Fear, Appeasement, and the Effectiveness of Deterrence

Ron Gurantz, Air War College
Alexander V. Hirsch, California Institute of Technology

Governments often fear the future intentions of their adversaries. In this article 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 US 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 have explored many such mechanisms, including "threats that leave something to chance" (Nalebuff 1986; Powell 1987; Schelling 1966), 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 article, 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 his adversary's future intentions affects his 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 when bargaining 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

Alexander V. Hirsch (avhirsch@hss.caltech.edu) is a professor of political science at the California Institute of Technology, Pasadena, CA 91125. Ron M. Gurantz (ron.gurantz@us.af.mil) is an assistant professor of strategy at the Air War College, Maxwell Air Force Base, Montgomery, AL 36112.

An online appendix with supplementary material is available at http://dx.doi.org/10.1086/691054.

1. See Carter and Perry (1999); KCNA News Agency (2003); Wit, Poneman, and Gallucci (2004).

The Journal of Politics, volume 79, number 3. Published online May 18, 2017. http://dx.doi.org/10.1086/691054
 0022-3816/2017/7903-0020$10.00

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 that she is unappeasably belligerent.[2] 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 first 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."

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 in equilibrium, 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 US 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" (US Department of State 1986, 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 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 actually choose to do so! Whether this will be 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.

2. Throughout the article we refer to the defender as "he" and the challenger as "she."

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, a literature on bargaining under complete information with endogenous power shifts already answers 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 how 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 that control 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 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 article 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 crisis 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 include "probabilistic threats" that increase the chance that 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 (Fearon 1994; Schelling 1966; 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 that 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

3. Our model is closer to the latter two in that the defender refuses to concede because of consequences beyond the immediate stakes.

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 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 data set 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 1988; Huth and Russett 1984).

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 crisis takes place to the crisis itself.[5] For example, the strength of the US's interests on the broader European continent are used as a proxy for the US's stakes in a crisis over West 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 (Hirshleifer 1991; Powell 1996).

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 uncertainty about the challenger, 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 McManus (2015) 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

4. These included demands in northern Iran, the Turkish Straits, Trieste, Berlin, and Tripolitania.

5. Huth (1988) and Huth and Russett (1984) 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.

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-shift literature studies states' strategic military investments (Debs and Monteiro 2014; Slantchev 2011). 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 analyzes what happens when states fear that their adversaries are unappeasably belligerent. Several works analyze how this possibility can induce a "spiral" of fear that causes peace to unravel (Baliga and Sjostrom 2009; Chassang and i Miquel 2010). Alternatively, Acharya and Grillo (2015) analyze 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 $x^1 \in \{a, \varnothing\}$ 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 ($x^1 = \varnothing$), 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 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

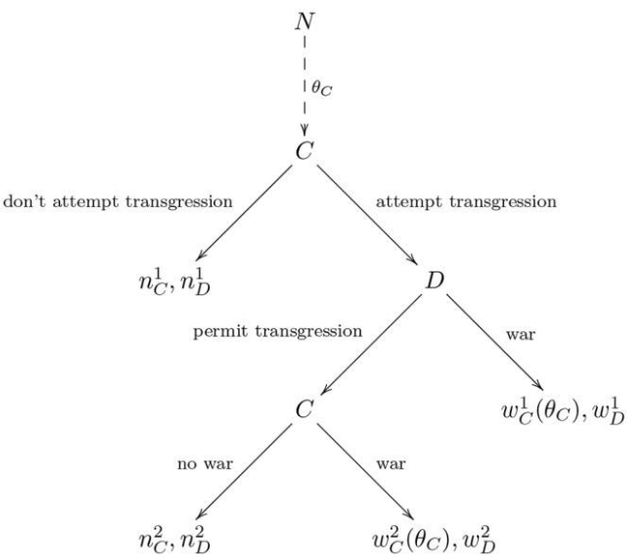

Figure 1. Model

capture that preference we denote the defender's payoff as n_D^t if the game ends with peace in period t and w_D^t if the game ends with war in period t, and assume that:

1. Allowing the transgression makes him worse off in both peace ($n_D^2 < n_D^1$) and war ($w_D^2 < w_D^1$).
2. Allowing the transgression is strictly better than responding with war if the challenger will subsequently choose peace ($n_D^2 > w_D^1$).

Given these assumptions, the defender's optimal response to a transgression depends on his interim assessment of the probability β 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 $w_D^1 - w_D^2 > 0$ 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 $n_D^2 - w_D^1$ of a preventable war. It is easily shown that the defender will prefer to respond to the transgression with war whenever β exceeds a threshold $\bar{\beta}$, where

$$\bar{\beta} = \frac{n_D^2 - w_D^1}{(n_D^2 - w_D^1) + (w_D^1 - w_D^2)} \in (0, 1).$$

Crucially, $\bar{\beta} < 1$—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_C^t for peace in each period, but her payoffs from war $w_C^t(\theta_C)$ depend on a type $\theta_C \in \Theta \subset \mathbb{R}$ drawn by "nature" at the start of the game, where Θ is a closed interval. The defender's prior beliefs over the challenger's type are described by an atomless distribution with full support over Θ and CDF $F(\theta_C)$, and the challenger's payoffs n_C^t and $w_C^t(\theta_C)$ satisfy the following:

1. Successfully transgressing has a direct value if the game ends in peace ($n_C^2 - n_C^1 > 0$) and a military value if the game ends in war ($w_C^2(\theta_C) - w_C^1(\theta_C) > 0 \; \forall \; \theta_C \in \Theta$).
2. In each period t the challenger's war payoff $w_C^t(\theta_C)$ is continuous and strictly increasing in θ_C. In addition, there exists a unique challenger type $\bar{\theta}_C^t$ strictly interior to Θ that is indifferent between peace and war (i.e., $w_C^t(\bar{\theta}_C^t) = n_C^t$).

Our assumptions imply the following. First, all challenger types intrinsically value the transgression, that is, even absent a war. Second, a challenger's type θ_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 ($\theta_C < \bar{\theta}_C^t$) and war to peace ($\theta_C > \bar{\theta}_C^t$). Thus, there is always the possibility (however unlikely) that she prefers war to the status quo ($w_C^1(\theta_C) \geq n_C^1 \Leftrightarrow \theta_C \geq \bar{\theta}_C^1$). 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 $\theta_C \geq \bar{\theta}_C^2$ 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 appendix, available online.

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

- There exists a *no deterrence equilibrium*, in which the challenger always transgresses, and she is always permitted to do so, if and only if $\bar{\beta} \geq P(\theta_C \geq \bar{\theta}_C^2)$.
- 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 $P(\theta_1 < \bar{\theta}_C^1)$, if and only if $\bar{\beta} \leq P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$.

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 $P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$ 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, that is, $\bar{\theta}_C^2 \leq \bar{\theta}_C^1 \Leftrightarrow P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1) = 1$, 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 ($n_D^1 - w_D^1$), a low cost of allowing the transgression in both direct ($n_D^1 - n_D^2$) and military ($w_D^1 - w_D^2$) terms, and/or a sufficiently low probability that the challenger is belligerent $P(\theta_C \geq \bar{\theta}_C^t)$ in both periods.[6] 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.[7]

Figure 2 depicts the equilibrium correspondence when the defender's belief threshold $\bar{\beta}$ for responding with war is very

6. The 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.

7. Baliga 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.

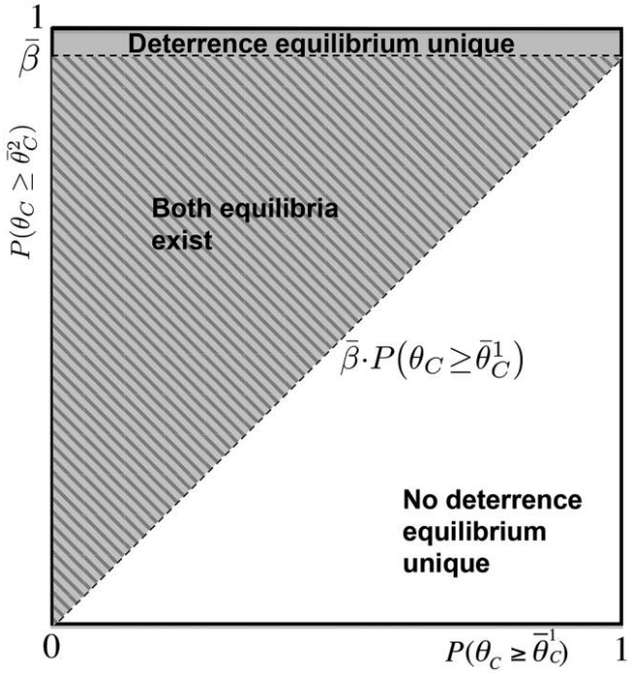

Figure 2. Equilibria

high (the area where corollary 1 holds is not indicated but is the region above a 45 degree line drawn from the origin). The defender's prior $P(\theta_C \geq \bar{\theta}_C^1)$ that the challenger prefers war to the status quo is on the x-axis, while the prior $P(\theta_C \geq \bar{\theta}_C^2)$ 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 $\bar{\beta}$ were made arbitrarily high.

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 his 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 $P(\theta_C < \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$ that appeasement is effective is below

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

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 ($\bar{\theta}_C^2 \leq \bar{\theta}_C^1$), but the challenger is actually peaceful ($\theta_C \leq \bar{\theta}_C^2$). 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 ($\theta_C \geq \bar{\theta}_C^1$).

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 $P(\theta_C < \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1)$ is sufficiently low. The stated condition ensures that the expected cost of fighting avoidable wars against appeasable challengers ($\theta_C \in [\bar{\theta}_C^1, \bar{\theta}_C^2]$) is outweighed by the expected benefits of deterring peaceful challengers ($\theta_C < \bar{\theta}_C^1$). It clearly holds when appeasement is impossible ($\bar{\theta}_C^2 < \bar{\theta}_C^1$), 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 θ_C or to remain ignorant. When the condition in proposition 2.2 holds, 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 his 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 chal-

lenger, 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 $w_C^2(\theta_C) - w_C^1(\theta_C)$, which is the challenger's gain in the event of war, and a direct value $n_C^2 - n_C^1$, which is her gain in the event of peace. We henceforth denote these as $\delta_C^m(\theta_C)$ and δ_C^d, 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 $\delta_C^m(\bar{\theta}_C^1) \geq \delta_C^d$.

A sufficient condition for the deterrence equilibrium to exist is thus that the military value of the transgression $\delta_C^m(\cdot)$ exceed its direct value δ_C^d to a challenger of type $\bar{\theta}_C^1$ who is initially indifferent between peace and war. For such a challenger type, $\delta_C^m(\bar{\theta}_C^1) \geq \delta_C^m$ 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., $P(\theta_C \geq \bar{\theta}_C^2 | \theta_C \geq \bar{\theta}_C^1) = 1$), 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 $P(\theta_C < \bar{\theta}_C^1)$ 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 δ_C^m 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 $\delta_C^m - \delta_C^d$.

The model thus predicts that the probability of deterrence is increasing in the difference $\delta_C^m - \delta_C^d$ between the military and direct value of the transgression to the challenger. The intuition is similar to corollary 3; the greater is $\delta_C^m - \delta_C^d$, 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.

8. See the appendix for details about figures 3 and 4.

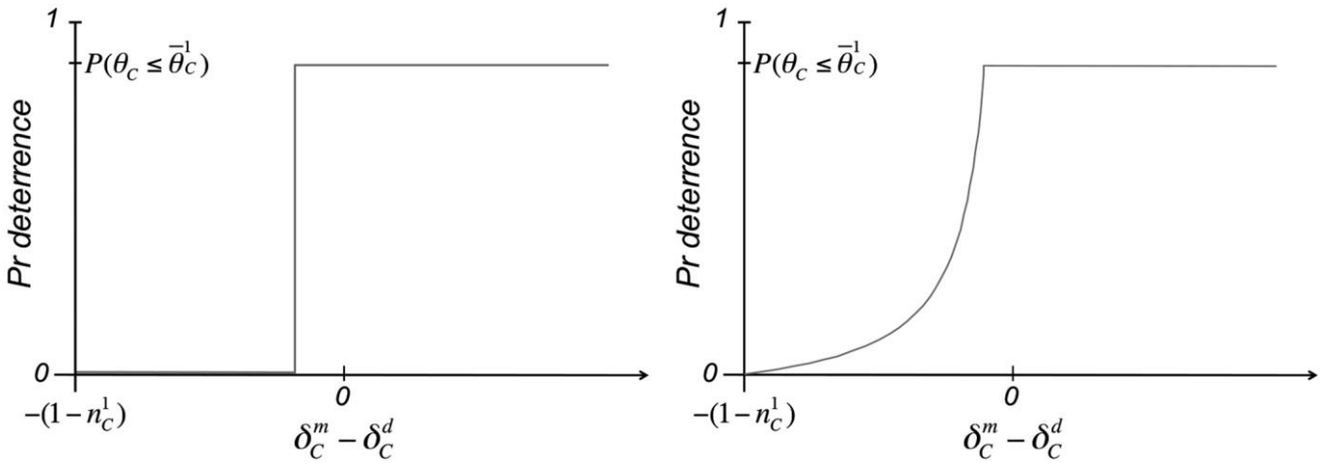

Figure 3. Probability of deterrence as function of $\delta_C^m - \delta_C^d$

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 nonmonotonic effect. The probability of deterrence first decreases due to the logic 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 trangression exceeds its direct value (even if both are minor), and the challenger does not in fact affirmatively 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, 123–24). The Soviet Union had some intrinsic interest in protecting its trade through the Straits, though the United States 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 United States had no obvious economic or political interests other than a small trade in tobacco, machinery, and vehicles (Kuniholm 1980, 65–66). The United States also anticipated that any war with the Soviet Union would be enormously costly despite the US nuclear monopoly, involving Soviet ground offensives across Europe and Asia and requiring major US ground operations (Ross 1996, 12–19, 31).

Turkey's primary value to the United States and the Soviet Union 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 United States and its allies (Leffler 1985, 814–15). 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 US 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.[9] In fact, President Truman was prepared to infer far-reaching ambitions from a Soviet attack in the face of a US 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, 192). Ultimately, it was the Soviet aversion

9. While 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).

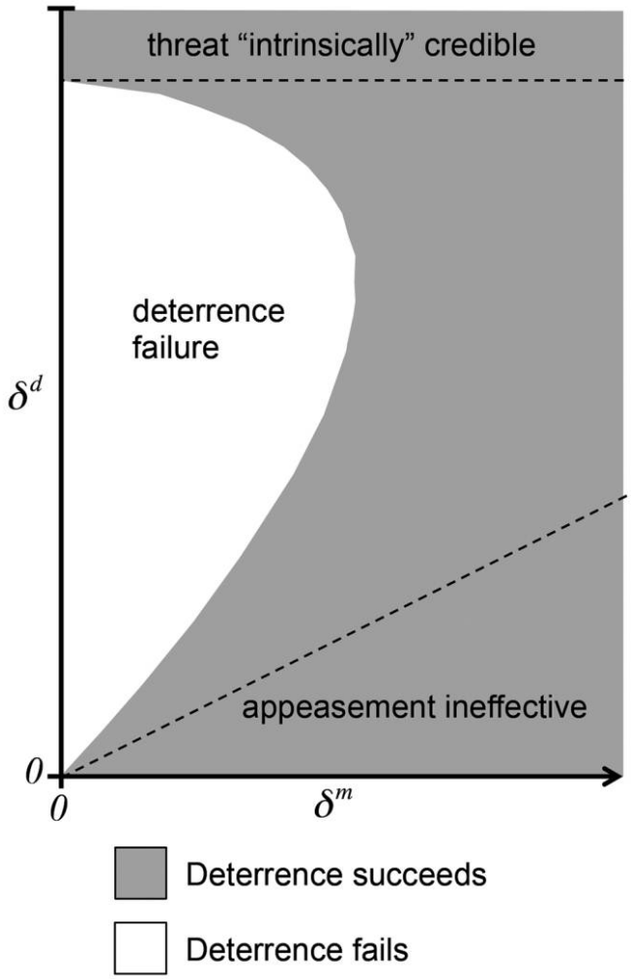

Figure 4. Probability of deterrence when challenger's gains equal to defender's costs.

to war that led them to back down in the face of a credible US threat (Mark 2005).[10]

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, 138–39; 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, 648–49). 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]

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 coethnics, and Germany justified its policy as one of national unification. The British

10. See the appendix for an expanded version of this case study.

11. See the appendix for discussion of the literature on Finland's decision to fight.

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 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 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 the share of the landmass she holds, so that war is attractive if her 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.

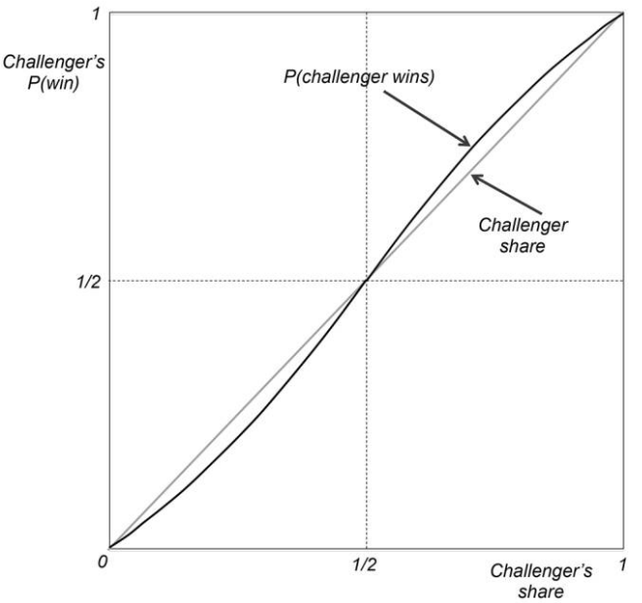

Figure 5. Probability that challenger wins in salami tactics extension

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 en-

dogenous "demand" of how far to advance. The game then proceeds as in the baseline model.

In the 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 conceding 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 (see, e.g., Lewis and Schultz 2003). In the 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 that 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 both 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 his intrinsic willingness to fight or "resolve," in order to shift attention away from reputation to fear. However, in the 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; for example, an ex ante appealing military adventure like the Soviet invasion of Afghanistan also risks a protracted and costly conflict.[12] In the 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. It also does not change the 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.[13]

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 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

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.

and eliminate the fear that sustains deterrence.[14] 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.[15] In these equilibria 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 article 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 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 United States 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 signaling, 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." For

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.

example, Kydd (1997) 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? 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.

ACKNOWLEDGMENTS

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.

REFERENCES

Acharya, Avidit, and Edoardo Grillo. 2015. "War with Crazy Types." *Political Science Research and Methods* 3 (2): 281–307.

Alt, James, Randall Calvert, and Brian Humes. 1988. "Reputation and Hegemonic Stability: A Game-Theoretic Analysis." *American Political Science Review* 82 (2): 445–66.

Baliga, Sandeep, and Tomas Sjostrom. 2008. "Strategic Ambiguity and Arms Proliferation." *Journal of Political Economy* 116 (6): 1023–57.

Baliga, Sandeep, and Tomas Sjostrom. 2009. "Conflict Games with Payoff Uncertainty." Unpublished manuscript.

Banks, Jeffrey S., and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica* 55 (3): 647–61.

Carter, Ashton B., and William J. Perry. 1999. *Preventive Defense: A New Security Strategy for America*. Washington, DC: Brookings Institution Press.

Carter, Ashton B., and William J. Perry. 2006. "If Necessary, Strike and Destroy." *Washington Post*, June 22.

Chassang, Sylvain, and Gerard Padro i Miquel. 2010. "Conflict and Deterrence under Strategic Risk." *Quarterly Journal of Economics* 125 (4): 1821–58.

Chiozza, Giacomo, and H. E. Goemans. 2004. "International Conflict and the Tenure of Leaders: Is War Still 'Ex Post' Inefficient?" *American Journal of Political Science* 48 (3): 604–19.

Danilovic, Vesna. 2001. "The Sources of Threat Credibility in Extended Deterrence." *Journal of Conflict Resolution* 45 (3): 341–69.

Danilovic, Vesna. 2002. *When the Stakes Are High: Deterrence and Conflict among Major Powers*. Ann Arbor: University of Michigan Press.

Debs, Alexandre, and Nuno P. Monteiro. 2014. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68 (1): 1–31.

Enthoven, Alain C. 1975. "U.S. Forces in Europe: How Many? Doing What?" *Foreign Affairs* 53 (3): 513–32.

Fearon, James. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88 (3): 577–92.

Fearon, James. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.

Fearon, James. 1996. "Bargaining over Objects That Influence Future Bargaining Power." Unpublished manuscript, Chicago, IL.

Fearon, James. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41 (1): 68–90.

Fey, Mark, and Kristopher W. Ramsay. 2011. "Uncertainty and Incentives in Crisis Bargaining." *American Journal of Political Science* 55 (1): 149–69.

Hirshleifer, Jack. 1991. "Appeasement: Can It Work?" *American Economic Review* 91 (2): 342–46.

Hopf, Ted. 1994. *Peripheral Visions: Deterrence Theory and American Foreign Policy in the Third World, 1965–1990*. Ann Arbor: University of Michigan Press.

Howard, Michael. 1984. *The Causes of Wars and Other Essays*. Cambridge, MA: Harvard University Press.

Huth, Paul. 1988. "Extended Deterrence and the Outbreak of War." *American Political Science Review* 82 (2): 423–43.

Huth, Paul. 1999. "Deterrence and International Conflict: Empirical Findings and Theoretical Debates." *Annual Review of Political Science* 2:25–48.

Huth, Paul, and Bruce Russett. 1984. "What Makes Deterrence Work? Cases from 1900 to 1980." *World Politics* 36 (4): 496–526.

Jackson, Matthew, and Massimo Morelli. 2007. "Political Bias and War." *American Economic Review* 97 (4): 1353–73.

Jakobson, Max. 1961. *The Diplomacy of the Winter War*. Cambridge, MA: Harvard University Press.

Kalinovsky, Artemy. 2009. "Decision-Making and the Soviet War in Afghanistan." *Journal of Cold War Studies* 11 (4): 46–73.

KCNA News Agency. 2003. "North Korea Warns 'Total War' if USA Attacks 'Peaceful' Plant." BBC Summary of World Broadcasts, February 6, 2003. Lexis-Nexis Academic: News.

Kuniholm, Bruce Robellet. 1980. *The Origins of the Cold War in the Near East*. Princeton, NJ: Princeton University Press.

Kydd, Andrew. 1997. "Game Theory and the Spiral Model." *World Politics* 49 (3): 371–400.

Kydd, Andrew H., and Roseanne W. McManus. 2017. "Threats and Assurances in Crisis Bargaining." *Journal of Conflict Resolution* 61 (2): 325–48.

Leffler, Melvyn P. 1985. "Strategy, Diplomacy, and the Cold War: The United States, Turkey, and NATO, 1945–1952." *Journal of American History* 71 (4): 807–25.

Lewis, Jeffrey B., and Kenneth A. Schultz. 2003. "Revealing Preferences: Empirical Estimation of a Crisis Bargaining Game with Incomplete Information." *Political Analysis* 11:345–67.

Luard, Evan. 1986. *War in International Society*. London: Routledge.

Mark, Eduard. 2005. "The Turkish War Scare of 1946." In Melvyn P. Leffler and David S. Painter, eds., *Origins of the Cold War: An International History*. 2nd ed. New York: Routledge.

Mercer, Jonathan. 1996. *Reputation and International Politics*. Ithaca, NY: Cornell University Press.

Mills, Walter, ed. 1951. *The Forrestal Diaries*. New York: Viking.

Nalebuff, Barry. 1986. "Brinkmanship and Nuclear Deterrence: The Neutrality of Escalation." *Conflict Management and Peace Science* 9:19–30.

Overy, Richard, and Andrea Wheatcroft. 1989. *The Road to War*. New York: Random House.

Powell, Robert. 1987. "Crisis Bargaining, Escalation, and MAD." *American Political Science Review* 81 (3): 717–35.

Powell, Robert. 1996. "Uncertainty, Shifting Power and Appeasement." *American Political Science Review* 90 (4): 749–64.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169–203.

Press, Daryl G. 2005. *Calculating Credibility: How Leaders Assess Military Threats*. Ithaca, NY: Cornell University Press.

Pritchard, Charles L. 2006. "No, Don't Blow It Up." *Washington Post*, June 23.

Ross, Steven T. 1996. *American War Plans, 1945–1950*. Portland, OR: Cass.

Schelling, Thomas. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Schwarz, Michael, and Konstantin Sonin. 2008. "A Theory of Brinkmanship, Conflicts and Commitments." *Journal of Law, Economics, and Organization* 24 (1): 163–83.

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64 (4): 627–60.

Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. New York: Cambridge University Press.

Snyder, Jack, and Erica D. Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105 (3): 1–20.

Soman, Appu Kuttan. 2000. *Double-Edged Sword: Nuclear Diplomacy in Unequal Conflicts: The United States and China, 1950–1958*. Westport, CT: Praeger.

Spring, D. W. 1986. "The Soviet Decision for War against Finland, 30 November 1939." *Soviet Studies* 38 (2): 207–26.

Trachtenberg, Marc. 1989. "Strategic Thought in America, 1952–1966." *Political Science Quarterly* 104 (2): 301–34.

Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis." *Security Studies* 21 (1): 3–42.

US Department of State. 1986. *Foreign Relations of the United States, 1952–1954: Germany and Austria*. Vol. 7, pt. 2. Washington, DC: United States Government Printing Office.

Van Evera, Stephen. 1999. *Causes of War*. Ithaca, NY: Cornell University Press.

Weinberg, Gerhard. 1980. *The Foreign Policy of Hitler's Germany: Starting World War II, 1937–1939*. Chicago: University of Chicago Press.

Wit, Joel S., Daniel B. Poneman, and Robert L. Gallucci. 2004. *Going Critical: The First North Korean Nuclear Crisis*. Washington, DC: Brookings Institution Press.

Zagare, Frank C. 2004. "Reconciling Rationality with Deterrence: A Reexamination of the Logical Foundations of Deterrence Theory." *Journal of Theoretical Politics* 16 (2): 107–41.