yet the strategy of the normal type of player 2 prescribed  $a_t^1 = D$ , so the first-mover's belief is again  $\mu_{\lfloor T/2 \rfloor}$ . Moreover, by construction  $\mu_{\lfloor T/2 \rfloor} < \mu^{\text{seq}}$  for all  $T < 2n^*$ . This completes the inductive argument, and thus the description of the normal

---

<sup>46</sup>Here  $\lfloor \cdot \rfloor$  denotes the round-down function.

players' strategies at periods  $T < 2n^*$ .

Now, at period  $T = 2n^*$ , conditional on the event  $E_T$  the first-mover (who is player 2, since  $T$  is even) has belief  $\mu_{n^*} > \mu^{\text{seq}}$ . By Lemma 1, player 2 then takes  $a_T^1 = H$ . Moreover, player 2 continues to take  $H$  whenever he is the first-mover, so long as the signal  $s_t$  always equals  $H$  when he is the second-mover, because such observations can only increase his belief that player 1 is bad. Thus, conditional on the event  $E_t$ , player 2's play starting in period  $T$  is the identical whether he is the normal type or the bad type. This implies that player 1's belief remains fixed at  $\mu_{n^*-1}$ , and therefore player 1 continues to take  $D$  in all periods  $t$  where she is the first mover, so long as the signal  $s_t$  always equals  $H$ . Almost surely, this behavior eventually leads to the signal taking value  $D$ , at which point the players' continuation strategies are determined as above, and eventually lead to coordination on  $(D, D)$  with probability 1.

## A.2 Proof of Proposition 4

Consider an arbitrary sequential equilibrium, and consider the event that both players are normal. Since  $\mu_0 < \mu^{\text{seq}}$ , player 1 takes  $a_1^1 = D$  in period 1. With probability  $1 - \pi$ ,  $s_1 = D$ , in which case player 2 takes  $a_2^1 = D$  in period 2. With probability  $\pi$ ,  $s_1 = H$ , in which case player 2's belief that player 1 is bad equals  $\rho_1$ . Since  $\rho_1 > \rho_2$  and we assume that  $\rho_2 > \mu^{\text{seq}}$ , we have  $\rho_1 > \mu^{\text{seq}}$ , so player 2 takes  $a_2^1 = H$  in period 1 following  $s_1 = D$  (by Lemma 1). Hence, if  $s_1 = D$  then  $s_2 = H$ , and in this event player 1's belief that player 2 is bad equals  $\rho_2$  (as explained in the text). Since  $\rho_2 > \mu^{\text{seq}}$ , player 1 then takes  $a_3^1 = H$  in period 3, and subsequently, with probability 1,  $s_t = H$  and  $(a_t^1, a_t^2) = (H, H)$  in every period, since each player's belief that the opponent is bad remains constant in periods where the opponent is known to take  $H$  regardless of their type. In total, whenever a misperception occurs in the first period, the first-mover takes  $H$  in every subsequent period. Moreover, if in addition  $\rho_2 > \mu^{\text{sim}}$ , in this event the second-mover also takes  $H$  in every period other than first (by the second part of Lemma 1).

## References

- [1] Acemoglu, Daron, and Alexander Wolitzky. "Cycles of conflict: An economic model." *American Economic Review* 104.4 (2014): 1350-67.
- [2] Acharya, Avidit, and Kristopher W. Ramsay. "The calculus of the security dilemma." *Quarterly Journal of Political Science* 8.2 (2013): 183-203.
- [3] Aghion, Philippe, et al. "Optimal learning by experimentation." *The review of economic studies* 58.4 (1991): 621-654.
- [4] Avenhaus, Rudolf, Bernhard Von Stengel, and Shmuel Zamir. "Inspection games." *Handbook of game theory with economic applications* 3 (2002): 1947-1987.
- [5] Axelrod, Robert, *The Evolution of Cooperation*. Vol. 5145. Basic Books (AZ), 1984.
- [6] Bagwell, Kyle. "Commitment and Observability in Games." *Games and Economic Behavior* 8.2 (1995): 271-280.
- [7] Baliga, Sandeep, and Tomas Sjöström. "Arms races and negotiations." *The Review of Economic Studies* 71.2 (2004): 351-369.
- [8] Baliga, Sandeep, and Tomas Sjöström. "Strategic ambiguity and arms proliferation." *Journal of political Economy* 116.6 (2008): 1023-1057.
- [9] Baliga, Sandeep, and Tomas Sjöström. "The strategy of manipulating conflict." *American Economic Review* 102.6 (2012): 2897-2922.
- [10] Baliga, Sandeep, Ethan Bueno De Mesquita, and Alexander Wolitzky. "Deterrence with imperfect attribution." *American Political Science Review* 114.4 (2020): 1155-1178.
- [11] Ball, Ian and Jan Knoepfle, "Should the Timing of Inspections be Predictable?" *Working paper (2023)*.
- [12] Banerjee, Abhijit V. "The economics of rumours." *The Review of Economic Studies* 60.2 (1993): 309-327.
- [13] Besley, Timothy, and Marta Reynal-Querol. "The legacy of historical conflict: Evidence from Africa." *American Political Science Review* 108.2 (2014): 319-336.
- [14] Blackwell, David. "The Comparison of Experiments" in *Proceedings, Second Berkeley Symposium on Mathematical Statistics and Probability*, e.d Jerzy Neyman, 93-102. University of California Press: Berkeley. 1951.
- [15] Blattman, Christopher. *Why We Fight: The Roots of War and the Paths to Peace*. Penguin, 2022.
- [16] Brito, Dagobert L., and Michael D. Intriligator. "Conflict, war, and redistribution." *American political science review* 79.4 (1985): 943-957.



- [17] Butterfield, Herbert. *History and human relations*. (1951).
- [18] Coe, Andrew J., and Jane Vaynman. "Why arms control is so rare." *American Political Science Review* 114.2 (2020): 342-355.
- [19] Coy, Peter. "What Game Theory Says about China's Strategy," *New York Times*, Sept. 13, 2021.
- [20] Chassang, Sylvain, and Gerard Padró I. Miquel. "Conflict and deterrence under strategic risk." *The Quarterly Journal of Economics* 125.4 (2010): 1821-1858.
- [21] Cremér, Jacques. "Arm's Length Relationships," *Quarterly Journal of Economics* 110 (2): 275-295.
- [22] Debs, Alexandre, and Jessica Chen Weiss. "Circumstances, domestic audiences, and reputational incentives in international crisis bargaining." *Journal of Conflict Resolution* 60.3 (2016): 403-433.
- [23] Debs, Alexandre, and Nuno P. Monteiro. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68.1 (2014): 1-31.
- [24] DellaVigna, Stefano, et al. "Cross-border media and nationalism: Evidence from Serbian radio in Croatia." *American Economic Journal: Applied Economics* 6.3 (2014): 103-132.
- [25] Denti, Tommaso, Massimo Marinacci, and Aldo Rustichini. "Experimental Cost of Information." IGER, Università Bocconi, 2021.
- [26] Dilmé, Francesc, and Daniel F. Garrett. "Residual deterrence." *Journal of the European Economic Association* 17.5 (2019): 1654-1686.
- [27] Easley, David, and Nicholas M. Kiefer. "Controlling a stochastic process with unknown parameters." *Econometrica: Journal of the Econometric Society* (1988): 1045-1064.
- [28] Eeckhout, Jan, Nicola Persico, and Petra E. Todd. "A theory of optimal random crack-downs." *American Economic Review* 100.3 (2010): 1104-1135.
- [29] Fearon, James D. "Domestic political audiences and the escalation of international disputes." *American political science review* 88.3 (1994): 577-592.
- [30] Fearon, James D. "Rationalist explanations for war." *International organization* 49.3 (1995): 379-414.
- [31] Fearon, James D. "Bargaining, enforcement, and international cooperation." *International organization* 52.2 (1998): 269-305.
- [32] Fearon, James D. "Two states, two types, two actions." *Security studies* 20.3 (2011): 431-440.
- [33] Fearon, James D. "Coups, police shootings, and nuclear war." *Working paper* (2020).

- [34] Fey, Mark, and Kristopher W. Ramsay. "Mutual optimism and war." *American Journal of Political Science* 51.4 (2007): 738-754.
- [35] Fey, Mark, and Kristopher W. Ramsay. "When is shuttle diplomacy worth the commute? Information sharing through mediation." *World Politics* 62.4 (2010): 529-560.
- [36] Fudenberg, Drew, David Levine, and Eric Maskin. "The folk theorem with imperfect public information." *Econometrica* 62.5 (1994): 997-1039.
- [37] Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. "Crime and social interactions." *The Quarterly journal of economics* 111.2 (1996): 507-548.
- [38] Glaser, Charles L. "Political consequences of military strategy: Expanding and refining the spiral and deterrence models." *World politics* 44.4 (1992): 497-538.
- [39] Glaser, Charles L. *Rational theory of international politics*. Princeton University Press, 2010.
- [40] Goltsman, M., Hörner, J., Pavlov, G., & Squintani, F. (2009). Mediation, arbitration and negotiation. *Journal of Economic Theory*, 144(4), 1397-1420.
- [41] Hartlyn, Jonathan. *The Politics of Coalition Rule in Colombia*. Cambridge University Press: New York. 1988.
- [42] Herz, John H. "Idealist internationalism and the security dilemma." *World politics* 2.2 (1950): 157-180.
- [43] Hobbes, Thomas, *Leviathan*. 1651.
- [44] Hörner, Johannes, Massimo Morelli, and Francesco Squintani. "Mediation and peace." *The Review of Economic Studies* 82.4 (2015): 1483-1501.
- [45] Horowitz, Donald L. *Ethnic groups in conflict*, updated edition with a new preface. Univ of California Press, 2000.
- [46] Jackson, Matthew O., and Massimo Morelli. "Strategic Militarization, Deterrence and Wars." *Quarterly Journal of Political Science* 4.4 (2009): 279-313.
- [47] Jervis, Robert. *Perception and misperception in international politics*. Princeton University Press, 1976.
- [48] Jervis, Robert. "Cooperation under the security dilemma." *World politics* 30.2 (1978): 167-214.
- [49] Kaplan, Robert D. *Asia's cauldron: The South China Sea and the end of a stable Pacific*. Random House Trade Paperbacks, 2015.
- [50] Kapon, Sam. "Dynamic amnesty programs." *American Economic Review* 112.12 (2022): 4041-75.

- [51] Keohane, Robert O., and Robert O. Keohane. *After hegemony*. Princeton university press, 2005.
- [52] Kydd, Andrew. "Game theory and the spiral model." *World Politics* 49.3 (1997): 371-400.
- [53] Kydd, Andrew. "Trust, reassurance, and cooperation." *International Organization* 54.2 (2000): 325-357.
- [54] Kydd, Andrew. "Which side are you on? Bias, credibility, and mediation." *American Journal of Political Science* 47.4 (2003): 597-611.
- [55] Kydd, Andrew H. *Trust and mistrust in international relations*. Princeton University Press, 2005.
- [56] Kydd, Andrew H. "When can mediators build trust?." *American Political Science Review* 100.3 (2006): 449-462.
- [57] Kydd, Andrew H. *International relations theory*. Cambridge University Press, 2015.
- [58] Lagunoff, Roger, and Akihiko Matsui. "Asynchronous choice in repeated coordination games." *Econometrica: Journal of the Econometric Society* (1997): 1467-1477.
- [59] Lazear, Edward P. "Speeding, terrorism, and teaching to the test." *The Quarterly Journal of Economics* 121.3 (2006): 1029-1061.
- [60] Mailath, George J., and Larry Samuelson. "Who wants a good reputation?." *The Review of Economic Studies* 68.2 (2001): 415-441.
- [61] Matsushima, Hitoshi. "On the theory of repeated games with private information: Part I: anti-folk theorem without communication." *Economics Letters* 35.3 (1991): 253-256.
- [62] Mearsheimer, John J. *The tragedy of great power politics*. WW Norton & Company, 2001.
- [63] Meirowitz, Adam, and Anne E. Sartori. "Strategic uncertainty as a cause of war." *Quarterly Journal of Political Science* 3.4 (2008): 327-352.
- [64] Meirowitz, Adam, Massimo Morelli, Kristopher W. Ramsay, and Francesco Squintani. "Dispute resolution institutions and strategic militarization." *Journal of Political Economy* 127.1 (2019): 378-418.
- [65] Michalopoulos, Stelios, and Elias Papaioannou. "The long-run effects of the scramble for Africa." *American Economic Review* 106.7 (2016): 1802-1848.
- [66] Myerson, Roger B. "Optimal coordination mechanisms in generalized principal-agent problems." *Journal of mathematical economics* 10.1 (1982): 67-81.
- [67] Nunn, Nathan, and Leonard Wantchekon. "The slave trade and the origins of mistrust in Africa." *American Economic Review* 101.7 (2011): 3221-3252.

- [68] Ortner, Juan, and Sylvain Chassang. "Making corruption harder: Asymmetric information, collusion, and crime." *Journal of Political Economy* 126.5 (2018): 2108-2133.
- [69] Oye, Kenneth A. *Cooperation under anarchy*. Princeton University Press, 1986.
- [70] Pei, Harry, and Bruno Strulovici. "Crime Aggregation, Deterrence, and Witness Credibility." *Working paper* (2021).
- [71] Phelan, Christopher. "Public trust and government betrayal." *Journal of Economic Theory* 130.1 (2006): 27-43.
- [72] Powell, Robert. *Nuclear deterrence theory: The search for credibility*. Cambridge University Press, 1990.
- [73] Powell, Robert. "Uncertainty, Shifting Power, and Appeasement." *American Political Science Review* (1996a): 749-764.
- [74] Powell, Robert. "The inefficient use of power: Costly conflict with complete information." *American Political science review* 98.2 (2004): 231-241.
- [75] Powell, Robert. "War as a commitment problem." *International organization* 60.1 (2006): 169-203.
- [76] Ramsay, Kristopher W. "Information, uncertainty, and war." *Annual Review of Political Science* 20 (2017): 505-527.
- [77] Rauchhaus, Robert W. "Asymmetric information, mediation, and conflict management." *World Politics* 58.2 (2006): 207-241.
- [78] Ravid, Doron. "Ultimatum Bargaining with Rational Inattention." *American Economic Review* 110.9 (2020): 2948-63.
- [79] Rohner, Dominic, Mathias Thoenig, and Fabrizio Zilibotti. "War signals: A theory of trade, trust, and conflict." *Review of Economic Studies* 80.3 (2013): 1114-1147.
- [80] Rohner, Dominic, Mathias Thoenig, and Fabrizio Zilibotti. "Seeds of distrust: Conflict in Uganda." *Journal of Economic Growth* 18 (2013): 217-252.
- [81] Rousseau, Jean-Jacques, *A Lasting Peace through the Federation of Europe*, translated by C. E. Vaughan (1782).
- [82] Safford, Frank, and Marco Palacios. *Colombia: Fragmented Land, Divided Society*. Oxford University Press: New York. 2002.
- [83] Schelling, Thomas C. *The Strategy of Conflict*. Harvard University Press, Cambridge, 1960.
- [84] Schelling, Thomas C. *Arms and influence*. Yale University Press, New Haven, 1966.
- [85] Slantchev, Branislav L., and Ahmer Tarar. "Mutual optimism as a rationalist explanation of war." *American Journal of Political Science* 55.1 (2011): 135-148.

- [86] Smith, Alastair, and Allan Stam. "Mediation and peacekeeping in a random walk model of civil and interstate war." *International Studies Review* 5.4 (2003): 115-135.
- [87] Thucydides, *The Project Gutenberg eBook of The History of the Peloponnesian War*, translated by Richard Crawley (2003).
- [88] Tsebelis, George. "The abuse of probability in political analysis: The Robinson Crusoe fallacy." *American Political Science Review* 83.1 (1989): 77-91.
- [89] Tuchman, Barbara W. *The Guns of August: The Outbreak of World War I*. Random House, 1962.
- [90] Van Evera, Stephen. *Causes of War: Power and the Roots of Conflict*. Cornell University Press, 1999.
- [91] Varas, Felipe, Iván Marinovic, and Andrzej Skrzypacz. "Random inspections and periodic reviews: Optimal dynamic monitoring." *The Review of Economic Studies* 87.6 (2020): 2893-2937.
- [92] Waltz, Kenneth N. *Theory of international politics*. Waveland Press, 1979.
- [93] Wolitzky, Alexander. "Unobserved-Offers Bargaining," *American Economic Review*, Forthcoming (2022).
- [94] Yildiz, Muhamet. "Waiting to persuade." *The Quarterly Journal of Economics* 119.1 (2004): 223-248.