Fear, Appeasement, and the Effectiveness of Deterrence$^1$

Ron Gurantz$^2$ and Alexander V. Hirsch$^3$

September 19, 2016

$^1$We thank Robert Trager, Barry O’Neill, Tiberiu Dragu, Kristopher Ramsay, Robert Powell, Douglas Arnold, Mattias Polborn, participants of the UCLA International Relations Reading Group, Princeton Q-APS International Relations Conference, and Cowbell working group, as well as the editor and three anonymous reviewers for helpful comments and advice.

$^2$University of California, Los Angeles. Department of Political Science, 4289 Bunche Hall, Los Angeles, CA 90095-1472. Phone: (310) 825-4331. Email: rongurantz@ucla.edu.

$^3$Professor of Political Science, Division of the Humanities and Social Sciences, California Institute of Technology, MC 228-77, Pasadena, CA 91125. Email: avhirsch@hss.caltech.edu.
Abstract

Governments often fear the future intentions of their adversaries. In this paper we show how this fear can make deterrent threats credible under seemingly incredible circumstances. We consider a model in which a defender seeks to deter a transgression with both intrinsic and military value. We examine how the defender’s fear of the challenger’s future belligerence affects his willingness to respond to the transgression with war. We derive conditions under which even a very minor transgression effectively “tests” for the challenger’s future belligerence, which makes the defender’s deterrent threat credible even when the transgression is objectively minor and the challenger is ex-ante unlikely to be belligerent. We also show that fear can actually benefit the defender by allowing her to credibly deter. We apply the model to analyze a series of historical cases, and show the robustness of our results to a variety of extensions.
A central question in the study of deterrence has been how threats can be credible when they are meant to defend interests that do not immediately appear to be worth fighting over. For example, in 1954-55 the Eisenhower administration prepared for war and even raised the possibility of using nuclear weapons in response to a Communist Chinese attack on the sparsely populated island of Quemoy, ultimately deterring Chinese aggression (Soman 2000). More recently, the government of North Korea threatened war in response to both economic sanctions and an airstrike on their nuclear plant, and evidence suggests that the United States government took these threats seriously.¹

Indeed, many of the most significant Cold War crises were over stakes that were relatively insignificant compared to the costs and consequences of nuclear war. Consequently, much of classical deterrence theory was developed to understand how the United States could credibly threaten to use (possibly nuclear) force “even when its stakes were low” (Danilovic 2001). Deterrence scholars explored many such mechanisms, including “threats that leave something to chance” (Schelling 1966; Nalebuff 1986; Powell 1987), limited commitments like “trip-wire” forces and public speeches (Fearon 1994; Schelling 1966; Slantchev 2011), and reputation (Alt, Calvert and Humes 1988; Sechser 2010).

In this paper, we formally explore a novel mechanism that can explain how the threat of a major war can credibly deter an objectively minor transgression. Our model does not rely on commitment devices, or on concerns about a defender’s reputation. Instead, we study how a defender’s fear about her adversary’s future intentions affects her willingness to fight in response to a minor transgression. We derive conditions under which even a minor transgression can signal hostile intentions that go far beyond the immediate stakes of a crisis, making major war a rational response by the defender. The adversary will then be deterred if it wants to avoid signaling these intentions and provoking a major war. The model produces surprising results that previous theories have been unable to identify: that the defender can be better off fearing its adversary’s intentions rather than knowing they are benign; that a condition that leads to war when bargaining under complete information leads to deterrence.

under incomplete information; and that the relationship between the military and intrinsic value of a transgression, rather than the size of either individual value, determines deterrence credibility.

**Main Result** The model incorporates two features of international crises that are prominent in the international relations literature, but rarely studied in deterrence models: a defender’s uncertainty about a challenger’s intentions, and endogenous power shifts (Fearon 1996; Powell 2006). In the model there is a potential transgression with both direct value to the challenger if there is peace, and military value if there is war. The defender prefers to allow the transgression if it would lead to peace, but he is uncertain about the challenger’s intentions, and fears she is unappeasably belligerent. We use the term “fear” to refer to the defender’s belief that the challenger may be belligerent with positive probability; this encompasses both that he is uncertain about the challenger’s intentions, and that he entertains the possibility that she affirmatively desires war.

We show that combining these ingredients can produce credible deterrence under seemingly incredible circumstances: when the challenger is very unlikely to be unappeasably belligerent, the transgression is incredibly minor, and the threatened response is a major and costly war. Why does this happen? Intuition suggests that a peaceful defender would allow a minor transgression if the challenger is unlikely to exploit it in a future war. But this intuition ignores a key element: that a credible threat of war affects what the defender can infer from a transgression. Specifically, a challenger who transgresses in the face of a credible threat of war reveals that she prefers triggering war to accepting the status quo. This revelation can lead a defender to infer that war is inevitable, inducing him to initiate it immediately rather than waiting to first be weakened by a small power shift. If the challenger believes the defender to be using this logic, then she will indeed be deterred from transgressing unless she actually desires war, fulfilling the defender’s expectations. In our mechanism, the challenger’s reaction to the deterrent threat is thus part of what sustains its credibility: “types” of challengers against whom a defender would not want to fight are “screened out.”

---

2Throughout the paper we refer to the defender as “he” and the challenger as “she.”
We next explore the conditions under which this logic is most likely to produce credible deterrence. The key is to consider what transgressions effectively “test” for the inevitability of war. We show that a transgression’s absolute size is not what makes it a good test, in contrast to a large literature arguing that the credibility of deterrent threats derives from the “stakes” involved (Danilovic 2002; Zagare 2004). Rather, what matters is the extent to which allowing the transgression would fail to appease an already belligerent challenger. The reason is that the defender can already infer the challenger’s initial belligerence from observing the transgression itself. The implication is surprising: the less likely that allowing the transgression will appease an already-belligerent challenger, the more likely deterrence will work. When allowing a trangression cannot appease an already-belligerent challenger, then deterrence can always work: regardless of how minor the transgression or how costly a war.

An example helps to both clarify the logic, and demonstrate the mechanism’s relevance to a historical deterrence scenario. During the early Cold War, the United States feared the Soviet Union intended to launch a full-scale war against Western Europe and the United States. A 1952 National Security Council report on possible U.S. responses to Soviet aggression against West Berlin begins by asserting that “control of Berlin, in and of itself, is not so important to the Soviet rulers as to justify involving the Soviet Union in general war” (FRUS 1952-54 VII, 1268-69). Thus, the report reasons that the Soviet Union will only attack West Berlin if they “decide for other reasons to provoke or initiate general war,” and that the United States would therefore “have to act on the assumption that general war is imminent.” In other words, an invasion of Berlin must imply that the Soviet Union both expects to trigger a wider war and affirmatively desires it, rather than implying that they think they can conquer Berlin without war. Since an invasion would imply that a wider war is imminent, the United States was to respond with “full implementation of emergency war plans,” thereby fulfilling the United States’ commitment to fight in the event of an invasion.

Can Fear Be Beneficial? In the deterrence mechanism we present, the defender’s willingness to fight and ability to deter are driven by his fear of the adversary’s intentions. Without this fear
the challenger could not be influenced by the “signaling” implications of her actions. This raises an unusual possibility: might the defender actually benefit from being incompletely informed about the challenger’s intentions? To answer this question, we compare our baseline model to a slight variant in which the defender is informed up-front about the challenger’s “type” rather than attempting to infer it from her actions. We show that an informed defender is strictly less able to deter than a fearful one, since the ability to deter is rooted in fear. In addition, the defender may indeed be better off remaining ignorant and fearful, and thus may actually choose to do so! Whether this is the case again depends on how effective is the transgression at appeasing an already belligerent challenger. If it is not very effective or entirely ineffective, then the downside risk of fear – that it will result in an avoidable war against an appeasable challenger – is outweighed by the deterrence benefits.

Which Actions Sustain Deterrence? The model shows that deterrence can succeed when intuition suggests that it should fail, and links this success to the anticipated ineffectiveness of appeasement. To derive testable empirical implications about deterrence success thus requires answering a simple question: what sort of transgression is least effective at appeasing a belligerent challenger?

Usefully, the literature on bargaining under complete information with endogenous power shifts has already answered this question: it is one whose military value to the challenger exceeds its direct value (Fearon 1996; Schwarz and Sonin 2008). The reason is that allowing such a transgression will only increase an already-belligerent challenger’s appetite for war. The bargaining literature shows that when the challenger’s initial belligerence under the status quo is known, such a condition results in inefficient war, or the gradual elimination of the defender. Combined with our analysis, this implies that when the challenger’s belligerence is merely feared – even if only with infinitesimal probability – a transgression with this property can always be effectively deterred with a credible threat of war, regardless of how costly the war or minor the transgression.

Lastly, we generalize this result to show that deterrence is more likely the greater is the difference between a transgression’s military and direct values to the challenger. The substantive implication
is that it is not only, or even mainly, the size of a transgression that matters for deterrence. Equally important is the relationship between its military and direct values. Thus, empirical studies of deterrence controlling for the “interests at stake” may be flawed because they fail to properly measure, or separately control for, these values and the relationship between them (Huth 1999). For example, our results suggest that a challenger may treat a defender’s threat to fight for a barren rock as credible, if the rock yields even a minor strategic advantage. The defender can reason that if the challenger is already belligerent, allowing her to occupy the rock will only make her (a bit) more so. Conversely, a challenger may also discount a defender’s threat to fight for a valuable population center because conceding it might reasonably appease her. We apply this reasoning to three brief case studies and show how it can help explain otherwise-puzzling variation in crisis outcomes.

The paper proceeds as follows. We first discuss related literature and motivate our model. Next, we present the model and derive results. We then present three brief case studies showing that the relationship between a transgression’s military and direct values can help predict variation in crises outcomes. Next, we discuss robustness of our results to several changes to the information structure, game sequence, and bargaining protocol. In particular, we show our insights hold in a fully “rationalist” extension in which war is costly and the challenger may make a successive series of small demands (Fearon 1995). Finally, we summarize and pose questions for future research.

Related Literature

Deterrence Theory The academic study of deterrence began with the recognition that the United States faced a credibility problem in the Cold War due to the catastrophic nature of nuclear war (Trachtenberg 1989). Theorists have developed a variety of mechanisms that explain why states may be willing to fight a war whose costs seem disproportionate to the stakes involved. These mechanisms included “probabilistic threats” that increase the chance war will break out through uncontrollable events, “commitment devices” that make it more difficult for a state to back down such as audience
costs or “trip-wires,” and reputational considerations (Schelling 1966; Fearon 1994; Sechser 2010). The first mechanism allows states to make threats that are “proportionate” to the stakes, while the latter two increase the cost of concession beyond the immediate stakes of the crisis.\(^3\)

These theories have been very influential, but each faces theoretical and empirical problems in explaining credible threats of catastrophic war. The idea of “probabilistic threats” is that states can manipulate the chance a catastrophic war breaks out through processes they don’t control (Powell 1987). However, case study evidence suggests that governments always face a moment when they have the discretion of whether or not to escalate to full-scale war (Howard 1984; Luard 1986). In his study of over 500 years of major conflict, Luard (1986) writes “it is impossible to identify a single case in which it can be said that a war started accidentally: in which it was not, at the time when war broke out, the deliberate intention of at least one party that war should take place.”

With respect to audience costs, scholars have struggled to identify historical cases where they have played a major role (Snyder and Borghard 2011; Trachtenberg 2012). The logic of trip-wire mechanisms requires a government to initiate a catastrophic war rather than abandon a few thousand troops, which stretches credulity; Enthoven (1975) writes that the Soviet Union would not believe “the United States would be willing to risk the destruction of more than 100 million Americans merely because a small number of American troops in Europe were threatened.” Even Schelling (1966, 47), in his well-known description of the trip-wire force in West Berlin, ultimately relies on a reputational mechanism, writing that the United States would respond to an attack because the troops represented “the pride, the honor, and the reputation of the United States government.”

The logical difficulties of trip-wire explanations also extend to reputational ones; Mercer (1996), Danilovic (2002) and others argue that it is illogical for states to endanger their core interests by starting a major war to maintain a reputation for defending those same core interests. More

\(^3\)Our model is closer to the latter two in that the defender refuses to concede because of consequences beyond the immediate stakes of the crisis.
Importantly, scholars have failed to find empirical evidence for the core property of reputational theories – that states’ past crisis behavior influences future deterrence crises. In studies of pre-WWII diplomacy and the Cold War, respectively, Press (2005) and Hopf (1994) find that backing down in a crisis did not seem to influence enemy expectations about behavior in future crises. Similarly, across a large number of cases Huth and Russett (1984) and Danilovic (2002) find that whether or not a defender stood firm in past crises does not predict whether he will successfully deter in later crises.

**Inherent Credibility and Deterrence** These logical and empirical weaknesses have led some scholars to argue that the emphasis on explaining “disproportionate threats” has been misguided; in fact, most credible threats are actually “proportionate” most of the time. Specifically, Danilovic (2002) and others argue that the credibility of a defender’s deterrent threats derives primarily from their actual stakes in a crisis. To support this proposition, Danilovic (2002) develops an empirical measure of the defender’s stakes in a dataset of historical deterrence crises, and indeed finds that her measure strongly predicts deterrence success. Earlier work also finds evidence that the value of the stakes influences crisis outcomes (Huth and Russett 1984; Huth 1988).

However, this approach too has theoretical and empirical limitations. If it were true that the credibility of deterrent threats was based solely on the contemporaneous stakes in a crisis, then all states would be vulnerable to “salami tactics” in which their adversaries carefully calibrate each successive demand to be below their threshold for fighting. But historically salami tactics have failed, such as the Soviet Union’s attempt to make piecemeal gains following World War II.4

Empirically, a key limitation of such studies lies in how they actually measure a defender’s stakes in a crisis. Roughly speaking, this is done by imputing a defender’s interests in the broader region in which a crises takes place to the crisis itself.5 For example, the strength of the U.S.’s interests on the broader European continent are used as a proxy for the U.S.’s stakes in a crisis over West

---

4These included demands in northern Iran, the Turkish Straits, Trieste, Berlin, and Tripolitania.
5Huth and Russett (1984) and Huth (1988) use the strength of a state’s relationship with a protege to proxy for the stakes in crises involving the protege, but no particulars about what is under dispute.
Berlin. While this exercise provides useful evidence that states sometimes approach crises over small stakes by thinking about the larger context, it is unable to help us understand exactly why and when they do so. Instead, by effectively assuming that some small stakes are actually large, this approach underestimates the empirical prevalence of credible disproportionate threats. Moreover, it is unable to explain variation in crisis outcomes occurring within a broader region where interests are strong.

A final limitation of the “inherent credibility” approach is that it assumes a fixed relationship between the value of an asset under dispute, and the willingness of a defender to fight for it. However, to the extent that an asset has military value, that value is only instrumental for defending other presumably valuable assets in a war. A defender’s willingness to fight over such an asset should therefore logically depend on his beliefs about his adversary’s future intentions, an approach consistent with previous studies of appeasement (Powell 1996; Hirshleifer 1991).

Our model parsimoniously captures these subtleties and their impact on the credibility of deterrent threats. It provides a theoretical link between the deterrence of minor transgressions and the larger issues at stake, generates new predictions by differentiating the military from the inherent value of an asset, and allows for variation in and uncertainty about a challenger’s future intentions.

**Deterrence and International Relations Theory** By developing a deterrence model with both endogenous power shifts and challenger-side uncertainty, we are also able to connect deterrence to literatures in international relations where these factors have been more prominent. Endogenous shifts in military power have been widely studied. An early example is Powell (1996), who examines states’ responses to salami tactics. More recently, Kydd and McMamus (2014) study how endogenous power shifts create incentives for states to make costly commitments in the form of assurances.

One major branch of this literature studies bargaining over objects or actions that are both intrinsically and strategically valuable, like the transgression in our model (Fearon 1996; Schwarz and Sonin 2008). A core dilemma in these works (and ours) is that allowing such actions may appease an adversary, but will also make them stronger. The key distinction is that these works assume a
challenger whose belligerence under the status quo is known (i.e. complete information), while we assume a challenger whose belligerence is merely feared (perhaps with infinitesimal probability). These works find that it is difficult or impossible to maintain stable settlements under a specific condition: if the available peaceful arrangements shift the military balance toward the challenger more quickly than they increase her payoff from peace (Fearon 1996; Schwarz and Sonin 2008). This resembles the unappeasability condition in our model for successful deterrence. Thus, one interpretation of our result is that a condition resulting in war or unstable settlements under complete information results in deterrence and a fearful peace with incomplete information.

A second branch of the power shifts literature studies states’ strategic military investments (Slantchev 2011; Debs and Monteiro 2014). Like the transgression in our model, military investments shift military power. But unlike the transgression, they are intrinsically costly rather than beneficial; this presents adversaries with a very different inference problem. Slantchev (2011) studies the case of a state whose preferences are unknown by her adversary (like our model), but in which the adversary has no opportunity to “preempt” a power shift (unlike our model) – he analyzes a defender’s incentive to signal strength with costly military investments. Debs and Monteiro (2014) consider the case of a state whose preferences are known (unlike our model) but whose military investments are unknown (also unlike our model) – they analyze how a defender’s fear that his adversary has militarized can lead him to initiate preventive war even absent conclusive evidence.

Finally, our work relates to an entirely different literature that also considers what happens when states fear that their adversaries may be unappeasably belligerent. Several works analyze how this possibility induce a “spiral” of fear that causes peace to unravel (Baliga and Sjostrom 2009; Chassang and Miquel 2010). Alternatively, Acharya and Grillo (2014) focus on how an adversary may exploit this fear by taking actions that attempt to mimic a “crazy type.”
The Model

The model is a two-period game played between a challenger (C) and a defender (D).

**Sequence** In the first period, the challenger chooses whether or not to attempt a transgression $x^1 \in \{a, \emptyset\}$ that has both *direct value* to her in the event that peace prevails, and *military value* in the event that war breaks out. The transgression could represent any number of prohibited actions that would shift the military balance toward the challenger, but also benefit her if her intentions vis-a-vis the defender were ultimately peaceful; it therefore presents the defender with an inference problem about the challenger’s true intentions. Such actions could include occupying territory belonging to the defender or a protégé, enacting sanctions, or developing valuable scientific technology that could be weaponized like nuclear capability. The challenger’s attempt to transgress is observable to the defender, and thus could also be interpreted as making a demand of the defender to allow it.

If the challenger does not attempt to transgress ($x^1 = \emptyset$), then the game ends with peace. If she does ($x^1 = a$), then the defender may either allow the transgression ($y^1 = n$) or resist it ($y^1 = w$). To make credible deterrence as difficult as possible, we assume that the challenger’s act presents the defender with a fait accompli; to resist the transgression means war. If the defender allows the challenger to transgress, then the game proceeds to a second period. In the second period, the challenger’s payoffs are assumed to be higher in the event of either peace or war as a result of having successfully transgressed, and the defender’s are assumed to be lower. The challenger then decides whether to enjoy her direct gains and end the game peacefully ($x^2 = n$), or herself initiate war under the more favorable military balance ($x^2 = w$). The sequence of the game is depicted in Figure 1.

**Defender’s Payoffs** Unlike reputational models of deterrence credibility, we assume that the defender’s payoffs are common knowledge. Moreover, he has a known preference for appeasement. To capture that preference we denote the defender’s payoff as $n^t_D$ if the game ends with peace in period $t$ and $w^t_D$ if the game ends with war in period $t$, and assume that:
1. allowing the transgression makes him worse off in both peace ($n_2^D < n_1^D$) and war ($w_2^D < w_1^D$),

2. allowing the transgression is strictly better than responding with war if the challenger will subsequently choose peace ($n_2^D > w_1^D$).

Given these assumptions, the defender’s optimal response to a transgression depends on his interim assessment of the probability $\beta$ that a challenger who has attempted to transgress will initiate war even after being allowed to do so. If war is inevitable, then he prefers to avoid the cost $w_1^D - w_2^D > 0$ of allowing an unappeasably belligerent challenger to transgress, which captures the (potentially small) endogenous shift in military power. However, if allowing the transgression would actually appease the challenger, then he prefers to do so and avoid the cost $n_2^D - w_1^D$ of a preventable war.

It is easily shown that the defender will prefer to respond to the transgression with war whenever $\beta$ exceeds a threshold $\tilde{\beta}$, where $\tilde{\beta} = \frac{n_2^D - w_2^D}{(n_2^D - w_1^D) + (w_1^D - w_2^D)} \in (0, 1)$. Crucially, $\tilde{\beta} < 1$ – that is, if war is truly inevitable, then the defender prefers war sooner to war later regardless of its cost.

The defender’s dilemma in our model is thus closely related to Powell’s (1996) analysis of “salami
tactics,” which we further explore in the Robustness section. The defender is vulnerable to exploitation by the challenger because a small transgression is below his known threshold for war. However, his fear that the challenger’s intentions may in fact be far reaching, and his preference for war sooner rather than war later if it is to be inevitable, may sometimes lead him to respond with war.

**Challenger’s Payoffs** Because the defender never intrinsically prefers to fight to prevent the transgression, the key factor sustaining his willingness to do so must be his fear that the challenger seeks to strengthen herself for a future war. To model this fear, we assume that the challenger has fixed and known payoffs \( n^t_C \) for peace in each period, but her payoffs from war \( w^t_C(\theta_C) \) depend on a type \( \theta_C \in \Theta \subset \mathbb{R} \) drawn by “nature” at the start of the game, where \( \Theta \) is a closed interval. The defender’s prior beliefs over the challenger’s type are described by an atomless distribution with full support over \( \Theta \) and CDF \( F(\theta_C) \), and the challenger’s payoffs \( n^t_C \) and \( w^t_C(\theta_C) \) satisfy the following:

1. Successfully transgressing has a *direct value* if the game ends in peace \( (n^2_C - n^1_C > 0) \) and a *military value* if the game ends in war \( (w^2_C(\theta_C) - w^1_C(\theta_C) > 0 \forall \theta_C \in \Theta) \),

2. In each period \( t \) the challenger’s war payoff \( w^t_C(\theta_C) \) is continuous and strictly increasing in \( \theta_C \).

In addition, there exists a unique challenger type \( \bar{\theta}^t_C \), strictly interior to \( \Theta \) that is *indifferent* between peace and war (i.e. \( w^t_C(\bar{\theta}^t_C) = n^t_C \)).

Our assumptions imply the following. First, all challenger types intrinsically value the transgression, i.e., even absent a war. Second, a challenger’s type \( \theta_C \) indexes her willingness to fight. Third and most importantly, in each period \( t \) there is positive probability that the challenger prefers peace to war \((\theta_C < \bar{\theta}^t_C)\) and war to peace \((\theta_C > \bar{\theta}^t_C)\). Thus, there is always the possibility (however unlikely) that she prefers war to the status quo \((w^1_C(\theta_C) \geq n^1_C \iff \theta_C \geq \bar{\theta}^1_C)\). In addition, once the challenger has successfully transgressed, there is always the possibility that the challenger is a type against whom war is inevitable; formally, these are types \( \theta_C \geq \bar{\theta}^2_C \) who would unilaterally initiate war even after being allowed to transgress.
Although challenger types $\theta_C > \bar{\theta}_C^2$ are modeled as unilaterally initiating war, this outcome could also represent an unmodeled continuation game where the challenger makes an additional demand against which the defender is willing to fight. Interpreted as such, a number of rationales for the defender’s willingness to fight are possible; it could once again be driven by fear that war is inevitable in a future unmodeled period, his threat over the subsequent demand could be “intrinsically” credible as in perfect deterrence theory (Zagare 2004), he could fail to fully internalize the cost of war (Chiozza and Goemans 2004; Jackson and Morelli 2007), or war could result from a commitment problem (Powell 2006). Our baseline model is agnostic about the rationale so as not to confuse the main points. However, in the “salami tactics” extension of the model considered in the Robustness section, wars result from a mixture of fear and commitment problems. This extension demonstrates that the model’s results are robust to a scenario in which states bargain over many periods and war is costly, consistent with the bargaining literature (Fearon 1995).

Main Results

We now characterize equilibria and present main results; all proofs are located in the online Appendix.

**Proposition 1.** A pure strategy equilibrium of the model always exists.

1. There exists a no deterrence equilibrium, in which the challenger always transgresses, and she is always permitted to do so, i.f.f. $\tilde{\beta} \geq P(\theta_C \geq \bar{\theta}_C^2)$

2. There exists a deterrence equilibrium, in which (i) the defender always responds to the transgression with war, (ii) all types $\theta_C < \bar{\theta}_C^1$, who do not initially prefer war are deterred, and (iii) the probability of deterrence is $P(\theta_1 < \bar{\theta}_C^1)$, i.f.f. $\tilde{\beta} \leq P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$

When both pure strategy equilibria exist, there also exists a mixed strategy equilibrium, but the defender is best off in the deterrence equilibrium.
A pure strategy equilibrium thus always exists, and payoff-dominates any mixed strategy equilibrium for the defender. We therefore restrict attention to these. Pure strategy equilibria are of two types. The first is a “no deterrence equilibrium.” The challenger always attempts to transgress, and consequently the defender can infer nothing about the challenger simply by observing the transgression itself. He therefore decides how to respond on the basis of his prior $P(\theta_C \geq \bar{\theta}_C^2)$ that the challenger is sufficiently belligerent to initiate war after transgressing. If that prior $P(\theta_C \geq \bar{\theta}_C^2)$ is low and/or the defender’s belief threshold $\bar{\beta}$ for responding with war is high, then this equilibrium will exist. Recall that $\bar{\beta}$ is determined by the cost $n_D^2 - w_D^1$ of an avoidable war relative to the cost $w_D^1 - w_D^2$ of allowing an unappeasably belligerent challenger to transgress. These conditions accord with the standard logic for when deterrence should fail – when the cost of war is high relative to the defender’s “stakes,” and the challenger is very unlikely ex-ante to be belligerent.

The second type of pure strategy equilibrium, however, is a “deterrence equilibrium.” In this equilibrium the defender responds to the transgression with war. Consequently, the challenger is deterred from transgressing unless she is initially belligerent, in the sense of preferring war to the status quo (i.e. $\theta_C \geq \bar{\theta}_C^1$). This deterrence allows the defender to draw an inference from observing the transgression itself even if it is objectively minor – precisely that the challenger is initially belligerent. As a result, he decides whether to respond with war not on the basis of his prior $P(\theta_C \geq \bar{\theta}_C^2)$ that the challenger will initiate war after transgressing, but his posterior $P(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1)$ that allowing an already belligerent challenger to transgress will fail to appease her.

This simple observation is in fact our key insight. In the presence of fear that war may be inevitable, the primary factor determining the defender’s ability to credibly deter in equilibrium is not the cost of war, the severity of the transgression, or the initial probability that the challenge is belligerent. The reason is that when deterrence is actually effective, the defender can already infer the challenger’s initial belligerence from the transgression itself. Instead, the primary factor is actually the effectiveness of appeasement against an already belligerent challenger, encapsulated by
the probability $P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$ that an already-belligerent challenger will remain belligerent after transgressing. The implications of this simple insight are surprisingly strong.

**Corollary 1.** When allowing the transgression cannot appease an already belligerent challenger, i.e. $\bar{\theta}_C^2 \leq \bar{\theta}_C^1 \iff P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1) = 1$, then the deterrence equilibrium exists for all defender payoffs and probability distributions satisfying the initial assumptions.

Thus, when appeasement is impossible against a belligerent challenger, the deterrence equilibrium always exists. This is true even when the “no deterrence equilibrium” also exists because of a high cost of war $(n^1_D - n^1_D)$, a low cost of allowing the transgression in both direct $(n^1_D - n^2_D)$ and military $(w^1_D - w^2_D)$ terms, and/or a sufficiently low probability that the challenger is belligerent $P(\theta_C \geq \bar{\theta}_C)$ in both periods. The deterrence equilibrium remains because the defender can use the transgression (however minor) as a test of the challenger’s initial belligerence, knows that initial belligerence ensures future belligerence because appeasement is ineffective, and therefore prefers to respond with war upon observing the transgression. The challenger is thereby deterred unless she affirmatively prefers immediate war, fulfilling the defender’s expectations.

Figure 2 depicts the equilibrium correspondence when the defender’s belief threshold $\bar{\beta}$ for responding with war is very high (the area where Corollary 1 holds is not indicated, but is in the region above a 45 degree line drawn from the origin). The defender’s prior $P(\theta_C \geq \bar{\theta}_C)$ that the challenger

---

6The irrelevance of the defender’s cost of war for Corollary 1 partially depends on a literal interpretation of the second period. If it is instead interpreted as an unmodeled continuation game where the challenger attempts a transgression against which the defender is willing to fight, the equilibrium further requires that the defender prefer a war to ceasing to exist entirely. We address this point more fully in the “salami tactics” extension in the Robustness section.

7Baliga and Sjostrom (2008) also exhibits a qualitatively similar separating equilibrium when their assumption (3) fails and “crazy types” value weapons sufficiently highly – they reveal themselves by acquiring weapons and refusing inspections, and the defender sometimes attacks. However, their model uses reduced-form payoffs, and assumes outright that some defender types prefer to attack only if the challenger is crazy. We explicitly derive the defender’s desire to attack a “crazy type” from an anticipated power shift, and show that this logic will make any defender prefer to attack a crazy type. Importantly, their model also abstracts away from the payoff properties of a transgression itself, while our key results directly relate these properties to the effectiveness of deterrence.
prefers war to the status quo is on the x-axis, while the prior $P(θ_C ≥ \bar{θ}_C)$ that she would initiate war after transgressing is on the y-axis; both quantities are derived from the challenger’s underlying payoffs and the distribution over her type. The figure shows that the deterrence equilibrium can remain even when the probabilities that the challenger would be belligerent in either period are arbitrarily low, which can be seen by observing that the hatched triangle extends to the origin. Moreover, this property would persist even if the defender’s threshold for war $\bar{β}$ were made arbitrarily high.

Figure 2: Equilibria

The Benefits of Fear The preceding analysis shows that it can be the defender’s fear – rather than an intrinsic willingness to fight – that allows him to credibly deter. This suggests that the defender may actually benefit in expectation from his fear and uncertainty, and by implication may actually choose to remain ignorant of her adversary’s intentions. To examine whether this is true we compare the baseline model to a variant that is identical in every respect except that the challenger’s type $θ_C$ is revealed to the defender at the start of the game, which yields the following results.
Proposition 2. Suppose that the deterrence equilibrium prevails whenever it exists. Then,

1. the probability of deterrence would decrease if the defender knew the challenger’s type.

2. when the probability \( P(\theta_C < \bar{\theta}^2_C | \theta_C \geq \bar{\theta}^1_C) \) that appeasement is effective is below

\[
\min \left\{ 1 - \beta, \left( \frac{P(\theta_C < \bar{\theta}^1_C)}{P(\theta_C \geq \bar{\theta}^1_C)} \right) \cdot \left( \frac{n_D^1 - n_D^2}{n_D^2 - w_D} \right) \right\},
\]

the defender is better off in expectation not knowing the challenger’s type.

Proposition 2 first shows that the probability of deterrence always decreases when the challenger’s type is revealed to the defender. To see why, suppose for simplicity that appeasement is completely ineffective (\( \bar{\theta}^2_C \leq \bar{\theta}^1_C \)), but the challenger is actually peaceful (\( \theta_C \leq \bar{\theta}^2_C \)). If the defender were to learn this (and the challenger knew that she had), then the challenger would exploit the defender’s known preference for appeasement. However, if the defender remains ignorant, then he can credibly deter by maintaining his fear that the challenger is unappeasably belligerent (\( \theta_C \geq \bar{\theta}^1_C \)).

The second part of the Proposition shows that the defender’s uncertainty indeed sometimes benefits her in expectation. This is the case when the effectiveness of appeasement \( P(\theta_C < \bar{\theta}^2_C | \theta_C \geq \bar{\theta}^1_C) \) is sufficiently low. The stated condition ensures that the expected cost of fighting avoidable wars against appeasable challengers (\( \theta_C \in [\bar{\theta}^1_C, \bar{\theta}^2_C] \)) is outweighed by the expected benefits of deterring peaceful challengers (\( \theta_C < \bar{\theta}^1_C \)). It clearly holds when appeasement is impossible (\( \bar{\theta}^2_C < \bar{\theta}^1_C \)), and thus in this case the defender is unambiguously better off being ignorant. Counterintuitively then, the defender’s fear can actually be a source of strength, an insight that yields the following corollary.

Corollary 2. Suppose that at the start of the game, the defender could costlessly and publicly choose whether to learn the challenger’s type \( \theta_C \) or to remain ignorant. When the condition in Proposition 2.2 holds, then the best equilibrium for the defender involves remaining ignorant.
The deterrence benefits of fear may thus be sufficiently large that a rational defender would actually choose to remain fearful rather than learn her adversary’s intentions!

**The (In)effectiveness of Appeasement**

Our analysis demonstrates that in the presence of fear about a challenger’s intentions, the effectiveness of appeasement and the credibility of deterrence are really two sides of the same coin. Deterrence can be credible if appeasement would be ineffective against a belligerent challenger, even if the ex-ante probability of that belligerence is very low. Conversely, if appeasement could be effective then deterrence can be undermined, even if allowing the transgression is costly and the ex-ante probability that the challenger is belligerent is high.

We conclude by thus directly considering the question of what makes appeasement less effective, and consequently deterrence more effective. To answer this question we examine the payoff properties of the transgression itself. Recall that transgressing has both a military value $w^2_C(\theta_C) - w^1_C(\theta_C)$, which is the challenger’s gain in the event of war, and a direct value $n^2_C - n^1_C$, which is her gain in the event of peace. We henceforth denote these as $\delta^m_C(\theta_C)$ and $\delta^d_C$ respectively and ask how they influence the effectiveness of deterrence, which yields the following result.

**Corollary 3.** Appeasement is ineffective, and thus the deterrence equilibrium exists for all defender payoffs and probability distributions satisfying the initial assumptions, if and only if $\delta^m_C(\bar{\theta}^1_C) \geq \delta^d_C$.

Thus, a sufficient condition for the deterrence equilibrium to exist is that the military value of the transgression $\delta^m_C(\cdot)$ exceed its direct value $\delta^d_C$ to a challenger of type $\bar{\theta}^1_C$ who is initially indifferent between peace and war. For such a challenger type, $\delta^m_C(\bar{\theta}^1_C) \geq \delta^m_C$ means that the military gains from successfully transgressing increase her net benefit from war as much as the direct gains from transgressing reduce it. Since she initially weakly preferred war to peace, she and all types more belligerent than her must continue to prefer war to peace after transgressing. Allowing the transgression therefore cannot appease any type of challenger who was initially belligerent (i.e.
\[ P(\theta_C \geq \bar{\theta}_C \mid \theta_C \geq \bar{\theta}^1_C) = 1, \] which by Corollary 1 implies that the deterrence equilibrium exists.

The condition in Corollary 3 is familiar from the literature examining complete information bargaining with endogenous shifts in military power (Fearon 1996; Schwarz and Sonin 2008). To our knowledge, however, it is absent from the literature (either empirical or theoretical) on deterrence. The bargaining literature finds that similar conditions generally result in wars or the gradual elimination of one player. In contrast, we find that this condition can lead to a fearful peace with deterrence of even a very minor transgression with very high probability. Both predictions are rooted in the same property; allowing the transgression cannot appease a belligerent challenger. However, the distinction arises from the difference in assumptions about whether the challenger is initially belligerent. In the complete information setting, belligerence at the outset is assumed. In our model, the defender can believe that the challenger is very likely to be peaceful ex-ante; however, his fear that the challenger is unappeasably belligerent allows him to credibly deter.

**Deriving Empirical Implications** The preceding result takes us part-way toward extracting empirical implications by examining the properties of the transgression itself. However, the model still exhibits multiple equilibria, and our analysis cannot speak to how the players will form expectations about which one will prevail. We therefore proceed with the additional assumption that the deterrence equilibrium will prevail whenever it exists. With this, the probability that deterrence succeeds is 0 when the deterrence equilibrium does not exist, and is \( P(\theta_C < \bar{\theta}^1_C) \) when it does. This allows us to consider the probability of successful deterrence as a function of the challenger’s payoffs (holding those of the defender’s fixed), and yields the following empirical prediction.

**Proposition 3.** Suppose that (i) the deterrence equilibrium prevails whenever it exists, (ii) the transgression’s military value is equal to \( \delta^m_C \) for all challenger types, and (iii) the challenger’s first period payoffs are held fixed. Then the probability that deterrence is successful is increasing in \( \delta^m_C - \delta^d_C \).

The model thus predicts that the probability of deterrence is increasing in the difference \( \delta^m_C - \delta^d_C \).
between the military and direct value of the transgression to the challenger. The intuition is similar to Corollary 3; the greater is $\delta^m_C - \delta^d_C$, the more likely it is that appeasement will fail against a belligerent challenger, the more willing is the defender to respond with war conditional on deterrence failing, and the better able he is to deter. This effect is depicted in Figure 3; the left panel shows the probability of deterrence when the defender’s payoffs are fixed, while the right panel depicts the probability when the defender’s payoffs are initially drawn from a distribution.\(^8\).

Finally, observe that Proposition 3 varies the challenger’s values for transgressing while holding those of the defender fixed. However, in many applications it is reasonable to suppose that a transgression with greater direct or military value for the challenger is also one that imposes greater direct or military costs on the defender. This relationship affects predictions because a transgression with a greater direct value might more effectively appease, but is also more intrinsically worth fighting over. To better understand how this wrinkle would modify our results, Figure 4 considers a numerical example in which values to the challenger for transgressing are equal to the costs imposed on the defender. The figure demonstrates that the probability of deterrence is always increasing in the transgression’s military value (on the x-axis); however, increasing the transgression’s direct value (on the y-axis) has a non-monotonic effect. The probability of deterrence first decreases due to the logic

\[^8\text{See Supplemental Appendix for details about Figures 3 and 4.}\]
of Proposition 3, and then increases as the defender’s intrinsic willingness to fight dominates.

Case Studies

Our model thus predicts a range of outcomes – deterrence success, deterrence failure with war, and deterrence failure with appeasement – under conditions in which traditional empirical analyses of deterrence would simply predict failure; when the objective magnitude of a transgression is minor relative to the cost of war. It does so by directly considering states’ uncertainty and inferences about their adversaries’ future behavior, and by distinguishing the military from the direct value of a transgression. The effect of these features on crisis outcomes has been largely ignored in previous empirical analyses (see Huth 1999). We now present an example of each of these potential outcomes of a crisis, and discuss how the model helps to shed light on it.

Deterrence Success The model predicts deterrence success when the military value of a transgression exceeds its direct value (even if both are minor), and the challenger does not in fact affir-
matively prefer to fight a war. In 1946, the United States deterred the Soviet Union from military action against Turkey following Soviet demands that it be allowed to place bases on the Turkish Straits (Mark 2005, pp. 123-124). The Soviet Union had some intrinsic interest in protecting its trade through the Straits, though the U.S. and Turkey were willing to renegotiate the agreement governing this trade to satisfy Soviet demands. The United States, on the other hand, had almost no intrinsic interest in the Turkish Straits, or in Turkish independence more generally. The U.S. had no obvious economic or political interests other than a small trade in tobacco, machinery and vehicles (Kuniholm 1980, pp. 65-66). The United States also anticipated that any war with the Soviet Union would be enormously costly despite the U.S. nuclear monopoly, involving Soviet ground offensives across Europe and Asia and requiring major U.S. ground operations (Ross 1996, pp. 12-19, 31).

Turkey’s primary value to the USA and USSR was military. In the event of a general war, Turkish resistance to a Soviet offensive was meant to temporarily protect American access to the Suez Canal, the Persian Gulf, and air bases in Egypt from which the United States planned to bomb central Russia. The loss of these assets would have weakened the U.S. and its allies (Leffler 1985, pp. 814-815). The model suggests that the deterrence of a Soviet invasion was successful because Turkey had greater military value than direct value, and therefore could not possibly appease a Soviet Union intent on war. Given U.S. fears of Soviet ambitions, an invasion of Turkey could have easily been perceived as an informative signal of both the present and future belligerence of the Soviet Union.  

In fact, President Truman was prepared to infer far-reaching ambitions from a Soviet attack in the face of a U.S. commitment. When asked if he understood that the decision to defend Turkey may mean war, Truman responded that “we might as well find out whether the Russians were bent on world conquest now as in five or ten years” (Mills 1951, p. 192). Ultimately, it was the Soviet aversion to war that led them to back down in the face of a credible U.S. threat (Mark 2005).  

9While officials believed that the Soviet Union would not fight a major war to satisfy their expansionist ambitions, they were not entirely confident in this assessment (Mark 2005, 119, 129).  

10See the Supplemental Appendix for an expanded case study.
**Deterrence Failure and War**  The model predicts deterrence failure and war when the military value of a trangression exceeds its direct value, but the defender’s fears about the challenger’s intentions are actually realized – the challenger does in fact affirmatively prefer to fight a war rather than maintain the status quo. In 1939, Finland’s rejection of the Soviet Union’s relatively modest territorial demands for naval bases in the Gulf of Finland and territorial revisions on the Karelian Isthmus led to war between the two countries, despite the enormous costs Finland anticipated in a war with its much more powerful neighbor. The naval bases had greater military than inherent value due to their facilities and location, and Finland feared that the granting of bases to the Soviets would weaken them in a future war (Jakobson 1961, 138-139; Van Evera 1999, 188). The Karelian Isthmus was an intrinsically valuable territory, but Jakobson (1961, 139) notes that Finland was willing to concede on the Isthmus and made its stand on the issue of the bases.

The model thus suggests that Finland’s deterrent threat would be credible, and that deterrence would only fail if the Soviet Union affirmatively desired war to fulfill more ambitious goals. The Finnish government clearly feared Soviet intentions and believed that further demands or war would follow any concessions (Jakobson 1961, 133, 139; Sechser 2010, 648-649). Although Stalin’s intentions are not definitively known, he very likely desired the complete subjugation of Finland before the war even started. Plans to impose a Communist government on Finland were likely developed long before the war began, and the Soviet Union implemented similar plans against the Baltic states less than one year later (Spring 1986, 214). If indeed the Finnish threat was believed to be credible, then the Soviet invasion could easily have been interpreted as a signal of their future belligerence. While the literature has puzzled over why Finland fought rather than why deterrence failed, our model shows how these issues are inextricably intertwined; the inference that Finland drew from the failure of deterrence may explain their willingness to fight rather than appease.\(^{11}\)

\(^{11}\)See the Supplemental Appendix for discussion of the literature on Finland’s decision to fight.
Deterrence Failure and Appeasement  Lastly, the model predicts deterrence failure and appeasement if the value of a transgression is more intrinsic than military. Under these conditions a defender will reason that even a belligerent challenger may be successfully appeased by being allowed to transgress, and therefore allow the transgression even if the stakes are substantial.

This prediction helps to explain the Allies failure in deterring Hitler from annexing Austria and the Sudetenland prior to World War II. These territories were of great military value. They were wealthy, populous, and greatly contributed to Germany’s ability to continue military conquests into Central and Eastern Europe (Overy and Wheatcroft 1989, 47-50). In addition, Great Britain and France had good reason to fear that Germany’s ambitions would not stop with these territories. However, the territories’ high intrinsic value to Germany is precisely why Great Britain and France were willing to concede them. Both of these territories were heavily populated with German co-ethnics, and Germany justified its policy as one of national unification. The British leadership still entertained the possibility that German ambitions were limited and that an agreement could be reached that satisfied their grievances and avoided war (Weinberg 1980, 346). The notion that occupying Austria and the Sudetenland might satisfy a belligerent Germany was plausible, resulting in exploitation of the Allies’ known preference for appeasement and deterrence failure.

Robustness
We last discuss the robustness of our main results to a variety of common complexities studied in the international relations literature; details are in the online Appendix.

Salami Tactics  A large “rationalist” literature in international relations begins with the premise that war is costly (Fearon 1995). However, our baseline model assumes that the challenger sometimes makes a unilateral decision to fight, which appears to be inconsistent with this literature. An alternative interpretation is that this outcome represents an unmodeled continuation game where the challenger attempts an additional transgression against which the defender is willing to fight.
However, this raises the question of whether such a continuation game can be explicitly modeled using assumptions consistent with the rationalist literature. In the Supplemental Appendix we consider an extension of the model in which the challenger may attempt a series of successive transgressions each below the defender’s cost of war, and the defender always holds the decision to fight. This yields a game of “salami tactics” similar to Powell (1996). Because this structure vastly multiplies the space of potential parameters, for simplicity we consider a particular payoff structure.

The challenger and defender bargain and potentially fight over a landmass of size and value equal to 1. The challenger initially possesses at least half, and she can attempt to take more in a series of discrete steps each below the defender’s cost of war. Each successive transgression is thus an attempt to advance from one “threshold” on the landmass to the next. If the challenger attempts to advance in a period, the defender can respond by allowing it, or by fighting an all-or-nothing war in which the victor receives the entire landmass. In a war the challenger’s probability of victory is an increasing function of her share of the landmass. This function is depicted in Figure 5, and is assumed to have properties similar to our main condition for successful deterrence. First, in each period the challenger’s probability of victory (slightly) exceeds her share of the landmass she holds, so that war is attractive if his costs are low enough. Second, further advancement initially shifts the military balance toward the challenger (a little bit) more quickly than they increase her payoff from peace. Finally, the challenger’s cost of war is always strictly positive, but there is always some chance that it is low enough for her to prefer war to peace in that same period.

In this extension, we show that as long as the defender prefers fighting a war to ceasing to exist entirely, there are actually many equilibria sustained by the same logic as our deterrence equilibrium. The intuition is as follows. First, there is always a final threshold at which the defender is “intrinsically” willing to fight for reasons similar to Powell’s (1996) model of salami tactics; he anticipates that if the challenger advances beyond it, she will be unable to commit not to exploit salami tactics to eventually take the entire landmass. At any threshold prior to this one, there exists
an equilibrium in which the defender is willing to fight due to our logic: because the challenger expects the defender to respond to further advancement (however small) with war, the defender can infer in equilibrium that a challenger who attempts to advance further in fact desires war. Because advancing makes war relatively more attractive, the probability of appeasing an already-belligerent challenger by allowing further advancement is 0, and hence responding with war is optimal.

**An Endogenous Transgression**  In our baseline model, the magnitude of the challenger’s transgression is exogenous. However, in many settings this is the challenger’s choice. To explore the robustness of our insights to this possibility, we revisit the payoff environment of our “salami tactics” extension with an alternative game form. In the first period the challenger can make an endogenous “demand” of how far to advance. The game then proceeds as in the baseline model.

In the online Appendix, we show that there exists an equilibrium in which the defender responds to any strictly positive transgression, however small, with war. The rationale is similar to the baseline model. Even when the challenger can moderate her transgression, it remains true that conceeding to
most transgressions would increase an already-belligerent challenger’s payoff from war more than her payoff from peace. Only the largest of transgressions can potentially sate a belligerent challenger’s thirst for war, and against such transgressions the defender is intrinsically willing to fight.

**A Challenger Who Can Back Down**  In our baseline model the challenger presents the defender with a fait accompli; to resist the transgression means war. However, in many crisis bargaining models the defender’s resistance to initial aggression does not result in immediate war; instead, the challenger has an opportunity to first back down (e.g. Lewis and Schultz (2003)). In the online Appendix we show that introducing this ability actually makes the deterrence equilibrium even easier to sustain; the reason is that the defender can entertain the possibility that the challenger is actually bluffing when he observes an attempted transgression, which makes him only more willing to resist.

**Interdependent War Values**  Our baseline model assumes the defender’s payoffs are unaffected by the challenger’s type, which would be the case if he was uncertain about the challenger’s cost of war. However, in many crisis models the challenger has private information about factors affecting both parties’ war payoffs, such as the probability of victory (Fey and Ramsay 2011). Introducing this possibility complicates the defender’s inference; upon observing a transgression he can simultaneously infer that appeasement is less likely to work – making him more willing to fight – and that he would be weak in a war – making him less willing to fight. Nevertheless, Corollaries 1 and 3 hold unaltered – the deterrence equilibrium always exists when appeasement is impossible. However, when appeasement is possible, equilibria are more complex and can exhibit new and interesting patterns.

**A Defender with Private Information**  Our baseline model assumes away any private information possessed by the defender about her intrinsic willingness to fight or “resolve,” in order to shift attention away from reputation to fear. However, in the online Appendix we show that Corollaries 1 and 3 continue to hold when this possibility is introduced. Interestingly, we also find that introducing even a small possibility that the defender is intrinsically willing to fight can sometimes uniquely
select the deterrence equilibrium. Intuitively, the reason is that “deterrence begets deterrence” – more deterrence increases the defender’s interim assessment that a challenger who transgresses is unappeasable, which makes him more willing to respond with war, generates a higher probability that the transgression will provoke him, and thereby results in yet more deterrence.

A Transgression with Uncertain Consequences Our baseline model assumes that the transgression’s military consequences are known. However, in reality these consequences can be unpredictable; e.g., an ex-ante appealing military adventure like the Soviet invasion of Afghanistan also risks a protracted and costly conflict. In the online Appendix we consider a variant with some initial uncertainty about the transgression’s military benefit that is only resolved after it has taken place. We show that as long as there is not too much uncertainty, our main results are robust and can be restated as a function of the transgression’s expected military benefits. Intuitively, the reason is that uncertainty about the transgression’s consequences weakens – but does not eliminate – the ability to infer future belligerence from present belligerence. In also does not change fact that a higher relative military gain strengthens this inference. We also provide analogous conditions to those in the baseline model for the defender to benefit from his fear and uncertainty in expectation.

A Challenger Who Can Signal Finally, our baseline model assumes that the challenger may only “signal” her type through the transgression itself. This raises the concern that our results may not be robust to a challenger who can also send costly signals (Fearon 1997; Slantchev 2011). In the Supplemental Appendix we show that introducing this ability can both strengthen and weaken our results. Crucially, when the transgression’s military value exceeds its direct value, there always exists an equilibrium in which peaceful challengers cannot credibly signal their intentions and eliminate

---

12 The Soviet Union expected the invasion to be beneficial to the balance of power because it would prevent Afghanistan from becoming a Western ally or base (Kalinovsky 2009).

13 The chance that the transgression may “backfire” also suggests that a strategic defender may want to “bait” an unwitting challenger into committing it. This possibility lies outside our extension because the defender is assumed to be equally uncertain; however, exploring when a better-informed defender will undertake such baiting strategies would be an interesting avenue for future work.
the fear that sustains deterrence. Under this condition there is always the possibility that the challenger is “opportunistically belligerent” – that is, initially deterrable, but seeking to transgress to strengthen herself for a future war. An opportunistically-belligerent challenger necessarily values transgressing more a peaceful one, and so would always be willing to send any costly signal that a peaceful one would. This precludes the possibility that a peaceful challenger could “separate” herself with a costly signal. However, when our key condition fails and the possibility for successful appeasement exists, then for some parameter values costly signaling indeed causes deterrence to unravel even if it was possible in the baseline model. In equilibrium the challenger always sends a credible costly signal when she is peaceful, and the defender allows her to transgress. The defender is also weakly worse off not knowing the challenger’s intentions; the challenger transgresses no matter what, is allowed to do so if she signaled, and triggers a (sometimes avoidable) war if she did not.

Conclusion

This paper examines a model of deterrence where a defender is uncertain about a challenger’s intentions, and fears that she is unappeasably belligerent. We show that this fear can generate credible deterrence even when the probability of belligerence is arbitrarily small, and the value of the transgression being deterred is small relative to the cost of war. Unlike most previous studies of deterrence, we do not assume that the defender is sometimes intrinsically willing to fight, or that he has access to commitment devices that help him to do so. Instead, our mechanism relies on the inference that the defender can make from a transgressive act taken in the face of an expectation of war.

After illustrating this simple insight, we show that the defender’s fear can sometimes benefit him by allowing him to credibly deter at a negligible risk of avoidable wars. We also derive several empirical implications about deterrence credibility that are previously untested in the empirical

---

14 Formally, under this condition there always exists a universally divine equilibrium (Banks and Sobel 1987) in which deterrence occurs and there is no costly signaling.

15 This condition creates the possibility that the challenger values transgressing most when she is peaceful; when this is the case the challenger always transgresses in all universally divine equilibria.
literature. We show that transgressions that make effective “tests” are not ones that are objectively large, but ones that carry a high military value relative to their direct value; the reason is that allowing such transgressions cannot appease an already belligerent challenger. We argue that this insight helps illuminate specific historical episodes of deterrence success and failure.

The logic of our model can also help to explain contemporary episodes, such as North Korea's successful deterrence of an American air strike against their nuclear plant using the threat of a potentially suicidal war. How could such a threat be taken as credible? The available evidence suggests that both sides understand that North Korea is using certain actions as a test of the United States' intention to invade. Pyongyang’s 2003 warning that an air strike on their nuclear plant would lead to “total war” explicitly stated that such an attack would be viewed a precursor to invasion (KCNA News Agency 2003). In recommending an airstrike against a North Korean missile testing site, Carter and Perry (2006) wrote that the United States must warn North Korea that the attack would only be against a specific target. Pritchard (2006) responded that, despite the warning, Pyongyang might very well interpret the air strike as the “start of an effort to bring down [their] regime.” The incentive by North Korea to claim uncertainty about the United States’ intentions, as well as the incentive by the U.S. to claim sharply limited goals, both follow directly from our logic.

These incentives, however, also point to limits in our analysis and interesting avenues for future work. Much of the deterrence literature focuses on things that a defender can do – issue cheap talk claims, engage in costly signalling, employ commitment devices – to improve the credibility of his deterrent threats. Our analysis is closer in spirit to the “inherent credibility” approach (Zagare 2004) – we examine structural features of the environment outside the defender’s control that can sustain his credible deterrence. But the logic of our model and the North Korean case clearly suggest actions that the defender would like to “do” – to claim that he fears the challenger is unappeasably belligerent (even when he does not), and to claim that he is using the transgression as a test of that belligerence in order to select the deterrence equilibrium when it exists. Our model, however, is
insufficiently rich for such actions to affect equilibrium outcomes. Cheap talk cannot select equilibria, and there is nothing for the defender to signal.

The history of the deterrence literature, however, suggests a way forward. Classical deterrence theory conceives of the credibility of deterrence as rooted in an intrinsic willingness to fight. In order to understand how a defender can increase his credibility, subsequent theories assumed that a challenger was uncertain of that willingness. Our theory, in contrast, conceives of the credibility of deterrence as rooted in fear; thus, a way forward may be to assume that the challenger is uncertain of that fear. This sort of “higher-order uncertainty” has been considered in the study of the “spiral model”; Kydd analyzes a game in which a state is uncertain about what his enemy believes about him, and this complicates his ability to draw inferences from the enemy’s arming decisions – is the enemy aggressive, or just afraid? (Kydd 1997). In Kydd’s analysis, states wish to signal about their own intentions to improve their ability to draw inferences about their enemy’s intentions. Our model, however, suggests that it may also be fruitful to study when states wish to signal about their fear to improve their ability to deter. Such a modeling approach could potentially eliminate the issue of multiple equilibria in the model. Moreover, classical mechanisms for improving the credibility of deterrence could be understood in a new light. Under what conditions could cheap talk about fear increase the credibility of deterrence? When do costly signals most credibly communicate fear? Finally, how can a challenger credibly communicate limited aims and exploit a defender who prefers to appease? Exploring such questions may be a fruitful avenue for future work.

References

Acharya, Avidit and Edoardo Grillo. 2014. “War with Crazy Types.” Political Science Research and Methods, forthcoming.


Online Appendix – NOT FOR PUBLICATION

“Fear, Appeasement, and the Effectiveness of Deterrence”

August 24, 2016.

This Online Appendix is divided into three parts. Appendix A contains proofs of the main propositions and corollaries text, as well as notes on Figures 3 and 4. Appendix B contains an expanded version of the case study on the Turkish Straits Crisis of 1946, additional notes on the Russo-Finnish War case study, and an additional case study on the Taiwan Straits Crisis of 1954-1955. Appendix C analyzes model variants and proves results discussed in the Robustness section.
A Proofs of Main Results

Proof of Proposition 1  The defender’s strategy consists of a probability of responding to the transgression with war, which we denote $\alpha$. The challenger’s utility from not transgressing is $n_C^1$, and from transgressing is $\alpha \cdot w_C^1(\theta_C) + (1 - \alpha) \cdot \max\{w_C^2(\theta_C), n_C^2\}$. The latter is strictly increasing in $\theta_C$ and greater than $n_C^1$ for all $\alpha$ when $\theta_C = \bar{\theta}_C^1$. Thus, the challenger’s strategy must be to always transgress, or to transgress i.f.f her type is above a cutpoint $\bar{\theta}_C \leq \bar{\theta}_C^1$ at which she is indifferent between transgressing and not.

The necessary and sufficient conditions for existence of the two pure strategy equilibria ($\alpha^* = 0$ the no deterrence equilibrium, and $\alpha^* = 1$ the deterrence equilibrium) are described in the main text and straightforward to derive. There may also exist mixed strategy equilibria in which the defender responds with war with a strictly interior probability $\alpha^* \in (0, 1)$. For such an equilibrium to hold, the defender must be indifferent between responding with war and allowing the transgression. This requires that,

$$P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C) = \bar{\beta}$$

i.e. the defender’s posterior belief that the challenger will initiate war if allowed to transgress is equal to his threshold belief $\bar{\beta}$. The left hand side approaches $P(\theta_C \geq \bar{\theta}_C)$ as $\bar{\theta}_C$ approaches the lower bound of the type space, is equal to 1 at $\bar{\theta}_C = \bar{\theta}_C^2$, and is strictly increasing in between. Thus, a cutpoint satisfying (1) exists i.f.f. the no deterrence equilibrium exists ($P(\theta_C \geq \bar{\theta}_C) < \bar{\beta}$). We denote this cutpoint $\bar{\theta}_C^*$, which must be $< \bar{\theta}_C^2$.

We now check conditions such that there exists some $\alpha^* \in (0, 1)$ that induces the challenger to play the cutpoint strategy $\bar{\theta}_C^* < \bar{\theta}_C^2$. A necessary condition and sufficient condition is that this type be indifferent between transgressing and not, i.e. there exists an $\alpha^*$ s.t.

$$\alpha^* \cdot w_C^1(\bar{\theta}_C^*) + (1 - \alpha^*) \cdot n_C^2 = n_C^1.$$  

(2)
If \( \tilde{\theta}_C^r > \tilde{\theta}_C \), \( \iff P(\theta_C \geq \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C) < \tilde{\beta} \) (i.e. the deterrence equilibrium does not exist) then the condition cannot be satisfied since this would imply that both \( w_C^1(\tilde{\theta}_C^r) \) and \( n_C^1 \) are greater than \( n_C^1 \).

Conversely, if \( \tilde{\theta}_C^r < \tilde{\theta}_C \), then an \( \alpha^* \) satisfying (2) exists and is unique.

Thus, a unique mixed strategy equilibrium exists \( \iff \) both the no deterrence and deterrence equilibria exist, and the equilibrium strategies \( (\alpha^*, \tilde{\theta}_C^r) \) are uniquely characterized by (1) and (2). We now show that when there are multiple equilibria, i.e. \( \tilde{\beta} \in [P(\theta_C \geq \tilde{\theta}_C) \mid \theta_C \geq \tilde{\theta}_C^r), P(\theta_C \geq \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C)] \), the defender is strictly better off in the deterrence equilibrium than in either the no deterrence or mixed strategy equilibrium. The defender’s utility in the deterrence equilibrium is \( U^{de} = P(\theta_C < \tilde{\theta}_C) \cdot n_D + P(\theta_C \geq \tilde{\theta}_C) \cdot w_D \). His utility in the mixed strategy equilibrium is

\[
U^{ms} = P(\theta_C < \tilde{\theta}_C^r) \cdot n_D + P(\theta_C \geq \tilde{\theta}_C^r) \cdot (P(\theta_C < \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C^r) \cdot n_D^2 + P(\theta_C \geq \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C^r) \cdot w_D^2)
\]

\[
= P(\theta_C < \tilde{\theta}_C) \cdot n_D + P(\theta_C \geq \tilde{\theta}_C^r) \cdot w_D \quad \text{by def'n of } \tilde{\theta}_C.
\]

This is less than \( U^{de} \) since \( \tilde{\theta}_C^r < \tilde{\theta}_C \) by construction \( \rightarrow P(\theta_C < \tilde{\theta}_C^r) < P(\theta_C < \tilde{\theta}_C^r) \), and \( n_D^1 > w_D^1 \).

Finally, his utility in the no deterrence equilibrium is

\[
U^{nd} = P(\theta_C < \tilde{\theta}_C) \cdot n_D^2 + P(\theta_C \geq \tilde{\theta}_C) \cdot w_D^2.
\]

\[
= P(\theta_C < \tilde{\theta}_C) \cdot (P(\theta_C < \tilde{\theta}_C \mid \theta_C < \tilde{\theta}_C) \cdot n_D^2 + P(\theta_C \geq \tilde{\theta}_C \mid \theta_C < \tilde{\theta}_C) \cdot w_D^2) \quad < n_D^1 \text{ since } n_D^1 > n_D^2 > w_D^1 > w_D^2
\]

\[
+ P(\theta_C \geq \tilde{\theta}_C) \cdot (P(\theta_C < \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C) \cdot n_D^2 + P(\theta_C \geq \tilde{\theta}_C \mid \theta_C \geq \tilde{\theta}_C) \cdot w_D^2) \quad < w_D^1 \text{ since the deterrence equilibrium exists}
\]

\[
< P(\theta_C < \tilde{\theta}_C) \cdot n_D^1 + P(\theta_C \geq \tilde{\theta}_C) \cdot w_D^1 = U^{de} \quad \blacksquare
\]

**Proof of Proposition 2** If the defender knew the challenger’s type, then he would respond with war \( \iff \theta_C > \tilde{\theta}_C \), and thus the challenger would be deterred \( \iff \theta_C \in (\tilde{\theta}_C^r, \tilde{\theta}_C^l) \). The probability of deterrence would therefore be \( P(\theta_1 < \tilde{\theta}_C^l, \theta_1 \geq \tilde{\theta}_C^r) \). Now suppose first that the deterrence equilib-
rium exists when the challenger’s type is unknown; then the probability of deterrence is \( P(\theta_1 < \bar{\theta}_C) \), which is > the probability of deterrence \( P(\theta_1 < \bar{\theta}_C^1, \theta_1 \geq \bar{\theta}_C^2) \) when the challenger’s type is known.

Next suppose that the deterrence equilibrium does not exist when the challenger’s type is unknown, so that the probability of deterrence is 0. Then we must have \( \bar{\theta}_C^2 > \bar{\theta}_C^1 \), which implies that the probability of deterrence is also 0 when the challenger’s type is known.

Now we consider when the defender is better off not knowing the challenger’s type. This is clearly the case when \( \bar{\theta}_C^2 \leq \bar{\theta}_C^1 \); types \( < \bar{\theta}_C^2 \) are deterred when they otherwise would not be, and for all other types the outcome is identical. So suppose that \( \bar{\theta}_C^1 < \bar{\theta}_C^2 \), and observe that a necessary condition for the defender to be better off not knowing is that the deterrence equilibrium exists. Hence we must characterize when the defender prefers not knowing and the deterrence equilibrium to knowing the challenger’s type; this will be the case when

\[
P(\theta_C < \bar{\theta}_C^1) \cdot (n_D^1 - n_D^2) > P(\theta_C \in [\bar{\theta}_C^1, \bar{\theta}_C^2]) \cdot (n_D^2 - w_D^1),
\]

i.e. when the benefit \( n_D^1 - n_D^2 \) of deterring types \( < \bar{\theta}_C^1 \) exceeds the cost of preventable wars \( n_D^2 - w_D^1 \) against appeasable types \( \theta_C \in [\bar{\theta}_C^1, \bar{\theta}_C^2] \). It is simple to show that the conjunction of this condition and the deterrence equilibrium existence condition reduces to the condition in the Proposition.

**Proof of Corollary 2** For the purposes of expositional simplicity, we consider the game form in which the defender first chooses whether he will be informed or ignorant about the challenger’s type, and nature next draws that type. (This is strategically equivalent to the game form in which nature moves first – since nature is non-strategic – but allows us to simply refer to proper subgames in which the defender is ignorant vs. informed). The result then follows immediately; by Proposition 1 the best equilibrium for the defender in the case of multiplicity involves selecting the deterrence equilibrium in the “ignorant” subgame whenever it exists, and the first stage is then simply the defender choosing which game he wants to play according to the calculus in Proposition 2.
Proof of Corollary 3

\[ \delta_m^C (\bar{\theta}_C) \geq \delta_d^C \iff w_1^C (\bar{\theta}_C) \geq n_1^C + \left( \delta_d^C - \delta_m^C (\bar{\theta}_C) \right) \iff w_2^C (\bar{\theta}_C) \geq n_2^C \]
\[ \iff \bar{\theta}_C^2 \leq \bar{\theta}_C^1 \iff P (\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1) = 1 \]

Proof of Proposition 3  
Holding \( \bar{\theta}_C^1 \) (i.e. the challenger’s first period payoffs) fixed, the ineffectiveness of appeasement \( P (\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1) \) is decreasing in \( \bar{\theta}_C^2 \). Thus for the first claim, it suffices to show \( \bar{\theta}_C^2 \) is decreasing in \( \delta_m^C - \delta_d^C \). By assumption, \( n_2^C = n_1^C + \delta_d^C \) and \( w_2^C (\theta_C) = w_1^C (\theta_C) + \delta_m^C \)
\[ \text{and} \quad n_2^C = w_2^C (\bar{\theta}_C^2), \text{which together imply that} \quad n_1^C = w_1^C (\bar{\theta}_C^2) + (\delta_m^C - \delta_d^C). \]
This implies the desired property since \( w_1^C (\theta_C) \) is increasing in \( \theta_C \).

To show that the probability of deterrence is increasing in \( \delta_m^C - \delta_d^C \), note that with the assumed equilibrium selection and by Proposition 1, the probability of deterrence is 0 if \( P (\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1) < \bar{\beta} \) and \( P (\theta_C \geq \bar{\theta}_C^1) \) otherwise. Holding \( \bar{\beta} \) (the defender’s payoffs) fixed, the probability of deterrence is therefore (step-wise) increasing in \( P (\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1) \). Since this is increasing in \( \delta_m^C - \delta_d^C \), the result is shown.

Construction of Figures 3 and 4  
In Figure 3, the right panel depicts \( G \left( \frac{1 - F (\bar{\theta}_C)}{1 - F (\bar{\theta}_C)} \right) : F (\bar{\theta}_C^1) \)
where \( G (\bar{\beta}) \) is the induced probability distribution over \( \bar{\beta} \). To generate both figures we assume that \( w_1^C (\theta_C) = \theta_C \), the transgression’s military and direct cost to the defender’s are equal to .1, and the challenger’s type is uniformly distributed on [0, 1]. The left panel assumes that the defender’s benefit from peace is \( n_1^D - w_1^D = .55 \), while the right panel assumes it is uniformly distributed on [−.1, 1].

Figure 4 is generated by assuming that \( w_1^C (\theta_C) = \theta_C \) with \( \theta_C \sim U[0, 1.1] \), and both the defender’s benefit \( n_1^D - w_1^D \) and lowest type of challenger’s benefit \( n_1^C - w_1^C (0) \) for avoiding war is .5.
B  Case Study Notes

Extended Case Study: The Turkish Straits Crisis of 1946

We examine in depth a particular instance of deterrence success upon which the model sheds light; the crisis over control of the Turkish Straits that occurred between the United States and the Soviet Union in the early days of the Cold War. In 1945 and 1946, the Soviet Union repeatedly demanded that Turkey allow it to place bases on the Turkish Straits (Kuniholm 1980). These demands, coupled with extensive Soviet military preparations in the Balkans, led American officials to prepare for armed aggression against Turkey. In August 1946, President Truman decided that the United States would fight a full-scale war to defend Turkey in the event of Soviet invasion. Although this commitment was never announced publicly, it was communicated to Stalin through multiple channels. Upon learning about Truman’s decision, Stalin reversed course (Mark 2005, pp. 123-124). We argue that the model we present helps explain why the United States was willing to fight, and why the Soviet Union found this threat credible.

Incentives  The Turkish Straits Crisis contained the elements that are essential to our model: the defender’s unwillingness to fight for the immediate stakes, the fear and uncertainty about the challenger’s intentions, and the defender’s preference to fight sooner rather than later. The United States had limited economic and political ties to Turkey and attached little intrinsic value to the Turkish Straits or Turkish independence.\textsuperscript{16} The Soviet Union, on the other hand, had a direct interest in controlling the Straits. It sought military control of the Straits for the purposes of protecting trade and denying their use by hostile powers, objectives that American officials sympathized with.\textsuperscript{17}

Furthermore, decision makers anticipated that any war with the Soviet Union would be enor-

\textsuperscript{16}The U.S. had no obvious economic or political interests other than a small trade in tobacco, machinery and vehicles (Kuniholm 1980, pp. 65-66). The administration later claimed an interest in democratization, but this was merely rhetoric to justify the alliance to the public (Kayaoglu 2009).

\textsuperscript{17}See DeLuca 1977 pp. 511-514; \textit{FRUS 1946 VII}, 827-829; and Leffler (1985) pp. 809-810.
mously costly, despite the U.S. monopoly in atomic weapons. The military anticipated that the Red
Army would launch offensives in Europe, the Middle East and Asia, that bombing would to be slow
to produce results, and that ground operations would eventually be necessary (Ross 1996, pp. 12-19,
31). This combination of low stakes and a costly war makes the United States’ decision to fight for
Turkey puzzling. In fact, before the U.S. began to fear a general war with the Soviets, it made no
plans for Turkey’s defense despite believing that an attack was likely (Mark 1997, 398)

Fear  By 1946, American officials had come to believe that the Soviet Union desired to dominate
the Eurasian continent and eventually the world (Ross 1996, p. 3, 7). These ambitions did not
necessarily imply that war would occur: most intelligence and military assessments assumed the
Soviet Union was practical enough to avoid a destructive war with the United States and would
accept the status quo (Mark 1997, 397). However, officials could not be perfectly confident in this
assessment, and entertained the possibility that the Soviets would initiate general war in pursuit of
their objectives. For example, in a meeting on June 12, 1946, President Truman speculated that the
Soviet Union might start a war to divert public unrest, Secretary of the Navy Forrestal argued they
might start a war if external circumstances were favorable for completing the “world revolution,”
and Admiral Leahy responded that the Soviets were simply unpredictable (Mark 2005, p. 119, 129).

If such a war were to occur, it is clear that the United States would have preferred that the
fighting begin before losing Turkey. In the war plans, Turkey was to be the first line of defense
against a Soviet advance toward strategically vital areas of the Middle East, the loss of which would
weaken the U.S. and its allies. Turkish resistance to a Soviet offensive would protect American access
to the Suez Canal, the Persian Gulf, and air bases in Egypt from which the United States planned
to bomb central Russia (Leffler 1985, pp. 814-815).

Some officials believed that a successful transgression would not only strengthen the Soviet Union,
but increase the Soviet appetite for general war. For example, Ambassador Edwin Wilson argued
that, if the Soviet Union were allowed to overrun Turkey, they would be unable to resist the temp-
tation to advance toward the Suez Canal and the Persian Gulf. He wrote that “once this occurs, another world conflict becomes inevitable” because of the military advantages the Soviets would then have against the West (FRUS 1946 VII, p. 819, 822). This is similar to the unappeasability condition in the model: conquering Turkey would increase Soviet military strength and make general war more attractive, outweighing any pacifying effect from satisfying the Soviet demands over Turkey itself. Therefore, a concession could not appease the Soviet government if it was already belligerent.

**Inference and Deterrence** The Turkish Straits Crisis thus contained the incentive structure and uncertainty about challenger intentions that are essential to our model. Historical accounts of the crisis also suggest that, following the U.S. decision to defend Turkey, the Truman Administration was prepared to infer far-reaching Soviet ambitions from their willingness to attack Turkey in the face of a U.S. commitment, which is the key inference that sustains deterrence in equilibrium.

Most officials believed that the Soviet Union would be deterred by a U.S. commitment because it was generally thought that they wanted to avoid a major war (Mark 1997, 399). Conversely, officials appear to have believed that the Soviet Union would only invade Turkey if it did desire such a war. Undersecretary of State Dean Acheson said he believed that the Soviet Union would most likely be deterred, but he also argued the United States would “learn whether the Soviet policy includes an affirmative provision to go to war now” if deterrence failed.18 This is the key inference in the deterrence equilibrium; the defender can learn of the challenger’s preference for war from her willingness to transgress in the face of a deterrent threat. It is also clear that this inference sustained the U.S. willingness to initiate war following a Soviet attack. President Truman, when asked if he understood the decision to defend Turkey may mean war, responded that “we might as well find out whether the Russians were bent on world conquest now as in five or ten years” (Mills 1951, p. 192). Truman was prepared to infer far-reaching, long-term ambitions for world conquest from a Soviet willingness to challenge a credible U.S. commitment to Turkey.

It is unclear whether the Soviets were deterred because they understood that American decision makers would interpret invasion as evidence of an intent to initiate war. As the first postwar crisis where the Soviets attempted to control an area where they did not already have a presence at the end of WWII, it seems likely that Stalin would have realized that invading Turkey would appear to the Americans as a dangerous new direction in Soviet policy, and that both parties ultimately came to understand the act of invasion as focal for revealing Soviet intentions. Although the invasion didn’t occur, this episode dramatically reshaped American perceptions of Soviet intentions, and Soviet Foreign Minister V.M. Molotov later admitted that they had overreached (Mark 1997, 414).

**Alternative Explanations** While other deterrence mechanisms may have also been relevant, some fail to explain key features of the crisis. It is possible that reputational concerns drove decision-making, and that the United States felt it had to demonstrate its resolve to its allies and the Soviet Union. However, the primary fear in losing Turkey was never reputational, it was strategic. Government officials and military estimates repeatedly emphasized that the major concern in the crisis was that losing Turkey would disadvantage the United States in a future war with the Soviet Union. Truman himself, when told that a commitment to Turkey may mean war, pulled out a map and lectured his advisors about the strategic importance of the Middle East (Acheson 1969, 196).

Other commonly cited mechanisms in the deterrence literature do not seem to explain the outcome of the crisis. Any explanation involving audience costs would require threats to have been made publicly. While the United States did dispatch a naval force to the Mediterranean, it never publicly announced its decision to defend Turkey, instead communicating with the Soviet Union privately and downplaying the crisis in public (Trachtenberg 2012, 24-25). In addition, there was no obvious commitment device that would have automatically engaged the United States in a conflict, such as military forces stationed in Turkey as a “trip-wire.” Finally, there was nothing probabilistic about the Americans’ threat that “left something to chance.” On the contrary, Truman clearly asserted that, if the Soviet Union invaded, he would follow the recommendation to defend Turkey “to the
Notes on Russo-Finnish War Case Study, 1939

In the literature on this case, the puzzle is not why deterrence failed, but why Finland was willing to fight rather than concede. Of particular interest is how the model here contributes to this existing debate. Consistent with the model, a common interpretation of Finland’s decision to fight is that Finland feared that the granting of bases to the Soviets would weaken them in a future war (Jakobson 1961, 138-139; Van Evera 1999, 188). Sechser (2010) objects to this interpretation, arguing that these concessions were of limited military value and therefore can’t explain Finland’s willingness to fight against its far more powerful neighbor. Our model provides an explanation. It is clear that these military bases had more military value than intrinsic value, which would ensure they could not appease an already belligerent Soviet Union. While the case is not often examined from the perspective of deterrence, the information about Soviet intentions that Finland gained from the deterrence failure explains why they decided to fight rather than appease, thus demonstrating an important link between the circumstances of a deterrence failure and the outbreak of war.

Extra Case Study of Deterrence Success: Taiwan Straits Crisis, 1954-55

In 1954-55, the United States successfully deterred a Chinese invasion of the tiny island of Quemoy. Control of the island had some minor intrinsic value for the Communists: while it was a small island of mostly farmers and fisherman, Communist control would stop the Nationalists’ occasional harassment of the mainland and merchant and fishing fleets near Amoy (Soman 2000, 120-121; Chang 1988, 99). However, its primary value was military: it housed a large contingent of Nationalist forces, blocked military deployments from Amoy toward Taiwan, and contained radar installations that helped defend Taiwan (Soman 2000, 120-125; Zhang 1992, 207-208). George and Smoke (1974) argue that control of the island had some “prestige” value. However, Communist propaganda during
this period focused mostly on Taiwan, and the attacks on Quemoy were launched soon after the beginning of a massive propaganda campaign about the liberation of Taiwan from the Nationalists. Quemoy’s value in relation to the defense of Taiwan – as well as the importance of the U.S. fear of a Chinese invasion of Taiwan – can be seen in the fact that both Dulles and Eisenhower voiced a willingness to abandon Quemoy were China to pledge not to invade Taiwan (Wang 2011, 155; FRUS 1955-1957 II, pp. 146, 439). Nevertheless, the military importance of the island was not great, and certainly was not worth fighting a war over on its own. The island was distant from Taiwan and difficult to defend, and a Chinese invasion of Taiwan would have been extremely difficult with or without Quemoy due to American naval superiority in the Taiwan Straits. As a result, there were disagreements about its value within the Eisenhower Administration and between the U.S. and British governments (Zhai 1994, 159-161; Chang 1988, 100; Wang 2011, 172-173). The relatively small value of the island is also apparent in Eisenhower’s worries about the difficulty of explaining to the American people why it would be worth starting a war with China over this small outpost (Zhai 1994, 159).

The model suggests that the deterrence of a Chinese invasion was successful because the island had greater military value than direct value, and therefore could not possibly appease a China intent on war. Given U.S. fears that China intended aggression against Taiwan, an invasion of Quemoy could have easily been perceived as an informative signal of both the present and future belligerence of the Communists (Zhang 1992, 193-194). The logic of the model, that attacking Quemoy would have signalled an affirmative desire for war because U.S. threats were credible, seemed to be operating. Communist propaganda acknowledged that the United States would defend Quemoy if attacked, and this was noted by American officials (Wang 2011, 180). In the face of a credible U.S. threat, the Communists backed down because of their desire to avoid war with the United States (Sheng, 487). Notably, the effectiveness of the U.S. threat over Quemoy is in contrast to the U.S. unwillingness to defend other island groups in the region like the Tachens, which had no military value for protecting
Taiwan, and which were promptly evacuated when attacked by Communist forces (Chang 1988, 102).

Extra Bibliography


C Analysis of Model Variants in Robustness Section

C.1 Robustness to Salami Tactics and Endogenous Demands

In this section we consider robustness to two alternative bargaining protocols – a) extending the sequence so that the game resembles a model of salami tactics, and b) endogenizing the demand made by the challenger. In the former extension the defender always has the final move in each period over whether to fight or concede.

Rather than fully solve out general versions of these games, we present two examples illustrating that our basic insight holds in these variants. Both examples are constructed from the following payoff environment for a finite period game of conflict over a landmass of size and value equal to 1. In both variants there is no discounting and no “flow” payoffs – payoffs are based on the holdings of the landmass in the period in which the game ends.

**Payoff Environment** Suppose a challenger and a defender jointly occupy a landmass of size and value equal to 1. Say the *advantaged* party at time $t$ is that which holds a majority of the landmass, and let $\delta_t$ denote the *excess* holdings of the advantaged party in period $t$ above $\frac{1}{2}$. If a war occurs in period $t$, the probability the advantaged party wins is:

$$p(\delta_t) = \left(\frac{1}{2} + \delta_t\right) + \phi(\delta_t)$$

where $\phi(\delta_t) = \frac{2t(1-2\delta)}{Z}$ and $Z$ is very large.\(^{19}\) Also suppose that the defender’s cost of war is commonly known to be $c_D \geq \frac{1}{4}$. The challenger’s type $\theta_C$ is unknown and uniformly distributed over $\theta_C \sim U \left[-\frac{1}{4}, 0\right]$, and her cost of war is $c_C = -\theta_C$. ■

The challenger’s probability of victory in a war as a function of her position is depicted in Figure 5 in the main text. We now present the first extension.

\(^{19}\)We require at least $Z > 6$ for $p(\delta_t)$ to be strictly increasing in $\delta_t$. 13
Extension 1 (Salami Tactics). Consider a $T \geq 3$ period game, and a $T+1$-length series of increasing values $0 < \delta_1 \ldots < \delta_T = \frac{1}{2}$. Each $\delta_t$ represents the challenger’s excess holdings above $\frac{1}{2}$ if the game advances to period $t$. Assume that the challenger is initially advantaged ($\delta_1 > 0$), and that the increments of advancement $\delta_t - \delta_{t-1}$ are less than the defender’s cost of war $c_D$ for all $t < T$. In each period $t$, the challenger decides whether or not to attempt to advance from $\delta_t$ to $\delta_{t+1}$. If she doesn’t attempt to advance the game ends. If she does attempt to advance, the defender chooses whether or not to respond with war, and the challenger’s probability of victory is $p(\delta_t)$.

In this extension there are a finite number of exogenously fixed positions to which the challenger can advance, each advancement represents a transgression of fixed size, and each increment $\delta_t - \delta_{t-1}$ of advancement short of possessing the entire landmass is less than the defender’s cost $c_D \geq \frac{1}{4}$ of war. Thus, the defender is always vulnerable to “salami tactics.” When $\delta_1 < \delta_2 < c_D < \delta_3$, the game essentially reduces to the baseline model; the reason is that the defender knows she will be unable to credibly resist any advancement beyond $\delta_2$, anticipates that conceding at $\delta_2$ will result in a concession of size $1 - \delta_2 > c_D$, and is therefore there willing to fight.\footnote{When $\delta_1 < \delta_2 < c_D < \delta_3$, the game maps to the baseline model by letting the defender’s payoffs be $n_D^{1} = \frac{1}{2} - \delta_1$, $n_D^{2} = \frac{1}{2} - \delta_2$, $w_D^{1} = (\frac{1}{2} - \delta_1) - \phi(\delta_1) - c_D$, $w_D^{2} = (\frac{1}{2} - \delta_2) - \phi(\delta_2) - c_D$, the challenger’s payoffs be $n_C^{1} = \frac{1}{2} + \delta_1$, $n_C^{2} = \frac{1}{2} + \delta_2$, $w_C^{1} = (\frac{1}{2} + \delta_1) + \phi(\delta_1) + \theta_C$, $w_C^{2} = (\frac{1}{2} + \delta_2) + \phi(\delta_2) + \theta_C$, and $\theta_C \sim U[-\frac{1}{4}, 0]$.}

The set of equilibria for this extension satisfies the following proposition.

**Proposition C.1.** If $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$, then for any $t^*$ such that $c_D < \frac{1}{2} - \delta_t^*$ $\iff$ $\delta_t^* < \frac{1}{2} - c_D$, there exists an equilibrium in which all types of challengers advance to $\delta_t^*$, only challengers with cost $c_C < \phi(\delta_t^*)$ attempt to advance to $\delta_t^* + 1$, and the defender always responds with war. If $c_D > \frac{1}{2}$ then in any equilibrium the challenger occupies the entire landmass.

**Proof:** We first show the desired equilibrium when $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$ and $\delta_t^* < \frac{1}{2} - c_D$. Define $\bar{t}$ as the period with the largest $\delta_{\bar{t}}$ strictly less than $\frac{1}{2} - c_D$, and observe that $\delta_{\bar{t}} < \frac{1}{4}$. Now consider the following strategy profile. After all histories, in periods $t \in [t^*, \bar{t}]$ the defender responds to
advancement with war, and the challenger only attempts to advance if $c_C < \phi(\delta_t)$. In all other periods the defender never responds with war, and the challenger always advances. This profile produces the desired equilibrium outcomes and the challenger is best responding. So we must show that the defender doesn’t wish to deviate.

Consider first a period $t < t^*$ in which the challenger attempts to advance. To get to this period the challenger must have advanced to $t$ and the defender must have always permitted it. So this is on equilibrium path, if the defender plays his equilibrium strategy of again permitting advancement then the challenger will advance all the way to $\delta_t^*$ before advancing triggers war, and the defender’s expected payoff is:

$$
\left( \frac{1}{2} - \delta_t^* \right) - P(c_C < \phi(\delta_t^*)) \left( \phi(\delta_t^*) + c_D \right).
$$

In words, the defender’s equilibrium expected holdings are $\frac{1}{2} - \delta_t^*$, with probability $P(c_C < \phi(\delta_t^*))$ war occurs in period $t^*$, and when this occurs the defender suffers the challenger’s excess military advantage $\phi(\delta_t^*)$ and the cost of war $c_D$. If instead the defender responds with war in period $t$, his payoff is $(1 - p(\delta_t)) - c_D = \left( \frac{1}{2} - \delta_t - \phi(\delta_t) \right) - c_D$, which is $< \text{eqn. (C.1)}$ i.f.f.

$$
c_D > \frac{((\delta_t^* + \phi(\delta_t^*)) - (\delta_t + \phi(\delta_t)))}{(1 - P(c_C < \phi(\delta_t^*)))} - \phi(\delta_t^*).
$$

Since $\phi(\delta_t) \to 0 \forall \delta_t$ as $Z \to \infty$, the r.h.s. approaches $\delta_t^* - \delta_t < \frac{1}{4}$ (since $\delta_t > 0$ and $\delta_t^* \leq \delta_t < \frac{1}{4}$) as $Z \to \infty$. So since $c_D \geq \frac{1}{4}$ there exists a $Z$ sufficiently large such that the inequality is satisfied for all $t < t^*$. Intuitively, we can scale down the excess military advantage function $\phi(\delta_t)$ by increasing $Z$ sufficiently so that the calculation essentially reduces to whether the cost of war exceeds the foregone share of the landmass from allowing the challenger to advance from $t$ all the way to $t^*$. This will always be true since (by assumption) the cost of war exceeds the challenger’s excess holdings in the period where war occurs ($\delta_t^* < c_D$).

Now consider a period $t \geq \bar{t}$ in which the challenger attempts to advance. This is off path, but
we do not need beliefs about the challenger’s type since if she is allowed to advance the strategies are for her to continue to advance and the defender to permit it. So if the defender allows advancement in $t$ the challenger will eventually possess the entire landmass and the defender’s payoff will be 0. If instead he responds with war his payoff is $\left(\frac{1}{2} - \delta_t - \phi(\delta_t)\right) - c_D$. Since $\delta_t < \frac{1}{2} - c_D$ and $\delta_t > \frac{1}{2} - c_D \forall t > \bar{t}$, for $Z$ sufficiently large it will be optimal for the defender to respond with war in $\bar{t}$ but not in $t > \bar{t}$. In words, at $\bar{t}$ the remaining landmass just exceeds the defender’s cost of war, so he will respond with war knowing that should he allow advancement he will also allow it in all future periods. For $t > \bar{t}$, the challenger is already sufficiently advanced that letting her take the remaining landmass is optimal.

Finally, consider a period $\hat{t} \in [t^*, \bar{t})$ in which the challenger attempts to advance and the defender is supposed to respond with war. The challenger already advanced in period $t^*$ expecting to trigger war. So the defender infers in equilibrium that her cost $c_C < \phi(\delta_{t^*})$, the threshold in the first period $t^*$ in which she advanced expecting war. If she is allowed to again advance in period $\hat{t}$ to period $\hat{t} + 1$, a further attempt to advance in $\hat{t} + 1$ will provoke war. Anticipating this, the challenger will once again advance i.f.f. $c_C < \phi(\delta_{\hat{t} + 1})$. Recall that $\delta_{\hat{t} + 1} \leq \delta_{\bar{t}} < \frac{1}{4}$ and $\phi(\delta_t)$ is increasing over $[0, \frac{1}{4}]$, so $\phi(\delta_{t^*}) < \phi(\delta_{\hat{t} + 1})$. In words, the region of the landmass is s.t. advancement makes war relatively more attractive to the challenger. So the defender can infer that a challenger who advanced to period $\hat{t}$ expecting war will again advance in period $\hat{t} + 1$ even though it will trigger war for sure. So responding with war in $\hat{t}$ is optimal, since permitting advancement will only weaken the defender in the inevitable war.

We last argue that when $c_D \geq \frac{1}{2}$, equilibrium requires the challenger to occupy the entire landmass. Suppose the defender’s strategy involves responding to further advancement with war with strictly positive probability in any period $t \geq 1$. This would yield utility $\left(\frac{1}{2} - \delta_t - \phi(\delta_t)\right) - c_D$ which is $< 0$ since $\delta_1 > 0$. If she were instead to deviate to always allowing advancement in every period, her utility would be $\geq 0$; at worst the challenger’s strategy will involve occupying the entire landmass.
(recall that the defender holds the final decision to fight). Thus, equilibrium requires the defender to always permit advancement. Equilibrium then must also requires the challenger to attempt advancement in every period regardless of her type since it will always be permitted, and the unique equilibrium outcome is that she will occupy the entire landmass.

We now consider the second extension, in which the challenger makes an endogenous “demand” $\delta_2$ of how far to advance. As in the baseline model, in this example the defender can allow a positive demand or respond with war, and if she advances the challenger can exploit her gains afterward by unilaterally initiating war.

**Extension 2 (Endogenous Transgression).** Consider the following $T = 2$ period game. In period 1 the challenger’s excess holdings are $\delta_1 > 0$, and she can attempt to advance to some $\delta_2 \in [\delta_1, \frac{1}{2}]$ of her choosing. The defender can permit the advancement or respond with war. If he permits it, then the game proceeds to the second period, and the challenger decides whether to unilaterally initiate war or enjoy her gains. After either choice the game ends.

The set of equilibria in this extension satisfy the following proposition.

**Proposition C.2.** If $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$ and the challenger’s excess share $\delta_1$ under the status quo is less than $\frac{1}{4} - \frac{c_D}{2}$, then there exists an equilibrium in which the defender responds to a strictly positive demand $\delta_2 \in (\delta_1, \frac{1}{2}]$, however small, with war. If $c_D \geq \frac{1}{2}$, then in any equilibrium the challenger demands the entire landmass, i.e. $\delta_2 = 1$, and it is accepted.

**Proof:** We first show the desired equilibrium when $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$; we construct an equilibrium where all demands are on-path. Challengers with cost $c_C > \phi(\delta_1)$ demand the status quo ($\delta^*_2(c_C) = \delta_1$), it is accepted, they do not initiate war in period 2, and the game ends. All challengers with cost $c_C \leq \phi(\delta_1)$ mix identically over all positive demands $\delta_2 \in (\delta_1, \frac{1}{2}]$ and the defender always responds with war. Should any such demand be accepted (off path), challengers with cost $c_C < \phi(\delta_2)$ unilaterally initiate war in the second period.
To see this is an equilibrium, consider first the defender’s strategy. If he sees no demand ($\delta_2 = \delta_1$), he infers that the challenger will initiate war in the second period with probability 0 and so maintaining the status quo is optimal. Should he see a positive demand ($\delta_2 > \delta_1$), he can infer that the challenger’s cost is below $\phi(\delta_1)$ but no more, since all such challengers mix identically over all positive demands. If the demand he receives satisfies $\delta_2 \in (\delta_1, \frac{1}{2} - \delta_1)$, then $\phi(\delta_1) < \phi(\delta_2)$, and since challengers with cost $c_C < \phi(\delta_2)$ will unilaterally initiate war in period 2, the probability of appeasing an already belligerent challenger by accepting such a demand is 0. Thus responding with war is optimal. If instead $\delta_2 \in [\frac{1}{2} - \delta_1, \frac{1}{2}]$, then even if allowing the demand would appease the challenger for sure the defender prefers to respond with war, since accepting such a demand will leave the defender with no more than $\delta_1$, while responding with war leaves him with $(\frac{1}{2} - \delta_1 - \phi(\delta_1)) - c_D > \delta_1$ when $\delta_1 < \frac{1}{4} - \frac{c_D}{2}$ for sufficiently large $Z$.

To see that the challenger wishes to play her equilibrium strategy, first note that period 2 strategies are straightforwardly optimal since the challenger is the last mover. In period 1, any positive demand will provoke war, and all challengers with cost $c_C \leq \phi(\delta_1)$ prefer war to the status quo. So such challengers are indifferent between all positive demands and are willing to mix according to the equilibrium strategy. Finally, challengers with cost $c_C > \phi(\delta_1)$ prefer the status quo to war and so making a 0 demand $\delta_2 = \delta_1$ is optimal.

We last argue that when $c_D \geq \frac{1}{2}$, equilibrium requires the challenger to occupy the entire landmass. If she demands $\delta_2 = 1$ and it is accepted, then in the second period it is optimal to end the game with peace regardless of her type. In the first period, the defender will thus accept such a demand, since it will yield utility 0, while war yields utility $(\frac{1}{2} - \delta_1 - \phi(\delta_1)) - c_D < 0$ for $c_D \geq \frac{1}{2}$. Finally, in any equilibrium the challenger must demand $\delta_2 = 1$ in the first period; all other demands will yield strictly lower utility regardless of the defender’s response, or her own anticipated strategy in the second period. □
Proposition C.2 further demonstrates that our result is not an artifact of having a fixed size of the transgression. When the status quo division is sufficiently close to an even division, there exists equilibria in which the defender responds to any positive demand, however small, with war. The logic is again identical to the two-period model. At the status quo, the challenger expects the defender to respond to any positive demand with war. Hence, the defender can infer in equilibrium that a challenger who makes such a demand desires war under the status quo. Because the challenger’s probability of victory \( p(\delta_t) \) is such that advancements \( \delta_2 \in (\delta_1, \frac{1}{2} - \delta_1) \) make war relatively more attractive, the probability of appeasing an already-belligerent challenger by permitting such an advancement is 0. Alternatively, while advancements \( \delta_2 \in [\frac{1}{2} - \delta_1, \frac{1}{2}] \) have some hope of successful appeasement, they are so large that the defender prefers to suffer the cost of war.

### C.2 Robustness to challenger backing down

**Proposition C.3.** Consider an alternative game \( \Gamma' \) form in which the challenger can back down in the first stage if the defender resists. Whenever the deterrence equilibrium exists in the original game \( \Gamma \) it also exists in \( \Gamma' \).

**Proof:** In \( \Gamma' \) the deterrence equilibrium takes the following form; the defender always resists, the challenger is deterred unless she prefers immediate war, and when she transgresses she also fights upon resistance. Now if the defender always resists, challenger types \( \theta_C \geq \bar{\theta}_1 \) still prefer to transgress because the defender will resist and they will then proceed with war. Challenger types \( \theta_C < \bar{\theta}_1 \) cannot get away with the transgression because the defender always resists, can back down upon encountering resistance, and are therefore indifferent between transgressing and not; they are thus willing to play the required strategy of not transgressing. Upon observing a transgression the defender therefore continues to infer that the challenger is of type \( \theta_C \geq \bar{\theta}_1 \), and in this case resisting is equivalent to unilaterally initiating war himself; his incentives and inferences are unchanged and
he is therefore willing to carry out his equilibrium strategy. ■

C.3 Game with interdependent war values

Suppose that both players’ payoffs in the event of war depend on the challenger’s type \( \theta_C \in \Theta \subset \mathbb{R} \) that is unknown to the defender but known to the challenger, where \( \Theta \) is an interval and \( \theta_C \) has a prior distribution \( f(\theta_C) \) with full support over \( \Theta \). The challenger’s type is therefore to be interpreted as a state of the world that affects both players’ payoffs over which the challenger has private information. Our notation and assumptions for the challenger’s payoffs are unchanged. For the defender, we now express the dependence of his war payoff on the challenger’s type using \( w^t_D(\theta_C) \), and make the following slightly-modified assumptions.

1. For all challenger types, allowing the transgression makes the defender strictly worse off in both peace \( (n^2_D < n^1_D) \) and war \( (w^2_D(\theta_C) < w^1_D(\theta_C) \ \forall \theta_C) \).

2. For all challenger types, allowing the transgression is strictly better than responding with war if the challenger will subsequently choose peace \( (n^2_D > w^1_D(\theta_C) \ \forall \theta_C) \).

Note that our defender assumptions jointly imply that the defender strictly prefers peace to war in each \( t \) for every type of challenger. Moreover, conditional on defender assumptions (1) – (2), any arbitrary dependence of the defender’s war payoff \( w^t_D(\theta_C) \) on the challenger’s type can be accommodated. However, it is natural to assume that \( w^t_D(\theta_C) \) is weakly decreasing in \( \theta_C \), i.e., a more belligerent challenger means a weaker defender. Our setup is not completely without loss of generality because it cannot capture when the challenger is privately informed about factors affecting the defender’s war payoffs but not her own; however, it is sufficiently general to capture private information about the probability of victory.

**Challenger Incentives** In the second period, the challenger transgresses i.f.f. \( \theta_C \geq \bar{\theta}_C^2 \). In the first period, challengers of type \( \theta_C \geq \bar{\theta}_C^1 \) always transgress. Challengers of type \( \theta_C < \bar{\theta}_C^1 \) transgress
i.f.f.,
\[ \alpha \cdot w_C^1(\theta_C) + (1 - \alpha) \cdot \max \{ n_C^2, w_C^2(\theta_C) \} \geq n_C^1. \]

For each such type, there exists a unique interior probability \( \hat{\alpha}(\theta_C) \) that would make them indifferent between transgressing and not, and given that probability the challenger would play a cutpoint strategy at \( \theta_C \). It is simple to verify that for \( \theta_C \leq \tilde{\theta}_C^1 \), \( \hat{\alpha}(\theta_C) \) is always well defined, strictly increasing in \( \theta_C \), strictly interior to \((0, 1)\), and \( \hat{\alpha}(\tilde{\theta}_C^1) = 1 \).

**Defender’s Incentives** Suppose that the challenger uses a threshold for transgressing equal to \( \hat{\theta}_C \). Then upon observing a transgression, the defender’s payoff from war is
\[
\int_{\hat{\theta}_C}^{\infty} w_D^1(\theta_C) \cdot \frac{f(\theta_C)}{1 - F(\theta_C)} d\theta_C
\]
and from appeasement is,
\[
\int_{\hat{\theta}_C}^{\max\{\hat{\theta}_C, \tilde{\theta}_C^2\}} n_D^2 \cdot \frac{f(\theta_C)}{1 - F(\theta_C)} d\theta_C + \int_{\max\{\hat{\theta}_C, \tilde{\theta}_C^2\}}^{\infty} w_D^2(\theta_C) \cdot \frac{f(\theta_C)}{1 - F(\theta_C)} d\theta_C.
\]
Hence she will prefer to respond to the transgression with war i.f.f.
\[
\int_{\max\{\hat{\theta}_C, \tilde{\theta}_C^2\}}^{\infty} (w_D^1(\theta_C) - w_D^2(\theta_C)) \cdot \frac{f(\theta_C)}{1 - F(\theta_C)} \geq \int_{\hat{\theta}_C}^{\max\{\hat{\theta}_C, \tilde{\theta}_C^2\}} (n_D^2 - w_D^1(\theta_C)) \cdot \frac{f(\theta_C)}{1 - F(\theta_C)} d\theta_C
\]
Now it is straightforward to show that the condition above is satisfied i.f.f.
\[
\tilde{\beta}(\hat{\theta}_C) \leq P(\theta \geq \tilde{\theta}_C^2 | \theta \geq \hat{\theta}_C), \quad (C.2)
\]
where

\[
\bar{\beta}(\hat{\theta}_C) = \frac{n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in [\hat{\theta}_C, \bar{\theta}_C^2]\right]}{(n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in [\hat{\theta}_C, \bar{\theta}_C^2]\right]) + E\left[w_D^1(\theta_C) - w_D^2(\theta_C) \mid \theta_C \geq \bar{\theta}_C^2\right]}
\]  

(C.3)

Intuitively, \(n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in [\hat{\theta}_C, \bar{\theta}_C^2]\right]\) is the benefit from appeasement conditional on the challenger being appeasable. Similarly, \(E\left[w_D^1(\theta_C) - w_D^2(\theta_C) \mid \theta_C \geq \bar{\theta}_C^2\right]\) is the benefit from preemptive war conditional on the challenger being unappeasable. Finally, as before \(P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \hat{\theta}_C\right)\) is the interim probability that the challenger is unappeasable.

Now note the following. First, \(\bar{\beta}(\hat{\theta}_C)\) is strictly interior to \([0, 1]\) for any value of \(\hat{\theta}_C\) by our payoff assumptions, since appeasement is beneficial when it is possible \((\theta_C \in [\hat{\theta}_C, \bar{\theta}_C^2])\) and war early is better than war later when it is not \((\theta_C > \bar{\theta}_C^2)\). Second, \(\bar{\beta}(\hat{\theta}_C)\) is weakly increasing in \(\hat{\theta}_C\) in the natural case where a belligerent challenger is “bad news” for the defender (i.e. \(w_D^1(\theta_C)\) is decreasing in \(\theta_C\)) since then \(E\left[w_D^1(\theta_C) \mid \theta_C \in [\hat{\theta}_C, \bar{\theta}_C^2]\right]\) is decreasing in \(\hat{\theta}_C\). Third and as in the baseline model, \(P\left(\theta \geq \bar{\theta}_C^2 \mid \theta \geq \hat{\theta}_C\right)\) is increasing in \(\hat{\theta}_C\) – that is, the defender’s interim assessment that appeasement will be ineffective is higher when the challenger uses a higher threshold for transgressing.

**Equilibrium Characterization** Applying the analysis above, we now have the following complete equilibrium characterization.

**Proposition C.4.** Equilibria of the model with interdependent values are as follows.

- The deterrence equilibrium exists i.f.f.

\[
\bar{\beta}(\hat{\theta}_C^1) \leq P\left(\theta \geq \bar{\theta}_C^2 \mid \theta \geq \hat{\theta}_C^1\right)
\]
• The no deterrence equilibrium exists i.f.f.

\[ P (\theta_C \geq \bar{\theta}_C^1) \leq \bar{\beta} (-\infty) \]

• A mixed strategy equilibrium in which the challenger uses threshold \( \hat{\theta}_C^* < \min \{ \bar{\theta}_C^1, \bar{\theta}_C^2 \} \) exists i.f.f

\[ \bar{\beta} (\hat{\theta}_C^*) = P (\theta \geq \bar{\theta}_C^2 \mid \theta \geq \hat{\theta}_C^*) \]

In the equilibrium, the defender responds to the transgression with war with probability \( \alpha^* = \hat{\alpha} (\hat{\theta}_C^*) \).

The most important observation from the above characterization is the following: because \( \bar{\beta} (\hat{\theta}_C) \) is interior for all \( \hat{\theta}_C \) (meaning that war sooner is better than war later), our basic insight holds unaltered. When appeasement is ineffective \( (\bar{\theta}_C^2 \leq \bar{\theta}_C^1) \), the deterrence equilibrium exists for all distributions over the challenger’s type \( \theta_C \) and functions \( w^{D} (\theta_C) \) mapping the challenger’s type into the defender’s payoff from war that satisfy the initial assumptions. Thus, Corollaries 1 and 3 continue to hold unaltered with interdependent values.

Other more subtle patterns of equilibria can occur with interdependent values. Because \( \bar{\beta} (\hat{\theta}_C) \) can be steeply increasing in \( \hat{\theta}_C \) rather than constant, it is no longer the case that the mixed strategy equilibrium can only exist when both pure strategy equilibria exist. Many different scenarios can occur, including an odd number of mixed strategy equilibria combined with an even number of pure strategy equilibria (including none), and a single pure strategy equilibrium combined with an even number of mixed strategy equilibria.

Intuitively, the reason for this multiplicity of equilibria is that a higher threshold for transgressing by the challenger has two countervailing effects. First, it makes the defender less willing to appease because her interim assessment of the probability that the challenger is unappeasable is higher.
Second, it makes the challenger more willing to appease because inferring the challenger is a higher type also means that war is worse, making appeasement more attractive if it can be effective. These countervailing effects can then generate multiple equilibria: with higher thresholds, the defender can find appeasement less likely to be effective, but simultaneously more desirable if it would be effective.

C.4 Game with two-sided uncertainty

The defender is now assumed to have a type $\theta_D$ upon which his war payoffs in each period depend, so we write $w^t_D(\theta_D)$ to express this dependence. We maintain the assumption that payoffs in peace for both players are fixed and common knowledge, and make new assumptions on the defender’s type that mirror those of the challenger. Specifically, $\theta_D$ also belongs to an interval, has some prior distribution $g(\theta_D)$ with full support, and is distributed independently of $\theta_C$. Thus, war values are private and each side’s uncertainty may be interpreted as about the opponent’s cost of war. We modify the assumptions the defender’s payoffs as follows:

1. For all defender types, allowing the transgression makes the defender strictly worse off in both peace ($n^2_D < n^1_D$) and war ($w^2_D(\theta_D) < w^1_D(\theta_D)$ $\forall \theta_D$).

2. In each period $t$ the defender’s war payoff $w^t_D(\theta_D)$ is continuous and strictly increasing in $\theta_D$.

In addition, there exists a unique defender type $\bar{\theta}_t^D$ that is indifferent between peace and war in period $t$.

3. The benefit $w^1_D(\theta_D) - w^2_D(\theta_D) > 0$ of war sooner vs. war is weakly increasing in the defender’s type.

The first assumption extends the properties of the transgression to a setting where the defender’s payoffs can vary, and the second mirrors the assumptions made on the challenger’s type. Importantly, it implies that with strictly positive probability the defender’s threat is “inherently” credible in that he is willing to go to war solely to prevent the transgression. Formally, for both players let $\tilde{\theta}^{x,t}_i$
denote a player indifferent between peace in period $s$ and war in period $t$ – since $n^2_D < n^1_D$ we have $\bar{\theta}^{2,1}_D < \bar{\theta}^{1,1}_D$ and types in between are willing to fight a war over the transgression.

The third assumption ensures that types who are overall more belligerent are also weakly more willing to go to war for preemptive reasons, and is necessary for the existence of cutpoint strategies. Finally, since the defender may unilaterally wish to initiate war in both periods, we augment the first period with a final stage in which the defender can start a war even if the challenger chooses not to transgress. It is unnecessary to augment the second period with a similar stage because any defender type who would unilaterally initiate war in the second stage would also initiate war in the first stage and end the game.

**Challenger Incentives**  Challenger incentives are identical to the game with interdependent war values except for the following distinction – because the defender may now be of a type $\theta_D \geq \bar{\theta}^1_D$ who would start a war whether or not the challenger attempts to transgress, $\alpha$ now denotes the probability that transgressing would *provoke* an otherwise peaceful challenger to start a war. If the defender uses a cutpoint strategy $\hat{\theta}_D \leq \bar{\theta}^1_D$ for responding to the transgression, then in equilibrium $\alpha = \frac{G(\bar{\theta}^1_D) - G(\hat{\theta}_D)}{G(\bar{\theta}^1_D)}$.

**Defender’s Incentives**  The defender’s war payoffs now depend on her type $\theta_D$; moreover, because types are independent the threshold $\hat{\theta}_C$ that the challenger uses for transgressing only affects her payoffs through the interim assessment $\beta$ that the challenger would initiate war after being allowed to transgress. He therefore prefers to respond to the transgression with war when $\beta \geq \bar{\beta}(\theta_D)$, where

$$\bar{\beta}(\theta_D) = \frac{n^2_D - w^1_D(\theta_D)}{(n^2_D - w^1_D(\theta_D)) + (w^2_D(\theta_D) - w^3_D(\theta_D))}. \quad (C.4)$$

It is simple to verify that for $\theta_D \in [0, \bar{\theta}^{2,1}_D)$ (where $\bar{\theta}^{2,1}_D$ is the defender type indifferent between immediate war and successful appeasement) the function $\bar{\beta}(\theta_D)$ is strictly interior to $[0,1]$ and
decreasing (by assumption 3). The latter property ensures that the defender always plays a cutpoint strategy, and we can therefore also work with the inverse function $\bar{\theta}_D(\beta) = \beta^{-1}(\beta)$ denoting the defender type indifferent between appeasement and war when his interim assessment is $\beta$.

**Equilibrium Characterization**

**Proposition C.5.** Equilibria of the model with two-sided uncertainty are as follows.

- **The deterrence equilibrium exists i.f.f.**
  \[ \bar{\beta}(-\infty) \leq P(\theta_C \geq \bar{\theta}_C^2 | \theta \geq \bar{\theta}_C^1) \]

- **The no deterrence equilibrium exists i.f.f.**
  \[ P(\theta_D \in [\bar{\theta}_D(P(\theta_C \geq \bar{\theta}_C^2)) , \bar{\theta}_D^1] | \theta_D \leq \bar{\theta}_D^1) \leq \hat{\alpha}(-\infty) \]

- **An interior equilibrium with challenger threshold $\hat{\theta}_C^* < \min \{ \bar{\theta}_C^1, \bar{\theta}_C^2 \}$ exists i.f.f.**
  \[ P(\theta_D \in [\bar{\theta}_D(P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \hat{\theta}_C^*) ) , \bar{\theta}_D^1] | \theta_D \leq \bar{\theta}_D^1) = \hat{\alpha}(\hat{\theta}_C^*) \]

  or equivalently
  \[ \frac{G(\bar{\theta}_D \left( \frac{1-F(\bar{\theta}_C^2)}{1-F(\hat{\theta}_C^*)} \right)) - G(\bar{\theta}_D^1)}{G(\bar{\theta}_D^1)} = \hat{\alpha}(\hat{\theta}_C^*) \]

  In the equilibrium, the challenger transgresses when $\theta_C \geq \hat{\theta}_C^*$ and the defender responds with war i.f.f. $\theta_D \geq \bar{\theta}_D \left( P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \hat{\theta}_C^*) \right) = \bar{\theta}_D \left( \frac{1-F(\bar{\theta}_C^2)}{1-F(\hat{\theta}_C^*)} \right)$.

  Again, the most important observation from the above characterization is that because $\bar{\beta}(\hat{\theta}_C^*)$ is interior for all $\hat{\theta}_C$ (meaning that war sooner is better than war later), our basic insight again
holds unaltered. When appeasement is ineffective \((\bar{\theta}_C^2 \leq \bar{\theta}_C^1)\), the deterrence equilibrium exists for all distributions over the challenger’s type \(\theta_C\) and defender’s type \(\theta_D\) that satisfy the initial assumptions, and Corollaries 1 and 3 hold unaltered.

As with interdependent war values other more subtle patterns of equilibria can also occur. Intuitively, the reason is that deterrence begets deterrence – a higher threshold for transgressing \((\hat{\theta}_C)\) generates a higher interim assessment \(\frac{1-F(\hat{\theta}_C^2)}{1-F(\hat{\theta}_C^1)}\) that the challenger is unappeasable, generating a lower threshold \(\bar{\theta}_D\left(\frac{1-F(\hat{\theta}_C^2)}{1-F(\hat{\theta}_C^1)}\right)\) for the defender to respond with war, a higher probability \(G\left(\bar{\theta}_D\left(\frac{1-F(\hat{\theta}_C^2)}{1-F(\hat{\theta}_C^1)}\right)\right) - G(\bar{\theta}_D^1)\) that the defender will be provoked by an attempted transgression, and thus more deterrence. Under some conditions this dynamic can set off a “deterrence spiral” where the challenger is very unlikely to be unappeasable ex-ante yet the deterrence equilibrium is unique – a sufficient condition for this occurring is the standard condition that the “no deterrence” equilibrium be unstable and the slope of the challenger’s best response function

\[
\hat{\alpha}^{-1}\left(\frac{G\left(\bar{\theta}_D\left(\frac{1-F(\hat{\theta}_C^2)}{1-F(\hat{\theta}_C^1)}\right)\right) - G(\bar{\theta}_D^1)}{G(\bar{\theta}_D^1)}\right)
\]

be greater than 1 (where \(\hat{\alpha}^{-1}(\alpha)\) denotes the well-defined inverse of \(\hat{\alpha}(\theta_C)\)).

### C.5 Game with uncertainty about transgression’s military consequences

We now consider robustness to ex-ante uncertainty about the military consequences of the transgression. Suppose that the challenger’s second period war payoff is \(w_C^2(\theta) - \sigma\varepsilon\), where \(\varepsilon\) is a symmetric random variable with mean 0, variance 1, and atomless full support (so that \(\sigma\varepsilon\) has variance \(\sigma^2\)). Let \(H(\varepsilon)\) denote the CDF describing the distribution of \(\varepsilon\). The realization of \(\varepsilon\) is initially unknown to both players, and then publicly revealed following a successful transgression. This captures the idea that the realized military benefits \((w_C^2(\theta_C) - \sigma\varepsilon) - w_C^1(\theta_C)\) of the transgression to the challenger
may be more or less than the initial expected value \( w_C^2(\theta_C) - w_C^1(\theta_C) = \delta^m(\theta_C) \).

**Challenger Incentives** In the second period, the challenger initiates war i.f.f. \( \varepsilon \leq \frac{w_C^2(\theta_C) - n_C^2}{\sigma} \). So from an ex-ante perspective, a challenger with type \( \theta_C \) will initiate war in the second period with probability \( H \left( \frac{w_C^2(\theta_C) - n_C^2}{\sigma} \right) \).

In the first period it is easily verified that challengers of type \( \theta_C \geq \bar{\theta}_1^C \) always transgress. Challengers of type \( \theta_C < \bar{\theta}_1^C \) transgress i.f.f.,

\[
\alpha \cdot w_C^1(\theta_C) + (1 - \alpha) \int \max \left\{ n_C^2, w_C^2(\theta_C) - \sigma\varepsilon \right\} h(\varepsilon) d\varepsilon \geq n_1^C
\]

It is also easily shown that for each such type, there exists a unique interior probability \( \hat{\alpha}(\theta_C) \) that would make them indifferent between transgressing and not, and given that probability the challenger would play a cutpoint strategy at \( \theta_C \). Finally, it is simple to verify that for \( \theta_C \leq \bar{\theta}_1^C \), \( \hat{\alpha}(\theta_C) \) is always well defined, strictly increasing in \( \theta_C \), strictly interior to (0, 1), and \( \hat{\alpha}(\bar{\theta}_1^C) = 1 \).

**Defender Incentives** Defender incentives are as in the baseline model; he prefers to respond to an attempted transgression with war i.f.f. his interim belief \( \beta \) is \( \beta = \frac{n_D^2 - w_D^1}{(n_D^2 - w_D^1) + (w_D^1 - w_D^2)} \). It is also helpful to denote \( \bar{\theta}_C \) as the challenger type against whom the defender would be exactly indifferent to responding to the transgression with war; i.e., \( H \left( \frac{w_C^2(\bar{\theta}_C) - n_C^2}{\sigma} \right) = \bar{\beta} \). Note that no such type exists in the baseline model because the challenger’s probability of initiating war after transgressing increases discontinuously from 0 to 1 when \( \theta_C \) crosses \( \bar{\theta}_1^C \).

We must also characterize the defender’s interim beliefs about the probability a challenger who transgressed will initiate war in the second period when she is uncertain of the challenger’s type; given the analysis in the preceding section we may restrict attention to cutpoint strategies by the challenger. When the challenger plays a cutpoint strategy of transgressing i.f.f. \( \theta_C \geq \hat{\theta}_C \), it is straightforward to see that the defender’s interim assessment of this probability is \( E_{\theta_C} \left[ H \left( \frac{w_C^2(\theta_C) - n_C^2}{\sigma} \right) \right] \mid \theta_C \geq \hat{\theta}_C \].
It is also easy to show that the derivative of this probability w.r.t. the challenger’s cutpoint $\hat{\theta}_C$ is

$$f(\hat{\theta}_C) \int_{\theta_C \geq \hat{\theta}_C} \left( H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right) - H\left( \frac{w^2_C(\hat{\theta}_C) - n^2_C}{\sigma} \right) \right) \frac{f(\theta_C)}{1 - F(\hat{\theta}_C)} d\theta_C.$$  

This is > 0 since $\frac{w^2_C(\theta_C) - n^2_C}{\sigma}$ is strictly increasing in $\theta_C$. The defender’s interim assessment is thus smoothly increasing in the challenger’s cutpoint $\hat{\theta}_C$, approaches the prior $E_{\theta_C} \left[ H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right) \right]$ as $\theta_C \to \inf \Theta$, and approaches $\lim_{\theta_C \to \infty} H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right)$ as $\theta_C \to \sup \Theta$.

**Equilibrium Characterization** Using the above characterizations, equilibrium is as follows.

**Proposition C.6.** With uncertainty about the transgression’s military value to the challenger,

1. there exists a **no deterrence equilibrium**, in which the challenger always transgresses, and she is always permitted to do so, i.f.f.

   $$\tilde{\beta} \geq E_{\theta_C} \left[ H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right) \right]$$

2. there exists a **deterrence equilibrium**, in which (i) the defender always responds to the transgression with war, (ii) all types $\theta_C < \bar{\theta}_1^C$ who do not initially prefer war are deterred, and (iii) the probability of deterrence is $P(\theta_1 < \bar{\theta}_1^C)$, i.f.f.

   $$\tilde{\beta} \leq E_{\theta_C} \left[ H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right) \mid \theta_C \geq \bar{\theta}_1^C \right]$$

3. a mixed strategy equilibrium exists i.f.f. both the no deterrence and deterrence equilibria exist, and is uniquely characterized by both a challenger cutpoint $\hat{\theta}_C^*$ satisfying

   $$\tilde{\beta} = E_{\theta_C} \left[ H\left( \frac{w^2_C(\theta_C) - n^2_C}{\sigma} \right) \mid \theta_C \geq \hat{\theta}_C^* \right]$$
and a defender probability of responding to the transgression with war equal to \( \alpha^* = \hat{\alpha} \left( \hat{\theta}_C \right) \).

When the mixed strategy equilibrium exists the defender is best off in the deterrence equilibrium.

To summarize, equilibrium incentives and conditions are effectively identical to the baseline model, except that the idiosyncratic shock \( \sigma \varepsilon \) “smooths out” the defender’s interim assessment of the probability that the challenger will initiate war if allowed to transgress. The proof is simply the assembly of the preceding analysis except for the payoff dominance of the deterrence equilibrium when there are multiple equilibria – this is straightforward to show using an identical argument as in Proposition 1 but with the defender’s interim assessments suitably modified.

**Results**  We now restate analogs to our main results in this variant of the model.

**Proposition C.7.** With uncertainty about the transgression’s military value to the challenger,

1. if \( \bar{\theta}_C^\sigma \leq \bar{\theta}_C^1 \) then the deterrence equilibrium exists.

2. if absent uncertainty the deterrence equilibrium exists strictly \( (\bar{\beta} > P (\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)) \), then it also exists with uncertainty for sufficiently small \( \sigma \). Consequently, if \( \bar{\theta}_C^2 < \bar{\theta}_C^1 \) \( \iff \) \( \delta_C^m (\bar{\theta}_C^1) > \delta_C^d \) then the deterrence equilibrium exists for sufficiently small \( \sigma \).

3. under the assumptions stated in Proposition 3, the probability that deterrence is successful is increasing in \( \delta_C^m - \delta_C^d \).

4. if the deterrence equilibrium prevails whenever it exists, then the probability of deterrence would decrease if the defender knew the challenger’s type. In addition, the defender is better off not knowing the challenger’s type i.f.f. either \( \bar{\theta}_C^\sigma \leq \bar{\theta}_C^1 \), or \( \bar{\theta}_C^\sigma > \bar{\theta}_C^1 \) and

\[
\int_{\bar{\theta}_C^1}^{\bar{\theta}_C^0} \left( \bar{\beta} - H \left( \frac{w_C^2 (\theta_C) - \tilde{n}_C^2}{\sigma} \right) \right) f (\theta_C) d\theta_C 
\leq
\]

30
Proposition C.7 shows that analogues of our main results hold in this variant of the model; for the purposes of exposition the proof is deferred to the end of this section. The reason is simple. Introducing uncertainty about the consequences of the transgression (captured by the size of $\sigma$) weakens the extent to which future belligerence can be inferred from present belligerence. However, it does not change the fact that a higher relative military gain strengthens this inference, nor the defender’s fundamental incentives. Consequently, our results are not knife edge – introducing a little bit of uncertainty about the transgression’s consequences does not perturb them, but with enough uncertainty the conditions for deterrence may fail to hold.

Specifically, the proposition first provides an analogue condition \( \bar{\theta}_C \sigma \leq \bar{\theta}_1 C \) to the existence condition \( \bar{\theta}_C^2 \leq \bar{\theta}_1 C \) for the deterrence equilibrium in the baseline model; that the challenger type at which the defender switches from preferring appeasement to preferring war be weakly less than the challenger type indifferent between peace and war in the first period. Absent uncertainty about the transgression’s consequences this type is exactly $\bar{\theta}_2 C$ – that is, the challenger type indifferent between peace and war in the second period. The reason is that absent uncertainty this is challenger type at which appeasement switches from being an effective to ineffective strategy. With uncertainty however, this type is a more complex function of the uncertainty and underlying payoffs.

The second part of the proposition states that introducing a little bit of uncertainty about the transgression’s consequences does not perturb the deterrence equilibrium; that is, if deterrence works without uncertainty, then it will also work with a little bit of uncertainty. This implies that deterrence always works when the transgression’s expected military value exceeds its direct value ($\delta_m^C - \delta_d^C > 0$) even when a little bit of uncertainty is introduced. Relatedly, the third part of the proposition states that with uncertainty it remains true that the probability of successful deterrence is increasing in $\delta_m^C - \delta_d^C$. Both with and without this uncertainty, a challenger who prefers war to peace in the first
period is ex-ante more likely to prefer war after transgressing the larger is this difference.

The last part of the proposition states that knowing the challenger’s type still reduces the probability of deterrence. With such knowledge the challenger can only successfully deter when $\hat{\theta}_C^\sigma < \bar{\theta}_C^1$ (where $\hat{\theta}_C^\sigma = \bar{\theta}_C^2$ when $\sigma = 0$), and moreover can only deters types $\theta_C \in [\hat{\theta}_C^\sigma, \bar{\theta}_C^1]$. However, when she lacks this knowledge, then under these conditions the deterrence equilibrium exists, and in it she deters all types $\theta_C \leq \bar{\theta}_C^1$. Finally, the proposition states analogous conditions to those in the baseline model for the defender to be better off lacking knowledge of the challenger’s type; this is the conjunction of the deterrence equilibrium existing, and the benefits of deterring types $\theta_C \leq \bar{\theta}_C^1$ outweighing the costs of fighting wars against types $\theta_C \in [\bar{\theta}_C^1, \bar{\theta}_C^2]$ that she’d prefer to appease.

**Proof of Proposition C.7 (Part 1)** Observe that

$$\bar{\beta} = H\left(\frac{w_C^2(\theta_C^\sigma) - n_C^2}{\sigma}\right) \leq E_{\theta_C} \left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \left| \theta_C \geq \bar{\theta}_C^1\right]\right] \leq E_{\theta_C} \left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \left| \theta_C \geq \bar{\theta}_C^1\right]\right],$$

so the deterrence equilibrium exists. The first equality follows from the definition of $\bar{\theta}_C^\sigma$ and the remaining inequalities from previous arguments.

**(Part 2)** Comparing the conditions for the deterrence equilibrium in Propositions 1 and C.6, it suffices to show that $\lim_{\sigma \to 0} \left\{ E_{\theta_C} \left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \left| \theta_C \geq \bar{\theta}_C^1\right]\right]\right\} = P(\theta_C \geq \bar{\theta}_C^1 | \theta_C \geq \bar{\theta}_C^1)$. Observe that $P(\theta_C \geq \bar{\theta}_C^1 | \theta_C \geq \bar{\theta}_C^1) = \int_{\theta_C \geq \bar{\theta}_C^1} 1_{\theta_C \geq \bar{\theta}_C^1} \cdot \frac{f(\theta_C)}{1-F(\bar{\theta}_C)} d\theta_C$ and

$$E_{\theta_C} \left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \left| \theta_C \geq \bar{\theta}_C^1\right]\right] = \int_{\theta_C \geq \bar{\theta}_C^1} H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \frac{f(\theta_C)}{1-F(\bar{\theta}_C)} d\theta_C$$

The desired result then follows from the dominated convergence theorem because $H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \leq 1$ $\forall (\theta_C, \sigma)$ and converges pointwise almost everywhere to $1_{\theta_C \geq \bar{\theta}_C^1}$ (since $\lim_{\sigma \to 0} \left\{ H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \right\} = 0$ for $w_C^2(\theta_C) < n_C^2 \iff \theta_C < \bar{\theta}_C^2$ and $\lim_{\sigma \to 0} \left\{ H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \right\} = 1$ for $w_C^2(\theta_C) > n_C^2 \iff \theta_C > \bar{\theta}_C^2$).
Last the stated implications of $\bar{\theta}_{C}^{2} \leq \bar{\theta}_{C}^{1}$ follow from the above and $P (\theta_{C} \geq \bar{\theta}_{C}^{2} \mid \theta_{C} \geq \bar{\theta}_{C}^{1}) = 1 > \bar{\beta}$.

(Part 2) Given the assumptions in the proposition and holding $\bar{\theta}_{C}^{1}$ (i.e. the challenger’s first period payoffs) fixed, the probability of deterrence is 0 if $E_{\theta_{C}} \left[ H \left( \frac{w_{C}^{1}((\theta_{C}) - n_{D}^{1})}{\sigma} + \frac{(\delta_{m}^{1} - \delta_{d}^{1})}{\sigma} \right) \mid \theta_{C} \geq \bar{\theta}_{C}^{1} \right] < \bar{\beta}$ and $P (\theta_{C} \leq \bar{\theta}_{C}^{1})$ otherwise. The quantity on the l.h.s. is self-evidently increasing in $\delta_{m}^{1} - \delta_{d}^{1}$; hence the result is shown.

(Part 3) If the defender knew the challengers type, he would respond with war i.f.f. $H \left( \frac{w_{C}^{2}(\theta_{C}) - n_{D}^{2}}{\sigma} \right) \geq \bar{\beta}$, implying that the challenger would be deterred i.f.f. $\theta_{C} \in (\bar{\theta}_{C}^{2}, \bar{\theta}_{C}^{1})$. The probability of deterrence would therefore be $P (\theta_{1} < \bar{\theta}_{C}^{1}, \theta_{1} \geq \bar{\theta}_{C}^{2})$. Now suppose first that the deterrence equilibrium exists when the challenger’s type is unknown; then the probability of deterrence is $P (\theta_{1} < \bar{\theta}_{C}^{1})$, which is $> \beta$ the probability of deterrence $P (\theta_{1} < \bar{\theta}_{C}^{1}, \theta_{1} \geq \bar{\theta}_{C}^{2})$ when the challenger’s type is known. Next suppose that the deterrence equilibrium does not exist when the challenger’s type is unknown, so that the probability of deterrence is 0. Then by Part 1 we must have $\bar{\theta}_{C}^{2} > \bar{\theta}_{C}^{1}$, which implies that the probability of deterrence is also 0 when the challenger’s type is known.

Next we consider when the defender is better off not knowing the challenger’s type. This is clearly the case when $\bar{\theta}_{C}^{2} \leq \bar{\theta}_{C}^{1}$; types $< \bar{\theta}_{C}^{2}$ are deterred when they otherwise would not be, and for all other types the outcome is identical. So suppose that $\bar{\theta}_{C}^{1} < \bar{\theta}_{C}^{2}$, and henceforth for notational simplicity denote $H \left( \frac{w_{C}^{2}(\theta_{C}) - n_{D}^{2}}{\sigma} \right) = p (\theta_{C})$. First observe that a necessary condition for the defender to be better off not knowing is that the deterrence equilibrium exists; if it does not, then with uncertainty the challenger will always transgress and the defender will always appease, but lacking uncertainty the defender can identify the most warlike types $\theta_{C} \geq \bar{\theta}_{C}^{2}$ and fight them early. We can write the existence condition in Propositions C.6.2 in a more usable form as:

$$\int_{\bar{\theta}_{C}^{2}}^{\infty} \left( (1 - p(\theta_{C})) n_{D}^{2} + p(\theta_{C}) w_{D}^{2} \right) \frac{f(\theta_{C})}{1 - P(\theta_{C})} d\theta_{C} \leq 0$$

$$\iff \int_{\bar{\theta}_{C}^{2}}^{\bar{\theta}_{C}^{2}} (\bar{\beta} - p(\theta_{C})) \ f(\theta_{C}) \ d\theta_{C} \leq \int_{\bar{\theta}_{C}^{2}}^{\infty} (p(\theta_{C}) - \bar{\beta}) \ f(\theta_{C}) \ d\theta_{C}$$

(C.5)
where the second line comes from observing that the ex-ante net benefit of appeasement \((1 - p(\theta_C)) n_D^2 + p(\theta_C) w_D^2 - w_D^1\) can be rewritten as \((n_D^2 - w_D^2) (\tilde{\beta} - p(\theta_C))\).

Last we must characterize when the defender prefers not knowing and the deterrence equilibrium to knowing the challenger’s type. Comparing the two scenarios, outcomes are the same when \(\tilde{\theta}_C > \tilde{\theta}_C^\sigma\) (war in the first period), there is a deterrence benefit of not knowing when \(\tilde{\theta}_C \leq \tilde{\theta}_C^1\), and there is a cost of fighting early wars when \(\theta_C \in [\tilde{\theta}_C^1, \tilde{\theta}_C^\sigma]\). To prefer not knowing therefore requires that:

\[
\int_{\tilde{\theta}_C^1}^{\tilde{\theta}_C^\sigma} \left((1 - p(\theta_C)) n_D^2 + p(\theta_C) w_D^2\right) f(\theta_C) d\theta_C \leq \int_{0}^{\tilde{\theta}_C^1} \left(n_D^1 - ((1 - p(\theta_C)) n_D^2 + p(\theta_C) w_D^2)\right) f(\theta_C) d\theta_C
\]

\[
\iff \int_{\tilde{\theta}_C^1}^{\tilde{\theta}_C^\sigma} (\tilde{\beta} - p(\theta_C)) f(\theta_C) d\theta_C \leq \int_{0}^{\tilde{\theta}_C^1} \left(\tilde{\beta} \left(n_D^1 - n_D^2\right) + p(\theta_C)\right) f(\theta_C) d\theta_C,
\]

where the second line comes from using the previously observed equality and observing that \(\left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1}\right) = \tilde{\beta} \left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1}\right)\). The conjunction of equations C.5 and C.6 then yields the condition in the Proposition.

### C.6 A challenger who can send a costly signal

In this section we consider whether the possibility of costly signaling prior to game play will undermine the deterrence equilibrium in the baseline model. For the purposes of exposition all proofs are deferred to the end of the section.

Consider a variant of the model in which the challenger, when she transgresses, can also burn utility \(c \geq 0\) (of her choosing) in order to signal information about her type. Let \(\alpha(c)\) denote the probability the defender responds to the transgression with war conditional on a costly signal of \(c\).

We first state and prove the main result – that when the military value of the transgression strictly exceeds its direct value, the previously-characterized deterrence equilibrium of our model is robust in the sense of satisfying universal divinity (Banks and Sobel 1987). There is thus a strong sense in which the ability to send costly signals does not undermine our main result.
Proposition C.8. If \( \bar{\theta}_C^2 < \bar{\theta}_C^1 \iff \delta_C^m (\bar{\theta}_C^1) > \delta_C^d \), then there exists a universally divine equilibrium in which

- the challenger never sends a costly signal, and transgresses i.f.f. \( \theta_C \geq \bar{\theta}_C^1 \)
- the defender responds to the transgression with war (\( \alpha (c) = 1 \)) for all \( c < \max_{\theta_C \geq \bar{\theta}_C^1} \{ \delta_C^m (\theta_C) \} \).

We now heuristically explain why when \( \bar{\theta}_C^2 < \bar{\theta}_C^1 \), there exists a universally divine equilibrium in which peaceful challengers cannot use costly signals to “separate” themselves and induce the defender to allow the transgression. Consider a strategy profile in which deterrence occurs and the challenger never sends the costly signal; the defender’s inferences upon observing the transgression but no signal (\( c = 0 \)) are therefore as in the original model, while transgressing and signaling (\( c > 0 \)) is “off path.” What must the defender believe about the challenger’s intentions if he were to observe that she transgressed but also sent a costly signal \( c > 0 \)? Roughly, universal divinity requires that he believe the signal to have been sent by the challenger type for whom successfully transgressing is most profitable, and respond accordingly. Crucially, this cannot be a peaceful type (\( \theta_C < \bar{\theta}_C^2 \)). The reason is that when appeasement is impossible, there is some chance that the challenger is “opportunistically belligerent” (\( \theta_C \in (\bar{\theta}_C^2, \bar{\theta}_C^1) \)) – that is, initially deterrable, but would start a war if allowed to transgress. Since an opportunistically-belligerent challenger necessarily places a strictly greater value \( w_C^2 (\theta_C) - n_C^1 \) on successfully transgressing than a peaceful one \( n_C^2 - n_C^1 = \delta_C^d \), the defender must respond with war, and sending such a signal cannot be profitable for the challenger.

While the introduction of costly signaling does not cause the deterrence equilibrium to unravel when our main condition holds, the same cannot be said when the condition fails. The following proposition characterizes sufficient conditions for deterrence to collapse in all universally divine equilibria when the possibility for costly signaling is introduced.

Proposition C.9. If \( \bar{\theta}_C^1 < \bar{\theta}_C^2 \iff \delta_C^d > \delta_C^m (\bar{\theta}_C^1) \) and \( \delta_C^d > \max_{\theta_C \geq \bar{\theta}_C^1} \{ \delta_C^m (\theta_C) \} \), then in every universally divine equilibrium the challenger always transgresses.
Thus, when appeasement is possible ($\bar{\theta}_C^1 < \bar{\theta}_C^2$), the potential for costly signaling can undermine the deterrence equilibrium. The reason is that the value $\delta_C^d = n_C^1 - n_C^1$ that peaceful challengers place on successfully transgressing is necessarily higher than the value $n_C^2 - w_C^1(\theta_C)$ that belligerent-but-appeasable challengers place on it. It is thus at least possible that a peaceful challenger places a higher value on transgressing than any unappeasably belligerent one ($\delta_C^d > \max_{\theta_C \geq \bar{\theta}_C^2} \{ \delta_C^m(\theta_C) \}$) and can therefore “separate” themselves with costly signaling. An additional interesting implication of Propositions C.8 and C.9 is that when the transgression’s military value is constant for all challenger types (as considered in Proposition 3), then with costly signaling there exists a universally divine equilibrium with deterrence if and only if $\delta_C^m \geq \delta_C^d$.

To conclude, we provide an example in which (i) the deterrence equilibrium exists in the baseline model, but (ii) with costly signaling deterrence unravels in every universally divine equilibrium.

**Example:** Suppose that $w_C^1(\theta_C) = \theta_C$, $\delta_C^m(\theta_C) = \delta_C^m$ $\forall \theta_C$, $\theta_C \sim U[0, \theta_C^{\max}]$ where $\theta_C^{\max} > \bar{\theta}_C^1 = n_C^1$, and $\delta_C^d \in (\delta_C^m, \delta_C^m + (1 - \bar{\beta}) \cdot (\theta_C^{\max} - n_C^1))$.

Here the deterrence equilibrium exists in the baseline model since $\beta = 1 - \frac{\delta_C^d - \delta_C^m}{\theta_C^{\max} - n_C^1} \geq \bar{\beta}$, but it unravels with costly signaling since $\delta_C^d > \delta_C^m$. The following strategy profile in which peaceful challengers separate themselves satisfies universal divinity.

- **(Challenger)** Always transgress, and $c(\theta_C) = 1_{\theta_C \leq \bar{\theta}_C^2} \cdot \delta_C^m$.

- **(Defender)** $\alpha(c) = 1_{c < \delta_C^m}$ (respond with war i.f.f. the costly signal is strictly less than $\delta_C^m$).

The equilibrium is depicted in Figure 6. In it, peaceful and appeasable challengers $\theta_C \leq \bar{\theta}_C^2$ transgress and send a costly signal barely high enough ($c = \delta_C^m$) to distinguish themselves from unappeasably belligerent ones, and are therefore always allowed to transgress. Unappeasably belligerent challengers $\theta_C > \bar{\theta}_C^2$ transgress without signaling, and always trigger a war.

---

21The figure uses $\delta_C^m = .2 < \delta_C^d = .3$, $\theta^{\max} = 1$, and $n_C^1 = .5$. 

---
Figure 6: Equilibrium with Costly Signaling

Proof of Proposition C.8 Let \( \tilde{\delta} = \max_{\theta_C \geq \tilde{\theta}_C} \{ \delta^m_C (\theta_C) \} \). Now consider a strategy profile-belief pair \((\sigma, \mu)\) (where \(\mu(T|c)\) denotes the defender’s interim-belief that \(\theta_C\) is in the set \(T\) conditional on observing the transgression and costly signal \(c\)) that satisfies (i) the conditions of the proposition, (ii) \(\mu \left( \arg\max_{\theta_C \geq \tilde{\theta}_C} \{ \delta^m_C (\theta_C) \} \mid c \right) = 1 \ \forall c < \tilde{\delta}\), and (iii) \(\alpha(c)\) is a best response to \(\mu(\cdot|c)\) \(\forall c \geq \tilde{\delta}\). We argue that any pair \((\sigma, \mu)\) is a universally divine equilibrium.

We first introduce some additional useful notation. Let \(\Pi(\theta_C, \alpha)\) denote the net benefit (excluding signaling costs) of a challenger type \(\theta_C\) deviating from her strategy in \(\sigma\) to transgressing and sending a signal that induces the defender to respond with war with probability \(\alpha\) (so \(\Pi(\theta_C, \alpha(c)) - c\) is the total net benefit of deviating to transgressing with a signal \(c\)). Then

\[
\Pi(\theta_C, \alpha) = \begin{cases} 
(1 - \alpha) \delta^d_C - \alpha (n_C^1 - w^1_C (\theta_C)) & \text{for } \theta_C \leq \tilde{\theta}_C^2 \\
(1 - \alpha) (w^2_C (\theta_C) - n_C^1) - \alpha (n_C^1 - w^1_C (\theta_C)) & \text{for } \theta_C \in [\tilde{\theta}_C^2, \tilde{\theta}_C^1] \\
(1 - \alpha) \delta^m_C (\theta_C) & \text{for } \theta_C \geq \tilde{\theta}_C^1
\end{cases}
\]
It is easily verified that (i) $\Pi(\theta_C, \alpha)$ is strictly decreasing in $\alpha \forall \theta_C$, and (ii) $\forall \alpha, \Pi(\theta_C, \alpha)$ is weakly increasing in $\theta_C$ over $\theta_C \leq \bar{\theta}_C^2$ and strictly increasing in $\theta_C$ over $\theta_C \in [\bar{\theta}_C^1, \bar{\theta}_C^2]$. This furthermore implies that $\Pi(\bar{\theta}_C^1, \alpha) > \Pi(\theta_C, \alpha) \forall \theta_C < \bar{\theta}_C^2$.

The proof now proceeds in two steps. First, we argue that $(\sigma, \mu)$ is a Perfect Bayesian Equilibrium. Second, we argue that $(\sigma, \mu)$ survives the application of the universal divinity refinement, and is hence a universally divine equilibrium. To see that $(\sigma, \mu)$ is a PBE, observe that since $\bar{\theta}_C^2 < \bar{\theta}_C^1$, the deterrence equilibrium exists absent the potential for costly signaling, strategies and beliefs are as in the deterrence equilibrium on equilibrium path, and the defender’s actions and beliefs satisfy sequential rationality by construction off-path. It remains only to show that no challenger type wishes to deviate to transgressing with a costly signal $c > 0$. Signals $0 < c < \delta$ are clearly strictly worse than transgressing with no signal since they are costly but yield no increase in the probability the transgression will be allowed. Signals $c \geq \delta$ are unprofitable since by previous observations about $\Pi(\theta_C, \alpha(c))$ we have $\Pi(\theta_C, \alpha(c)) \leq \max_{\theta_C \geq \bar{\theta}_C^1} \{\Pi(\theta_C, \alpha(c))\} = (1 - \alpha(c)) \cdot \max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\} \leq \delta \leq c$.

Finally, to see $(\sigma, \mu)$ survives universal divinity, we argue that it survives the iterative application of the NWBR signaling criterion, which is a strengthening of D2 (Fudenberg and Tirole 1991, pp. 454). First observe that all $\alpha \in [0, 1]$ are in the set of defender mixed best responses to the original type space $\Theta$. Second, observe that for $c \geq \delta$, $\Pi(\theta_C, \alpha) - c \leq 0 \forall (\theta_C, \alpha)$; since no type can be made strictly better off sending such signals for any value of $\alpha$, no type may be eliminated through the application of NWBR; the associated beliefs in $(\sigma, \mu)$ therefore satisfy universal divinity.

Last consider $c \in (0, \delta)$. For any pair $\hat{\theta}_C \in \arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$ and $\theta_C \notin \arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$ we have that $\Pi\left(\hat{\theta}_C, 0\right) - c = \delta - c > 0$, so there exists a mixed best response ($\alpha = 0$) that would make type $\hat{\theta}_C$ strictly benefit from the deviation. In addition, $\Pi\left(\hat{\theta}_C, \alpha\right) > \Pi(\theta_C, \alpha) \forall \alpha > 0$. Hence any $\alpha$ that would make $\theta_C$ indifferent to sending $c$ (which must be $> 0$) would make $\hat{\theta}_C$ strictly prefer to send $c$, and $\theta_C$ may be pruned. Therefore beliefs $\mu\left(\arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\} \mid c\right) = 1$ result from the first application of NWBR, and moreover yield a unique mixed best response $\alpha = 1$ that is exactly
the defender’s strategy in our profile. Since no type profits from deviating in this profile, further applications of NWBR cannot further restrict beliefs, and the profile satisfies universal divinity.

**Proof of Proposition C.9** Let $\tilde{\delta} = \max_{\theta_C \geq \bar{\theta}_C^2} \{\delta_{\theta_C}(\theta_C)\} < \delta_C^d$. Now suppose the conditions hold and consider a universally divine equilibrium. Transgressing and not signaling ($c = 0$) strictly dominates not transgressing for all $\theta_C > \bar{\theta}_C^1$, so all such types must transgress. Next consider types $\theta_C \leq \bar{\theta}_C^1 < \bar{\theta}_C^2$. We argue that for any $\hat{c} \in (\delta, \delta_C^d)$, universal divinity implies $\alpha(\hat{c}) = 0$ (the defender will always allow the transgression), which implies that transgressing and sending $\hat{c}$ yields a net benefit of $\delta_C^d - \hat{c} > 0$ over not transgressing for such types, which implies that they also must transgress in equilibrium.

Observe that for types $\theta_C \geq \bar{\theta}_C^2$, transgressing and sending the costly signal $\hat{c}$ is strictly dominated by transgressing and sending no signal; the *best* payoff the former could yield is $w_C^2(\theta_C) - \hat{c}$ (if it results in the transgression always being allowed) while the *worst* payoff the latter could yield is $w_C^1(\theta_C)$ (if it results in the transgression never being allowed), and $w_C^1(\theta_C) > w_C^2(\theta_C) - \hat{c} \iff \hat{c} > \delta_{\theta_C}(\theta_C)$. Thus, if $\hat{c}$ is on equilibrium path then only challengers types $\theta_C < \bar{\theta}_C^2$ may be sending $\hat{c}$. If $\hat{c}$ is off equilibrium path then (the first iteration of) universal divinity eliminates types $\theta_C \geq \bar{\theta}_C^2$ from the defender’s beliefs since they would benefit from the deviation for no defender responses, while types $\theta_C < \bar{\theta}_C^1$ would benefit for some defender responses. In either case this is sufficient to imply that the defender’s best response must be to allow the transgression.