Mechanism Design Goes to Nuclear War

Brenton Kenkel[*] Peter Schram[†]

August 30, 2024

Abstract

Crises in the nuclear era are commonly framed as "contests of nerves," where actors compete by raising the background risk of a nuclear exchange until one side lacks the resolve to continue and backs down. But this framing may be too reductive: in practice, actors deploy a range of coercive capabilities that both alter the risk of escalation and shape political outcomes. How do these limited coercive capabilities shape outcomes in nuclear crises? We analyze the "contests of nerves" framework, finding broadly that more resolved actors will take greater escalation risks and perform better in conflict. We also analyze a "contests of capabilities" framework, showing that when a state's resolve also shapes its willingness to compete at lower levels, more resolved actors may engage in less risky or less decisive measures. We use a game-free methodology to study how the underlying military fundamentals affect crisis behavior in settings with autonomous escalation risk across a wide variety of bargaining games.

[*] Associate Professor, Department of Political Science, Vanderbilt University. brenton.kenkel@vanderbilt.edu

[†] Assistant Professor, Department of Political Science, Vanderbilt University. peter.schram@vanderbilt.edu

International politics rarely goes according to plan. History is replete with accidents, idiosyncrasies, and unintended consequences altering the scope and context of crises, leading to instability and escalation. Perhaps nowhere is this instability better acknowledged and integrated than in the scholarship on nuclear deterrence. Because deliberate first use of strategic nuclear weapons against a capable nuclear armed opponent is tantamount to suicide, states cannot easily use these weapons to directly threaten their opponents or as a direct means of deterrence (Schelling 1966). Instead, nuclear weapons shape politics through the possibility of unintended use: in the nuclear era, states enter into crises or limited engagements knowing that their actions and their adversaries' actions carry the risk of missteps, miscalculations, or inadvertent escalation (Brodie 1966; Schelling 1966, 1980; Powell 1989, 2015; Posen 2014). To describe these new strategic dynamics, scholars have classified crises and limited engagements in the nuclear era as "contests of nerves" or exercises in "brinkmanship," where states strategically use this background escalation risk over the course of the engagement (Schelling 1966, 1980; Powell 1989, 2015). In these conflicts that leave something to chance, adversaries now seemingly stand eyeball to eyeball, hoping that their rivals are less willing to run risks of a general war and will blink first before disaster strikes.

For decades, the scholarship on nuclear brinkmanship has examined and formalized these settings (Snyder 1965; Jervis 1976; Powell 1989, 1990). But, much of what occurs in the international politics of the nuclear era does not resemble these stylized "contests of nerves." Consider the West's support for Ukraine in its war against Russia. By keeping Russia in a continued state of war, Russia remains on edge, thus opening the possibility that a human error or a faulty missile detection system results in an inadvertent escalation (Paul et al. 1990; Sagan 1994; Perrow 2011; Posen 2014). And by running military supply chains through NATO countries, through mistake or malice, a Russian General may one day strike within a NATO state, thus raising the unintended risk of a NATO-Russia war (Posen 2014). In short, support to Ukraine comes with risks: at any time, this proxy war could evolve into a more general or even nuclear war between NATO countries and Russia. But despite these (even nuclear) escalation risks, the West is not arming Ukraine because it is more resolved than Russia or because the West thinks that Russia will, at some point, deem the escalation risks too pressing and retreat. Supplying Ukraine is not "rocking the boat" or playing a game of "chicken" with Russia. Rather, through arming, Ukraine persists, continues to challenge the Russian militarily, and strains the inner politics of the Russian state. From the perspective of Western leaders, the political benefits of arming Ukraine have been enough to justify continued support, despite the background risks of greater war.

Western support for Ukraine is just one example highlighting an important shortcoming in our understanding of limited war and escalation. For decades, scholars have carefully studied "contests of nerves," where actors generate risks within crises, hoping that their opponents are less resolved and will not accept the risk of a catastrophic escalation (Snyder 1965; Jervis 1976; Schelling 1980, 1966; Powell 1989, 1990). While the contests of nerves framing has allowed scholars to make valuable progress in the theoretical study of nuclear deterrence, this framework is quite reductive—in practice, in the crises of the nuclear era, actors use their conventional or low-level capabilities to generate escalation risks *and* shape the politics on the ground. As examples, the Berlin Airlift (1948), the Soviet Union's response to the Hungarian Revolution (1959), and the US's support to Afghan Mujahideen during the Russian invasion of Afghanistan (1979-1989) all carried some form of escalation risk,[1] but the conventional or irregular capabilities were the pivotal factors shaping how the respective Cold War crises played out. In these crises and others, bargaining outcomes depend not only on the adversaries' resolve, but also on their capabilities.

To establish a more general theoretical framing and results, we proceed in two parts. First, we formally define the "contests of nerves" theoretical framing. In contests of nerves, actors engage in a deterrence or bargaining game while taking some action that raises or lowers the

[1] While there was no risk of a nuclear exchange in the Berlin Airlift case and close to zero risk of nuclear escalation in the Afghanistan case, both crises carried the risk of accidents or mishaps necessitating an escalation to a conventional engagement between Soviet and Western forces.

likelihood of a nuclear exchange—for example, by exiting or not exiting a crisis. This framing matches seminal models of nuclear deterrence like Nalebuff (1986) or Powell (1989). We then demonstrate that the equilibria to these contests of nerves share a similar underlying pattern: when an actor is more resolved, this actor will take riskier actions, and will, in expectation, end up with a greater payoff.[2] Importantly, to make these sweeping claims, we conduct a game-free analysis of these contests of nerves, along the along the lines of previous mechanism design research (Banks 1990; Fey and Ramsay 2011; Akçay et al. 2012; Spaniel 2020; Liu 2021). Doing so allows us to verify that, regardless of how bargaining, deterrence, or escalation happens in these models and regardless of how actors play the game (i.e., whether they play truthfully, bluff, signal, or posture), these contests of nerves will always have the feature of more resolved actors behaving in riskier ways and attaining, in expectation, better outcomes.

Second, we analyze a novel and more general theoretical framing for crises with stochastic escalation risk: "contests of capabilities." Formally, in these contests of capabilities, actors engage in a deterrence or bargaining game while selecting costly actions—like arming Ukraine or executing the Berlin Airlift—that shift political outcomes and generate escalation risks.[3] In these contests of capabilities, we assume that a state's resolve influences their willingness or abilities to conduct war at multiple levels. As one justification for this assumption, if a country possesses a technological capacity that would allow it to do well in a nuclear war (strong command and control, sophisticated technology for missile delivery, powerful tools for left-of-launch attacks, etc.), then that country also plausibly has the tools to do well in a conventional war. In these contests of capabilities, actors with greater resolve may be less likely to take the actions that risk such a war and may even do worse within the game.

[2] We model resolve as an actor's private willingness to go to the escalated-war option, as approximated by that actor's utility from this option. Note that while Powell (1990, 42-43) critiques some past discussions of resolve in verbal theories, the result we describe here is also found in every model with probabilistic escalation in Powell (1990).

[3] In our formalization, contests of resolve are a special case of contests of capabilities, in which the direct costs of the limited actions do not vary with a player's resolve.

Once more, rather than presenting a single model and presenting a single model that has this feature, we identify when these patterns arise in a wide range of possible contests of capabilities.

A key driver of our results is that a state's resolve—its private willingness to risk a nuclear strike—might also affect its effectiveness in using alternative policy choices, for better or worse. How a state's resolve affects its payoffs from alternate policy choices is critical: the private signal a state receives about its willingness to risk nuclear war also affects that state's willingness to engage in costly, low-level policy instruments that may present their own risk of escalation to war. For example, suppose that US military communication networks were compromised or that B-52 or B-2 bombers contained design flaws or vulnerabilities. Because this would undermine both conventional and nuclear capabilities, in this example, the US would be less resolved to risk nuclear or conventional engagements.

On the other hand, sometimes actors that are more resolved to engage at one level may be less inclined to engage at others. On the other hand, actors like North Korea may be willing to bid up nuclear escalation risks, but their ability to fight an effective conventional war or their ability to conduct special forces operations, limited airstrikes, third-party support to rebel groups, or sanctions is quite limited. Here, states like North Korea may be more resolved (in the nuclear sense) but less willing to engage in other forms of costly, coercive politics. The question of whether a state's resolve is associated with a greater or lower willingness to conduct lower forms of conflict is an empirical one, and its answer varies across cases and contexts—but these linkages undoubtedly exist, and their effects on the outcomes of crisis bargaining have not been systematically examined. We demonstrate that this relationship critically alters how an actor's resolve shapes their conflict behavior and their outcomes.

In contests of capabilities, the effect of resolve on the risk of inadvertent escalation is determined by two key features of the underlying military and strategic setting. On one side of the equation is the relationship between resolve and the marginal cost of the limited policy instruments that generate risk: the change in the additional direct cost required to go from low to high limited capability as the state's resolve increases. When these marginal costs decrease with resolve, or even when they only slightly increase, it becomes easier to support an equilibrium in which more resolved types run greater escalation risks, as in a traditional contest of nerves. The other key factor is the relationship between limited conflict instruments and the risk of escalation to a broader war. If the risk curve is flat, then a state's choice among limited policy instruments will be mainly determined by their direct costs. Resolve becomes the determinant factor when the risk curve is steeper, with small increases in the limited capability leading to much larger risks of inadverent escalation.

When a full-scale war would be catastrophic, such as a nuclear exchange with mutually assured destruction, inadvertent escalation might be the only path to war in equilibrium. In other contexts, however, states with high enough resolve may be willing to deliberately fight a full-scale war instead of resorting to limited conflict. Using the same methodology as in our analysis of inadvertent escalation, we also analyze the relationship between resolve and deliberate war in contests of capabilities. In any given strategic setting, the effect of resolve on the purposeful outbreak of war does not necessarily go in the same direction as its effect on accidental escalation via limited conflict. This is because the relationship between resolve and deliberate war is a function of the absolute level of escalation risk (rather than the effect of limited conflict on it) and the effect of resolve on the absolute cost (rather than the marginal cost) of limited policy options.

This paper makes three primary contributions. First, this paper offers new insights into the literature on nuclear deterrence theory. We present and analyze a general theoretical framework derived from how scholars have conceptualized the brinkmanship setting in the past. While these game-form free analyses have been applied to crisis bargaining models or to flexible-response crisis bargaining models (Banks 1990; Fey and Ramsay 2011; Kenkel and Schram 2024), those analysis do not speak to settings with a stochastic escalation risk.

Through our analysis of contests of nerves, we are able to offer sweeping results on the relationship between resolve, escalation risks, and outcomes that do not rely on any one game form. While results like these exist occurred in individual models, we are the first to establish their ubiquity across this general class of games. Then, we demonstrate that under a less restrictive modeling framework, the contests of capabilities framing, these general relationships may break down. In doing so, we are able to speak to a much broader class of real-world crises, and we are able to redefine how resolve matters in crises with escalation risks. As we discuss, our framework encompasses a broad class of different modeling choices, including those implemented in Powell (2015) and Schram (2024) (see section 5).

Second, while it has been useful to motivate this paper through the lens of nuclear deterrence theory, our results speak to a much broader class of settings. Contests of capabilities, as explored here, could also describe third party support to terror or insurgent groups, where the terror group could go too far and create a war between the sponsor and rival. They also could characterize conflict between rival drug cartels, interacting in the shadow of a government intervention. And they could describe an incident of economic or political repression that risks sparking mass protests, coups, or insurgency, pitting the political elites against its population.

Third, this research expands a recent line of work on crisis bargaining and deterrence in which states are assumed to have multiple coercive options available to respond to a threat (Schultz 2010; McCormack and Pascoe 2017; Coe 2018; Spaniel and Malone 2019; Qiu 2022; Baliga, Bueno de Mesquita and Wolitzky 2020; Schram 2021; Di Lonardo and Tyson 2022). The paper most closely related work to ours is (Kenkel and Schram 2024), which conducts a game-form free analysis of crises with multiple conflict options. Our key innovation is to assume that in this setting there is a stochastic risk of escalation associated with the use of a lower-level policy instrument. This modification generates new theoretical results, and allows us to describe a new and under-formalized set of empirical cases.

1 The Impact of Capabilities on Brinkmanship

Before we present our general results, we present stylized examples of the "contests of nerves" and the "contests of capabilities" settings. These game forms are intentionally sparse: these should be viewed as reduced-form representations that illustrate some of the intuition behind our general results. To preview what is to come, in section 5, we discuss our results in the context of the recent brinkmanship models in Powell (2015) and Schram (2024).

In our stylized contest of nerves game, we consider a Challenger (C) and Defender (D) in a dispute over some territory or policy (we will refer to this as the prize) whose value is normalized to 1. At the first stage of the game, Nature assigns whether D's war payoff is low ($\theta = \underline{\theta}$) or high ($\theta = \bar{\theta}$), with each outcome having positive probability. D observes their private type, while C only knows the prior probability that D is a high or low type. This private type represents D's resolve or D's willingness to risk war. In the nuclear deterrence setting, this type represents D's willingness to engage in a nuclear war, which is a function of D's capabilities, political motivations, and general hawkishness. While D may have no desire to intentionally select into a catastrophic nuclear exchange (as can be captured in the payoffs), D may be more or less willing to risk some probability of such an event.

At the second stage of the game, C selects whether to transgress ($t = 1$) or not ($t = 0$). If C does not transgress, then the game ends; if C does transgress, then D is able to respond.

Finally, if C previously transgressed, D can select some bargaining action or can deliberately go to the escalated war outcome. In this game form, if D does not want to go to war, D selects bargaining action $b_i \in \{b_1, b_2\}$. These bargaining actions could result in an autonomous risk of a war, but, conditional on war not occurring, they may be politically productive by giving D a greater share of the prize.

We summarize the game's payoffs in Table 1.

	$\underline{\theta}$ *D's Payoff*	$\bar{\theta}$ *D's Payoff*	*C's Payoff*
C does not transgress	1	1	0
C transgresses, *D selects* b_i	$R(b_i)W_D(\underline{\theta})+$ $(1-R(b_i))V_D(b_i)$	$R(b_i)W_D(\bar{\theta})+$ $(1-R(b_i))V_D(b_i)$	$R(b_i)W_C+$ $(1-R(b_i))(1-V_D(b_i))$
C transgresses, *D goes to war*	$W_D(\underline{\theta})$	$W_D(\bar{\theta})$	W_C

Table 1: **Contest of nerves** payoffs.

In this simple game, when C does not transgress, D attains the entirety of the prize. And, when D goes to war, both D and C receive their wartime payoffs denoted, W_C or $W_D(\theta)$. For D, the war payoff is a function of their type, with $W_D(\underline{\theta}) < W_D(\bar{\theta})$. Finally, when D engages in bargaining by setting some b_i, this will shape the bargained outcome $V_D(b_i)$ but will also generate some autonomous probability of war $R(b_i)$. Thus, bargaining here resembles how scholars have conceptualized brinkmanship crises in the past: staying in the crisis or undertaking certain moves could produce better settlement outcomes, but could also bear some risk of unintended escalation.

To illustrate how this game plays out, consider a numerical parameterization of the variables above. This is Figure 1. Under the selected parameters (see the Figure's caption), the contest of nerves has a simple structure. War is quite bad for both types of D ($W_D(\underline{\theta}) = -2$ and $W_D(\bar{\theta}) = -1$), so both types choose between bargaining options b_1 and b_2. For D, b_1 is less productive politically than b_2 (because $V_D(b_1) = 0.4$ and $V_D(b_2) = 0.75$), but b_1 comes with a lower autonomous risk of war ($R(b_1) = 0.05$ and $R(b_2) = 0.2$). Under these parameters, conditional on C transgressing, type $\underline{\theta}$ D's will select b_1 and type $\bar{\theta}$ D's will select b_2. Finally, faced with a 50%-50% gamble (because $Pr(\theta = \bar{\theta}) = 0.5$) between 0 and 0.52 and a sure-thing of 0, C will transgress.

In equilibrium, type $\bar{\theta}$ D's will take a greater escalation risk than type $\underline{\theta}$ D's. As intuition,

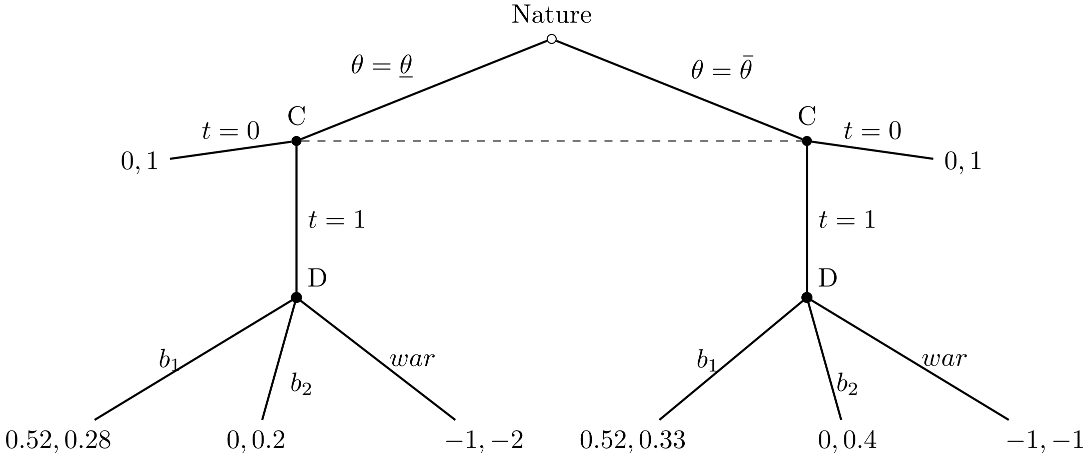

Figure 1: **A Contest of Nerves**: Greater resolve (θ) implies a greater likelihood of war.

C's payoffs are listed first. Note here that $\bar{\theta}$ has greater wartime payoffs. In equilibrium, C will transgress ($t = 1$), $\underline{\theta}$ D's will select bargaining action b_1, and $\bar{\theta}$ D's will select bargaining action b_2. Parameter values are $R(b_1) = 0.05$, $R(b_2) = 0.2$, $V_D(b_1) = 0.4$, $V_D(b_2) = 0.75$, $W_D(\underline{\theta}) = -2$, $W_D(\bar{\theta}) = -1$, $W_C = -1$, and $Pr(\theta = \bar{\theta}) = 0.5$.

type $\bar{\theta}$ D's do better in war (they have greater resolve), so they are more willing to take riskier bargaining moves to reach better settlement payoffs. More precisely, while type $\bar{\theta}$ D's select into a $R(b_2) = 0.2$ autonomous risk of war, they are willing to take this gamble because their upside (attaining $V_D(b_2) = 0.75$) outweighs their downside (attaining $W_D(\bar{\theta}) = -1$). In contrast, for type $\underline{\theta}$ D's, this downside (attaining $W_D(\underline{\theta}) = -2$) no longer makes the risk of b_2 worthwhile; type $\underline{\theta}$ D's play it safer by selecting b_1 (autonomous risk $R(b_1) = 0.05$). As we will demonstrate later in the paper, this relationship—where more resolved actors play riskier bargaining strategies—is generic to the contests of nerves class of games (see section 3).

Now consider a generalization to the contests of nerves framing: the "contests of capabilities" framework. Here, taking the productive political moves that generate autonomous risk now come with costs that are correlated with resolve. These actions could include sanctions, implementing a blockade, supplying allies through an airlift, offering third-party support to rebels, conducting special operations, or even carrying out a conventional war—what matters here is that these actions are politically productive, costly, and conducted in an environment where escalation to some higher level of conflict is still possible. We will refer to this class of costly, productive actions using the shorthand term “hassling” (see Schram (2021)), but in the nuclear deterrence, these limited actions could be a war using everything besides strategic nuclear weapons. We will assume that hassling costs could be positively or negatively correlated with the costs of the escalated conflict. In the nuclear context, if an actor is more capable within or more willing to risk a strategic nuclear conflict, this actor could also be more or less willing to engage at lower levels as well, with the specific empirical context determining the direction (more on this point below).

For ease, we present another stylized example of this more generalized setting. We keep the game very similar: Nature still sets D's private type, C still chooses to transgress or not, and D still responds to this transgression. The main difference is that here D responds to C transgressing with war or one of two hassling levels, h_1 and h_2 (rather than b_1 and b_2). These hassling actions still generate political benefits through V_D and still generate autonomous risk through R, which (we assume) are now functions of h_i; what's fundamentally new here is that D's hassling actions can have costs that are a function of D' type. We will denote these costs as $K(h_i, \theta)$.

We summarize the new payoffs to this game in Table 2.

	$\underline{\theta}$ *D's Payoff*	$\bar{\theta}$ *D's Payoff*	*C's Payoff*
C does not transgress	1	1	0
C transgresses, *D selects h_i*	$R(h_i)W_D(\underline{\theta}) - K(h_i, \underline{\theta}) + (1 - R(h_i))V_D(h_i)$	$R(h_i)W_D(\bar{\theta}) - K(h_i, \bar{\theta}) + (1 - R(h_i))V_D(h_i)$	$R(h_i)W_C + (1 - R(h_i))(1 - V_D(h_i))$
C transgresses, *D goes to war*	$W_D(\underline{\theta})$	$W_D(\bar{\theta})$	W_C

Table 2: **Contest of capabilities** payoffs.

This new framing allows us to consider how capabilities matter in the brinkmanship setting.

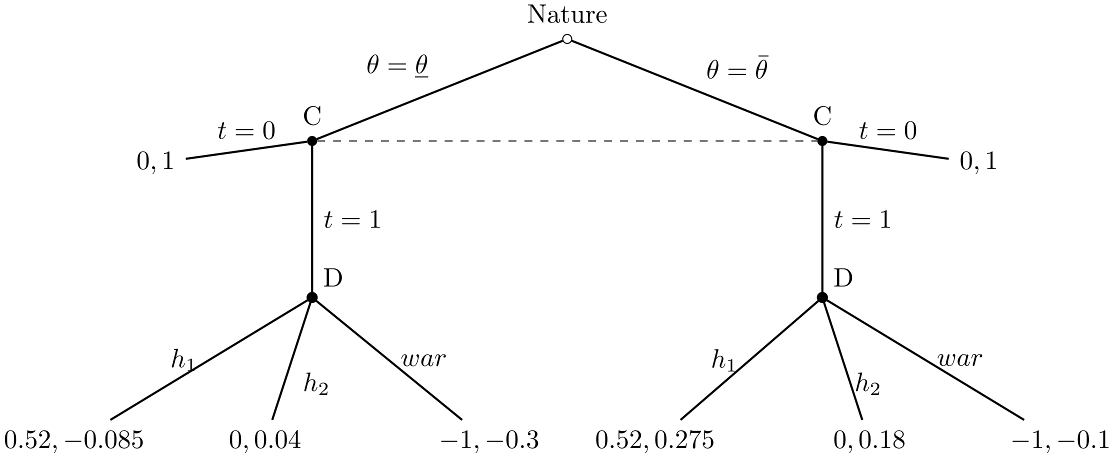

Figure 2: **A Contest of Capabilities**: Greater resolve (θ) implies a lower likelihood of war.

C's payoffs are listed first. Note here that $\bar{\theta}$ has greater wartime payoffs. In equilibrium, C will transgress ($t = 1$), $\underline{\theta}$ D's will select bargaining action b_1, and $\bar{\theta}$ D's will select bargaining action b_2. Parameter values are $R(h_1) = 0.05$, $R(h_2) = 0.2$, $V_D(h_1) = 0.4$, $V_D(h_2) = 0.75$, $W_D(\underline{\theta}) = -0.3$, $W_D(\bar{\theta}) = -0.1$, $W_C = -1$, $Pr(\theta = \bar{\theta}) = 0.5$, $K_D(h_1, \underline{\theta}) = 0.45$, $K_D(h_1, \bar{\theta}) = 0.1$, $K_D(h_2, \underline{\theta}) = 0.5$, $K_D(h_2, \bar{\theta}) = 0.4$.

In the previous stylized example, getting a better settlement offer (choosing b_2 rather than b_1) was analogous to being willing to pull a lever that generates better settlement payoffs at a greater stochastic risk. In practice, within crises, attaining a better settlement outcome may require conducting costly actions—actions like arming rebels, implementing or bypassing a blockade, or undertaking limited military actions. Here, choosing h_2 over h_1 generates different political settlements $V_D(h_i)$, different levels of autonomous risk $R(h_i)$, and different outright costs $K(h_i, \theta_i)$.

To illustrate what these new costs mean for equilibrium outcomes, we offer a numerical parameterization of the contest of capabilities game. This is Figure 2.

In the equilibrium here, the *more* resolved types go to war less. Why? Under the selected parameters, war is bad for D. Conditional on the game reaching D's decision node, both types of D choose between h_1 and h_2. Note that h_2 is more productive than h_1 (because $V_D(h_1) = 0.4$ and $V_D(h_2) = 0.75$), but h_2 comes both with greater autonomous escalation risk ($R(h_1) = 0.05$ and $R(h_2) = 0.2$) and greater inherent costs than h_1. For type $\bar{\theta}$ D's, selecting h_2 is quite costly relative to h_1 (because $K_D(h_1, \bar{\theta}) = 0.1$ and $K_D(h_2, \bar{\theta}) = 0.4$). This cost structure is enough to incentivise more resolved type $\bar{\theta}$ D's to select h_1 (the less risky option) in equilibrium. And, for type $\underline{\theta}$ D's, selecting h_2 is more costly than h_1, but not by much (because $K_D(h_1, \underline{\theta}) = 0.45$ and $K_D(h_2, \underline{\theta}) = 0.5$). This cost structure for less resolve type $\underline{\theta}$ D's is enough to incentivise the less resolve type to select h_2 (the more risky option) in equilibrium. Together, despite type $\bar{\theta}$ D's being (relatively) better at war and taking more aggressive hassling actions, type $\bar{\theta}$ D's will engage in lower-levels of hassling and incur lower risks of war than the less resolved type $\underline{\theta}$ D's.

The equilibrium in the contest of capabilities game is surprising: here the more resolved actors are more restrained. *Prime facie*, our treatment of resolve here seems natural: relative to less resolved types, the more resolved actor both does better in war and does better in conducting more aggressive hassling actions. But critically, resolve affects payoffs at all hassling levels; the counterintutive equilibrium behavior is driven by type $\bar{\theta}$ D's doing so much better at the lowest levels of hassling and type $\underline{\theta}$ D's incurring similar costs to both hassling levels.[4] Put another way, with this hassling cost structure, high-types are willing to tread lightly because it is so cheap for them to do so; meanwhile, low-types incur similar costs to both hassling actions, so they might as well take the action that generates more political benefits. We offer a general characterization of when this equilibrium phenomena occurs in subsection 4.1.

The two models we have presented here are deliberately sparse, so as to illustrate the simplest possible contests of nerves and contests of capabilities settings. Because these models lack many important elements of crisis bargaining and deterrence theory, such as pre-negotiation signaling and incentives to misrepresent, one might wonder whether the lessons we have drawn from them would carry over to richer, more realistic settings. To identify broad regularities in both classes of models, we employ the game-free methodology of Banks (1990)

[4] We will discuss this further, but formally, we are describing the $K(h_i, \theta)$ function as exhibiting increasing differences.

and Fey and Ramsay (2011), identifying properties that arise from foundational requirements of equilibrium rather than idiosyncratic features of any given game tree. We find, in fact, that the intuitive relationship between resolve and war risk in Figure 1 holds in every equilibrium of every game within the contests of nerves framework. And, we find that the counterintuitive relationship between resolve and war risk in Figure 2 can emerge in a wide range of games in the contests of capabilities setting.

2 Model Framework

To model contests of capabilities, we extend the formal definition of flexible-response crisis bargaining games developed by Kenkel and Schram (2024). A contest of capabilities is a game between a Challenger C and a Defender D, who are in a crisis over a prize whose value is normalized to 1.[5] The two sides bargain over the division of the prize. In the course of bargaining, each side may take costly low-level actions to influence the outcome of negotiations. What's novel here—and not considered in Kenkel and Schram (2024)—is that D's low-level action may generate a risk of accidental escalation to full-scale war.[6] If C and D do not agree on a negotiated settlement, then war occurs for certain. D has private information about its war payoff, which creates a friction that may result in bargaining failure in equilibrium.

2.1 Contests of Capabilities

The timing of a contest of capabilities is as follows. At the start of the game, Nature draws D's type θ from a commonly known distribution whose CDF is F_θ and whose support is $\Theta \subseteq \mathbb{R}$. Only D observes Nature's choice. The two players select bargaining strategies $b_C \in \mathcal{B}_C$ and $b_D \in \mathcal{B}_D$. Our analysis is agnostic as to the shape of the bargaining protocol.

[5] In section 5 below, we consider an alternative formulation where D's valuation of the prize is private information that varies across private types.

[6] The flexible-response crisis bargaining games studied by Kenkel and Schram (2024) are the special case of contests of capabilities in which this risk is identically zero.

Depending on the particular game form, these strategies might be simple proposals and responses, as in an ultimatum game, or more complex plans of offers and counteroffers. Alongside these baseline bargaining strategies, each player may take a costly action that directly shifts the payoffs from negotiations—and, in D's case, may generate a heightened risk of accidental escalation to war. We call C's action a "transgression" $t \in \mathcal{T} \subseteq \mathbb{R}_+$ and D's "hassling" $h \in \mathcal{H} \subseteq \mathbb{R}_+$.

The players' bargaining strategies determine whether the negotiation succeeds or fails. For any given game form g, there is a function $\pi^g : \mathcal{B}_C \times \mathcal{B}_D \to [0, 1]$ that describes the probability of agreement (i.e., neither player deliberately choosing war) given the players' bargaining strategies. In case of agreement, C pays a cost $K_C(t) \geq 0$ (increasing in t) for its transgressions, and D pays a potentially type-dependent cost $K_D(h, \theta) \geq 0$ (increasing in h) for its hassling.[7] In this case, the probability of accidental escalation to war is a strictly increasing function of D's hassling, denoted $R(h) \in [0, 1]$. We treat the costs of transgression and hassling, as well as the risk of accidental escalation, as underlying primitive features of the strategic interaction—not as features of a single particular game form g. If there is no accidental escalation to war, the players receive $V_C^g(t, h, b_C, b_D)$ and $V_D^g(t, h, b_C, b_D)$ respectively. Unlike the cost and risk functions, these are specific to a game form.

Like earlier game-free analyses of crisis bargaining (Fey and Ramsay 2011; Fey and Kenkel 2021; Kenkel and Schram 2024), we assume that both states have the option to unilaterally force a conflict. Formally, this amounts to assuming there exists $b_C^{\text{war}} \in \mathcal{B}_C$ such that $\pi^g(b_C^{\text{war}}, b_D) = 0$ for all $b_D \in \mathcal{B}_D$, as well as an analogous $b_D^{\text{war}} \in \mathcal{B}_D$. Reflecting the anarchic nature of international politics, this assumption ensures that neither state can be forced to accept a settlement that would leave it worse off than fighting.

[7] Because we conceptualize accidental escalation as arising from the coercive instrument itself, we assume these costs are paid even if accidental escalation occurs. For example, if a conventional war ends in an accidental nuclear exchange, the conventional war still carries costs (as was similarly formalized in Powell (2015)). Consequently, both players would prefer deliberate war—avoiding the costs of the low-level actions—over an agreement with a near-certain chance of accidental escalation.

The players' baseline war payoffs are solely a function of D's type, not their bargaining actions or low-level responses (see section 5 for an alternate treatment). We order the Defender's type space so that higher types are more resolved, i.e., the Defender's baseline war payoff function $W_D(\theta)$ is a strictly increasing function of θ. Meanwhile, the Challenger's baseline war payoff $W_C(\theta)$ is a non-increasing function of θ. If war occurs deliberately due to bargaining failure, then the players receive $W_C(\theta)$ and $W_D(\theta)$. If war occurs accidentally due to hassling-induced escalation, then they receive $W_C(\theta) - K_C(t)$ and $W_D(\theta) - K_C(h, \theta)$. Notice that transgressions and hassling are the only bargaining actions that may affect war payoffs, and they do so only for accidental escalation and only via the cost functions.

The Challenger's expected utility, given the bargaining strategies and the Defender's type, is given by the function

$$\begin{aligned} u_C^g(t, h, b_C, b_D \mid \theta) = \underbrace{\pi^g(b_C, b_D)}_{\text{agreement}} &\left[\overbrace{[1 - R(h)]V_C^g(t, h, b_C, b_D)}^{\text{no escalation}} + \overbrace{R(h)W_C(\theta)}^{\text{accidental escalation}} - K_C(t) \right] \\ &+ \underbrace{[1 - \pi^g(b_C, b_D)]}_{\text{disagreement}} W_C(\theta). \end{aligned}$$

Similarly, the Defender's expected utility function is

$$\begin{aligned} u_D^g(t, h, b_C, b_D \mid \theta) = \pi^g(b_C, b_D) &\Big[[1 - R(h)]V_D^g(t, h, b_C, b_D) + R(h)W_D(\theta) - K_D(h, \theta) \Big] \\ &+ [1 - \pi^g(b_C, b_D)]W_D(\theta). \end{aligned}$$

To close the definition of contests of capabilities, we place some additional assumptions on the model primitives. First, we assume that either player may refrain from low-level responses at no cost: $0 \in \mathcal{H} \cap \mathcal{T}$, $K_C(0) = 0$, and $K_D(0, \theta) = 0$ for all $\theta \in \Theta$. Second, we assume there is no risk of accidental escalation in the absence of hassling: $R(0) = 0$. Third, without loss of generality, we let $W_D(\theta) = \theta$ in the remainder of the analysis.

2.2 Game-Free Analysis

Our goal is to characterize patterns in the equilibria of contests of capabilities that hold across all game forms with the same underlying primitives, rather than being specific to a particular bargaining protocol (e.g., ultimatum game, alternating offers, etc.). Table 3 divides the model components into the primitive components and those that are specific to a particular game form. We will draw conclusions about the equilibrium outcomes of contests of capabilities solely as a function of the primitive components listed in the left-hand column. To this end, we adopt the mechanism design methodology of prior game-free analyses of crisis bargaining (Banks 1990; Fey and Ramsay 2009, 2011; Fey and Kenkel 2021; Liu et al. 2021; Kenkel and Schram 2024).

Underlying primitives	**Specific to game form**
D's type space: Θ	Bargaining actions: $\mathcal{B}_C$, $\mathcal{B}_D$
War payoffs: $W_C(\cdot)$, $W_D(\cdot)$	Actions → bargaining success: $\pi^g(\cdot)$
Low-level responses available: $\mathcal{T}$, $\mathcal{H}$	Actions → prize division: $V_C^g(\cdot)$, $V_D^g(\cdot)$
Cost functions: $K_C(\cdot)$, $K_D(\cdot)$	
Escalation risk: $R(\cdot)$	

Table 3: Classification of model components for a contest of capabilities.

As in similar analyses of models with one-sided incomplete information (e.g., Banks 1990; Fey and Kenkel 2021; Kenkel and Schram 2024), we characterize equilibrium outcomes for the player with private information, namely the Defender. Let $(t^*, h^*(\theta), b_C^*, b_D^*(\theta))$ be an equilibrium of a game form g, where D's strategies are written as functions of θ as different types may take different actions. We summarize the equilibrium via three functions of D's type. The first is the equilibrium probability of agreement for each Defender type:

$$\pi(\theta) = \pi^g(b_C^*, b_D^*(\theta)).$$

The second is the equilibrium amount of hassling, conditional on an agreement, for each Defender type:

$$h(\theta) = h^*(\theta).$$

The third is the equilibrium division of spoils going to the Defender, conditional on an agreement and no accidental escalation to war, for each type:

$$V_D(\theta) = V_D^g(t^*, h^*(\theta), b_C^*, b_D^*(\theta)).$$

We refer to these three functions $(\pi(\cdot), h(\cdot), V_D(\cdot))$ jointly as a direct mechanism. Rather than work with the complex set of all equilibria of all game forms, we will work with the set of direct mechanisms for contests of capabilities.

Given a direct mechanism, we can calculate the expected utility to each type of Defender as follows:

$$U_D(\theta) = \pi(\theta) \left[(1 - R(h(\theta)))V_D(\theta) + R(h(\theta))\theta - K_D(h(\theta), \theta)\right] + (1 - \pi(\theta))\theta.$$

Equally importantly, the direct mechanism gives us all we need to know to determine the payoff one Defender type would receive by deviating to another type's bargaining strategy. Consider a Defender whose true type is θ, but who mimics the bargaining strategy of another type θ'. This type receives the same lottery over agreement versus disagreement (probability $\pi(\theta')$ of agreement) and the same risk of accidental war in case of agreement (probability $R(h(\theta'))$) as the type it is mimicking. Additionally, if there is agreement and no accidental war, the mimicking type receives the same bargaining spoils, $V_D(\theta')$. But there are two key differences between the mimic's payoff and $U_D(\theta')$. First, in case of war (whether accidental or deliberate), the mimic receives its true private value θ—it does not become stronger or weaker on the battlefield just by adopting the bargaining strategy of a different type. Second, in case of agreement, the type-dependent component of the mimic's hassling cost reflects its true type; i.e., the mimic pays $K_D(h(\theta'), \theta)$. Altogether, then, the expected utility to type θ for adopting the bargaining strategy of type θ' is

$$\Phi_D(\theta' \mid \theta) = \pi(\theta') \left[(1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - K_D(h(\theta'), \theta)\right] + (1 - \pi(\theta'))\theta.$$

A direct mechanism is incentive compatible if no Defender type would strictly benefit from mimicking the bargaining strategy of a different type. Formally, the incentive compatibility condition is

$$U_D(\theta) \geq \Phi_D(\theta' \mid \theta) \qquad \text{for all } \theta, \theta' \in \Theta. \tag{IC}$$

The incentive compatibility condition is closely related to the equilibrium requirements of Bayesian games. Our game-free analysis depends critically on the revelation principle articulated by Myerson (1979): for every Bayesian Nash equilibrium of a Bayesian game, there exists a payoff-equivalent direct mechanism that is incentive compatible. Therefore, if some claim holds for all incentive compatible direct mechanisms for contests of capabilities, then the same claim is true for all equilibria of such contests. By analyzing the set of incentive compatible direct mechanisms, we can derive necessary conditions for equilibrium behavior without having to solve any specific game form.

In line with Fey and Ramsay (2011) and the subsequent mechanism design literature, we also impose a voluntary agreements condition—what economists would call a participation or individual rationality constraint—on the set of direct mechanisms we consider. Voluntary agreements holds when no Defender type is worse off than it would be from deliberately fighting:

$$\pi(\theta)\left[(1 - R(h(\theta)))V_D(\theta) + R(h(\theta))\theta - K_D(h(\theta), \theta)\right] \geq \pi(\theta)\theta \qquad \text{for all } \theta \in \Theta. \tag{VA}$$

Formally, voluntary agreements is a consequence of our assumption that both players have a bargaining action available that guarantees war. Substantively, this condition reflects the anarchic state of international politics, in which all agreements must be self-enforcing. Voluntary agreements hold trivially for any type that deliberately chooses war for certain, i.e., for which $\pi(\theta) = 0$. Additionally, if we have $\pi(\theta) = 0$ for at least one Defender type, then incentive compatibility implies voluntary agreements.

3 Contests of Nerves

We conceptualize a contest of nerves as a special case of a contest of capabilities, in which all Defender types have the same access to an instrument that generates exogenous escalation risks. In contests of nerves, Defender types may vary in their resolve—their expected value of fighting, and thus their willingness to risk war in order to receive a particular settlement at the bargaining table—but no type has an advantage over any other at pulling the levers that generate risk. In its simplest form, the risk of accidental war can be thought of as a pure brinkmanship measure, akin to rocking the boat in Schelling (1966, 90-91). This pure brinkmanship dynamic is formalized in models like Nalebuff (1986), Powell (1988), and Powell (1990).[8] Additionally, the contests of nerves framing can also describe settings where lower-levels actions may be inefficient, so long that these inefficiencies are uncorrelated with private type. Also, as discussed in section 5, this framework closely relates to the model in Powell (2015).

Formally, we define a contest of nerves as one in which the cost of each feasible low-level choice is constant across Defender types. In a contest of nerves, there exists a non-decreasing function $\kappa_D : \mathcal{H} \to \mathbb{R}_+$ such that

$$K_D(h, \theta) = \kappa_D(h) \qquad \text{for all } h \in \mathcal{H} \text{ and } \theta \in \Theta.$$

This includes as a special case "pure brinkmanship" scenarios, in which all types can generate accidental war risk costlessly: $\kappa_D(h) = 0$ for all $h \in \mathcal{H}$ (like in Nalebuff (1986) and Powell (1988). When the Defender's private information is about its war payoff, the distinction between the pure brinkmanship model and the contest of nerves turns out to be essentially

[8] Our framing differs from these models in our treatment of "resolve," which only depends on D's escalated war payoff. For example, in Powell (1988), resolve is a function of war payoffs, the payoffs from prevailing in the crisis, and the payoffs from conceding in the crisis; we choose a simpler treatment of resolve that does not rely on factors like the payoffs from dropping out of a crisis, which, in the bargaining setting, may be endogenous to the Defender's war payoff.

immaterial.[9]

Contests of nerves exhibit essentially the same patterns of behavior as in ordinary crisis bargaining games, where there are no low-level policy alternatives between peaceful settlement and all-out war. When accounting for both of the possible paths to conflict—deliberate war or accidental escalation—more resolved types of the Defender are more likely to end up at war in equilibrium. Additionally, more resolved types have higher equilibrium payoffs. The following proposition states these claims formally as properties of incentive compatible direct mechanisms for contests of nerves.[10]

Proposition 1. *In any equilibrium of a contest of nerves, the total probability of war and the Defender's equilibrium utility weakly increase with the Defender's resolve: if* $\theta' < \theta''$*, then* $\pi(\theta')[1 - R(h(\theta'))] \geq \pi(\theta'')[1 - R(h(\theta''))]$ *and* $U(\theta') \leq U(\theta'')$*.*

Proposition 1 holds because contests of nerves are, in fact, equivalent to ordinary crisis bargaining games at a deep level. Because the cost of the low-level option does not differ across types, any type that mimics the bargaining strategy of θ' receives exactly the same payoff in case of a peaceful outcome, namely $V_D(\theta') - \kappa_D(h(\theta'))$. This equivalence of settlement payoffs across types means that the monotonicity results from Banks (1990) apply to contests of nerves, so higher types of the Defender are more likely to go to war and have greater equilibrium expected utilities. By contrast, in the more general class of contests of capabilities that we study below, different Defender types might yield different payoffs from the same bargaining strategy, even conditional on the interaction ending peacefully. That is because the costs of the low-level policy may differ across types, leading to different overall settlement values under which the Banks (1990) results no longer apply (see Kenkel and Schram 2024).

[9] The same is not necessarily true when the Defender's private information concerns its prize valuation; see section 5 below.

[10] All proofs appear in Appendix A.

When full-scale war is sufficiently destructive, such as a nuclear exchange would be, it is plausible to suppose neither side would ever deliberately initiate a conflict (Brodie 1966; Schelling 1966; Powell 1990). In this case, we can obtain even stronger results about the relationship between the Defender's private resolve and equilibrium choices. Specifically, more resolved types engage in more brinkmanship and receive more favorable settlements at the bargaining table when the game ends peacefully.

Corollary 1. *In any equilibrium of a contest of nerves, if war never occurs deliberately ($\pi(\theta) = 1$ for all θ), then the probability of accidental war and the Defender's settlement value weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $R(h(\theta')) \leq R(h(\theta''))$ and $V_D(\theta') \leq V_D(\theta'')$.*

Altogether, in a contest of nerves in which no Defender type has a particular advantage or disadvantage at generating accidental war risk, outcomes are determined by resolve in a predictable way. Greater resolve implies a greater total risk of war, including a greater risk of accidental war when neither player would ever deliberately opt into conflict. But as we show below, this stark pattern does not necessarily hold in more general contests of capabilities, where we consider accidental war risk a byproduct of low-level policy responses whose cost or effectiveness is related to the Defender's resolve.

4 Contests of Capabilities

In the contest of capabilities framework, the Defender's private willingness to go to war is related to their ability or willingness to use limited instruments. This creates a potential new tradeoff that affects the Defender's equilibrium choices. Types may vary in their preferences for lower-level conflict not only due to the risk of full-scale war that these options generate, but also because of differences in their direct costs for using these instruments. Consequently, even when full-scale war never occurs on purpose, it is no longer certain that more resolved types take on a higher risk of accidental conflict.

Throughout this section, we restrict attention to direct mechanisms in which each Defender type is either certain to agree or certain to go to full-scale war: $\pi(\theta) \in \{0, 1\}$ for all $\theta \in \Theta$. This set of mechanisms corresponds to pure-strategy equilibria of contests of capabilities in which Nature's only moves are the initial assignment of the Defender's type and the risk of war $R(h)$ generated by the Defender's low-level policy choice (see Fey and Kenkel 2021). Given such a mechanism, we can partition the type space into those that reach agreement (with possible risk of accidental conflict) and those that deliberately fight a war: $\Theta = \Theta_1 \cup \Theta_0$, where $\Theta_1 = \{\theta \in \Theta \mid \pi(\theta) = 1\}$ and $\Theta_0 = \{\theta \in \Theta \mid \pi(\theta) = 0\}$. We relax this assumption, allowing for equilibria in which some Defender types mix, in subsection 4.3 below.

4.1 Probability of Accidental War

Unlike in the special case of the contest of nerves analyzed above, the probability of accidental war need not increase with the Defender's resolve in a contest of capabilities. To see why, consider the tradeoff between low and high hassling, and how it varies with the Defender's resolve. First, there will be a difference in the Defender's settlement value if accidental war does not occur.[11] The Defender's resolve is immaterial to the value of this difference. Second, higher hassling generates a greater risk of accidental conflict. This is the effect at the root of Corollary 1 above, leading more resolved types to be more tolerant of greater hassling. But now there is a third component to the tradeoff: the marginal cost of the higher value of hassling may differ with the Defender's resolve. In practical terms, some Defenders may be more or less willing to conduct low-level competition depending on their high-level resolve; for example, if a Defender is more hawkish and willing to risk a nuclear exchange, then this Defender could plausibly also be more willing to absorb the costs from a more expansive conventional conflict. Naturally, we cannot characterize the relationship between resolve and

[11] Intuitively, one might expect more hassling to yield more favorable terms. In fact, additional assumptions are required to guarantee this is the case. See Kenkel and Schram (2024).

equilibrium behavior without accounting for these marginal costs—and how they compare to the increase in the risk of accidental war.

If the marginal cost of hassling decreases with the Defender's resolve, then it is straightforward to see that more resolved types will run a greater risk of accidental war. In this case, compared to a less resolved Defender, a more resolved type gets the same benefit in case the agreement holds, is better off in case accidental war occurs, and pays less to go from low hassling to high hassling.

If instead more resolved Defender types face higher marginal costs of hassling (e.g., because investments in capabilities for total war crowd out the resources for lower-level instruments), then the tradeoff is harder to resolve. More resolved types are still better able to handle the risk of accidental war, but now they must pay more to demonstrate this resolve through low-level conflict. Ultimately, equilibrium behavior here comes down to the relative magnitude of (a) the effect of Defender resolve on the marginal cost of hassling and (b) the effect of hassling on the risk of accidental war. If the increase in the low-level instrument does not generate much risk, then we see the opposite pattern from the war of nerves, with more resolved Defenders investing less in low-level conflict and thus facing lower odds of accidental war.

The following proposition states sufficient conditions for the equilibrium probability of accidental war to increase or decrease with the Defender's resolve among all types that reach an agreement. The left-hand side of the equations in the proposition is, in essence, the effect of resolve on the marginal cost of greater hassling. The right-hand side is the effect of greater hassling on the risk of accidental conflict, weighted by the difference in war payoff between the more resolved and the less resolved type.

Proposition 2. *Consider an equilibrium of a contest of capabilities.*

(a) *The probability of accidental war weakly increases with the Defender's resolve if*

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] < [R(h'') - R(h')](\theta'' - \theta') \quad (1)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

(b) *The probability of accidental war weakly decreases with the Defender's resolve if*

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] > [R(h'') - R(h')](\theta'' - \theta') \quad (2)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

Equation 1 and Equation 2 are closely related to the single-crossing conditions that often arise in mechanism design and related economic settings (Milgrom and Shannon 1994; Ashworth and Bueno de Mesquita 2006). If K_D has global decreasing differences—i.e., the increase in cost between any two levels of hassling is always smaller for more resolved types—then Equation 1 must hold, and the probability of accidental war must behave as it does in a contest of nerves. On the other hand, if K_D has global increasing differences, then accidental war risks may still increase with resolve; a slim increase in marginal costs with θ is not enough to make Equation 2 hold.[12]

The key takeaway from Proposition 2 is that the direction of the relationship between resolve and accidental war depends on how resolve affects the marginal cost of hassling versus how hassling affects the risk. Can we say anything stronger about the shape of this relationship, such as how quickly the level of low-level activity varies with the Defender's resolve? Our next result, Proposition 3, answers this question in the negative. As long we can satisfy

[12] K_D need not have global increasing or decreasing differences. For example, consider the type space $\Theta = [0, 1]$, hassling space $h = \{0, 1, 2\}$, and cost function $K_D(0, \theta) = 0$, $K_D(1, \theta) = a + b\theta$ (where $0 < a < 1$ and $0 < b < 1 - a$), $K_D(2, \theta) = 1$. This function has increasing differences on $\{0, 1\}$, decreasing differences on $\{1, 2\}$, and constant differences on $\{0, 2\}$. Consequently, if the effect of hassling on the risk of accidental war is strong enough, there may simultaneously exist IC mechanisms with increasing hassling (on $\{1, 2\}$) and decreasing hassling (on $\{0, 1\}$).

the conditions on marginal effects set out in the previous proposition, we can design a game form to rationalize virtually any pattern of hassling.

To obtain the following result, we must impose slightly stronger technical conditions than in the baseline analysis. We assume a continuous type space, $\Theta = [\underline{\theta}, \overline{\theta}]$, and set of feasible low-level actions, $\mathcal{H} = [0, \overline{h}]$. We also assume that R is continuously differentiable and that K_D is twice differentiable. Together we refer to these as the differentiability assumptions.

Proposition 3. *Suppose the differentiability assumptions hold.*

(a) Let h^* be any absolutely continuous, weakly increasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \leq R'(h) \qquad \textit{for all } h \in h^*(\Theta),\ \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(b) Let h^{**} be any absolutely continuous, weakly decreasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \geq R'(h) \qquad \textit{for all } h \in h^{**}(\Theta), \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Though the formal statement is technical, this is a remarkable result. Proposition 3 tells us that the primitive features of the strategic environment—i.e., the players' types, the low-level actions available, the costs of those actions, and the risk of accidental war they generate—only determine whether the equilibrium degree of hassling (and accidental war) increases or decreases with the Defender's resolve. The precise magnitude of the increase or decrease may be quite specific to a particular bargaining game form. For example, suppose that K_D

has global decreasing differences (more resolved types face lower marginal costs of hassling), so that $\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} < 0$ for all h and θ. Then Proposition 3(a) implies that *every* increasing function $h^*(\theta)$ (subject to a continuity restriction) can be supported as equilibrium behavior. Whether there is a gradual increase, a sudden large spike, or no change at all depends only on the contingent features of the bargaining game, not on the underlying military fundamentals.

4.2 Occurrence of Deliberate War

We now consider how the choice of deliberate war varies with the Defender's resolve. By definition, the payoff from war is greater for more resolved types, increasing their incentive to opt for war. In the absence of low-level alternatives, this incentive results in a monotone increasing relationship between resolve and deliberate war (Banks 1990). However, in a more general setting, the effect of resolve on deliberate war depends on its relationship with the cost of low-level policy options. In particular, if more resolved types can also use low-level conflict more cheaply or effectively, then they may opt for limited conflict instead of war (Kenkel and Schram 2024).

While the relationship between resolve and hassling costs is important, the risk of accidental war from limited conflict also plays a role in the decision to fight war deliberately. If a limited policy instrument generates only an infinitesimal risk of accidental war, then the types with the most incentive to opt for it will be whichever ones have the lowest costs for that level of hassling. At the other extreme, if limited conflict carries an enormous risk of escalation, then resolve rather than costs will be the determining factor in the preference of each Defender type.

The following proposition summarizes the relationship between Defender resolve and the occurrence of deliberate war in the equilibria of contests of capabilities. If more resolved types pay higher (absolute) costs for low-level policy options, then less-resolved types settle and more-resolved types fight. However, when greater resolve also entails greater willingness or capability to hassle, then we must compare the magnitude of the cost effect to the degree of accidental war risk. If resolve only slightly reduces the hassling costs, or if the risk of accidental escalation is low, then we still have less-resolved types agreeing and more-resolved types fighting. But we yield the opposite pattern if hassling costs plunge quickly with θ or if escalation risks are high.

Proposition 4. *Consider an equilibrium of a contest of capabilities. Let Θ_1 denote the set of types that reach an agreement in equilibrium: $\Theta_1 \equiv \{\theta \in \Theta \mid \pi(\theta) = 1\}$.*

(a) Less-resolved Defender types reach agreement and more-resolved types deliberately choose war (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \leq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} < 1 - R(h') \tag{3}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

(b) Less-resolved Defender types deliberately choose war and more-resolved types reach agreement (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \geq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} > 1 - R(h') \tag{4}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

We briefly note the technical similarities and differences between Proposition 2 (conditions for the probability of accidental war to be monotone in the Defender's resolve) and Proposition 4. Both results work with differences in the cost function, $K_D(\cdot, \theta'') - K_D(\cdot, \theta')$, as well as in war payoffs, $\theta'' - \theta'$. Additionally, the risk of accidental war plays a role in both results. The probability of accidental war characterized in Proposition 2 is ultimately determined by a second-order comparison: the effect of Defender type on the *marginal* cost of hassling, compared to the marginal effect of hassling on escalation risk. By contrast, the occurrence of deliberate war characterized here in Proposition 4 depends more on first-order comparisons: the effect of Defender type on the *absolute* cost of hassling, versus the absolute level of escalation risk.

4.3 Probabilistic Deliberate War

Our baseline analysis concerns equilibria in which every Defender type either reaches agreement for certain or fights a deliberate war for certain. In such equilibria, the Defender can generate a limited risk of war only through choosing a corresponding level of hassling, not through mixed strategies. We now relax this restriction to consider equilibria in which we may have $\pi(\theta) \in (0,1)$ for some (or all) types of the Defender.

When we allow for a probabilistic occurrence of deliberate war, we find exceptions to some of the patterns characterized in our baseline analysis. For example, consider an environment in which the effect of resolve on hassling cost is negative (Equation 3 holds), and this effect becomes larger in magnitude at higher degrees of hassling (Equation 1 holds). Under these conditions, Proposition 2 would lead us to conclude that the probability of accidental war increases with the Defender's resolve, and we would infer from Proposition 4 that the same is true for the occurrence of deliberate war. However, these patterns are specific to equilibria in which the probability of deliberate war is exactly 0 or 1 for each Defender type. Proposition 5 below shows that the probability of accidental escalation may decrease with Defender resolve when the probability of deliberate war is locally increasing. Alternatively, if the probability of accidental escalation increases quickly enough, then the probability of deliberate war may decrease with Defender resolve.

For ease of characterization, we impose linearity assumptions that are even stronger than the differentiability assumptions of Proposition 3 above. We assume $\Theta = [\underline{\theta}, \overline{\theta}]$ and $\mathcal{H} = [0, 1]$.[13]

[13] Because the cost function and war payoff space are arbitrary in magnitude, the normalization $\overline{h} = 1$ is without loss of generality.

Additionally, we assume the cost function and risk function are linear in hassling: there exist a function $k : \Theta \to \mathbb{R}_{++}$ and a constant $r > 0$ such that $K_D(h, \theta) = k(\theta)h$ and $R(h) = rh$ for all $\theta \in \Theta$ and $h \in \mathcal{H}$. We also assume k is differentiable, and we denote $\underline{k} \equiv \min k'(\Theta)$ and $\overline{k} \equiv \max k'(\Theta)$.

Proposition 5. *Suppose the linearity assumptions hold and $r - 1 < \underline{k} \leq \overline{k} < r$.*

(a) Equation 1 and Equation 3 hold.

(b) Let π^* be any absolutely continuous, weakly decreasing function such that $\pi^*(\theta) > 0$ for all $\theta \in \Theta$, and let h^* be any absolutely continuous function that satisfies

$$\frac{dh^*(\theta)}{d\theta} \geq \frac{\frac{1}{r-\underline{k}} - h^*(\theta)}{\pi^*(\theta)} \cdot \frac{d\pi^*(\theta)}{d\theta}$$

for all $\theta \in \Theta$ at which h^* and π^* are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^*(\theta)$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(c) Let π^{**} be any absolutely continuous, weakly increasing function such that $\pi^{**}(\theta) > 0$ for all $\theta \in \Theta$, and let h^{**} be any absolutely continuous function that satisfies

$$\frac{dh^{**}(\theta)}{d\theta} \geq \frac{\frac{1}{r-\overline{k}} - h^{**}(\theta)}{\pi^{**}(\theta)} \cdot \frac{d\pi^{**}(\theta)}{d\theta}$$

for all $\theta \in \Theta$ at which h^{**} and π^{**} are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^{**}(\theta)$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Though it is important to understand these baseline exceptions to the patterns characterized in our main analysis, the takeaway here should not be that anything can happen. In substantively important contexts such as nuclear war where it is implausible that the Defender would ever deliberately opt for war, we cannot obtain these exceptions to Proposition 2. Additionally, the result here depends on the cost effect being in a tight range where it is small in magnitude. A larger negative or positive relationship between resolve and hassling cost would tighten the set of patterns that can be sustained in equilibrium.

5 Resolve as Prize Value

The analysis above considers a class of models that treat the Defender's resolve as their war payoff. Here, we consider an alternate formulation, more in line with recent nuclear brinkmanship models (e.g., Powell 2015; Schram 2024), in which the Defender's private resolve is represented by their value for the object at stake in the crisis. We are able to show that this alternate formulation can share similarities with the contest of capabilities framework, but, with some new model primitives (i.e. a option to quit and zero out), can also introduce new "outbidding" behavior.

To motivate the analysis, consider a simple model of brinkmanship based on Powell (2015).[14] C and D are in a crisis over a prize worth $\beta_C > 0$ to C and $\beta_D > 0$ to D. The prize is initially controlled by D. C's prize valuation is common knowledge, while D's is private information only known to D. The timing of the game is as follows:

1. Nature selects D's valuation β_D and reveals it to D.
2. C chooses a conventional arms level $p \in [\underline{p}, \overline{p}]$.
3. D chooses a risk level $r \in [0, \overline{r}(p)]$, where $\overline{r}(p) \in (0, 1)$ for each $p \in [\underline{p}, \overline{p}]$.
4. C may quit or continue with the challenge. If C quits, then C receives nothing and D receives the full prize.

[14] There are three differences between the model here and that of Powell (2015). First, we rule out C's initial option to end the game immediately by accepting the status quo. Any equilibrium in which this occurs involves no choices by D along the path of play, resulting in a trivial direct mechanism for our purposes. The second, related difference is that we normalize the costs k_C and k_D to zero, as both are sunk (and thus decision-irrelevant) once C opts not to accept the status quo. Third, we assume there is no baseline latent risk, i.e., $\underline{r}(p) = 0$ for all p. Our framework could be modified to incorporate non-zero baseline risk at the cost of some additional notation; we omit this possibility here for clarity of exposition.

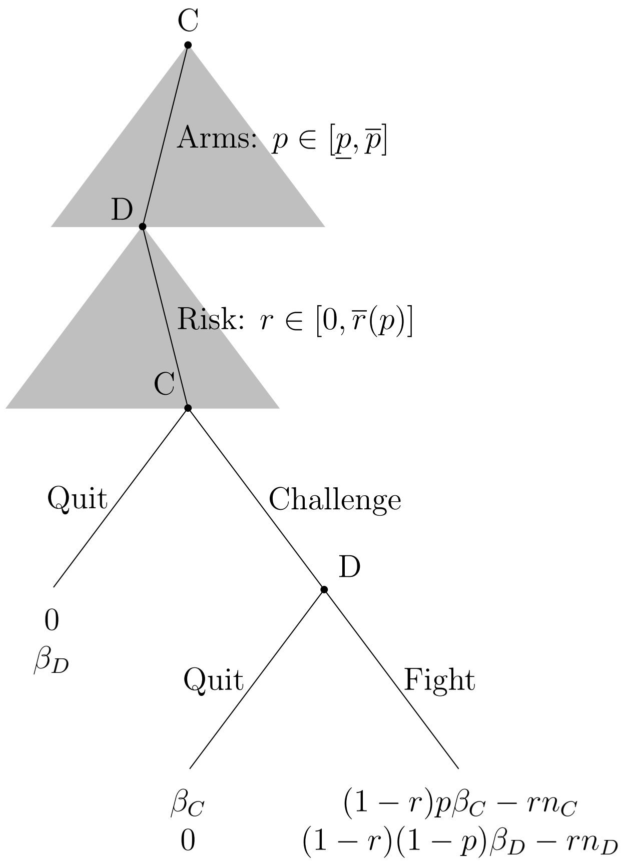
C
Arms: $p \in [\underline{p}, \overline{p}]$
D
Risk: $r \in [0, \overline{r}(p)]$
C
Quit
Challenge
D
0
β_D
Quit
Fight
β_C
0
$(1-r)p\beta_C - rn_C$
$(1-r)(1-p)\beta_D - rn_D$

Figure 3: Game tree for the motivating model where the Defender's type represents their war payoff (based on Powell 2015).

5. If C continued with the challenge, D may quit or fight. If D quits, then C receives and full prize and D receives nothing. If instead D fights, then with probability r a nuclear war occurs and each player receives $-n_i < 0$. Otherwise, C wins the prize with probability p and D wins it with probability $1-p$.

Figure 3 displays the game tree for this model of nuclear brinkmanship.

The baseline framework in section 2 captures some key features of this model, in particular an unintentional risk of nuclear escalation. However, there are also some key differences. The most visible is that D's type now represents their prize valuation instead of their war payoff. Additionally, even after D moves to generate a certain level of endogenous risk, either player may "shut off" that risk by conceding the issue. In this section, we consider a generalized framework that captures the features of the model in Figure 3. We show that the payoff structure in this new framework is essentially isomorphic to a special case in our original contest of capabilities, but that the possibility of shutting off the risk may generate distinct patterns in equilibrium outcomes. Nonetheless, certain outcome patterns can still arise only when risk is generated by a costly political process for D, rather than a costless lever as in the contest of nerves.

5.1 Direct mechanism

We consider a class of models with the same choice structure as in our baseline framework (introduced in section 2): C and D each choose bargaining actions, including low-level responses, which determine the distribution of spoils and the risk of war. The primitives differ from the baseline model in the following way:

- The war payoffs are fixed values $-n_C, -n_D < 0$.
- The value of the prize for D is a private type, $\beta_D \in [\underline{\beta}_D, \overline{\beta}_D]$.
- The hassling cost function, which we now denote $\kappa_D(h)$, is a function solely of hassling and not of D's type.

The change in the prize payoff structure necessitates a different definition of the direct mechanism than in our baseline analysis. We now model the *share* of the prize received by each type of D, via a function $S_D : [\underline{\beta}_D, \overline{\beta}_D] \to [0, 1]$. This may represent a share of the prize garnered from negotiations, or from conventional conflict as in the Powell (2015) model. Additionally, mirroring Powell (2015), we allow for the possibility that a state might "quit," ceding the prize to the other while zeroing out the risk of accidental war.[15] We include functions $Q_C, Q_D : [\underline{\beta}_D, \overline{\beta}_D] \to \{0, 1\}$ to capture these quitting outcomes. For simplicity in the exposition, we restrict $Q_C(\beta_D)+Q_D(\beta_D) \leq 1$; i.e., both states cannot simultaneously quit.

[15] We cannot fully capture this simply by setting $h(\beta_D) = 0$ when a player quits, as D may incur costs of hassling prior to either side quitting.

To avoid trivialities, we restrict attention to mechanisms in which $Q_C(\beta_D) = Q_D(\beta_D) = 0$ implies $S_D(\beta_D) \in (0, 1)$; i.e., we consider a state to have "quit" if it accepts a settlement with no value. Finally, the assumption of strictly negative war payoffs implies that neither side would start a nuclear war deliberately rather than quit, so we do not incorporate this possibility into the mechanism.

The direct mechanism dictates the payoff to D type β_D of reporting type β'_D. To save space in writing out the reporting function, let $\bar{S}_D(\beta_D)$ denote the final share received by each type of D, accounting for either state quitting:

$$\bar{S}_D(\beta_D) \equiv Q_C(\beta_D) + [1 - Q_C(\beta_D) - Q_D(\beta_D)]S_D(\beta_D).$$

This gives us a reporting function of

$$\Psi_D(\beta'_D \mid \beta_D) = \bar{S}_D(\beta'_D)\beta_D - R(h(\beta'_D))n_D - \kappa_D(h(\beta'_D)).$$

[Need to look into the proofs and see whether quitting actually zeroes out the autonomous risk, because this specification of the reporting function makes it look like the autonomous risk stays no matter what]

5.1.1 Application to motivating model

Let us return to the variant of the Powell (2015) brinkmanship model portrayed in Figure 3 above. A pure strategy equilibrium of the game consists of the following quantities and functions. C selects an arms level p^*. Each type of D responds to each possible arms level p with a risk level $r^*(p \mid \beta_D)$. C decides whether to quit after observing arms and risk; let $q^*_C(p, r) \in \{0, 1\}$ be an indicator for C's choice to quit. Finally, if C does not quit, then each type of D chooses whether to do so, denoted $q^*_D(p, r \mid \beta_D) \in \{0, 1\}$.

To adapt this model to our framework, we identify the "hassling" choice as D's selected level of risk. The hassling space is thus $\mathcal{H} = [0, \bar{r}(\bar{p})]$, with associated risk function $R(h) = h$ and cost function $\kappa_D(h) = 0$ (constant). Then given any equilibrium of the game, we can define an equivalent direct mechanism as follows:

- Hassling level: Set $h(\beta_D) = r^*(p^* \mid \beta_D)$ for all β_D.
- C quit: Set $Q_C(\beta_D) = q_C^*(p^*, r^*(p^* \mid \beta_D))$ for all β_D.
- D quit: Set

  $$Q_D(\beta_D) = [1 - q_C^*(p^*, r^*(p^* \mid \beta_D))]q_D^*(p^*, r^*(p^* \mid \beta_D) \mid \beta_D)$$

  for all β_D.
- Share of prize if neither quit: Set $S_D(\beta_D) = 1 - p^*$ (constant) for all β_D.

5.1.2 Incentive compatibility and participation constraints

As in the baseline analysis, any direct mechanism corresponding to an equilibrium must satisfy the incentive compatibility condition,

$$\Psi_D(\beta_D \mid \beta_D) \geq \Psi_D(\beta_D' \mid \beta_D) \qquad \text{for all } \beta_D, \beta_D' \in [\underline{\beta}_D, \overline{\beta}_D].$$

In the context of nuclear conflict, a direct mechanism may be incentive compatible yet substantively unlikely. For example, consider a direct mechanism in which $Q_C(\beta_D) = Q_D(\beta_D) = S_D(\beta_D) = 0$ and $h(\beta_D) = \overline{h}$ for all β_D, where $R(\overline{h}) \approx 1$. This trivially satisfies incentive compatibility, but seems unlikely to describe real-world crisis bargaining, as every type of D risks a near-certain nuclear war in order to attain none of the disputed good.

To rule out this sort of implausible equilibrium outcome, we only consider direct mechanisms that satisfy "participation" constraints ensuring that each state prefers the equilibrium over quitting (and, in D's case, not hassling). For D, the participation constraint amounts to each type garnering non-negative expected utility:

$$\Psi_D(\beta_D \mid \beta_D) \geq 0 \qquad \text{for all } \beta_D. \tag{IR-D}$$

It is more complicated to define a participation constraint for C, as the information C has when they decide to quit may vary across game forms. The weakest plausible participation constraint for C is an *ex ante* constraint of non-negative expected utility:

$$\mathbb{E}\left[\bar{S}_C(\beta_D)\beta_C - R(h(\beta_D))n_C\right] \geq 0, \tag{IR-C}$$

where the expectation is taken over the prior distribution of β_D and $\bar{S}_C$ is defined analogously to $\bar{S}_D$. At the other extreme is an *ex post* condition, stating that C yields non-negative expected utility for all type realizations of D:

$$\bar{S}_C(\beta_D)\beta_C - R(h(\beta_D))n_C \geq 0 \qquad \text{for all } \beta_D. \tag{IR-C$'$}$$

5.2 General results

In one sense, reformulating the type space as the prize value rather than as D's war payoff does not change its essential structure or features. Having a higher prize value is akin to having lower war costs or hassling costs, in that it increases D's willingness to run risks in order to achieve a better result at the negotiating table. On its own, then, this reformulation of the type space should not radically change the results of the analysis.

To justify this claim more formally, we can examine the structure of D's payoff function under the modified direct mechanism. A key property of Von Neumann-Morgenstern utility functions is the invariance to affine transformations. If we divide the reporting function $\Psi(\cdot \mid \beta_D)$ by the prize value β_D, we end up with a payoff function that looks akin to the one from our baseline model:

$$\frac{\Psi(\beta_D' \mid \beta_D)}{\beta_D} = \underbrace{\bar{S}_D(\beta_D')}_{V_D(\theta)} - R(h(\beta_D')) \underbrace{\frac{n_D}{\beta_D}}_{\theta} - \underbrace{\frac{\kappa_D(h(\beta_D'))}{\beta_D}}_{K(h,\theta)}.$$

The payoff structure thus should not cause any substantive difference from our baseline contest of capabilities model. This logic is the basis of the next proposition, which states that any equilibrium of the modified model in which neither state quits has an equivalent representation in our original framework.

Proposition 6. *In the model where D's type is prize value, if $Q_C(\beta_D) = Q_D(\beta_D) = 0$ for all β_D, then the direct mechanism is isomorphic to a contest of capabilities in the baseline framework in which Equation 1 is satisfied. If additionally $\kappa_D(h(\beta_D)) = 0$ for all β_D, then it is isomorphic to a contest of nerves.*

An equilibrium of the prize-value model in which neither state ever quits is, in this sense, equivalent to a contest of capabilities that satisfies the decreasing differences condition, Equation 1. Because D would never deliberately provoke war per our assumption that $-n_D < 0$, Proposition 2 then implies that the hassling level and the probability of accidental war weakly increase with D's prize value. Moreover, if risk is generated without any direct cost to D, as in the example model based on Powell (2015), then Proposition 1 and Corollary 1 imply that D's expected utility and settlement value also increase with β_D. From this standpoint, modeling resolve as prize value simply leads to a special (and in fact relatively restrictive) case of the contest of capabilities.

However, Proposition 6 only covers the case in which neither player quits in equilibrium. This rules out certain strategic behaviors related to brinkmanship, where D is prepared to run a high risk of nuclear disaster that is not ultimately realized on the path of play because C would find it intolerable. Once we allow for the possibility of quitting, we obtain a pattern of results that is potentially distinct from the contest of capabilities. Low types of D quit, incurring no costs while receiving nothing. Medium types do not quit but also do not induce C to quit, so any risks they generate are realized on the path of play. The highest types prepare to run a high enough risk—at a strictly greater cost than all low and medium types—to induce C to quit. Essentially, D here can "outbid" C's tolerance for risk, resulting D attaining the prize.

Proposition 7. *In the model where D's type is prize value, there exist $\tilde{\beta}, \hat{\beta} \in [\underline{\beta}_D, \overline{\beta}_D]$ such that:*

(a) For all $\beta_D < \tilde{\beta}$, $Q_D(\beta_D) = 1$ and $\kappa_D(h(\beta_D)) = 0$.

(b) For all $\beta_D \in (\tilde{\beta}, \hat{\beta})$, $Q_C(\beta_D) = Q_D(\beta_D) = 0$. $h(\beta_D)$ and $S_D(\beta_D)$ are weakly increasing on this interval of types.

(c) For all $\beta_D > \hat{\beta}$, $Q_C(\beta_D) = 1$. There exists $\hat{\kappa}$ such that $\kappa_D(h(\beta_D)) < \hat{\kappa}$ for all $\beta_D < \hat{\beta}$ and $\kappa_D(h(\beta_D)) = \hat{\kappa}$ for all $\beta_D > \hat{\beta}$.

In the general setting where type is prize value and the players have the option to quit, we have a nonmonotone (∩-shaped) probability of accidental war as a function of D's type. This is true even though we saw above that the model effectively satisfies the decreasing differences condition (Equation 1), which in our baseline model implied a non-decreasing probability of accidental war among types that do not start a war deliberately.

The nonmonotonicity of the chance of nuclear accidents is a key feature of the model in Schram (2024). But Proposition 7 also shows how it cannot arise in models like our example based on Powell (2015), in which D can generate nuclear risk costlessly. Part (c) of the proposition states that the cost incurred by types that induce C to quit must be *strictly* greater than the costs incurred by types that quit or settle in equilibrium. This cannot happen in a model with costless risk generation—unless all types of D induce C to quit, in which case the probability of accidental war is constant rather than nonmonotone. The upshot is that the nature of the process that generates nuclear risk continues to matter even when we consider a different conceptualization of resolve and allow states to zero out risk by "quitting." Key relationships between resolve and the risk of nuclear accidents arise only when risk is generated by low-level policy options whose attractiveness or capability varies across Defender types.

6 Conclusion

We analyze a class of crisis bargaining models in which states may employ limited policy options short of full-scale war that nonetheless generate a risk of accidentally triggering such a war. Using a mechanism design methodology that allows us to study all equilibria of all such games, we find that their outcomes critically depend on the relationship between a state's private resolve and its willingness and/or ability to use these limited policy instruments. If there is no relationship—i.e., if a state's access to limited conflict is independent of its willingness to engage in full-scale war—then we recover the traditional brinkmanship pattern in which more resolved states engage in more risky limited conflict. However, when the marginal cost of riskier limited policies increases quickly enough with a state's resolve, there are equilibria with the opposite pattern, in which the least resolved types are the most likely to experience accidental escalation. Depending on the technology of limited conflict and its relationship with a state's war payoffs, different bargaining games may lead to completely different patterns of accidental war, even with the same underlying military fundamentals. Our results highlight the complexity of the strategic relationship between resolve, conventional capabilities, and inadvertent escalation.

References

Akçay, Erol, Adam Meirowitz, Kristopher W. Ramsay and Simon A. Levin. 2012. "Evolution of Cooperation and Skew Under Imperfect Information." *Proceedings of the National*

Academy of Sciences 109(37):14936–14941.

Ashworth, Scott and Ethan Bueno de Mesquita. 2006. "Monotone comparative statics for models of politics." *American Journal of Political Science* 50(1):214–231.

Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzky. 2020. "Deterrence with Imperfect Attribution." *American Political Science Review* 114(4):1155–1178.

Banks, Jeffrey S. 1990. "Equilibrium Behavior in Crisis Bargaining Games." *American Journal of Political Science* 34(3):599–614.

Brodie, Bernard. 1966. *Escalation and the Nuclear Option.* Princeton University Press.

Cobzaş, Ştefan, Radu Miculescu and Adriana Nicolae. 2019. *Lipschitz Functions.* Springer.

Coe, Andrew J. 2018. "Containing Rogues: A Theory of Asymmetric Arming." *Journal of Politics* 80(4):1197–1210.

Di Lonardo, Livio and Scott Tyson. 2022. "Deterrence and Preventive Sanctions." *Working Paper* .

Fey, Mark and Brenton Kenkel. 2021. "Is an Ultimatum the Last Word on Crisis Bargaining?" *Journal of Politics* 83(1):87–102.

Fey, Mark and Kristopher W. Ramsay. 2009. "Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types." *Review of Economic Design* 13(3):233.

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

Jervis, Robert. 1976. *Perception and misperception in international politics.* Princeton University Press.

Kenkel, Brenton and Peter Schram. 2024. "Uncertainty in Crisis Bargaining with Multiple Policy Options." Forthcoming in *American Journal of Political Science.*
URL: https://doi.org/10.1111/ajps.12849

Liu, Linqun. 2021. "Domestic Constraints in Crisis Bargaining." Typescript, University of Chicago.
URL: https://sites.google.com/site/liqunliu90/home

Liu, Liqun et al. 2021. Domestic Constraints in Crisis Bargaining. Technical report.

McCormack, Daniel and Henry Pascoe. 2017. "Sanctions and Preventive War." *Journal of Conflict Resolution* 61(8):1711–1739.

Milgrom, Paul and Chris Shannon. 1994. "Monotone comparative statics." *Econometrica: Journal of the Econometric Society* pp. 157–180.

Milgrom, Paul and Ilya Segal. 2002. "Envelope Theorems for Arbitrary Choice Sets." *Econometrica* 70(2):583–601.

Myerson, Roger B. 1979. "Incentive Compatibility and the Bargaining Problem." *Econometrica* 47(1):61–73.

Nalebuff, Barry. 1986. "Brinkmanship and nuclear deterrence: The neutrality of escalation." *Conflict Management and Peace Science* 9(2):19–30.

Paul, Derek, Michael D Intriligator, Paul Smoker et al. 1990. *Accidental Nuclear War: Proceedings of the Eighteenth Pugwash Workshop on Nuclear Forces*. Dundurn.

Perrow, Charles. 2011. *Normal accidents*. Princeton university press.

Posen, Barry R. 2014. *Inadvertent escalation*. Cornell University Press.

Powell, Robert. 1988. "Nuclear brinkmanship with two-sided incomplete information." *American Political Science Review* 82(1):155–178.

Powell, Robert. 1989. "Nuclear Deterrence and the Strategy of Limited Retaliation." *American Political Science Review* 83(2):503–519.

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

Powell, Robert. 2015. "Nuclear Brinkmanship, Limited War, and Military Power." *International Organization* 69(3):589–626.

Qiu, Xiaoyan. 2022. "State Support for Rebels and Interstate Bargaining." *American Journal of Political Science* .

Sagan, Scott D. 1994. "The perils of proliferation: Organization theory, deterrence theory, and the spread of nuclear weapons." *International Security* 18(4):66–107.

Schelling, Thomas C. 1966. *Arms and influence*. Yale University Press.

Schelling, Thomas C. 1980. *The strategy of conflict*. Harvard university press.

Schram, Peter. 2021. "Hassling: How States Prevent a Preventive War." *American Journal of Political Science* 65(2):294–308.

Schram, Peter. 2024. "Conflicts that Leave Something to Chance: Establishing Brinkmanship through Conventional Wars." *Working Paper* .

Schultz, Kenneth A. 2010. "The Enforcement Problem in Coercive Bargaining: Interstate Conflict over Rebel Support in Civil Wars." *International Organization* 64(2):281–312.

Snyder, Glenn H. 1965. The Balance of Power and the Balance of Terror. In *World in Crisis*, ed. Fredrick H. Hartmann. The Macmillan Company chapter 20, pp. 180–191.

Spaniel, William. 2020. “Power Transfers, Military Uncertainty, and War.” *Journal of Theoretical Politics* 32(4):538–556.

Spaniel, William and Iris Malone. 2019. “The Uncertainty Trade-off: Reexamining Opportunity Costs and War.” *International Studies Quarterly* 63(4):1025–1034.

Appendix

Contents

A Proofs of Named Results

A.1 Proof of Proposition 1

Proposition 1. *In any equilibrium of a contest of nerves, the total probability of war and the Defender's equilibrium utility weakly increase with the Defender's resolve: if* $\theta' < \theta''$*, then* $\pi(\theta')[1 - R(h(\theta'))] \geq \pi(\theta'')[1 - R(h(\theta''))]$ *and* $U(\theta') \leq U(\theta'')$*.*

Proof. We prove the result by showing that there is a payoff-equivalent direct mechanism in an ordinary crisis bargaining game that satisfies the incentive compatibility conditions of Banks (1990). For Banks (1990), a direct mechanism is defined by a function $x : \Theta \to \mathbb{R}$ giving settlement values and a function $p : \Theta \to [0, 1]$ giving the probability of war. Given such a mechanism, the expected utility to type θ for mimicking the bargaining strategy of type θ' is given by

$$\tilde{\Phi}_D(\theta' \mid \theta) = p(\theta')\theta + (1 - p(\theta'))x(\theta').$$

Now consider a direct mechanism for a contest of nerves that satisfies our incentive compatibility condition, (IC), and define the following functions:

$$x(\theta) = V_D(\theta) - \frac{\kappa_D(h(\theta))}{1 - R(h(\theta))},$$
$$p(\theta) = 1 - \pi(\theta)\left[1 - R(h(\theta))\right].$$

For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned}\tilde{\Phi}_D(\theta' \mid \theta) &= p(\theta')\theta + (1 - p(\theta'))x(\theta') \\ &= (1 - [1 - R(h(\theta'))]\, \pi(\theta'))\, \theta + \pi(\theta')\, [1 - R(h(\theta'))]\, V_D(\theta') - \pi(\theta')\kappa_D(h(\theta')) \\ &= \pi(\theta')\, [(1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - \kappa_D(h(\theta'))] + (1 - \pi(\theta'))\theta \\ &= \Phi_D(\theta' \mid \theta).\end{aligned}$$

Incentive compatibility of the original direct mechanism therefore implies incentive compatibility of the Banks (1990) mechanism (x, p). The first claim of the proposition then follows from Lemma 1 of Banks (1990), and the second follows from his Lemma 4. □

A.2 Proof of Corollary 1

Corollary 1. *In any equilibrium of a contest of nerves, if war never occurs deliberately ($\pi(\theta) = 1$ for all θ), then the probability of accidental war and the Defender's settlement value weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $R(h(\theta')) \leq R(h(\theta''))$ and $V_D(\theta') \leq V_D(\theta'')$.*

Proof. The first claim is immediate from Proposition 1, setting $\pi(\theta') = \pi(\theta'') = 1$. To prove the second claim, observe that the function $\frac{\kappa_D}{1-R}$ is weakly increasing in h, as κ_D and R are both non-decreasing in h. Following the proof of Proposition 1, Lemma 2 of Banks (1990) implies

$$V_D(\theta'') - V_D(\theta') \geq \frac{\kappa_D(h(\theta''))}{1 - R(h(\theta''))} - \frac{\kappa_D(h(\theta'))}{1 - R(h(\theta'))}.$$

Because R is strictly increasing, $R(h(\theta'')) \geq R(h(\theta'))$ implies $h(\theta'') \geq h(\theta')$, so the RHS of the above expression is non-negative. □

A.3 Proof of Proposition 2

Proposition 2. *Consider an equilibrium of a contest of capabilities.*

(a) The probability of accidental war weakly increases with the Defender's resolve if

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] < [R(h'') - R(h')](\theta'' - \theta') \quad (1)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

(b) The probability of accidental war weakly decreases with the Defender's resolve if

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] > [R(h'') - R(h')](\theta'' - \theta') \quad (2)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

Proof. We will prove the first claim; the proof of the second is analogous. Take any $\theta', \theta'' \in \Theta_1$ such that $\theta' < \theta''$, and suppose $h(\theta') > h(\theta'')$. To economize on notation in the remainder of the proof, define $h' \equiv h(\theta')$, $V' \equiv V_D(\theta')$, and $R' \equiv R(h(\theta'))$; and let h'', V'', and R'' be defined analogously. Incentive compatibility for θ' implies

$$(1-R')V' + R'\theta' - K_D(h',\theta') \geq (1-R'')V'' + R''\theta' - K_D(h'',\theta'),$$

which is equivalent to

$$(R'-R'')\theta' - [K_D(h',\theta') - K_D(h'',\theta')] \geq (1-R'')V'' - (1-R')V'.$$

Similarly, incentive compatibility for θ'' implies

$$(1-R'')V'' + R''\theta'' - K_D(h'',\theta'') \geq (1-R')V' + R'\theta'' - K_D(h',\theta''),$$

which is equivalent to

$$(1-R'')V'' - (1-R')V' \geq (R'-R'')\theta'' - [K_D(h',\theta'') - K_D(h'',\theta'')] .$$

Combining the incentive compatibility conditions and rearranging terms gives

$$[K_D(h',\theta'') - K_D(h'',\theta'')] - [K_D(h',\theta') - K_D(h'',\theta')] \geq (R'-R'')(\theta''-\theta').$$

Because $h' > h''$ and $\theta'' > \theta'$, this implies that Equation 1 does not hold. □

A.4 Proof of Proposition 3

The proof of the proposition follows a series of lemmas. The method of proof is similar to other envelope theorem analyses (e.g., in Banks 1990; Kenkel and Schram 2024).

A.4.1 Envelope theorem

Lemma A.1. *Suppose the differentiability assumptions hold. For any IC direct mechanism in which all types settle, we have*

$$U_D(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t),t)}{\partial t} \right] dt \tag{A.1}$$

for all $\theta \in \Theta$.

Proof. (IC) implies $U_D(\theta) = \sup_{\theta' \in \Theta} \Phi_D(\theta' \mid \theta)$ for all $\theta \in \Theta$. The differentiability assumptions imply that $\frac{\partial \Phi_D(\theta'|\theta)}{\partial \theta}$ exists for all $\theta, \theta' \in \Theta$. Corollary 1 of Milgrom and Segal (2002) then implies

$$\begin{aligned} U_D(\theta) &= U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left. \frac{\partial \Phi_D(\theta' \mid t)}{\partial t} \right|_{\theta'=t} dt \\ &= U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t),t)}{\partial t} \right] dt \end{aligned}$$

for all $\theta \in \Theta$, as claimed. □

A.4.2 Value of settlement

Lemma A.2. *Suppose the differentiability assumptions hold, and consider a direct mechanism in which $\pi = 1$. For all $\theta \in \Theta$, Equation A.1 is satisfied if and only if*

$$V_D(\theta) = \frac{U(\underline{\theta}) - R(h(\theta))\theta + K_D(h(\theta),\theta) + \int_{\underline{\theta}}^{\theta}\left[R(h(t)) - \frac{\partial K_D(h(t),t)}{\partial t}\right] dt}{1 - R(h(\theta))}. \tag{A.2}$$

Proof. Immediate from setting

$$(1 - R(h(\theta)))V_D(\theta) + R(h(\theta))\theta - K_D(h(\theta),\theta) = U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta}\left[R(h(t)) - \frac{\partial K_D(h(t),t)}{\partial t}\right] dt$$

and solving for $V_D(\theta)$. □

A.4.3 Global incentive compatibility

Lemma A.3. *Suppose the differentiability assumptions hold. Consider a direct mechanism in which $\pi = 1$, $V_D(\theta)$ satisfies Equation A.2 for all $\theta \in \Theta$, and h is absolutely continuous. For all $\theta \in \Theta$, $\Phi_D(\cdot \mid \theta)$ is differentiable almost everywhere, with*

$$\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} = h'(\theta')\left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'),\theta)}{\partial h} - \frac{\partial K_D(h(\theta'),\theta')}{\partial h}\right)\right] \tag{A.3}$$

at each point of differentiability.

Proof. For all $\theta, \theta' \in \Theta$,

$$\begin{aligned}\Phi_D(\theta' \mid \theta) &= (1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - K_D(h(\theta'),\theta)\\ &= (1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta' - K_D(h(\theta'),\theta')\\ &\quad + R(h(\theta'))(\theta - \theta') - [K_D(h(\theta'),\theta) - K_D(h(\theta'),\theta'\\ &= U_D(\theta') + R(h(\theta'))(\theta - \theta') - [K_D(h(\theta'),\theta) - K_D(h(\theta'),\theta')].\end{aligned}$$

As R and K_D are Lipschitz (via their continuous differentiability), absolute continuity of h implies $R(h(\cdot))$ and $K_D(h(\cdot),\cdot)$ are absolutely continuous and thus differentiable almost everywhere (Cobzaş, Miculescu and Nicolae 2019). Additionally, Lemma A.2 implies that U_D is absolutely continuous as well. Therefore, $\Phi_D(\cdot \mid \theta)$ is differentiable almost everywhere, with

$$\begin{aligned}\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= U_D'(\theta') + R'(h(\theta'))h'(\theta')(\theta - \theta') - R(h(\theta')) - \frac{\partial K_D(h(\theta'),\theta)}{\partial h}h'(\theta') \\ &\quad + \frac{\partial K_D(h(\theta'),\theta')}{\partial h}h'(\theta') + \frac{\partial K_D(h(\theta'),\theta')}{\partial \theta'} \\ &= h'(\theta')\left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'),\theta)}{\partial h} - \frac{\partial K_D(h(\theta'),\theta')}{\partial h}\right)\right]\end{aligned}$$

at each point of differentiability. □

A.4.4 Proof of proposition

Proposition 3. *Suppose the differentiability assumptions hold.*

(a) Let h^* be any absolutely continuous, weakly increasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \leq R'(h) \qquad \text{for all } h \in h^*(\Theta),\ \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(b) Let h^{**} be any absolutely continuous, weakly decreasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \geq R'(h) \qquad \text{for all } h \in h^{**}(\Theta), \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Proof. We prove the first claim; the proof of the second is analogous. Take any $V_0 \in \mathbb{R}$, set $V_D(\underline{\theta}) = V_0$,[1] and then define $V_D(\theta)$ according to Equation A.2 for all $\theta \in (\underline{\theta}, \overline{\theta}]$. For any $\theta \in \Theta$ and almost all $\theta' \in [\underline{\theta}, \theta)$, Lemma A.3 implies

$$\begin{aligned}\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= h'(\theta')\left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'),\theta)}{\partial h} - \frac{\partial K_D(h(\theta'),\theta')}{\partial h}\right)\right] \\ &= h'(\theta')\left[R'(h(\theta'))(\theta - \theta') - \int_{\theta'}^{\theta} \frac{\partial^2 K_D(h(\theta'),t)}{\partial h \partial t}\, dt\right] \\ &\geq h'(\theta')\left[R'(h(\theta'))(\theta - \theta') - \int_{\theta'}^{\theta} R'(h(\theta'))\, dt\right] \\ &= 0.\end{aligned}$$

[1] A task for future work is to identify a sharp condition under which (VA) holds as well.

Therefore, $\Phi_D(\theta \mid \theta) \geq \Phi_D(\theta' \mid \theta)$ for all $\theta' < \theta$. Similarly, for almost all $\theta' \in (\theta, \overline{\theta}]$,

$$\begin{aligned}
\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'), \theta)}{\partial h} - \frac{\partial K_D(h(\theta'), \theta')}{\partial h} \right) \right] \\
&= h'(\theta') \left[\left(\frac{\partial K_D(h(\theta'), \theta')}{\partial h} - \frac{\partial K_D(h(\theta'), \theta)}{\partial h} \right) - R'(h(\theta'))(\theta' - \theta) \right] \\
&= h'(\theta') \left[\int_\theta^{\theta'} \frac{\partial^2 K_D(h(\theta'), t)}{\partial h \partial t} \, dt - R'(h(\theta'))(\theta' - \theta) \right] \\
&\leq h'(\theta') \left[\int_\theta^{\theta'} R'(h(\theta')) \, dt - R'(h(\theta'))(\theta' - \theta) \right] \\
&= 0.
\end{aligned}$$

Therefore, $\Phi_D(\theta \mid \theta) \geq \Phi_D(\theta' \mid \theta)$ for all $\theta' > \theta$, which combined with the prior result implies that the direct mechanism satisfies (IC). □

A.5 Proof of Proposition 4

Proposition 4. *Consider an equilibrium of a contest of capabilities. Let Θ_1 denote the set of types that reach an agreement in equilibrium: $\Theta_1 \equiv \{\theta \in \Theta \mid \pi(\theta) = 1\}$.*

(a) Less-resolved Defender types reach agreement and more-resolved types deliberately choose war (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \leq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} < 1 - R(h') \tag{3}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

(b) Less-resolved Defender types deliberately choose war and more-resolved types reach agreement (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \geq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} > 1 - R(h') \tag{4}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

Proof. We prove the first claim; the proof of the second is analogous. Consider a direct mechanism that satisfies (IC) in which a less-resolved Defender type deliberately chooses war and a more-resolved type reaches agreement—i.e., $\pi(\theta') = 0$ and $\pi(\theta'') = 1$, where $\theta' < \theta''$. Incentive compatibility for θ' implies

$$\theta' \geq (1 - R(h(\theta'')))V_D(\theta'') + R(h(\theta''))\theta' - K_D(h(\theta''), \theta'),$$

while incentive compatibility for θ'' implies

$$(1 - R(h(\theta'')))V_D(\theta'') + R(h(\theta''))\theta'' - K_D(h(\theta''), \theta'') \geq \theta''.$$

Combined these imply

$$\frac{K_D(h(\theta''), \theta') - K_D(h(\theta''), \theta'')}{\theta'' - \theta'} \geq 1 - R(h(\theta'')).$$

Therefore, Equation 3 does not hold. □

A.6 Proof of Proposition 5

Proposition 5. *Suppose the linearity assumptions hold and* $r - 1 < \underline{k} \leq \overline{k} < r$.

(a) Equation 1 and Equation 3 hold.

(b) Let π^* *be any absolutely continuous, weakly decreasing function such that* $\pi^*(\theta) > 0$ *for all* $\theta \in \Theta$*, and let* h^* *be any absolutely continuous function that satisfies*

$$\frac{dh^*(\theta)}{d\theta} \geq \frac{\frac{1}{r-\underline{k}} - h^*(\theta)}{\pi^*(\theta)} \cdot \frac{d\pi^*(\theta)}{d\theta}$$

for all $\theta \in \Theta$ *at which* h^* *and* π^* *are differentiable. There is an incentive compatible direct mechanism in which* $\pi(\theta) = \pi^*(\theta)$ *and* $h(\theta) = h^*(\theta)$ *for all* $\theta \in \Theta$.

(c) Let π^{**} *be any absolutely continuous, weakly increasing function such that* $\pi^{**}(\theta) > 0$ *for all* $\theta \in \Theta$*, and let* h^{**} *be any absolutely continuous function that satisfies*

$$\frac{dh^{**}(\theta)}{d\theta} \geq \frac{\frac{1}{r-\overline{k}} - h^{**}(\theta)}{\pi^{**}(\theta)} \cdot \frac{d\pi^{**}(\theta)}{d\theta}$$

for all $\theta \in \Theta$ *at which* h^{**} *and* π^{**} *are differentiable. There is an incentive compatible direct mechanism in which* $\pi(\theta) = \pi^{**}(\theta)$ *and* $h(\theta) = h^{**}(\theta)$ *for all* $\theta \in \Theta$.

Proof. Claim (a). Under the linearity assumptions, Equation 1 is equivalent to

$$\frac{k(\theta'') - k(\theta')}{\theta'' - \theta'} < r.$$

The assumption that $k' < r$ ensures that this holds. Meanwhile, Equation 3 is equivalent to

$$\frac{k(\theta'') - k(\theta')}{\theta'' - \theta'} > r - \frac{1}{h'}.$$

Because $\max \mathcal{H} = 1$, the assumption that $k' > r - 1$ ensures that this holds for all $h' \in \mathcal{H}$.

Preliminary to claims (b) and (c). We omit the proofs that U_D must satisfy a local incentive compatibility condition given by the envelope theorem and that V_D can always be chosen to satisfy this condition given $V_D(\underline{\theta})$; these are analogous to the proofs of Lemma A.1 and Lemma A.2 above. In the general case where $\pi(\theta) \in [0,1]$, the envelope condition is

$$\begin{aligned} U_D'(\theta) &= \left.\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta}\right|_{\theta'=\theta} \\ &= \pi(\theta)\left[R(h(\theta)) - \frac{\partial K_D(h(\theta),\theta)}{\partial \theta}\right] + 1 - \pi(\theta). \end{aligned}$$

Under the linearity assumptions, this simplifies further to

$$U_D'(\theta) = \pi(\theta)h(\theta)\left[r - k'(\theta)\right] + 1 - \pi(\theta).$$

We now obtain the derivative of Φ_D with respect to the reported type θ', which will allow us to verify global incentive compatibility. For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned} \Phi_D(\theta' \mid \theta) &= \pi(\theta')\left[(1 - rh(\theta'))V_D(\theta') + rh(\theta')\theta - k(\theta)h(\theta')\right] + (1 - \pi(\theta'))\theta \\ &= U_D(\theta') + [1 - \pi(\theta') + r\pi(\theta')h(\theta')](\theta - \theta') - \pi(\theta')h(\theta')[k(\theta) - k(\theta')]. \end{aligned}$$

Let h and π be absolutely continuous. $\Phi_D(\cdot \mid \theta)$ is absolutely continuous and thus differentiable almost everywhere, per the same argument as in the proof of Lemma A.3 above. At each point of differentiability,

$$\begin{aligned} \frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= 1 - \pi(\theta') + r\pi(\theta')h(\theta') - \pi(\theta')h(\theta')k'(\theta') \\ &\quad + [r(\pi h)'(\theta') - \pi'(\theta')](\theta - \theta') - [1 - \pi(\theta') + r\pi(\theta')h(\theta')] \\ &\quad - (\pi h)'(\theta')[k(\theta) - k(\theta')] + \pi(\theta')h(\theta')k'(\theta') \\ &= [r(\pi h)'(\theta') - \pi'(\theta')](\theta - \theta') - (\pi h)'(\theta')[k(\theta) - k(\theta')] \\ &= (\pi h)'(\theta')\left[r(\theta - \theta') - k(\theta) + k(\theta')\right] - \pi'(\theta')(\theta - \theta') \\ &= \left((\pi h)'(\theta')\left[r - \frac{k(\theta) - k(\theta')}{\theta - \theta'}\right] - \pi'(\theta')\right)(\theta - \theta'). \end{aligned} \tag{A.4}$$

Note that $r - \frac{k(\theta)-k(\theta')}{\theta-\theta'} \in [r - \overline{k}, r - \underline{k}] \subseteq (0,1)$ for all distinct $\theta, \theta' \in \Theta$.

Claim (b). Suppose π is weakly decreasing and that $h'(\theta') \geq \left[\frac{1}{r-\underline{k}} - h(\theta')\right]\frac{\pi'(\theta')}{\pi(\theta')}$ for all θ' at which h and π are differentiable. Because $\pi'(\theta') \leq 0$, this implies that for all $\theta \in \Theta$,

$$(\pi h)'(\theta') = \pi'(\theta')h(\theta') + \pi(\theta')h'(\theta') \geq \frac{\pi'(\theta')}{r - \underline{k}} \geq \frac{\pi'(\theta')}{r - \frac{k(\theta)-k(\theta')}{\theta-\theta'}},$$

so the first term in parentheses in Equation A.4 is non-negative. Therefore, for all $\theta, \theta' \in \Theta$ such that $\Phi_D(\theta' \mid \theta)$ is differentiable in θ', we have that $\theta > \theta'$ implies $\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} \geq 0$ and $\theta < \theta'$ implies $\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} \leq 0$. Because $\Phi_D(\cdot \mid \theta)$ is absolutely continuous, this implies that the direct mechanism satisfies (IC).

Claim (c). Suppose π is weakly increasing and that $h'(\theta') \geq \left[\frac{1}{r-\bar{k}} - h(\theta')\right] \frac{\pi'(\theta')}{\pi(\theta')}$ for all θ' at which h and π are differentiable. Because $\pi'(\theta') \geq 0$, this implies that for all $\theta \in \Theta$,

$$(\pi h)'(\theta') = \pi'(\theta')h(\theta') + \pi(\theta')h'(\theta') \geq \frac{\pi'(\theta')}{r-\bar{k}} \geq \frac{\pi'(\theta')}{r - \frac{k(\theta)-k(\theta')}{\theta-\theta'}}.$$

The first term in parentheses in Equation A.4 is again non-negative, and the proof of incentive compatibility follows as in the last step. □

A.7 Proof of Proposition 6

Proposition 6. *In the model where D's type is prize value, if $Q_C(\beta_D) = Q_D(\beta_D) = 0$ for all β_D, then the direct mechanism is isomorphic to a contest of capabilities in the baseline framework in which Equation 1 is satisfied. If additionally $\kappa_D(h(\beta_D)) = 0$ for all β_D, then it is isomorphic to a contest of nerves.*

Proof. Take an incentive compatible mechanism that satisfies (IR-D). Define a corresponding contest of capabilities as follows:

- Map prize-value types β_D into war-payoff types $\theta(\beta_D)$ as follows:

$$\theta(\beta_D) = \frac{-n_D}{\beta_D}.$$

  Because $-n_D < 0$, this is a strictly increasing (and thus invertible) function.

- Define the hassling cost function $K_D(h, \theta)$ as

$$K_D(h, \theta) = \frac{\theta \kappa_D(h)}{-n_D}.$$

  Notice that this is weakly decreasing in θ for $\theta < 0$, so Equation 1 is satisfied. It is constant in θ if $\kappa_D = 0$, in which case the transformed model is a contest of nerves per our definition.

Now define a direct mechanism in the baseline framework where $\tilde{\pi}(\theta) = 1$, $\tilde{h}(\theta) = h(-\frac{n_D}{\theta})$, and $\tilde{V}_D(\theta) = S_D(-\frac{n_D}{\theta})$. Observe that $\tilde{\Phi}_D(\cdot \mid \theta(\beta_D)) \propto \Psi_D(\cdot \mid \beta_D)$ for all β_D: for any

$\theta' \in [\theta(\underline{\beta}_D), \theta(\overline{\beta}_D)]$,

$$\begin{aligned}
\tilde{\Phi}_D(\theta' \mid \theta(\beta_D)) &= (1 - R(\tilde{h}(\theta')))\tilde{V}_D(\theta') + R(\tilde{h}(\theta'))\theta(\beta_D) - K_D(\tilde{h}(\theta'), \theta(\beta_D)) \\
&= (1 - R(\tilde{h}(\theta')))\tilde{V}_D(\theta') + \left[R(\tilde{h}(\theta')) + \frac{\kappa_D(\tilde{h}(\theta'))}{n_D}\right]\theta(\beta_D) \\
&= (1 - R(h(\beta'_D)))S_D(\beta'_D) + \left[R(h(\beta'_D)) + \frac{\kappa_D(h(\beta'_D))}{n_D}\right] \cdot \frac{-n_D}{\beta_D} \\
&= (1 - R(h(\beta'_D)))S_D(\beta'_D) - \frac{R(h(\beta'_D))n_D}{\beta_D} - \frac{\kappa_D(h(\beta'_D))}{\beta_D} \\
&= \frac{\Psi_D(\beta'_D \mid \beta_D)}{\beta_D}
\end{aligned}$$

where $\beta'_D \equiv -\frac{n_D}{\theta'}$. Incentive compatibility of (Q_C, Q_D, h, S_D) therefore implies incentive compatibility of $(\tilde{\pi}, \tilde{h}, \tilde{V}_D)$. Additionally, because $\theta(\beta_D) < 0$ for all β_D, the latter mechanism trivially satisfies voluntary agreements because the former satisfies (IR-D). □

A.8 Proof of Proposition 7

Proposition 7. *In the model where D's type is prize value, there exist $\tilde{\beta}, \hat{\beta} \in [\underline{\beta}_D, \overline{\beta}_D]$ such that:*

(a) For all $\beta_D < \tilde{\beta}$, $Q_D(\beta_D) = 1$ and $\kappa_D(h(\beta_D)) = 0$.

(b) For all $\beta_D \in (\tilde{\beta}, \hat{\beta})$, $Q_C(\beta_D) = Q_D(\beta_D) = 0$. $h(\beta_D)$ and $S_D(\beta_D)$ are weakly increasing on this interval of types.

(c) For all $\beta_D > \hat{\beta}$, $Q_C(\beta_D) = 1$. There exists $\hat{\kappa}$ such that $\kappa_D(h(\beta_D)) < \hat{\kappa}$ for all $\beta_D < \hat{\beta}$ and $\kappa_D(h(\beta_D)) = \hat{\kappa}$ for all $\beta_D > \hat{\beta}$.

Proof. First, consider types β'_D, β''_D such that $Q_D(\beta'_D) = 1$ and $Q_C(\beta''_D) = Q_D(\beta''_D) = 0$. Incentive compatibility for each of these types implies

$$S_D(\beta''_D)\beta''_D \geq R(h(\beta''_D))n_D + \kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \geq S_D(\beta''_D)\beta'_D,$$

which in turn implies $\beta''_D > \beta'_D$. Now consider a third type β'''_D such that $Q_C(\beta'''_D) = 1$. Incentive compatibility for β''_D and β'''_D implies

$$[1 - S_D(\beta''_D)]\beta'''_D \geq \kappa_D(h(\beta'''_D)) - \kappa_D(h(\beta''_D)) - R(h(\beta''_D))n_D \geq [1 - S_D(\beta''_D)]\beta''_D,$$

which in turn implies $\beta'''_D \geq \beta''_D$. This proves that the type space can be partitioned into (potentially empty) intervals in which the lowest types of D quit, the highest types induce C to quit, and in between neither state quits.

The claim in (a) that $Q_D(\beta_D) = 1$ implies $\kappa_D(h(\beta_D)) = 0$ is immediate from (IR-D).

To prove the claim in (b) that h and S_D are weakly increasing, consider types $\beta'_D, \beta''_D \in (\tilde{\beta}, \hat{\beta})$ such that $h(\beta'_D) < h(\beta''_D)$. Incentive compatibility for both types implies

$$\begin{aligned}
&[S_D(\beta''_D) - S_D(\beta'_D)]\beta''_D \\
&\qquad \geq [R(h(\beta''_D)) - R(h(\beta'_D))]n_D + \kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \\
&\qquad\qquad \geq [S_D(\beta''_D) - S_D(\beta'_D)]\beta'_D.
\end{aligned}$$

$h(\beta'_D) < h(\beta''_D)$ implies that the middle term of the above expression is strictly positive, so the first inequality implies $S_D(\beta''_D) > S_D(\beta'_D)$. This in turn implies $\beta''_D > \beta'_D$.

Finally, to prove the claims about $\hat{\kappa}$ in (c), consider types $\beta'_D, \beta''_D > \hat{\beta}$. Incentive compatibility for these two types implies

$$\kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \geq 0 \geq \kappa_D(h(\beta'_D)) - \kappa_D(h(\beta''_D)),$$

so $\kappa_D(h(\beta'_D)) = \kappa_D(h(\beta''_D)) = \hat{\kappa}$. Additionally, incentive compatibility for any $\beta_D < \hat{\beta}$ implies

$$S_D(\beta_D)\beta_D - R(h(\beta_D))n_D - \kappa_D(h(\beta_D)) \geq \beta_D - \hat{\kappa}.$$

Because $S_D(\beta_D) < 1$, this implies $\hat{\kappa} > \kappa_D(h(\beta_D))$. □