Political Instability and the Failure of Deterrence*

Livio Di Lonardo† Scott A. Tyson‡

Abstract

We develop a model of an international crisis between a country seeking to maintain a peaceful status quo (Defender), and a potential aggressor (Attacker). We introduce a key novelty, namely, that Attacker's leadership is politically insecure, and thus, may be unseated by domestic elites. Leaders and elites can each be hawkish types, who benefit from conflict, or dovish types, who prefer peace. We show that the ability to maintain peace through deterrence crucially depends on ideological cohesiveness between leaders and elites in Attacker countries. When there is ideological disagreement, we identify two novel mechanisms that lead the conventional logic of deterrence to break down. First, political instability can break the link between a leader's aggressive actions and Defender's retaliatory response. Second, political instability creates a commitment problem which ultimately leads doves to initiate international crises as a way to quell domestic conflicts. We show that asymmetric information between Defender and members of Attacker exacerbate these problems so severely that Defender would be better off committing to a strategy of complete inaction.

Keywords: Deterrence; Political Instability; Foreign Policy
JEL: D74, D82

*We thank Muhammet Bas, Eli Berman, Chris Bidner, Ethan Bueno de Mesquita, Micael Castanheira, Andrew Coe, Christian Davenport, Alex Debs, Eric Dickson, John Duggan, Tiberiu Dragu, Jon Eguia, Chris Fariss, Jim Fearon, Mark Fey, Scott Gehlbach, Hein Goemans, Tasos Kalandrakis, Leyla Karakas, Andrew Kydd, David Lake, Andrew Little, Jim Morrow, David Myatt, Roger Myerson, Gerard Padró i Miquel, Jack Paine, Carlo Prato, Adam Przeworski, Davin Raiha, Kris Ramsay, Peter Rosendorff, Mehdi Shadmehr, Jake Shapiro, Ken Schultz, Branislav Slantchev, Al Slivinski, Alastair Smith, Randy Stone, Jessica Sun, Roya Talibova, Richard Van Weelden, Stephane Wolton, Cathy Wu, participants at the 2015 APSA Meeting, the 2016 SPSA Meeting, seminar audiences at Bocconi, LSE, Michigan, NYU, Princeton, University of Texas at Austin, Rochester, Warwick, seminar participants at the 3^{rd} Annual Formal Theory and Comparative Politics Conference, the Inaugural Great Lakes Political Economy Conference, and the Formal Models of International Relations Conference for invaluable comments and discussions.

†Assistant Professor, Department of Social and Political Sciences and Carlo F. Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University. Contact information: livio.dilonardo@unibocconi.it

‡Assistant Professor, Department of Political Science, University of Rochester, and Research Associate, W. Allen Wallis Institute of Political Economy, University of Rochester. styson2@ur.rochester.edu.

1 Introduction

For over 30 years, deterrence represented the cornerstone of U.S. foreign policy (Schelling 1960, 1966; Freedman 2013), but when the Cold War came to an abrupt close, U.S. policymakers had to rethink their approach. In 1995, the U.S. Strategic Command issued a memorandum titled "Essentials of Post-Cold War Deterrence," which acknowledged that deterrence had become harder to apply and stressed the need to better understand the conditions that make deterrence work. The memorandum also argued that deterrence remained essential to U.S. foreign policy, as is indicated by its central role in both the Bush Doctrine as well as the national security policy of the Trump administration.[1]

The theoretical foundations of deterrence are fundamentally strategic, and their study was influential in the early development of game theory (Ellsberg 1961, 1968; Schelling 2006; Myerson 2009). The classical deterrence problem proceeds through three stages, provocation, retaliation, and escalation, and has been formulated as an incentive design problem. In particular, Defender (principal) wants to dissuade Attacker (agent) from taking provocative actions, and to do so, Defender needs to convey that she will retaliate following provocation. Furthermore, retaliation needs to be costly enough that provocation ceases to be desirable for Attacker, and this gives rise to the conventional logic of deterrence: as long as retaliation is credible and costly enough, there will be no provocation in the first place. This perspective has yielded important insights on the credibility of retaliatory threats (Schelling 1960; Powell 1989, 1990, 1985; Gurantz and Hirsch 2017), effective communication between parties (Bueno de Mesquita, Morrow and Zorick 1997; Baliga and Sjöström 2008), imperfect

[1]See the 2017 National Security Strategy of the United States of America: `secuirthttps://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905-2.pdf`, or "Trump's 'fire and fury' remark was improvised but familiar", Jeff Zeleny, Dan Merica and Kevin Liptak, CNN, `edition.cnn.com/2017/08/09/politics/trump-fire-fury-improvise-north-korea`, or "Trump Threatens 'Fire and Fury' Against North Korea if It Endangers U.S.", *New York Times*, Peter Baker and Choe Sang-Hunsee, `www.nytimes.com/2017/08/08/world/asia/north-korea-un-sanctions-nuclear-missile-united-nations`.

attribution of blame (Baliga, Bueno de Mesquita and Wolitzsky 2018), and the development and maintenance of military capacity (Schelling 1966, 1967; Zagare 2004).

Yet, studies of deterrence often ignore one of the most prominent features of political rule, namely that political leaders rarely enjoy perfect security in office.[2] Although leaders face different mechanisms of political turnover, they are all subject to the threat of removal. Even authoritarian leaders, who are not subject to formal accountability via competitive elections, often face threats from ambitious elites within the government. This has profound consequences on international politics, because new leaders may have a different resolve to fight wars (Wolford 2007, 2012), and because evidence suggests that political insecurity in autocracies fuels the emergence of international crises (Chiozza and Goemans 2011).

To assess the influence of political instability on deterrence, we develop a model that incorporates two novel features into the classical deterrence problem: (1) political insecurity of the attacker country's leadership, and (2) ideological disagreement between the leader and domestic political elites within Attacker. To capture political instability formally, we represent Attacker as being composed of two distinct decision-makers, Leader and Elite, where Leader relies on Elite's support to maintain power (e.g., Bueno de Mesquita, Smith, Siverson and Morrow 2003; Baliga, Lucca and Sjöström 2011). We model the presence of political disagreements by assuming that both Leader and Elite can be either a hawk, who finds conflict beneficial, or a dove, who supports the peaceful status quo.[3]

Our main model proceeds through four stages. In the first stage, Leader can take aggressive actions that provoke a crisis. In the second stage, which is novel to our model, Elite can seize power from Leader, regardless of what occurred in the first stage. In the next stage, Defender can take retaliatory actions that are costly for all players to endure, such as

[2]The importance of political instability was stressed in the 1995 memorandum, and has also been noted by traditional international relations scholars (e.g., Jervis 1979).

[3]We abstract from what drives their political disagreement, which can arise from previous military investments (Jackson and Morelli 2009) or audience costs (Ashworth and Ramsay 2010).

military strikes or economic sanctions. In the final stage, if a conflict has been provoked, the ruler of the attacker country (Leader or Elite) can escalate the conflict, potentially to war.

In our first result we formalize the conventional logic of deterrence, and show that it holds as long as there is no political instability or ideological disagreement within Attacker. We do this by considering three distinct scenarios. In the first, we assume that Leader is hawkish and perfectly secure (c.f., Wagner 1992; Signorino and Tarar 2006). In the second, Leader is perfectly secure but her type is privately known. In the third scenario, we suppose that Leader relies on the support of Elite (political instability is present), and it is commonly known that both Leader and Elite are hawks (ideological disagreement is absent). In each of these three scenarios, the conventional logic of deterrence—no Leader type takes provocative actions because of the threat of retaliation—follows from two incentive conditions. The first is the *credibility constraint*, which ensures that retaliation is incentive compatible for Defender following provocation. The second is the *capability constraint*, which guarantees that retaliation is costly enough to dissuade Attacker's leader from taking provocative actions.

We then show that the introduction of political instability and ideological disagreement together undermine the conventional logic of deterrence. In particular, the actors who compose the attacker country are now involved in two different conflicts: one between themselves at the domestic level and another at the international level with Defender. Although these conflicts are distinct, they become strategically linked due to a novel tradeoff faced by elites in ideologically divided attacker countries. In particular, domestic elites who want to seize power must consider that doing so may trigger retaliation. This tradeoff gives rise to a novel incentive condition, the *salience constraint*, which outlines when retaliation is severe enough that the shared desire of avoiding retaliation becomes a more salient political issue than domestic disputes between Leader and Elite. Thus, the salience constraint outlines when deterrence policies create a common ground among domestic political factions who would otherwise clash.

Our main results show that when the credibility and salience constraints are satisfied, the presence of political instability and ideological disagreement cause the conventional logic of deterrence to fail. Depending on the political position of doves within attacker countries, deterrence can fail in two distinct ways. First, when doves are not initially in a leadership position, their ability to unseat a hawkish Leader undermines the credibility of retaliatory threats. When the individual who provokes a conflict, Leader, is replaced by a dovish Elite (who will not escalate the conflict), then the defender country no longer wants to retaliate. Leader can thus avoid the costs associated with retaliation, and the direct connection between provocation and retaliation is broken. We call this friction the *instability curse*, since the presence of a dovish Elite, which seemingly should be a force for peace, turns out to be a source of deterrence failure.

Second, when doves hold a leadership position, they provoke a crisis as a way of motivating domestic support from hawkish elites. Without provocation, there is no threat of an escalating conflict, and because a new hawkish ruler would not be a threat, Defender is not concerned with Attacker's domestic politics. To maintain power, a dovish Leader needs to coerce Defender into protecting her interests, which she accomplishes by provoking a crisis, thus ensuring that escalation is a concrete threat if a hawk seizes power. Consequently, Defender retaliates following a leadership change in the attacker country. Hawkish elites, anticipating this turn of events, forgo efforts to seize power from dovish leaders when the threat of retaliation is credible and salient. We call this the *deterrence curse*, as deterrence fails precisely because Leader manipulates the threat of retaliation *to deter* hawkish elites from challenging her power.

We complete our analysis by introducing asymmetric information between Defender and members of attacker countries. We show that Defender cannot learn whether the *ruler* of Attacker poses a threat from the actions of Leader and Elite alone. Because of this uncertainty, the possibility that the ruler might be a dove weakens the credibility of retaliatory

threats, and undermines deterrence. Members of the attacker country, who want to avoid retaliation, are aware that Defender's uncertainty discourages retaliation, and thus choose their actions so as to maintain this uncertainty when the credibility and salience constraints are satisfied. These incentives to block information imply that when deterrent threats are credible and salient, there are two pure-strategy equilibria, and deterrence fails in both of them. In one, hawkish leaders provoke a crisis but are immediately removed, regardless of the type of domestic elites. In the other, all leader types provoke an international crisis, and all elite types support them. We assess the severity of these deterrence failures by focusing on Defender's welfare and consider an alternative strategy where defender commits to complete inaction, thus never using retaliation in response to provocation. In an interesting contrast with conventional accounts of deterrence where commitment to aggression typically improves deterrence, we show that Defender is better off committing to inaction than in either equilibrium with asymmetric information. Although without the threat of retaliation, attacker countries composed purely of hawks will provoke and escalate crises, removing the threat of retaliation removes a salient political dimension between domestic factions that would otherwise prompt more aggressive actions overall.

Our model extends the standard crisis bargaining framework that has become a workhorse model of coercive diplomacy in international relations (e.g., Powell 1987; Morrow 1989; Kilgour and Zagare 1991; Wagner 1992; Fearon 1994; Schultz 1998; Smith 1998; Signorino 1999; Schultz 2001; Signorino and Tarar 2006; Chapman and Wolford 2010; Fey and Ramsay 2011; Sechser and Fuhrmann 2013; Fey, Meirowitz and Ramsay 2013). Our model builds on this framework by incorporating the possibility that the attacker country is not a unitary actor, but comprises two factions vying for control. Domestic politics, which plays a key role in our theory, is an important factor of international conflict. Schultz (2001) studies a defender country with a perfectly secure incumbent, and highlights how opposition parties can enhance deterrence because their supportive pronouncements signal an incumbent's resolve.

Highlighting the role of transparency, Kurizaki (2007) examines how crisis outcomes depend on whether countries can keep international crises secret from domestic audiences. While these models introduce domestic conflicts within the defender country, our model looks at domestic conflicts within the attacker, and shows how political instability affects the defender's ability to prevent crises through deterrent threats.

Our results contribute to a large influential literature on deterrence (Schelling 1960, 1966, 1967; Powell 1989, 1990, 1985). More recently, Gurantz and Hirsch (2017) show how the threat of a major war can deter even minor transgressions when there is a possibility that Attacker prefers war to the status quo, making a future war inevitable. Baliga, Bueno de Mesquita and Wolitzsky (2018) focus on imperfect attribution of provocation, demonstrating that when it is hard to detect an attack or identify the perpetrator, deterrence becomes harder to achieve, and moreover, improving detection or attribution in isolation can hinder deterrence even further. Baliga and Sjöström (2008) examine "deterrence by doubt" where a country considering the development of new military technologies benefits most by engendering uncertainty that leaves Defender unsure of the current military balance.[4] Kydd (2000) interprets deterrence as investments today that positively affect bargains struck tomorrow, while Kydd and McManus (2017) formally distinguish deterrence from assurance and how these differences affect diplomatic relations. Zagare and Kilgour (1993) analyze a model where both Defender and Attacker may be hawkish, and find that under two-sided incomplete information deterrence equilibria exist for any positive probability that both players are hawkish.[5]

Deterrence is also a critical component in "spiral" models that focus on how the creation of arms leads to mutual distrust among countries that can prompt an arms race (Jervis

[4]Meirowitz and Sartori (2008) show that states may have incentives to create informational asymmetries (using mixed strategies) which may lead to a conflict.

[5]Nalebuff (1991) studies implications of equilibrium refinements with respect to signaling resolve in a deterrence scenario.

1978; Kydd 1997; Baliga and Sjöström 2004). Baliga and Sjöström (2012) show how hawkish provocateurs can manipulate fears by sending cheap-talk messages to raise the likelihood of conflict. Chassang and Padró i Miquel (2010) use an exit game to analyze how deterrence is affected by strategic risk (i.e. uncertainty over actions that persists in equilibrium).

2 The Model

The classical deterrence problem depicts an interaction between two countries: a Defender, D, and an Attacker, A. We depart from standard models of deterrence and assume that while Defender is a single decisionmaker, Attacker is comprised of two decisionmakers, an incumbent leader, L, and an elite, E. Each member of Attacker can be one of two types, a *hawk* or a *dove*. Hawks desire (or are willing to tolerate) conflict with Defender, whereas doves prefer to avoid conflict. Moreover, the ideological disagreement between hawks and doves is not limited to international issues, but also involves domestic concerns. The game proceeds in four stages: (1) a provocation stage; (2) a domestic politics stage that decides who rules Attacker; (3) a retaliation stage; and (4) an escalation stage in which the ruler of Attacker (either L or E) can escalate the conflict (potentially to war), or back down, thereby ending the international crisis.

Actions and Timing: In the first stage of the game, Leader in Attacker chooses whether to take a provocative action, denoted by $m = 1$, or maintain the status quo, $m = 0$. In the second stage, Elite decides whether to support Leader, denoted by $s = 1$, or seize control of A, denoted by $s = 0$. If E seizes control from L, then E becomes the *ruler*, whereas if E supports L, then L remains the ruler.

In the third stage, and before the ruler of Attacker can escalate the crisis, Defender has the opportunity to retaliate against A, regardless of whether Leader or Elite is the ruler of A. We denote D's decision to retaliate by $r = 1$, and the decision to concede to A's provocation

against the status quo by $r = 0$. In the last stage of the game, if Leader provoked a conflict, and regardless of whether D chose to retaliate, the ruler of Attacker (either L or E), can choose to escalate conflict with D, denoted by $x = 1$, or back down, $x = 0$. Escalation can include, but need not be limited to, the decision to start a war.[6]

Types and Payoffs: We denote the type of actor $j \in \{L, E\}$ by $\theta_j \in \{0, 1\}$, where $\theta_j = 1$ indicates that $\theta_j = 1$ is a hawk and $\theta_j = 0$ indicates that j is a dove. The *ideological composition* of A is the pair $\Theta = (\theta_L, \theta_E)$. When both L and E are hawks (i.e. $\theta_L = \theta_E = 1$), we say that A is *ideologically cohesive*, and we denote this by $\Theta = \Theta_C$. When θ_L and θ_E are different, we say that A is *ideologically divided*. An ideologically divided Attacker with a hawkish Leader (i.e. $\theta_L = 1$ and $\theta_E = 0$), is denoted by $\Theta = \Theta_h$, whereas an ideologically divided A with a dovish Leader (i.e. $\theta_L = 0$ and $\theta_E = 1$), is denoted by $\Theta = \Theta_d$.[7]

Doves do not benefit from conflict, whereas hawks benefit from conflict directly or, equivalently, from the rewards granted by hawks' political constituency (via, e.g., audience costs). Member i's utility is additively separable between the payoff from provocation, $\pi_1(\theta_i)$, and the payoff from escalation, $\pi_2(\theta_i)$. To reflect the difference in attitudes toward conflict between hawks and doves, we need $\pi_t(1) > \pi_t(0)$, and for convenience, we drop the dependence on type unless needed to avoid confusion, letting $\pi_t(1) = \pi_t > 0 = \pi_t(0)$ for all t. For Defender, provocation yields a cost $\lambda_1 > 0$, and an escalated conflict yields a cost $\lambda_2 > 0$.

Retaliation imposes costs on both Defender and members of Attacker's government. For Defender, retaliatory actions can involve diplomatic costs, the committal of military forces, lost gains from trade, all of which are reflected by the cost parameter $c > 0$. The benefit for Defender is that retaliatory actions reduce the disutility associated with an escalated conflict by $q \in (0, 1)$, where q represents the effectiveness of sanctions and military operations. Retaliation also imposes a cost $k > 0$ on members of Attacker, depending on how the

[6] Allowing x to be continuous would have no effect on our results.

[7] We show in Appendix C that introducing the case where L and E are doves has no effect on the results.

domestic politics stage played out. Specifically, if E has supported L, then both L and E pay a cost k, whereas if E has seized power from L, then only E is affected by retaliation.[8] The cost k is meant to reflect the influence of sanctions (military or economic) as well as the diplomatic costs associated with an escalating international incident.

The utility of Defender, as a function of the provocation decision, $m \in \{0,1\}$, the support decision, $s \in \{0,1\}$, the retaliation decision, $r \in \{0,1\}$, and the escalation decision, $x \in \{0,1\}$, is given by

$$U_D(m,s,r,x) \equiv -m\Big(\lambda_1 + (1-rq)x\lambda_2\Big) - rc.$$

The ideological disagreement between hawks and doves is not confined to their attitude towards conflict with Defender. In addition, Leader and Elite disagree on domestic issues, and hence, care about the ideological type of the actor who ends up being the ruler, which we denote by θ_R. To capture this feature, we assume that member i of Attacker receives a domestic political payoff given by

$$u(\theta_i;\theta_R) = \theta_i \cdot \theta_R + (1-\theta_i)(1-\theta_R),$$

which is equal to 1 whenever $\theta_i = \theta_R$, and 0 otherwise.

A power struggle within Attacker, along with determining who ultimately holds power, reduces the domestic political payoffs received by the new ruler of Attacker. After seizing power from Leader, Elite's utility from domestic concerns is $\delta u(\theta_i;\theta_R)$, where $\delta \in (0,1)$ represents the inefficiency of internal conflict.[9] Additionally, when L loses power, her payoff is determined only by the provocation decision, while she does not receive a payoff from

[8] Assuming that L avoids the cost of retaliation when she loses power is not essential for our results. With complete information this has no impact, and with incomplete information, we would obtain the same results through a similar, but more elaborate, equilibrium refinement argument than is presented.

[9] Domestic power struggles have no direct effect on payoffs from international politics, which is not necessary for our results, but simplifies the analysis.

Defender's retaliation decision, E's escalation decision, or domestic alignment.

To summarize, the utility of Elite is given by

$$U_E(m,s,r,x;\theta) \equiv \underbrace{\Big(s(1-\delta)+\delta\Big)u(\theta_E;\theta_R)}_{\text{Domestic Concerns}} + \underbrace{\theta_E \cdot \Big(m\Big(\pi_1(\theta_E)+(1-rq)x\pi_2(\theta_E)\Big)\Big)}_{\text{International Concerns}} - \underbrace{rk}_{\text{Retaliation Cost}},$$

and the utility of Leader is given by

$$U_L(m,s,r,x;\theta) \equiv \underbrace{s\cdot(u(\theta_L;\theta_R)-rk)}_{\text{Domestic Concerns}} + \underbrace{\theta_L \cdot \Big(m\Big(\pi_1(\theta_L)+s(1-rq)x\pi_2(\theta_L)\Big)\Big)}_{\text{International Concerns}}.$$

An equilibrium in our model is composed of (1) a provocation strategy for L given the ideological composition of Attacker; (2) a support strategy for E given L's provocation choice and the ideological composition of A; (3) a retaliation strategy for D, given the choices of L and E, and a (possibly degenerate) system of beliefs regarding the ideological composition of A; (4) an escalation strategy for the ruler in the last stage of the game, depending on her type, and the provocation, support, and retaliation choices.

In our framework, *deterrence succeeds* when L, anticipating D's retaliation, does not take provocative actions. To make it harder for deterrence to fail, we restrict attention to equilibria in which a dovish Leader, when indifferent, does not take provocative actions. Interestingly, we will show that in spite of having no direct incentive to do so, dovish types provoke a crisis with Defender because doing so creates strategic advantages.[10] We also focus our analysis on cases where the escalation of conflict is more important than domestic concerns for hawks, i.e., $\pi_2 > 1$ since the utility from domestic politics is normalized to 1.

[10]It is straightforward to consider an extension of our model in which provocation entails an opportunity cost and show that doves would strictly prefer not to provoke unless there is a large enough positive indirect benefit from doing so.

2.1 Comments on the Model

Before moving on to our analysis, we briefly discuss the structure of our model, some of our modeling choices, and how our model fits into canonical models of international conflict.

Timing and structure. Our model captures scenarios where a conflict between two countries occurs across successive stages. Between provocation and escalation there is a window of opportunity to retaliate, which mitigates the consequences of an escalated conflict. This structure mirrors canonical deterrence models and reflects a variety of conflict scenarios, including military situations that exhibit first-strike advantages, which was a prominent theme during the Cold War, confrontations stemming from the development of new weapon technologies, which is an important factor in India-Pakistani relations, and territorial conflicts such as the Six-Day War and many historical European conflicts where geographic contiguity is important (e.g., Carter 2010; Abramson and Carter 2016).

Our game ends with Attacker country's decision to escalate a conflict, perhaps even to war. Our main results are unchanged if we extend the horizon and allow for multiple opportunities to provoke a crisis, as long as all provocative actions could be escalated. If instead the game's time horizon is artificially truncated so that some provocative actions could not be escalated, then such provocative actions would not be deterrable, since retaliation in never incentive compatible without the threat of escalation.

In our analysis, political instability within Attacker emerges between the provocation and retaliation stages. Alternatively, one could consider a game where domestic political conflicts get resolved prior to the potential emergence of an international crisis, or one in which leaders enjoy complete security and domestic challenges only emerge after potential international crises have been settled. In either case, because L is perfectly secure during international crises, domestic political concerns cannot affect the conventional logic of deterrence. Since our analysis seeks to clarify the conditions under which political instability influences the logic of deterrence, we focus only on cases where this potential is present.

Political Turnover. If E chooses to challenge L for control of Attacker, then E seizes power with certainty. We make this seemingly stark assumption for two reasons. First, a more realistic alternative where E's challenge might fail will not change the substantive content of our results. Second, and more importantly, our main results will identify the conditions under which Elite in an ideologically divided country will choose *not to seize power in spite of the fact that she cannot fail*. To highlight the strategic incentives that produce this result, we ignore other frictions that dampen E's incentive to seize power.[11] Thus, absent international considerations, an ideological disagreement between hawks and doves leads to a power struggle, whereas ideological cohesion leads to political support. Our results would be identical if we simply assumed that when indifferent, Elite supports Leader.

We do not explicitly consider the possibility of bargaining between members of Attacker as a way of avoiding domestic conflicts. In Appendix B we show how including such bargaining in our framework would not affect our results. But since this inclusion involves complications that distract from our substantive focus, we leave it out of our main analysis.

Punishment and denial. At the conceptual level, Snyder (1961) emphasizes two distinct channels by which retaliation influences the incentives of potentially aggressive countries in the context of international politics. The first, *deterrence by denial*, refers to the use of retaliation to prevent hawks from obtaining the benefits associated with aggressive actions. The second channel, *deterrence by punishment*, refers to the use of retaliatory strikes to impose punitive costs on potentially aggressive actors (see also Schelling 1967). Our framework features both. Deterrence by punishment is represented by the cost term k that enters directly into the payoff of any member still politically involved in Attacker, and deterrence by denial is captured via the parameter q, which partially denies hawks the benefit they receive from conflict. Unless $q = 1$, deterrence cannot be achieved exclusively through denial.

[11] However, recall that internal conflicts within Attacker produce a payoff loss of δ, thus ensuring that Elite is never indifferent between supporting and challenging Leader.

There are two alternative, and equivalent, interpretations of how retaliation can affect the payoffs of countries A and D. First, retaliation could completely eliminate, with probability q, the threat Attacker poses to Defender's interests. Alternatively, the parameter q can represent the extent to which retaliation by Defender reduces the impact of escalation on the status quo, affecting the payoffs of both countries A and D.

Relation to crisis bargaining games. Our framework closely resembles a crisis bargaining game, which is the workhorse model of coercive diplomacy in international relations (e.g., Powell 1987; Morrow 1989; Fearon 1994; Schultz 1998). The only difference is that our model (technically) allows Defender to retaliate even following no provocation, whereas the standard crisis bargaining game only allows retaliation following provocation. However, because it is not incentive compatible in our model for Defender to retaliate when Attacker has not taken provocative actions, this difference is not consequential.

3 The Conventional Logic of Deterrence

We start by formalizing the *conventional logic of deterrence*, which is defined as a scenario in which it is incentive compatible for: (1) the defender to retaliate following provocation, and (2) no type of Leader chooses provocative actions. We focus on the classical deterrence problem, formulated as a complete information game that unfolds in three stages: provocation, retaliation, and escalation. Importantly, Attacker is comprised of only one decisionmaker who is a hawkish Leader that does not face any domestic challenges to her hold on power.

In the classical deterrence benchmark, absent provocation, Defender has no incentive to retaliate because escalation cannot materialize. Instead, if Leader has taken provocative actions, since she is a hawk and will escalate the conflict, retaliatory actions are sequentially rational for Defender if

$$\underbrace{-\lambda_1 - \lambda_2}_{\text{No Retaliation}} \leq \underbrace{-\lambda_1 - (1-q)\lambda_2 - c}_{\text{Retaliation}},$$

which rearranges to

$$c \leq q\lambda_2 \equiv c^*. \tag{1}$$

We refer to this condition as the *credibility constraint*, because when satisfied, threats of retaliation are credible. The threshold c^* characterizes the set of retaliation costs D is willing to incur in anticipation of escalation (following A's provocation). This set grows larger with the disutility D suffers from an escalated conflict, λ_2, and with the effectiveness of retaliation in reducing the costs of escalation, q. If retaliation had no effect on the D's cost from escalation, i.e. $q = 0$, retaliation would never be credible as long as c is positive.

When $c \leq c^*$, Leader anticipates that provocation will be followed by retaliation. In light of this, provocative actions are not worthwhile for L whenever

$$\underbrace{1}_{\text{No Provocation}} \geq \underbrace{1 + \pi_1 + (1-q)\pi_2 - k}_{\text{Provocation}},$$

which rearranges to

$$k \geq \pi_1 + (1-q)\pi_2 \equiv k^*. \tag{2}$$

We refer to this condition as the *capability constraint*, with k^* characterizing the set of retaliation costs incurred by members of Attacker that are severe enough to make a hawkish Leader want to avoid conflict. This set grows larger with a hawk's utility to both provocation, π_1, and escalation, π_2. When the capability constraint is satisfied, provocation is not incentive compatible, and hence the threat of retaliation is capable of deterring conflict. An increase in q reduces the benefit of escalated conflict for Leader, effectively scaling downward the capability constraint. This means that denial and punishment are effectively substitutes in ensuring that retaliation is capable of generating deterrence.

In contrast to the classical deterrence problem, where L's hold on power is perfectly secure and everything is commonly known, our main model has three additional features:

political insecurity, ideological disagreement, and asymmetric information. It is important to understand how each of these factors individually affects the conventional logic of deterrence. We start by adding asymmetric information and political insecurity separately to the classical deterrence problem. Thus, we consider two new games: (1) a game where L's hold on power is perfectly secure but L's type is privately known, and the probability L is a hawk is $\mu \in (0, 1)$; (2) a game where L's hold on power is not secure, but it is commonly known that both L and E are hawks. We next show that in each of these two games, just like in the classical deterrence game, the conventional logic of deterrence holds.

Proposition 1 (The Conventional Logic of Deterrence) *Consider the game:*

(I) L is commonly known to be a hawk, and her hold on power is perfectly secure, i.e. E's choice is exogenously fixed to $s = 1$; or

(II) L's type is privately known, and her hold on power is perfectly secure, i.e. E's choice is exogenously fixed to $s = 1$; or

(III) E chooses whether to support L, and the ideological composition of A is commonly known to be cohesive, Θ_C.

Then, there exists a unique equilibrium, and it exhibits the conventional logic of deterrence, i.e. D retaliates if and only if L provokes, and no type of Leader provokes if and only if the credibility constraint, $c \leq c^*$, and the capability constraint, $k \geq k^*$, are satisfied.

The analysis of the classical deterrence problem above showed that the conventional logic of deterrence holds as long as the capability and credibility constraints are satisfied. The second case, where L is perfectly secure but her type is privately known, is a costly signaling game between Leader (sender), and Defender (receiver). The unique equilibrium reflects the conventional logic of deterrence, since all types (hawks and doves) choose not to take provocative actions as long as (1) and (2) are satisfied. Since a dovish Leader whose hold on

power is perfectly secure has no incentive to provoke, provocation serves as a fully revealing signal that Leader is a hawk. Anticipating escalation, if the credibility constraint is satisfied, D has an incentive to retaliate following provocation, and if the capability constraint is satisfied, this threat of retaliation pushes even a hawkish Leader not to take provocative actions. Thus the presence of asymmetric information when leaders have a perfectly secure hold on power does not alter the conventional logic of deterrence.

In the last game, Leader relies on Elite's support to maintain power. However, since both L and E are hawks, a change in power within Attacker does not help Defender. If Leader has provoked a crisis, even if Elite were to seize power, Defender would face an escalated conflict. No matter who holds power in Attacker, D still responds to provocation with retaliation, as long as the credibility constraint is satisfied. An internal conflict then cannot prevent retaliation, and would only reduce E's payoff from domestic political issues from 1 to δ. This leads E to support L regardless of L's provocation choice. Consequently, anticipating the downstream choices of both Elite and Defender, L must decide between provocation, which entails retaliation, or foregoing conflict. As long as the credibility and the capability constraints are satisfied, L abstains from initiating a conflict, implying that the conventional logic of deterrence holds.

4 Ideological Disagreement and Deterrence

We now introduce ideological disagreement between members of Attacker and show how it undermines the conventional logic of deterrence. To highlight the role of ideological disagreement, we focus on cases where the ideological composition of Attacker is commonly known.

Recall from above that Defender will retaliate if and only if Leader provoked a crisis, a hawkish ruler controls Attacker, and the credibility constraint is satisfied. Proceeding

backwards, the next step is the domestic politics stage which consists of Elite's decision. Recall that there are two ways Attacker can be ideologically divided in our model: (1) Leader is a hawk and Elite is a dove ($\Theta = \Theta_h$), and (2) Leader is a dove and Elite is a hawk ($\Theta = \Theta_d$). Consider the latter case. If E's decision does not affect whether D retaliates, the ideological disagreement between L and E is the only relevant dimension determining E's choice, which will be to seize power because L is a different type. If, instead, E's support prevents retaliation, then E will support a dovish L if and only if

$$\underbrace{\pi_1}_{\text{Support}} \geq \underbrace{\pi_1 + \delta + (1-q)\pi_2 - k}_{\text{Seizing Power}},$$

which, after rearranging, holds as long as

$$k \geq \delta + (1-q)\pi_2 \equiv k^{**}. \tag{3}$$

We refer to Condition (3) as the *salience constraint*, because when this condition holds, avoiding retaliation is a more salient political issue among members of Attacker than clashes driven by domestic differences. The salience constraint differs from the capability constraint because it emerges from E's decision calculus rather than L's incentives when anticipating retaliation. When Leader's hold on power is insecure, the salience constraint generates the key strategic force in our model.

Lemma 1 *Let the credibility constraint, $c \leq c^*$, and the salience constraint, $k \geq k^{**}$, be satisfied. Then,*

(i) when Attacker is ideologically cohesive, $\Theta = \Theta_C$, E seizes power if and only if doing so prevents retaliation;

(ii) when Attacker is ideologically divided, $\Theta = \Theta_h$ or $\Theta = \Theta_d$, E supports L if and only if support prevents retaliation.

Lemma 1 establishes how the threat of retaliation changes the attitude of Elite towards domestic disputes when the credibility and salience constraints hold. If either of these two constraints are not satisfied, or if Elite's choice cannot influence the retaliation decision, then E's decision depends only on domestic political considerations, implying that Elite seizes power in an ideologically divided Attacker and supports Leader in an ideologically cohesive Attacker. In contrast, when both constraints hold, Elite is willing to choose the opposite action if this is the only way to avoid retaliation.

We first consider the case where Leader is a hawk and Elite is a dove.

Proposition 2 (Instability Curse) *Let Θ_h be the commonly known ideological composition of A. Then, there is a unique equilibrium where*

(i) D retaliates if and only if L has provoked a crisis;

(ii) the dovish E seizes power regardless of L's provocation decision; and

(iii) the hawkish L provokes a crisis,

if and only if the credibility constraint, $c \leq c^*$, is satisfied.

Proof: The decision rule for D follows from the derivation of the credibility constraint, (1). Since D retaliates if a hawk is the ruler, support is never a weak best-response for a dovish E. Moving to the provocation stage, taking provocative actions is a best response for L if and only if $\pi_1(\theta_L) \geq 0$, which holds since L is a hawk. ■

Proposition 2 identifies the first mechanism that undermines the conventional logic of deterrence, and that we refer to as the *instability curse*. Because L is a hawk, both domestic political concerns and international pressures lead E to seize power. However, once the dovish E has become the ruler, Defender knows that escalation will not materialize. Without the threat of escalation, retaliation is not credible, and L, anticipating that she cannot retain power *regardless of her provocation choice*, provokes a conflict to reap the short-term benefit

π_1. It is important to stress that in this scenario L does not lose power because she took provocative actions, but rather, takes provocative actions precisely because she anticipates that she cannot retain power, and thus will not endure Defender's retaliation.

At first glance, when Leader is hawkish and Elite is dovish, one might expect instability to benefit Defender. But because a dovish actor can take power from a hawkish one, D can avoid escalation without having to deter provocation. In this way political instability creates a friction that ultimately leads the conventional logic of deterrence to fail because it breaks the direct connection between retaliation and provocation.[12]

We next consider the case where Leader is a dove and Elite is a hawk.

Proposition 3 (Deterrence Curse) *Let Θ_d be the commonly known ideological composition of A. Then, there is a unique equilibrium where*

(i) D retaliates if and only if L provokes and E seizes power;

(ii) the hawkish E supports L if and only if L provokes; and

(iii) the dovish L provokes a conflict,

if and only if the credibility constraint, $c \leq c^*$, and the salience constraint, $k \geq k^{**}$, are satisfied.

Proof: Following the derivation of the credibility constraint, (1), Defender only retaliates when escalation will happen in the last stage. When Leader is dovish and Elite is hawkish, the only way for escalation to occur is if Leader provokes and Elite seizes power. Moving backward, if L has not provoked, then by Lemma 1, E seizes power from L, implying that L receives a payoff of 0. Instead, following provocation, E supports L if

$$\pi_1 \geq \pi_1 + \delta + (1-q)\pi_2 - k,$$

[12]Consequences of this kind of friction have been examined in the context of electoral fraud (Gehlbach and Simpser 2015) and repression (Tyson 2018).

which is equivalent to the salience constraint, (3). Moving to the provocation stage, L has a strict incentive to provoke only if doing so motivates E to support her, which from above, holds if and only if the credibility and salience constraints are satisfied. ■

When the credibility and salience constraints hold a dovish Leader can remain in power *only* if she provokes a conflict with Defender, even though she does not benefit directly from provocative actions. That is, *the dovish Leader's provocation of a conflict is purely a tactic that serves the purpose of deterring a domestic challenge*. We call this novel mechanism through which deterrence can fail the *deterrence curse*, and it is the result of two key ingredients.

The first ingredient of the deterrence curse is a novel commitment problem. If L does not provoke a crisis, D has no interest in whether Attacker's ruler is a hawk or a dove. Without provocation there is no risk of escalation, and thus, no reason to retaliate. However, if D could commit to retaliate after the hawkish Elite has seized power from the dovish Leader, the threat of retaliation would *deter* the hawkish Elite from taking power, and thus, a dovish Leader would have no incentive to take provocative actions. But, because D will not retaliate in the absence of the threat of an escalated conflict, a dovish L is forced to provoke a crisis to coerce D into having a stake in the domestic politics of Attacker.

The second component that gives rise to the deterrence curse is precisely Defender's ability to deter undesirable actions, and it is encapsulated in the credibility and salience constraints. If either the credibility or salience constraint do not hold, then the threat of retaliation would not deter the hawkish Elite from seizing power. Retaliation would either not be a credible threat (if the credibility constraint is not satisfied), or it would not be severe enough to make the international conflict with D more salient than the domestic conflict with L (if the salience constraint is not satisfied). The dovish Leader then could not use the threat of retaliation to obtain the hawkish Elite's support, and she would have no reason to provoke a crisis. In this way, the potential to deter becomes a curse for Defender.

Introducing an ideological disagreement within Attacker leads all types of leaders, hawks and doves, to provoke a crisis.[13] This means that when L's hold on power is not perfectly secure, *irrespective of the nature of the ideological disagreement*, the conventional logic of deterrence fails. Interestingly, in both the instability and deterrence curses, it is the presence of doves, which seemingly should be an asset for Defender, that undermines deterrence. While in Propositions 2 and 3, provocation occurs because Attacker is ideologically divided, the presence of doves ensures that conflict is not ultimately escalated.

5 Asymmetric Information and Deterrence

We now introduce asymmetric information to explore the influence of the *potential* for ideological division within Attacker on the conventional logic of deterrence. We focus on three possible ideological compositions of Attacker: (1) L and E are hawks ($\Theta = \Theta_C$), which occurs with probability μ_C; (2) L is a hawk and E is a dove ($\Theta = \Theta_h$), which occurs with probability μ_h; and (3) L is a dove and E is a hawk ($\Theta = \Theta_d$), which occurs with probability μ_d.[14] The ideological composition of Attacker is common knowledge among Leader and Elite, and hence, the only asymmetry in information is between Defender and members of Attacker. This implies that the model in this section is a signaling game with two commonly informed senders (L and E) and a receiver (Defender).

We focus on pure-strategy Perfect Bayesian equilibria that satisfy the Intuitive Criterion of Cho and Kreps (1987).[15] To keep our analysis simple, we focus on cases where c is sufficiently high that Defender would not find retaliation incentive compatible ex ante. That

[13] That these equilibria are exhaustive follows by our requirement that doves do not provoke when indifferent, and follows formally from Lemma A.1 in the appendix.

[14] We have not included the case in which both the Leader and Elite are doves, because although realistic, as we show in Appendix C, it does not add much to our study. In any equilibrium, a dovish L would not provoke a crisis and would be supported by a dovish E.

[15] We do not use the Intuitive Criterion to rule out any equilibria. Yet, since we find equilibria where behavior does not fully reveal types, it is sensible to check such equilibria against a common refinement. Our equilibria also survive D1 from Banks and Sobel (1987).

is, if Defender did not observe the provocation and support decisions, she would prefer *not* to retaliate. Formally, this corresponds to restricting attention to $c \geq (1-\mu_d)(1-q)\pi_2 \equiv \underline{c}$, which also ensures that there are equilibria in pure strategies.

We call an equilibrium *fully revealing* when Defender learns both L's and E's type, *partially revealing* when Defender learns the type of either L or E, and *nonrevealing* when Defender cannot infer anything about the ideological composition of Attacker from the actions of L and E. With asymmetric information regarding the ideological composition of Attacker, anticipating Defender's response, members of Attacker have incentives to misrepresent information.

Lemma 2 *If the credibility constraint, $c \leq c^*$, and salience constraint, $k \geq k^{**}$, are satisfied, then there is no fully revealing equilibrium. Moreover, in any equilibrium, if a hawkish L has taken provocative actions, a hawkish E and a dovish E choose the same action.*

When both the credibility and salience constraints are satisfied, provocation by a hawkish Leader creates an incentive for a hawkish Elite to mimic a dovish Elite. To see this, suppose that only a dovish Elite supports following provocation. For this to be part of an equilibrium, Lemma 1 implies that support prevents retaliation. But if this is the case, a hawkish Elite has an incentive to deviate, supporting Leader, and avoiding the costs associated with domestic political conflict. Consider the opposite scenario, where only a hawkish Elite supports Leader, while a dovish Elite does not. Then, Defender knows that any ruler who maintained control of Attacker is a hawk, and consequently, retaliates following support. But again, this provides a hawkish Elite with an incentive to deviate, seizing power from Leader. Together, these imply that in any equilibrium where L has provoked a crisis, there is an incentive by some Elite types to strategically misrepresent what they know about the ideological composition of Attacker.

We next characterize equilibria in the game with asymetric information.

Proposition 4 *If the credibility constraint, $c \leq c^*$, and salience constraint, $k \geq k^{**}$, are satisfied, then there are two distinct equilibria:*

(I) (Partially Revealing): If $\pi_1 \geq 1$, then there exists a $\hat{c} \in [\underline{c}, c^*]$ such that

- *Defender retaliates following provocation and support, and does not retaliate otherwise;*
- *Elite does not support Leader following provocation, otherwise, Elite supports only if $\theta_L = \theta_E$; and*
- *Leader provokes a conflict with Defender only if she is a hawk.*

(II) (Nonrevealing): For all values of $c \in [\underline{c}, c^*]$,

- *Defender retaliates only after provocation followed by a seizure of power;*
- *Elite always supports Leader following provocation, and otherwise, supports only if $\theta_L = \theta_E$; and*
- *Leader provokes a conflict, regardless of her type.*

Moreover, when $c < \hat{c}$ or $\pi_1 < 1$, the latter is the unique equilibrium.

In the first part of Proposition 4 we characterize a partially revealing equilibrium, where a hawkish L provokes a crisis and a dovish L does not. As in the instability curse, when Attacker is ideologically divided and led by a hawk, L anticipates that she will lose power regardless of her provocation choice. The political turnover from a hawkish Leader to a dovish Elite removes Defender's incentive to retaliate, but then, following provocation from a hawkish Leader, a hawkish Elite chooses the same action as a dovish E, strategically misrepresenting her type (highlighted by Lemma 2). By projecting an image of internal division, the hawkish Elite in a cohesive Attacker keeps Defender "guessing" about whether

Attacker truly represents a threat. When c is sufficiently high (i.e. $c \geq \hat{c}$), this uncertainty is enough to dissuade retaliation, constituting an equilibrium when π_1 is sufficiently high.[16]

The equilibrium associated with the instability curse has important implications regarding the advantages of supporting (seemingly) dovish, or moderate, factions against rogue states, international rivals, and insurgent or terrorist organizations. Recall that in our model, Elite can seize power if she desires, suggesting that even if elites could be empowered costlessly, such policies may be of limited use because of the *strategic* incentives that empowering moderate factions engender (when combined with deterrence).

The second part of Proposition 4 characterizes the equilibrium associated with the deterrence curse. Recall that a dovish Leader, who otherwise does not benefit from conflict, sees provocation as a tool for political survival. In this equilibrium every type of Leader is able to maintain power by provoking a crisis, and Defender cannot learn anything about the ideological composition of Attacker from the provocation and support decisions. In particular, D does not know whether Leader is a hawk, seeking to escalate the conflict, or a dove, provoking a conflict purely for domestic political reasons, and this uncertainty is enough to discourage Attacker from retaliating.

With asymmetric information, deterrence fails, but on top of this, some crises will escalate beyond provocation. Moreover, because some attacker countries are led by doves, and hawkish elites mimic their dovish counterparts (following provocation) Defender is prevented from identifying who poses a true threat of escalation. Thus, the presence of doves, which should be advantageous for Defender, actually weakens the credibility of retaliation, undermining deterrence. It is important that in the classical deterrence problem with asymmetric information, Defender was also blocked from knowing Attacker's type because both hawks and doves did not provoke a crisis (when the credibility and capability constraints

[16]It is worth noting that Defender is not "fooled" by the appearance of division, but rather, uncertainty regarding the ideology of Elite hinders D's resolve to retaliate.

are satisfied). In that game, however, the inability to learn Attacker's type was a fortunate byproduct of successful deterrence, whereas in the game with political instability and ideological disagreement, this is the *source* of deterrence failure.

To assess the severity of the deterrence failures highlighted by our model, we conclude our analysis by assessing the welfare of Defender in the two equilibria of the game with asymmetric information and comparing them with Defender's welfare if she were to commit to never using retaliation in response to provocation, a strategy we refer to as *inaction*. Although there is a cost to inaction, namely that D must tolerate escalated conflicts from countries where doves are absent (or not in politically influential positions), there are benefits from suppressing the curses that arise from the credibility and salience constraints.

Proposition 5 *Suppose that the credibility constraint, $c \leq c^*$, and salience constraint, $k \geq k^{**}$, are satisfied. Defender is better off in equilibrium 4.I, associated with the instability curse, than in equilibrium 4.II, associated with the deterrence curse. Defender is at least as well off with inaction than equilibrium 4.I, associated with the instability curse, and is strictly better off with inaction than equilibrium 4.II, associated with the deterrence curse.*

Let us first compare the two equilibria described in Proposition 4. The equilibrium associated with the deterrence curse $(4.II)$, is uniformly worse for Defender than the equilibrium associated with the instability curse, $(4.I)$. Recall that when D is more willing to use force to eliminate threats, namely, when the cost of retaliation is relatively low (i.e. $c < \hat{c}$), the equilibrium associated with the deterrence curse is unique. This welfare comparison suggests that having the resolve to use aggressive foreign policy can be self-defeating because it can push dovish actors to support hawkish Leaders, and more surprisingly, it can prompt doves to engage in activities (e.g., provoking crises) that are normally associated with hawks.

To assess the welfare consequences of inaction, we must consider the decisions of Leader and Elite in an ideologically divided Attacker when retaliation is off the table. Recall that,

by Lemma 1, when E's decision is not pivotal in D's retaliation choice, E supports L only when A is cohesive, and seizes power otherwise. Thus, ideological divisions within Attacker lead to a change of rulership, and ideological cohesion leads to political security. In terms of outcomes there are three possibilities: (1) if Attacker is cohesive, there will be an escalated conflict; (2) if Attacker is divided and initially led by a hawk, there will be provocation but not escalation; and (3) if Attacker is divided and initially led by a dove, there will be neither provocation nor escalation.

Observing that the outcome path with inaction is identical to that from Equilibrium $4.I$, Defender is at least as well off with inaction as in Equilibrium $4.I$, associated with the instability curse. More surprisingly, Defender is strictly better off with inaction than in Equilibrium $4.II$. The logic of the deterrence curse enables both hawkish and dovish Leaders to maintain control of Attacker, but as a result, more countries remain under the control of hawks. Moreover, even in countries where dovish Leaders are in control, doves find the provocation of a crisis an attractive tool to fend off threats from hawks. When D takes retaliation off the table by committing to inaction, the deterrence curse breaks. Thus, inaction removes a salient political issue that pushes hawks and doves together, specifically, their shared desire to avoid retaliation by D. In the absence of a potential conflict with Defender, *doves will not support hawkish Leaders nor will they provoke international crises.*

6 Conclusion

In this article, we formalize the conventional logic of deterrence and explore its strategic foundations. We show how this logic crucially relies on the political security of leaderships in attacker countries, or the ideological agreement among different factions within it. Political instability within an attacker country can undermine deterrence in two ways, neither of which have been highlighted previously, and that relate to special kinds of incentive frictions

special to political settings. First, when leaders anticipate that they cannot maintain power, this severs the direct connection between the provocation of a conflict and the negative consequences that follow from retaliatory actions, which we label the instability curse. Second, when dovish factions in attacker countries face the threat of removal by hawkish factions, Defender cannot commit to protecting the interests of the doves unless there is a concrete threat that hawks could escalate conflict after taking power. This creates an incentive for politically vulnerable doves to provoke international disputes as a way to make Defender concerned with their political survival.

In the modern world where political instability has become a prevalent and important feature of international politics, our results highlight a potential weakness in many international treaties and security alliances that heavily rely on the conventional logic of deterrence (see, e.g., Benson, Meirowitz and Ramsay (2014) for examples), by showing that this logic relies on assumptions which are particularly restrictive. Our results also suggest that policies hoping to take advantage of the presence of "friendly" factions within potentially aggressive rivals can be a strategic disadvantage if such reliance is not accompanied by a firm commitment to take future retaliatory actions out of consideration. Without such a commitment, the threat of retaliation creates a salient political dimension that acts to unite otherwise opposing factions.

Finally, when considering investments into arsenals with the intention of creating a deterrent threat, for example, by investing in new military technologies or weapons of mass destruction, our results offer a word of caution. Efforts that focus on the development (or maintenance) of a capable arsenal, perhaps motivated by the motto "peace through strength", ultimately rely on the soundness of the conventional logic of deterrence. Our results show that when the leadership of attacker countries are not perfectly secure, enhancing military capacity might render the defender-vs-attacker conflict more salient than domestic disputes, thus leading to worse outcomes than if an arsenal was never created.

A Appendix

Proof of Proposition 1: The proof has three parts.

(I) Follows by the argument in the text.

(II) We begin by observing that when $s = 1$, it is never a strict best-response for a dove to take a provocative action. This implies that in any equilibrium, the posterior probability D places on L being a hawk following provocation is 1. From this, D retaliates if and only if the credibility constraint is satisfied. A hawk then prefers not to provoke if and only if the capability constraint is satisfied.

(III) Since $\Theta = \Theta_C$, conflict will escalate, i.e. $x^* = 1$, regardless of whether L or E is the ruler. Moving to the retaliation stage, since the ruler is a hawk, D will retaliate if $m = 1$ and the credibility constraint is satisfied. Moving backwards to the domestic politics stage, following $m = 0$, E chooses to support since $\delta < 1$. Instead, if $m = 1$, E chooses to support if,

$$\pi_1 + 1 + \pi_2(1 - q) - k \geq \pi_1 + \delta + \pi_2(1 - q) - k \tag{A.1}$$

which, since $\delta < 1$, is always true. Moving to the provocation stage, L, anticipating that she will remain the ruler, chooses $m = 0$ if the capability constraint is satisfied.

■

Proof of Lemma 1: Let $H(s)$ be an indicator function which takes the value 1 if and only if D retaliates, and which depends on E's choice, s. Suppose that D's retaliation decision does not depend on E's choice, that is $H(0) = H(1)$. Then, E need only consider whether L is a hawk or a dove. If $\theta_L = \theta_E$, then since $\delta < 1$, E strictly prefers to support. In contrast, if $\theta_L \neq \theta_E$, E's utility from domestic concerns is equal to δ if she seizes power, and 0 otherwise. Since $\delta > 0$, E strictly prefers to seize power.

Suppose now that D's retaliation decision depends on E's choice, implying that $H(0) \neq H(1)$. For the first part of the result, suppose A is cohesive ($\Theta = \Theta_C$). There are two cases to consider:

(i) If support triggers retaliation, i.e. $H(0) = 1 - H(1) = 0$, then E strictly prefers to support when

$$1 + (1 - q)\pi_2 - k > \delta + \pi_2$$

which is equivalent to

$$k \leq 1 - \delta - q\pi_2, \tag{A.2}$$

thus violating the salience constraint.

(ii) If support prevents retaliation, i.e. $H(1) = 1 - H(0) = 0$, then E strictly prefers to support when

$$1 + \pi_2 > \delta + (1 - q)\pi_2 - k,$$

which always holds.

Consider next when A is divided ($\Theta = \Theta_h$ or $\Theta = \Theta_d$). Let $V(m, s, r, x^*(\Theta) \mid \theta_E)$ be the indirect expected utility of Elite. For E, support yields

$$-k \cdot H(1),$$

while seizing power yields,

$$H(0) \cdot [V(1, 0, 1, x^*(\Theta) \mid \theta_E) - k] + (1 - H(0))V(1, 0, 0, x^*(\Theta) \mid \theta_E).$$

There are two cases to consider:

(i) If support triggers retaliation, i.e. $H(0) = 1 - H(1) = 0$, then E strictly prefers to

support when

$$V(1, 0, 0, x^*(\Theta) \mid \theta_E) < -k$$

which never holds.

(ii) If support prevents retaliation, i.e. $H(1) = 1 - H(0) = 0$, then E strictly prefers to support if and only if

$$V(1, 0, 1, x^*(\Theta) \mid \theta_E) < k,$$

which holds for both doves and hawks if and only if the salience constraint is satisfied.

Putting the arguments above together establishes the result. ■

We next present three results which are useful for the rest of the proofs.

Lemma A.1 *A dovish L provokes only if E supports.*

Proof: The indirect utility of a dovish L who does not provoke is 0, whereas the indirect utility of a dovish L is strictly positive when she provokes only if E supports. Since the dovish L provokes only if she has a strict incentive to do so, it must be that E supports following $m = 1$. ■

Lemma A.2 *In any equilibrium, a hawkish L in a divided Attacker takes provocative actions.*

Proof: Suppose to the contrary that there exists an equilibrium in which a hawkish L in a divided A chooses not to provoke. Then by Lemma 1, L expects to lose control of Attacker, thereby receiving a payoff of 0. Since $\pi_1 > 0$, for not provoking to be optimal, two things must occur: (1) E must support, thus leaving the hawkish L in control; and (2) D must retaliate. However, in such an equilibrium, E's support does not prevent retaliation, contradicting Lemma 1. ■

Lemma A.3 *If the credibility constraint holds, then there exists a fully revealing equilibrium if and only if* $k < 1 - \delta - q\pi_2 \equiv \underline{k}$

Proof: We proceed to establish a contradiction. Suppose there is a equilibrium in which the pair Θ is fully revealed to Defender and $k \geq \underline{k}$. Denote the outcome mapping $\sigma(\Theta) = (m, s)(\Theta)$, and observe that an equilibrium is fully revealing if and only if the mapping σ is one-to-one, meaning that the strategy profile is fully revealing on the path of play.

The proof proceeds in three steps: (1) we show that there cannot be a fully revealing equilibrium where $\sigma(\Theta_h) = (1, 1)$, and thus by Lemma A.1, it has to be that $\sigma(\Theta_h) = (1, 0)$; (2) we show that the fully revealing strategy profile that yields the path, $\sigma(\Theta_C) = (0, 1)$, $\sigma(\Theta_h) = (1, 0)$, and $\sigma(\Theta_d) = (0, 0)$, cannot constitute an equilibrium; and (3) we show that the only remaining fully revealing outcome path, $\sigma(\Theta_C) = (1, 1)$, $\sigma(\Theta_h) = (1, 0)$, and $\sigma(\Theta_d) = (0, 0)$, is an equilibrium if and only if $k < \underline{k}$.

Step 1: Suppose there exists a fully revealing equilibrium where $\sigma(\Theta_h) = (1, 1)$. This implies that the D's best response is to retaliate after observing $\sigma = (1, 1)$ since L has provoked a crisis and a hawk is the ruler of A. Therefore, a dovish E, by supporting when $\Theta = \Theta_h$, cannot avoid retaliation, and thus by Lemma 1, E seizes power. Moreover, by Lemma A.2, L in Θ_h chooses $m = 1$ in any equilibrium, and thus, $\sigma(\Theta_h) = (1, 0)$ in any fully revealing equilibrium.

Step 2: Consider the fully revealing strategy profile that yields the outcome path, $\sigma(\Theta_C) = (0, 1)$, $\sigma(\Theta_h) = (1, 0)$, and $\sigma(\Theta_d) = (0, 0)$. According to this profile, D's optimal response after observing $\sigma = (1, 0)$ is not to retaliate. Consider E's best response to a deviation by L to $m = 1$ when $\Theta = \Theta_C$, which depends on D's response to the off-the-equilibrium outcome path $(1, 1)$.

Suppose that following the outcome path $(1, 1)$, D chooses not to retaliate. Then following provocation, E chooses $s = 1$ since $\delta < 1$. But this implies that L has a profitable deviation, since by provoking a crisis she maintains E's support and does not suffer retaliation.

Suppose next that following the outcome $(1,1)$, D chooses to retaliate. Then, E weakly prefers to seize power following provocation if

$$\delta + \pi_2 \geq 1 + (1-q)\pi_2 - k, \tag{A.3}$$

which is equivalent to

$$k \geq 1 - \delta - q\pi_2 = \underline{k}. \tag{A.4}$$

When satisfied, if support invites retaliation, E strictly prefers to seize power following provocation. Since D will not retaliate following $(1,0)$ in the proposed equilibrium, L has a profitable deviation to provoke, and hence the proposed strategy profile cannot constitute an equilibrium. For the case where E is indifferent between supporting and seizing power, so $k = \underline{k}$, if E chooses to support following provocation, thus triggering retaliation, L still has a profitable deviation to provoke, since she receives a payoff of $\pi_1 + 1 + (1-q)\pi_2 - \underline{k}$, whereas by not provoking, she receives a strictly lower payoff of 1. Combining this argument with Lemma 1, the outcome path $\sigma = (0,1)$ cannot follow from a cohesive A in any fully revealing equilibrium when $k \geq \underline{k}$.

Step 3: Given the above argument, for an equilibrium to be fully revealing the image of the types Θ_C and Θ_d, according the mapping σ, must be $(1,1) \times (0,0)$ and $\sigma(\Theta_h) = (1,0)$. Therefore, a fully revealing equilibrium can take only one of two forms:

(i) $\sigma(\Theta_C) = (0,0)$, $\sigma(\Theta_h) = (1,0)$, and $\sigma(\Theta_d) = (1,1)$, which contradicts Lemma 1 since E in a cohesive A always supports L following no provocation.

(ii) $\sigma(\Theta_C) = (1,1)$, $\sigma(\Theta_h) = (1,0)$, and $\sigma(\Theta_d) = (0,0)$. The outcome $\sigma(\Theta_C) = (1,1)$ reveals that both L and E are hawks and will thus escalate the conflict. Hence, Defender, after observing $\sigma = (1,1)$ will retaliate. However, if the hawkish E, in opposition to the prescribed strategy profile, seizes power following provocation, then E will be the

ruler and D will conclude that E is a dove since it expects $\sigma = (1, 0)$ only from Θ_h. Thus, Defender, will not retaliate. This deviation will not be optimal for E only when $k < \underline{k}$.

■

Proof of Lemma 2: Since $\underline{k} < k^*$, Lemma A.3 implies that there is no fully revealing equilibrium when the credibility constraint, (1), and salience constraint, (3), hold.

To establish the second part we argue by contradiction and suppose that there is an equilibrium in which different types of E choose different actions after a hawkish L takes provocative actions. Since, by Lemma A.1, a dovish L only provokes if E supports following provocation, there are two cases to consider:

(i) Suppose that $s^*(\Theta_h \mid m = 1) = 0$, and hence $s^*(\Theta_C \mid m = 1) = 1$. This implies that D, after observing provocation and a seizure of power, concludes that a dovish E rules A. Consequently, D does not retaliate, and the hawkish E in a cohesive A has a profitable deviation to seize power, contradicting that $s^*(\Theta_C \mid m = 1) = 1$.

(ii) Suppose that $s^*(\Theta_h \mid m = 1) = 1$ and hence $s^*(\Theta_C \mid m = 1) = 0$. In this case, by Lemma 1, it must be that when $\Theta = \Theta_h$, E's support prevents retaliation. This implies that D, after observing $m = 1$ and $s = 1$, does not retaliate. But this then implies, with Lemma 1, that E in a cohesive A will support, contradicting $s^*(\Theta_C \mid m = 1) = 0$.

■

Proof of Proposition 4: *Part (I)*: By Lemma A.2, the hawkish L in a divided A takes provocative actions, and so the following strategy

$$m^*(\theta_L) = \begin{cases} 1 & \text{if } \theta_L = 1 \\ \\ 0 & \text{if } \theta_L = 0, \end{cases} \tag{A.5}$$

is the only strategy in which L can reveal her type. The dovish L of a divided A does not provoke, and thus by Lemma 1, the hawkish E seizes power.

Following provocation by the hawkish L, Lemma 2 establishes that both the hawkish and dovish E choose the same action. Suppose first that $s^*(\theta_C \mid m = 1) = s^*(\Theta_h \mid m = 1) = 1$, i.e. both support. In this case, Defender concludes that A is controlled by a hawk with probability 1, and thus retaliates. But then the dovish E's decision to support L does not prevent retaliation, contradicting Lemma 1, and hence $s^*(\theta_C \mid m = 1) = s^*(\Theta_h \mid m = 1) = 0$ in any equilibrium where L's type is revealed to be a hawk following provocation.

Since Lemma A.1 establishes that a dovish L provokes only if doing so motivates E to support, $\sigma = (1, 0)$ cannot follow when $\Theta = \Theta_d$. Thus, following the outcome path $\sigma = (1, 0)$, Defender retaliates if and only if

$$\frac{\mu_C}{\mu_C + \mu_h}(1 - q)\lambda_2 - c \geq \frac{\mu_C}{\mu_C + \mu_h}\lambda_2.$$

Thus, after rearranging, Defender will not retaliate if and only if

$$c \geq \frac{\mu_C}{\mu_C + \mu_h} q\lambda_2 \equiv \hat{c}.$$

We must next show that the dovish L does not want to provoke. By Lemma A.1, we know that a dovish L will only provoke if E supports, which by Lemma 1, is only incentive compatible if support prevents retaliation. Thus, D cannot retaliate after observing $\sigma = (1, 1)$ in equilibrium. If this is the case, then since $\delta < 1$, E in Θ_C has an incentive to deviate and support, contradicting Lemma 2.

Finally, we must check that provocation is optimal for the hawkish L in a cohesive A. A hawkish L who provokes expects no Elite type to support. Therefore, the hawkish L who provokes receives a utility of π_1. If instead, L were to not provoke, then Lemma 1 implies that she would be supported by the hawkish E and receive a utility of 1. Provocation is a

best response whenever $\pi_1(\theta_L) \geq 1$.

Part (II): We first establish that there is only one possible nonrevealing strategy profile sustainable in equilibrium. A nonrevealing equilibrium is one in which $m^*(\Theta) = y$ for all Θ, and $s^*(\Theta \mid m = y) = z$ for all Θ. By Lemma A.2, we know that y must be 1, and thus we require that $s^*(\Theta \mid m = 1) = z$ for all Θ.[17] By Lemma A.1, the dovish L in a divided A only provokes if the hawkish E then supports, hence, it must be that $s^*(\Theta \mid m = 1) = 1$ for all Θ. Thus, the only nonrevealing equilibrium is one where every type of Leader provokes and every type of Elite supports.

Let $\gamma = Pr(\theta_E = 1|\sigma = (1, 0))$, so that, upon observing a seizure of power, D believes E is a hawk with probability γ. By Lemma 1, for E to support L in a divided A it must be that, following the outcome $\sigma = (1, 0)$, Defender retaliates. This is the case if

$$-\lambda_1 - \gamma(1 - q)\lambda_2 - c > -\lambda_1 - \gamma\lambda_2.$$

Rearranging, we obtain

$$\gamma > \frac{c}{q\lambda_2} \equiv \hat{\gamma}_c.$$

The credibility constraint, (1), ensures that $\hat{\gamma}_c \in (0, 1)$. Therefore, for all $\gamma \in (\hat{\gamma}_c, 1]$, E prevents retaliation by supporting, and hence, no E has an incentive to deviate to $s = 0$. Given that E supports following provocation and retaliation is avoided, no L has an incentive to deviate to $m = 0$.

We now show that given a $c \in [\underline{c}, c^*]$, each equilibrium with $m^*(\Theta) = 1$ and $s^*(\Theta \mid m = 1) = 1$ for all Θ, and $\gamma \in (\hat{\gamma}_c, 1]$ survives the Intuitive Criterion of Cho and Kreps (1987). To do this, we will show that every type of E is strictly better off in equilibrium, than by seizing power. The minimum payoff E can get by seizing power is $-k+V_E(1, 0, 1, x^*(\Theta) \mid \theta_E)$, which, by the salience constraint, (3), is smaller than the equilibrium payoff of E in a divided A,

[17] Notice that $s^*(\Theta \mid m = 0)$ need not be the same for all θ since it is off the equilibrium path.

which is 0. As such, there is no action and no Θ such that E is worse off in equilibrium than in the case of a deviation followed by retaliation. This establishes that for every c, there exists a nonempty set, $(\hat{\gamma}_c, 1]$, of off-the-path beliefs such that for every $\gamma \in (\hat{\gamma}_c, 1]$ the uninformative equilibrium survives the Intuitive Criterion. ■

Proof of Proposition 5: Recall the path of play in equilibrium $4.I$ associated with the instability curse: provocation is followed by a seizure of power when $\Theta = \Theta_C$ and $\Theta = \Theta_h$, while no provocation and a seizure of power occurs when $\Theta = \Theta_d$. By direct calculation, Defender's welfare in this case is

$$\mathcal{W}_{IC} \equiv -\mu_C(\lambda_1 + \lambda_2) - \mu_H\lambda_1. \tag{A.6}$$

The path of play in equilibrium $4.II$, associated with the deterrence curse, is where initiation and support occur for all $\Theta \in \{\Theta_C, \Theta_h, \Theta_d\}$. Defender's welfare in this case is

$$\mathcal{W}_{DC} \equiv -(\mu_C + \mu_H)(\lambda_1 + \lambda_2) - \mu_D\lambda_1. \tag{A.7}$$

Taking the difference

$$\begin{aligned} \mathcal{W}_{IC} - \mathcal{W}_{DC} &= -\mu_C(\lambda_1 + \lambda_2) - \mu_H\lambda_1 + (\mu_C + \mu_H)(\lambda_1 + \lambda_2) + \mu_D\lambda_1 \\ &= \mu_H\lambda_2 + \mu_D\lambda_1 > 0, \end{aligned} \tag{A.8}$$

establishing the first part.

The path of play under an inaction strategy is provocation and support when $\Theta = \Theta_C$, provocation and a seizure of power when $\Theta = \Theta_h$, and no provocation and a seizure of power when $\Theta = \Theta_d$. By direct calculation, Defender's welfare under the inaction strategy is

$$\mathcal{W}_0 \equiv -\mu_C(\lambda_1 + \lambda_2) - \mu_H\lambda_1. \tag{A.9}$$

For the first part:

$$\mathcal{W}_0 - \mathcal{W}_{IC} = -\mu_C(\lambda_1 + \lambda_2) - \mu_H\lambda_1 + \mu_C(\lambda_1 + \lambda_2) + \mu_H\lambda_1 = 0. \tag{A.10}$$

Combining this with (A.8) establishes that

$$\mathcal{W}_0 = \mathcal{W}_{IC} > \mathcal{W}_{DC}.$$

■

B Bargaining and the Curses of Deterrence

In this section we examine the possibility of including a bargaining stage between Leader and Elite within Attacker and explore how such a possibility affects our equilibria. Consider an extension in which L has the opportunity to make a private (not observed by D) take-it-or-leave-it offer to E. Moreover, whether this bargaining stage occurs before the provocation stage, or between the provocation stage and the domestic politics stage, is not consequential, so for convenience, we will consider the game where the bargaining stage happens between the provocation and domestic politics stages.[18]

Denote A's offer by v_m following provocation choice m, and recall that $m = 1$ means L takes provocative actions. We begin with an ideologically divided Attacker.

Lemma B.1 *Let Attacker be ideologically divided so that $\theta_L \neq \theta_E$, then in any equilibrium*

(i) following provocation, L's sequentially optimal offer is zero, i.e. $v_1^* = 0$;

(ii) following no provocation, L's sequentially optimal offer is $v_0^* = \delta$.

[18]Note that introducing a bargaining stage any later would leave the game unchanged as neither player could commit to honoring a proposal.

Proof: For the first part, by Lemma 1, Elite supports L if doing so prevents retaliation by Defender, in which case L's optimal offer is $v_1^* = 0$. Thus, we need only consider when E would avoid retaliation by choosing to seize power from L, while supporting L would trigger retaliation by Defender. In order for support to to be incentive compatible, L must propose a transfer to E that compensates for the value of holding power following an internal power struggle, δ, as well as compensation for the cost of incurring retaliation. Thus, L's proposal, v_1, must satisfy

$$v_1 - k \geq \delta.$$

Holding this at equality, L's utility is then

$$1 - k - v_1^* = 1 - \delta - 2k,$$

which must be nonnegative, contradicting the salience constraint. This argument establishes that no positive proposal will happen after provocation.

For the second part, when provocation has not occurred, Elite's payoff from seizing power is δ, and the payoff to support is 0. Consequently, L must give at least δ to E in order to ensure support. Thus, support is incentive compatible for Elite if Leader's proposal, v_0, satisfies,

$$v_0 \geq \delta,$$

which leaves $1 - \delta$ to Leader. ■

We next consider the case of an ideologically cohesive Attacker.

Lemma B.2 *Let Attacker be ideologically cohesive so that $\theta_L = \theta_E$, then*

(i) following provocation, L's sequentially optimal offer is $v_1^* = \delta + \pi_2 - (1 + (1-q)\pi_2 - k)$ if and only if $3\delta + \pi_2 \leq 2$, and $v_1^* = 0$ otherwise;

(ii) following no provocation, L's sequentially optimal offer is $v_0^* = 0$.

Proof: For the first part, by Lemma 1, E seizes power only when doing so prevents retaliation. If seizing power does not prevent retaliation, then L's optimal offer is $v_1^* = 0$. Otherwise, L must compensate E for support which includes the cost of retaliation she would otherwise avoid. The offer, v, then must satisfy

$$1 + \pi_1 + (1-q)\pi_2 - k + v \geq \delta + \pi_1 + \pi_2,$$

for Elite. Rearranging yields the expression

$$v_1^* = \delta + \pi_2 - (1 + (1-q)\pi_2 - k).$$

By substitution, Leader's payoff must satisfy

$$1 + (1-q)\pi_2 - k - [\delta + \pi_2 - (1 + (1-q)\pi_2 - k)] \geq 0,$$

which reduces to

$$2(1 + (1-q)\pi_2 - k) \geq \delta + \pi_2.$$

Rearranging, this holds if and only if

$$\frac{\delta + \pi_2}{2} \leq 1 + (1-q)\pi_2 - k,$$

which, after rearranging, holds if and only if

$$k \leq \frac{1}{2}(2 - \delta + (1-2q)\pi_2).$$

Recalling that $k^{**} = \delta + (1-q)\pi_2$, this set is nonempty if and only if

$$2(\delta + (1-q)\pi_2) \leq 2 - \delta + (1-2q)\pi_2.$$

Distributing gives

$$2\delta + 2\pi_2 - 2q\pi_2 \leq 2 - \delta + \pi_2 - 2q\pi_2,$$

which is true if and only if

$$3\delta + \pi_2 \leq 2.$$

The second part follows since $\delta < 1$. ■

We now consider the comparison with the equilibria under complete information.

Proposition B.1 *Consider the bargaining extension of the game where Θ is commonly known and suppose that $c \leq c^*$ and $k \geq \max\{k^*, k^{**}\}$. Then,*

(I) If $\Theta = \Theta_C$, the equilibrium path is identical;

(II) If $\Theta = \Theta_h$, the equilibrium path is identical provided that $\pi_1 \geq 1 - \delta$;

(III) If $\Theta = \Theta_d$, the equilibrium path is identical.

Proof: In each case we need to check that the incentives to provoke are not affected by the possibility of bargaining. The proof has three parts:

(I) Follows directly from Lemma B.2 and the proof of Proposition 1.

(II) Follows from Lemma B.1 by considering that provocation yields π_1 to L, while no provocation and an offer of $v_0^* = \delta$ yields $1 - \delta$ to L.

(III) Again, by Lemma B.1, provocation gives the dovish L a payoff of 1 while no provocation gives her a payoff of $1 - \delta$.

■

As can be seen from Proposition B.1, when both L and E are hawks, or when L is a dove and E is a hawk, the introduction of bargaining has no effect, and no positive offers are made, along the equilibrium path. This means that when Attacker is cohesive, the conventional logic of deterrence is not affected by the potential for bargaining. Additionally, when Attacker is divided and initially led by a dove, introducing the possibility of bargaining has no effect on the deterrence curse.

The only case where bargaining affects the equilibria we identify in the main model is when L is a hawk and E is a dove, and when the payoff of provocation is sufficiently small, i.e. $\pi_1 < 1 - \delta$. The instability curse followed in the main model because L in Θ_h was unable to maintain power, and thus, would provoke to reap her benefit of provocation, π_1. By introducing the possibility of avoiding domestic threats through bargaining, the instability curse holds only when the payoff L receives from keeping power, net what she must give up to keep it, cannot compensate for the benefit L receives from provocation.

Last we explore how the introduction of bargaining affects the equilibria when there is asymmetric information between members of Attacker and Defender.

Proposition B.2 *Consider the bargaining extension of the game where Θ is privately known among members of Attacker and suppose that $c \leq c^*$ and $k \geq k^{**}$. Then, the introduction of bargaining*

(I) has no affect on Equilibrium 4.I when $k \leq \frac{1}{2}(2 - \delta + (1 - 2q)\pi_2)$, which is holds if and only if $3\delta + \pi_2 \leq 2$;

(II) has no effect on Equilibrium 4.II.

Proof: First, in order to establish the result, we must check that the hawkish L when $\Theta = \Theta_h$, does not have an incentive to not provoke. No provocation followed by a proposal

of δ, is optimal for the hawkish Leader only when $\pi_1 < 1-\delta$. However, since the equilibrium associated with the instability curse exists only when $\pi_1 \geq 1$, this condition is satisfied. We need only ensure that following provocation by L, and a proposal of $v_1^* = \delta + \pi_2 - (1 + (1 - q)\pi_2 - k)$, E will not support. By Lemma B.2, this occurs only when $k \in (k^{**}, \frac{1}{2}(2 - \delta + (1 - 2q)\pi_2)$.

The second part follows by noticing that no provocation yields at most $1-\delta$ in a divided Attacker, while according to the strategy profile, and using Lemma B.1, provocation yields at least 1. ■

The results of this section show that if the benefits of provocation are not too high for hawks, then they may be willing to forego provocation and reach an agreement through bargaining with dovish elites. The avoidance of conflict, and the conditions under which such avoidance can be achieved in this case, correspond to those highlighted by Fearon (1995), where conflict can be avoided through bargaining as long as there are no information frictions or commitment problems (see also Blattman and Miguel 2010).

C Cohesive Dovish Countries

In this subsection we consider our model with the added possibility of a country where Leader and Elite are both doves. Formally, this corresponds to the case where $\theta_L = \theta_E = 0$.

Lemma C.1 *If L and E are doves, then L has a strictly dominant strategy to not provoke.*

Proof: For a dovish L the largest achievable payoff in the game is achieved when she maintains power. By Lemma 1, if E is also dovish, she will support unless seizing power prevents retaliation. Since not taking provocative actions prevents retaliation, L does not have a strict incentive to provoke. ■

Lemma C.1 establishes that the main analysis we present immediately extends to the case

that includes the possibility of Attacker who is cohesive and dovish. In particular, by iterated elimination of strictly dominated strategies, the case where L and E are doves can be eliminated from consideration and the analysis in the main text is unchanged. In the model with asymmetric information, the posterior beliefs need to be modified to adjust for the knowledge that provocation was not pursued by a cohesive dovish attacker country.

References

Abramson, Scott F and David B Carter. 2016. "The historical origins of territorial disputes." *American Political Science Review* 110(4):675–698.

Ashworth, Scott and Kristopher W Ramsay. 2010. "Should audiences cost? optimal domestic constraints in international crises." *manuscript, Princeton University* .

Baliga, Sandeep, David O Lucca and Tomas Sjöström. 2011. "Domestic political survival and international conflict: is democracy good for peace?" *The Review of Economic Studies* pp. 458–486.

Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzsky. 2018. "Deterrence with Imperfect Atribution." *Mimeo* pp. 1–40.

Baliga, Sandeep and Tomas Sjöström. 2004. "Arms races and negotiations." *The Review of Economic Studies* 71(2):351–369.

Baliga, Sandeep and Tomas Sjöström. 2008. "Strategic ambiguity and arms proliferation." *Journal of political Economy* 116(6):1023–1057.

Baliga, Sandeep and Tomas Sjöström. 2012. "The strategy of manipulating conflict." *The American Economic Review* 102(6):2897–2922.

Banks, Jeffrey S and Joel Sobel. 1987. "Equilibrium selection in signaling games." *Econometrica: Journal of the Econometric Society* pp. 647–661.

Benson, Brett V, Adam Meirowitz and Kristopher W Ramsay. 2014. "Inducing deterrence through moral hazard in alliance contracts." *Journal of Conflict Resolution* 58(2):307–335.

Blattman, Christopher and Edward Miguel. 2010. "Civil war." *Journal of Economic literature* 48(1):3–57.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph Siverson and James Morrow. 2003. *The Logic of Political Survival*. MIT University Press.

Bueno de Mesquita, Bruce, James D Morrow and Ethan R Zorick. 1997. "Capabilities, perception, and escalation." *American Political Science Review* 91(1):15–27.

Carter, David B. 2010. "The Strategy of Territorial Conflict." *American Journal of Political Science* 54(4):969–987.

Chapman, Terrence L and Scott Wolford. 2010. "International organizations, strategy, and crisis bargaining." *The Journal of Politics* 72(1):227–242.

Chassang, Sylvain and Gerard Padró i Miquel. 2010. "Conflict and deterrence under strategic risk." *The Quarterly Journal of Economics* 125(4):1821–1858.

Chiozza, Giacomo and Hein Erich Goemans. 2011. *Leaders and international conflict*. Cambridge University Press.

Cho, In-Koo and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102(2):179–222.

Ellsberg, Daniel. 1961. "The crude analysis of strategy choices." *The American Economic Review* 51(2):472–478.

Ellsberg, Daniel. 1968. The theory and practice of blackmail. Technical report RAND CORP SANTA MONICA CA.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American Political Science Review* 88(3):577–592.

Fearon, James D. 1995. "Rationalist explanations for war." *International organization* 49(3):379–414.

Fey, Mark, Adam Meirowitz and Kristopher W Ramsay. 2013. "Credibility and Commitment in Crisis Bargaining." *Political Science Research and Methods* 1(1):27–52.

Fey, Mark and Kristopher W Ramsay. 2011. "Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict." *American Journal of Political Science* 55(1):149–169.

Freedman, Lawrence. 2013. *Strategy: a history*. Oxford University Press.

Gehlbach, Scott and Alberto Simpser. 2015. "Electoral manipulation as bureaucratic control." *American Journal of Political Science* 59(1):212–224.

Gurantz, Ron and Alexander V Hirsch. 2017. "Fear, appeasement, and the effectiveness of deterrence." *The Journal of Politics* 79(3):000–000.

Jackson, Matthew O and Massimo Morelli. 2009. "Strategic Militarization, Deterrence and Wars." *Quarterly Journal of Political Science* 4(4):279–313.

Jervis, Robert. 1978. "Cooperation under the security dilemma." *World politics* 30(2):167–214.

Jervis, Robert. 1979. "Deterrence theory revisited." *World Politics* 31(2):289–324.

Kilgour, D Marc and Frank C Zagare. 1991. "Credibility, uncertainty, and deterrence." *American Journal of Political Science* pp. 305–334.

Kurizaki, Shuhei. 2007. "Efficient secrecy: Public versus private threats in crisis diplomacy." *American Political Science Review* 101(3):543–558.

Kydd, Andrew. 1997. "Game theory and the spiral model." *World Politics* 49(3):371–400.

Kydd, Andrew. 2000. "Arms races and arms control: Modeling the hawk perspective." *American Journal of Political Science* pp. 228–244.

Kydd, Andrew H and Roseanne W McManus. 2017. "Threats and assurances in crisis bargaining." *Journal of conflict resolution* 61(2):325–348.

Meirowitz, Adam and Anne E Sartori. 2008. "Strategic uncertainty as a cause of war." *Quarterly Journal of Political Science* 3(4):327–352.

Morrow, James D. 1989. "Capabilities, uncertainty, and resolve: A limited information model of crisis bargaining." *American Journal of Political Science* pp. 941–972.

Myerson, Roger B. 2009. "Learning from Schelling's strategy of conflict." *Journal of Economic Literature* 47(4):1109–1125.

Nalebuff, Barry. 1991. "Rational deterrence in an imperfect world." *World Politics* 43(03):313–335.

Powell, Robert. 1985. "The Theoretical Foundations of Strategic Nuclear Deterrence." *Political Science Quarterly* 100(1):75–96.

Powell, Robert. 1987. "Crisis bargaining, escalation, and MAD." *American Political Science Review* 81(3):717–735.

Powell, Robert. 1989. "Nuclear deterrence and the strategy of limited retaliation." *American Political Science Review* 83(2):503–519.

Powell, Robert. 1990. *Nuclear deterrence theory: The search for credibility*. Cambridge University Press.

Schelling, Thomas. 1966. *Arms and Influence*. Yale University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard university press.

Schelling, Thomas C. 1967. The strategy of inflicting costs. In *Issues in defense economics*. NBER pp. 105–127.

Schelling, Thomas C. 2006. "An astonishing sixty years: the legacy of Hiroshima." *The American economic review* 96(4):929–937.

Schultz, Kenneth A. 1998. "Domestic opposition and signaling in international crises." *American Political Science Review* 92(4):829–844.

Schultz, Kenneth A. 2001. *Democracy and coercive diplomacy*. Vol. 76 Cambridge University Press.

Sechser, Todd S and Matthew Fuhrmann. 2013. "Crisis bargaining and nuclear blackmail." *International organization* 67(1):173–195.

Signorino, Curtis S. 1999. "Strategic interaction and the statistical analysis of international conflict." *American Political Science Review* 93(2):279–297.

Signorino, Curtis S and Ahmer Tarar. 2006. "A unified theory and test of extended immediate deterrence." *American Journal of Political Science* 50(3):586–605.

Smith, Alastair. 1998. "International crises and domestic politics." *American Political Science Review* 92(3):623–638.

Snyder, Glenn H. 1961. *Deterrence and Defense: Toward a Theory of National Security.* Princeton University Press.

Tyson, Scott A. 2018. "The Agency Problem Underlying Repression." *Journal of Politics* 80(4):1297–1310.

Wagner, R Harrison. 1992. "Rationality and misperception in deterrence theory." *Journal of Theoretical Politics* 4(2):115–141.

Wolford, Scott. 2007. "The turnover trap: New leaders, reputation, and international conflict." *American Journal of Political Science* 51(4):772–788.

Wolford, Scott. 2012. "Incumbents, successors, and crisis bargaining: Leadership turnover as a commitment problem." *Journal of Peace Research* 49(4):517–530.

Zagare, Frank C. 2004. "Reconciling Rationality with Deterrence: A Re-examinination of the Logical Foundations of Deterrence Theory." *Journal of Theoretical Politics* 16(2):107–141.

Zagare, Frank C and D Marc Kilgour. 1993. "Asymmetric deterrence." *International Studies Quarterly* 37(1):1–27.