Journal of Conflict Resolution

The Shadow of Deterrence: Why capable actors engage in contests short of war

Journal:	*Journal of Conflict Resolution*
Manuscript ID	Draft
Manuscript Type:	Original Manuscript
Keywords:	bargaining, conflict, game theory, international alliance, international security, militarized disputes, militarized interstate disputes
Field:	political science, other
Abstract:	Recent trends increasingly place conflict in a ``gray zone'' between peace and war. Observers interpret gray zone conflicts as deterrence failures. New technologies or tactics---from cyber operations to ``little green men''---reduce the costs or increase the effectiveness of low-intensity aggression. But gray zone conflict could also reflect deterrence success. Credible prospects of retaliation encourage challengers to adopt less effective means of aggression. These dueling ``push-pull'' logics suggest contrasting conflict dynamics impacting stability. We develop a formal model that synthesizes both perspectives by analyzing deterrence success as variable, rather than dichotomous. In the model, the intensity of a challenger's provocation varies inversely with the credibility of the defender's deterrent threat. We empirically analyze Russian gray zone activity since the 1990s. Russia is more restrained, and less effective, against nations in or closer to NATO. The model suggests inherent trade-offs between stability and military potency in limiting the risk of escalation.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
05b_Model_Results.Rmd	

SCHOLARONE™
Manuscripts

longtable

For Peer Review

The Shadow of Deterrence: Why capable actors engage in contests short of war

Anonymized authors

May 25, 2021

Abstract

Recent trends increasingly place conflict in a "gray zone" between peace and war. Observers interpret gray zone conflicts as deterrence failures. New technologies or tactics—from cyber operations to "little green men"—reduce the costs or increase the effectiveness of low-intensity aggression. But gray zone conflict could also reflect deterrence success. Credible prospects of retaliation encourage challengers to adopt less effective means of aggression. These dueling "push-pull" logics suggest contrasting conflict dynamics impacting stability. We develop a formal model that synthesizes both perspectives by analyzing deterrence success as variable, rather than dichotomous. In the model, the intensity of a challenger's provocation varies inversely with the credibility of the defender's deterrent threat. We empirically analyze Russian gray zone activity since the 1990s. Russia is more restrained, and less effective, against nations in or closer to NATO. The model suggests inherent trade-offs between stability and military potency in limiting the risk of escalation.

For Peer Review

1 Introduction

In the wake of the downfall of Ukrainian President Viktor Yanukovych in February 2014, the Crimean Peninsula was invaded by "little green men," soldiers whose uniforms lacked insignia or other identifying information. The Kremlin formally annexed Crimea shortly thereafter. Protracted ground skirmishes, cyber campaigns, and "active measures" continue to plague Ukraine. Russia's intervention has emerged as a paradigmatic example of a technologically novel and politically efficient form of "hybrid warfare," designed to challenge the status quo without triggering broader conflict (Marten 2015; Lanoszka 2016). Similar tactics and imagery have emerged elsewhere, like Chinese "little blue men" eroding "red lines" in maritime East Asia (Green et al. 2017). According to former British Defense Secretary Michael Fallon (2017), "That is not a Cold War. It is a grey war. Permanently teetering on the edge of outright hostility. Persistently hovering around the threshold of what we would normally consider acts of war."

There is increasing concern that conventional conceptions of deterrence are inadequate to address burgeoning

threats in the gray zone (Holmes and Yoshihara 2017; Matisek 2017; Hicks and Friend 2019; Pettyjohn and Wasser 2019). Deterrence is typically believed to have failed if a challenger disrupts the status quo or resorts to military violence. Capable defenders, in turn, appear ill-equipped to respond to revisionism in the gray zone (Jackson 2017). As General Dunford (2016), Chairman of the United States Joint Chiefs of Staff, commented, "Our traditional approach is either we're at peace or at conflict. And I think that's insufficient to deal with the actors that actually seek to advance their interests while avoiding our strengths."

Rather than repudiating deterrence, however, gray zone challenges could actually reflect a certain respect for existing frameworks, both to avoid major war and preserve interests in common with status quo powers. The challenge of gray zone conflict may have more to do with how we think about deterrence, rather than how it is practiced. Deterrence not only shapes *whether* a challenge emerges but also *how*. Shaping other nations' behaviors could certainly prove valuable, even in the absence of peace and a full retention of the status quo. An enemy that engages "with one hand tied behind its back" to avoid triggering a larger contest is not fighting as effectively. Even if the challenger resorts to force and the defender does not intervene, fear of subsequent intervention by the defender could cause the challenger to adopt a more furtive (indecisive) military strategy.

The research on deterrence is vast and its theories heterogeneous (Huth 1999; Freedman 2004; Danilovic and Clare 2010; Quackenbush 2011). Yet there is relatively little work on choices about the *means* of deterrence compared to the tremendous literature on political *ends* (Carcelli and Gartzke 2017). We contribute to this emerging interest in the modality of deterrence across different "domains" by analyzing the drivers of different types of activity. We develop a formal model where a challenger and defender can select from variable intensities of limited conflict or choose to resolve the crisis with a decisive war. This model allows us to differentiate and assess two distinct causal logics for gray zone conflict, one driven by military innovation (i.e., reduced costs or more effective aggression) and the other by deterrence. We show that the challenger's choice varies with both the level of the deterrent threat posed by the defender (deterrence) and the challenger's ability to gray zone efforts at low cost (innovation). We then test key results of the model, finding empirical evidence that the magnitude and risk of NATO intervention shapes recent Russian uses of force. We also find evidence that Russia moderates the intensity of its efforts in response to the implicit credibility of the West's deterrence posture along an East-West gradient. These outcomes suggest Russian gray zone activity is predicted by deterrence, rather than only the logic of military innovation.

In the sections that follow, we first locate gray zone conflict in the broader literature on limited war (Section 2). We then analyze limited conflict using a formal model to illustrate the trade-offs that states face in deciding to enter into gray zone conflict or to go to war (Sections 3-5). Next, we assess our argument empirically, using

For Peer Review

data on Russian aggression, with an emphasis on cyber-enabled operations (Section 6) typically highlighted as evidence for the efficiency logic of gray zone conflict. We then revisit the game theoretical results to discuss the implications of our findings for critical deterrence and defense trade-offs (Section 7). Section 8 concludes.

2 Between Peace and War

Despite the analytical convenience offered by conceptualizing war and peace as discrete outcomes, there is nothing new about conflict that falls ambiguously between peace and war (Lebow 2010). There is a long history of, and a vast literature on, limited war (Kissinger 1955; Osgood 1969), salami slicing (Schelling 1966), low-intensity conflict (Turbiville 2002), hybrid wars (Lanoszka 2016), frozen conflict (Driscoll and Maliniak 2016), covert operations (Carson 2018; O'Rourke 2018), and hassling (Schram 2021).

Early Cold War writings on "weakening the enemy with pricks instead of blows" emphasized limited political objectives in the shadow of nuclear escalation (Hart 1954, 186). The Korean War seemed a then-underappreciated type of war fought to achieve political ends short of total victory with military means short of a total commitment (Osgood 1969). Contemporary treatments understood limited war as a conflict between actors that had the capacity to increase battlefield commitments but did not want to do so, creating a third option between major war and acquiescence (Brodie 1957; Kissinger 1957). Strategists introduced the "stability-instability paradox" to describe how disincentives for nuclear war, or even major conventional war, encourage conflict at lower levels of intensity, or in peripheral theaters (Jervis 1984; Sagan and Waltz 2003). There is some threshold above which any given threat becomes too costly to be both credible and effective, and non-credibility invites challenges. As Snyder (1965, 167) observes, nuclear weapons introduced a new uncertainty "concerning what types of military capability the opponent was likely to use and what degree of violence he was willing to risk or accept." Similarly, Powell (2015, 598) notes that "how much power the challenger brings to bear limits how much risk the defender can generate." As George and Smoke (1989, 173) explain, adversaries can "design around" deterrence by discovering new options that offer "an opportunity for gain while minimizing the risk of an unwanted response".

Yet, other wars are limited not by the risk of escalation, but rather by cost concerns. During the Cold War there were numerous decolonization struggles and proxy wars in the developing world. In these "low-intensity conflicts", the immediate adversaries tended to be irregular guerrilla forces rather than peer competitors (Galula 1964; Taber 1965). An insurgent might give all, but still not be able to give much. Guerrillas with rudimentary arsenals simply could not directly engage powerful security forces, and thus opted for indirect ambush and subversion as a matter of necessity. After the Cold War, as great power competition waned

For Peer Review

and the United States became embroiled in occupations abroad, there was a revival of interest in questions of counterterrorism and counterinsurgency (Nagl 2005; Kilcullen 2010). Yet a common theme involves the limited military capacity of at least one of the combatants. Asymmetric contests thus contrast starkly with a superpower opting to forego military effectiveness to control escalation.

The renewal of interest in low-intensity conflict between more capable competitors represents a return to the earlier theme. A common thread in definitions of gray zone conflict is that it involves "a carefully planned campaign operating in the space between traditional diplomacy and overt military aggression" employed by challengers with grand geopolitical ambitions and potent capabilities (Mazarr 2015b).[1] A number of practitioners and journalists highlight the worrying expansion of technologies by which low-intensity conflict can be practiced (Olson 2015).

Scholars similarly note the technological novelty of this phenomenon as a form of aggression against which the US is unprepared (Hoffman 2007; Thornton 2015; Brands 2016; Jackson 2016; Wirtz 2017; Hughes 2020). This pessimism concerning adversaries' ability to overthrow the existing military balance with "innovative doctrines or cunning strategies" (Goldstein 2017, 906) has even led some to advocate revamping deterrence to focus on threats from the gray zone (Tor 2015; Matisek 2017; Hicks and Friend 2019).

Even sceptics of cyber warfare highlight the expanded repertoire of means available for low-intensity conflict, especially online espionage and disinformation (Rid 2013; Jensen, Valeriano, and Maness 2019). Russia, and its intervention in Ukraine in particular, is paradigmatic. Russia uses novel forms of "hybrid warfare" to facilitate increased aggression against NATO and the West (Marten 2015; Lanoszka 2016). If technological advances drive gray zone conflict, then we might expect to see Russia engaging in it as often as possible. Covert cyber campaigns are argued prevalent because "deterrence capabilities would be rendered useless" (Farmanfarmaian 2021, 43). The common refrain is that aggressors are working around defenders' red lines to achieve coercive success without triggering escalation (Altman 2018). The policy prescription emerging from this conventional wisdom is that the United States—and other targets of gray zone aggression—must use more of the tools available at its disposable to re-assert deterrence and counter innovations in this unconventional battle space (Hicks and Friend 2019; McCarthy, Moyer, and Venable 2019).

We wish to push back on this growing consensus that an expanded repertoire of potent tools for engaging in conflict short of war means that gray zone conflict represents the new modal form of conflict. Indeed, militaries have been innovating novel technologies in *all* domains of warfare, at *all* levels of intensity. Many of the sharpest arrows remain in the quiver in most conflicts. Thus the choice to use only some of these

[1]There is much ambiguity concerning how practitioners define and interpret gray zone conflict. See Bragg (2017) and Janičatová and Mlejnková (2021).

For Peer Review

innovations, but not others, really represents a restriction, rather than an expansion, of the total means available. This shift in perspective, viewing gray zone conflict as a relative reduction rather than an absolute expansion of options, calls into question claims about its effectiveness, and even its novelty, below the threshold of war. On the contrary, the familiar logic of the stability-instability paradox may be playing out today at different, and usually lower, thresholds (Lindsay and Gartzke 2018). In its classic formulation, nuclear deterrence paradoxically encourages limited conventional war in peripheral regions where nuclear escalation is not credible. Today, the prospect of costly conventional war encourages provocation in cyberspace. The bad news about persistent conflict may thus be good news about restraint.

In the last decade of the Cold War, US Secretary of State George Shultz (1986), 204 offered a note of cautious optimism in this regard:

> The ironic fact is, these new and elusive challenges have proliferated, in part, because of our success in deterring nuclear and conventional war. Our adversaries know they cannot prevail against us in either type of war. So they have done the logical thing: they have turned to other methods. Low-intensity warfare is their answer to our conventional and nuclear strength a flanking maneuver, in military terms.

Below we develop Shultz's insight to explore whether modern gray zone conflict should be viewed as similarly reassuring or newly alarming.

3 Theoretical Intuition

The Euromaidan Revolution presented Russia with a new political reality: Kiev was abruptly realigning with the West. Almost as quickly, Russia set about altering this new "status quo."[2] Russia's actions in Ukraine were limited and, because Russia stands to benefit from continued access to Crimea and from instability in Ukraine, politically advantageous. We refer to this behavior as conducting "limited challenges," implicitly to the status quo. In response to Russian activity, NATO increased its presence in the Black Sea, reinforced its support for capacity-building in Ukraine, and has stepped up its presence and cooperation in other countries in Eastern Europe. This has helped Ukraine in countering the Russian-sponsored separatist movement in the East of the country.

The model captures these strategic dynamics. In the model, a challenger and a defender, experience a crisis. First, the challenger decides whether to accept the status quo and do nothing, conduct limited challenges,

[2]The notion of "status quo" is inherently contextual, and thus subject to perspective and interpretation. Here, we treat the status quo as whatever political conditions prevail at a given point in time, regardless of past changes.

For Peer Review

or escalate to war. If the challenger conducts a limited challenge, then the defender has an opportunity to respond with acceptance, escalation to war, or countering the challenge with the defender's own limited conflict. When the challenger engages in limited activity and the defender responds with a limited response (rather than accepting or going to war), we will say that the states are engaged in a gray zone conflict. Thus, gray zone conflict occurs when militarily capable challengers intentionally limit the intensity and capacity with which they conduct military operations, and the defender engages but chooses not to escalate to a decisive war. Narrowing the scope of gray zone conflict to militarily capable actors distinguishes the phenomenon from wars limited by means.[3] Our definition means that gray zone conflict must be preferred by both sides in a contest. Both actors have the capacity to escalate to a larger war but prefer not to, meaning that gray zone conflict is an equilibrium (Carson 2016, 2018). This treatment allows us to formally examine the phenomena and take hypotheses to data.

Our model highlights the trade-offs confronting actors when choosing varying degrees of force. An actor may forgo the most effective or decisive means when they are too costly, they are insufficiently committed, or some other more appealing alternative exists. While war can accomplish an aggressor's goals, it could also be unnecessary and inefficient if partial victories or *faits accomplis* can be achieved at lower cost (Altman 2018). By considering these trade-offs, our model offers insight into two central questions. First, why do states engage in gray zone conflict and, once they do, what explains variation in the intensity with which they pursue it? Second, when a defender faces the prospect of limited challenges, when does restraint benefit the defender?

Why do states engage in gray zone conflict? Challengers that attempt to alter the status quo through limited challenges must decide how aggressive to be. A more intensive challenge can yield greater political gains. But the challenger faces two possible constraints into selecting limited challenges: an *external deterrent constraint*, and an *internal efficiency constraint*. If the *external deterrent constraint* binds, the challenger is resolved, but will limit the level of force to attempt to avoid a greater war. Thus, when the deterrent constraint binds, the challenger is most reactive to the defender's willingness to go to war. Alternatively, if the *internal efficiency constraint* binds, the challenger selects relatively low levels of force based on its own internal cost-benefit calculations. With the internal constraint, the challenger is most sensitive to its own valuation of the stakes versus the costs of challenging the status quo.

The theoretical distinction here is empirically consequential. Contrary to conventional wisdom,[4] we show

[3]Many gray zone challenges involve a capable foreign patron and a limited capability proxy force (Plana 2020). Admittedly, reliance on proxies can complicate the deterrence calculus by obscuring attribution (Danilovic 2001). For analytical parsimony, we consider a target's allies as part of the targets capabilities, discounted by the level of commitment (or disunity) in an alliance (Quackenbush 2006; Sobek and Clare 2013).

[4]See Tor (2015), Brands (2016), Goldstein (2017) and many others cited earlier.

For Peer Review

that gray zone conflict is not the product of a newly expansive and potent military repertoire. Rather, it results from one of two distinct constraints on the military force a state can employ. When the challenger's internal efficiency constraint binds, the challenger freely chooses the scope and intensity of conflict that it believes is most efficient for accomplishing its objective. Here the challenger can pursue the optimal challenge without concern of provoking an escalation. When the challenger's external deterrent constraint binds, by contrast, the challenger must scale back from its optimal low-level challenge in order to avoid triggering a larger contest with the defender. In the latter case, the defender's willingness to go to war constrains the challenger to select from a set of less effective gray zone options. A key empirical implication is that the intensity of gray zone conflict limited by deterrence should vary inversely with the credibility of the defender's deterrent posture, where the defender is more willing to absorb the costs of war. We label this moderating effect the defender's "deterrence gradient," encouraging greater provocations in areas where the defender is less resolved but more limited efforts where it is more so.

When is a restrained ability to counter gray zone conflict opportunistic? Within the game, we will assume that the defender incurs "gray zone costs" to countering limited challenges. A defender with low gray zone costs is very effective at engaging threats in the gray zone; thus, the defender having lower gray zone costs makes gray zone conflict less efficient for the challenger, which can encourage the challenger to pursue options outside of gray zone conflict. This can produce mixed results for the defender state. On one hand, lower gray zone costs for the defender could be productive for the defender if, upon the challenger's low-level activity becoming ineffective, the challenger abandons the use of force altogether and accepts the status quo. On the other hand, lower gray zone costs for the defender could be counterproductive for the defender if, upon the challenger's low-level activity becoming ineffective, the challenger instead escalates to war. Ultimately, the challenger's response to changes in the defender's costs and ability to conduct gray zone conflict will be arbitrated by whether the challenger prefers war to the status quo (or vice-versa), which is a function of the resolve of the challenger.

The model integrates some features from the formal literature on endogenous power shifts and deterrence (Fearon 1997; Schultz 2010; Debs and Monteiro 2014; Gurantz and Hirsch 2017; Baliga, Mesquita, and Wolitzky 2020), specifically, a back-and-forth set of decisions between a challenger and a defender who select some form of conflict and then escalate (or not). While this paper makes simplifying assumptions regarding private information and hidden actions, it extends existing research by considering two competing actors, each with a continuum of policy options outside of declaring war or accepting peace. Far from just a technical flourish, this is critical for our results concerning the efficacy of a restrained ability to counter gray zone conflict. The model is thus most similar to those where politicians have flexible responses rather than just

For Peer Review

war or peace (Zagare and Kilgour 1998; Schultz 2010; Slantchev 2011; Powell 2015; McCormack and Pascoe 2017; Coe 2018; Spaniel 2019; Joseph 2021; Schram 2021), but is still unique in that it allows both states to select from a continuum of low-level options. Essentially, we push the decision-theoretic strands of classical deterrence theory towards embracing a broader repertoire of possible actions (Zagare 1996; Huth 1999).

4 Model and Equilibrium

4.1 Game Form

Two states, a challenger C and defender D, are in a crisis over a divisible asset with a normalized value of 1. At the onset, the states are presented with a "status quo" policy, and then they decide whether and how to escalate. As an intuition, the Euromaidan protests and overthrow of Yanukovych presented Russia with a new political reality, which can be thought of as the status quo policy in this model. Alternatively, the status quo policy could be viewed as an offered policy that arose through a bargaining process where a bargaining failure may have occurred.[5]

C moves first.[6] C either goes to war by setting $w_C = 1$, or sets $w_C = 0$. If C goes to war, the game terminates. If C sets $w_C = 0$, C also selects $g_C \in \mathcal{G}_C = \mathbb{R}_{\geq 0}$, where $g_C = 0$ is walking away from the crisis and accepting the status quo, and $g_C > 0$ is conducting some limited, costly military action that shifts the political status quo in favor of the challenger. Second, as long as C did not previously go to war, D can either escalate to war by setting $w_D = 1$, or not by setting $w_D = 0$ and selecting some gray zone response $g_D \in \mathcal{G}_D = \mathbb{R}_{\geq 0}$, with $g_D = 0$ implying that D does not respond to the limited challenge. When D selects a gray zone response $g_D > 0$, D is using its own costly military means to weaken the impact of C's limited challenge. After the challenger acts and the defender responds, the game terminates, and payoffs are realized. For convenience, payoffs are summarized in Table 1.

If C goes to war at the outset (setting $w_C = 1$), C and D receive expected payoffs of $U_C = \theta\rho_W - \kappa_C$ and $U_D = 1 - \rho_W - \kappa_D$, respectively. These payoffs are largely consistent with the treatment of war as a costly lottery (Fearon 1995, 1997). $\kappa_C > 0$ and $\kappa_D > 0$ are C's and D's costs from war. $\rho_W \in [0, 1]$ is C's likelihood

[5] We do not wish to discount how bargaining theory has revolutionized scholarship on conflict (Fearon 1995; Wagner 2000; Filson and Werner 2002; Smith and Stam 2004; Powell 2006; Fey and Ramsay 2011; Ramsay 2017; Spaniel 2019). While bargaining is relevant to the conflict process, we make our main point using a simpler model. In the appendix we offer a microfoundation for a (potential) bargaining failure by analyzing a model with information asymmetry and an endogenous ultimatum offer. The results and comparative statics are similar.

[6] An alternate game structure—where C and D simultaneously select a level of gray zone force and either actor can unilaterally initiate a war— yields an identical equilibrium. By virtue of the selected gray zone cost structures, there is no crowding-out like one could expect in a von Stackelberg duopoly game. If we assumed decreasing marginal costs to gray zone conflict over some set of possible gray zone challenges, the sequence of moves could matter and crowding out could occur.

For Peer Review

of winning in a war. The $\theta > 0$ term represents C's "resolve," or how much C cares about the asset in dispute.

If C sets $w_C = 0$ and $g_C = 0$, and D sets $w_D = 0$ and $g_D = 0$, this is equivalent to both states accepting the status quo $\rho_0 \in [0, 1]$, and C and D receive payoffs $U_C = \theta\rho_0$ and $U_D = 1 - \rho_0$. We assume that $\rho_0 < \rho_W$, which implies that C is potentially dissatisfied with the status quo, and for a great enough resolve (θ), or low enough costs of war (κ_C), C will choose to go to war.

Now consider all outcomes where war does not occur (including the status quo), which is when C sets $w_C = 0$ and $g_C \geq 0$, and D sets $w_D = 0$ and $g_D \geq 0$. Here C and D receive payoffs $U_C = \theta P(g_C, g_D) - \beta_C g_C^2$ and $U_D = 1 - P(g_C, g_D) - \beta_D g_D^2$. The function P represents the political outcome of gray zone conflict, and we assume functional form $P(g_C, g_D) = max\{min\{\rho_W, \rho_0 + g_C - g_D\}, \rho_0\}$. This functional form implies that P is weakly increasing in g_C and $-g_D$, and that P falls between ρ_0 and ρ_W inclusive. When $g_C = 0$ and $g_D = 0$, actors receive their status quo payoffs. By challenging the status quo ($g_C > 0$), C is shifting the status quo in its favor, but this can be dampened by D's selection of a gray zone response (when $g_D > 0$). That the final political settlement $P(g_C, g_D)$ must be weakly greater than the status quo ρ_0 and weakly less than ρ_W captures that gray zone challenges are limited; we are assuming that gray zone conflict cannot push C's final share of the asset to a level below what C is expected to attain from the status quo or above what C is expected to attain from war.[7] How C internalizes the final political outcome will depend on C's resolve, hence C's utility function having the $\theta P(g_C, g_D)$ term. Gray zone conflict is also costly to both actors. C pays costs $-\beta_C g_C^2$ for challenging and D pays $-\beta_D g_D^2$ for its gray zone conflict response, with $\beta_C > 0$ and $\beta_D > 0$.[8] When β_D (β_C) is high, then D's (C's) costs of gray zone conflict are greater. That each actor has their own gray zone conflict cost functions distinct from their costs of war is consistent with C's gray zone challenge being a limited action that relies on its own set of technologies rather than just a kind of lesser-war.[9]

Finally, when D initiates war after C engages in limited challenges (formally when $w_C = 0$, $g_C \geq 0$, and $w_D = 1$), C and D receive payoffs $U_C = \theta\rho_W - \kappa_C - \beta_C g_C^2$ and $U_D = 1 - \rho_W - \kappa_D$. Note that here C pays the costs of war as well as the costs of the limited challenge. Practically speaking, what is conducted for the purposes of gray zone conflict is different than what is conducted in a war, which makes the effort undertaken during gray zone conflict produce additional costs. Formally, these results would not change so long that a

[7] Challengers often do worse in wars than the status quo, but adverse outcomes are much less likely with a limited disputes. In part, this is a rationale for our assumption that gray zone contests are less costly than major war.

[8] We assume a simple quadratic loss function rather than an unspecified loss function to allow for explicit solutions throughout.

[9] We do not model gray zone conflict and war as part of one continuous form of conflict. The continuous-conflict-choice approach would be appropriate if gray zone conflict looked just like war, only a lesser degree of war. Yet gray zone conflict and war generally rely on different technologies. For example, gray zone operations often emphasize supporting third-party militants or conducting limited cyberattacks; war, on the other hand, utilizes conventional forces (Mazarr 2015a; Matisek 2017; Wirtz 2017).

For Peer Review

nonzero proportion of the costs of C's limited challenge carry through.[10]

Scenario	C's utility	D's utility
C initially initiates war $(w_C = 0)$	$\theta\rho_W - \kappa_C$	$1 - \rho_W - \kappa_D$
C and D select gray zone/accept status quo $(w_C = 0,\ g_C \geq 0,\ w_D = 0,\ g_D \geq 0)$	$\theta P(g_C, g_D) - \beta_C g_C^2$	$1 - P(g_C, g_D) - \beta_D g_D^2$
D escalates to war after C acts $(w_C = 0,\ g_C \geq 0,\ w_D = 1)$	$\theta\rho_W - \kappa_C - \beta_C g_C^2$	$1 - \rho_W - \kappa_D$

Table 1: Summarized payoffs for actors

The model above is designed to simply illustrate how gray zone conflict plays out, conditional on intuitive cost and benefit parameters. In the Appendix, we consider three extensions. First, we examine an extension where β_D is endogenous. The new model offers little insight beyond what is discussed in Section 7. Second, we include a model where limited challenges can probabilistically escalate. While this modification makes gray zone conflict less desirable to the challenger and generates new equilibrium conditions, the results are substantively identical. Third, we also include a model where the defender select the "status quo" policy as a bargained offer, and the defender has a private type. This third extension demonstrates that under a fairly standard micro-foundation, a rational D could make an offer to C where C responds to the offer as a potentially dissatisfied state. The key relationships identified in the primary model— internal efficiency and external deterrent constraints as key drivers of gray zone conflict outcomes—also exists in the third extension.

For Peer Review

4.2 Equilibrium Concepts and Assumptions

We limit our attention to pure strategy subgame perfect Nash equilibria.[11] We use asterisks to denote equilibrium behavior, for example: g_C^* and g_D^*.

We make one technical simplifying assumption, which we discuss formally in the Appendix. We limit analysis below to scenarios where, conditional on gray zone conflict occurring ($g_C^* \geq 0$ and $g_D^* \geq 0$, with at least one inequality holding strictly), the optimal limited challenge and response are such that the constraints on P do not bind (i.e. the final realized P is strictly less than ρ_W and strictly greater than ρ_0). This assumption eliminates the possibility that kinks in the P function drive our results and it prevents excessive casework.

[10] If C does not pay for gray zone investments when a war occurs, then a challenger that prefers war to their tolerated-by-D gray zone challenges may select a gray zone challenge that deliberately provokes D to initiate a war. This does not change the substance of the results, but it does introduce a new equilibrium.

[11] The focus on pure strategies eliminates edge cases where one player is indifferent over two actions and mixes.

4.3 Equilibria

The challenger decides whether the game ends with peace, war, or gray zone conflict. C's decision can be expressed in terms of C's resolve, or θ. When C has low resolve, C will accept the status quo. When C has high resolve, C will select into war. When C's resolve falls between the two, assuming gray zone conflict is cost-effective enough, C will conduct limited challenges.

There is nuance in how C conducts gray zone conflict. To illustrate this, consider a hypothetical setting where C can select any intensity of limited challenge, and D will respond with gray zone conflict. In this hypothetical, C faces an internal optimization, selecting a limited challenge based on its resolve over the issue and its costs for conducting limited challenges—essentially where marginal returns are equal to marginal costs. This limited challenge level based on C's internal cost-benefit analysis is determined by the challenger's *internal efficiency* constraints. Of course, in order not to select a limited challenge that would cause the defender to escalate to war, C's limited challenge is also bound by the defender's *external deterrent threat* constraint. When C's optimal limited challenge is less aggressive than a challenge that would provoke D to escalate, we say that C's internal efficiency binds. Otherwise, C will select a limited challenge tailored to make D refrain from war, and the deterrent threat binds. Which constraint binds is arbitrated by a technical condition—if $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, or not.

Proposition 1: *In equilibrium, the game will play out in the following manner.*

Case 1, D's Deterrent Threat Binds, $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*:*

- *1.A. C accepts the status quo (* $w_C^* = 0$ *and* $g_C^* = 0$*) if* $\theta \leq \frac{\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)}$ *and* $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$.
- *1.B. C initially initiates war (*$w_C^* = 1$*) if* $\theta > \frac{\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$ *and* $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$.
- *1.C. C selects into gray zone conflict and is constrained by D's deterrent threat (*$w_C^* = 0$ *and* $g_C^* = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*) otherwise.*

Case 2, C's Internal Efficiency Binds, $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*:*

- *2.A. C accepts the status quo (*$w_C^* = 0$ *and* $g_C^* = 0$*) if* $\theta \leq \frac{2\beta_C}{\beta_D}$ *and* $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$.
- *2.B. C initially initiates war (*$w_C^* = 1$*) if* $\theta > \frac{\kappa_C}{\rho_W - \rho_0 - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$ *and* $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$.
- *2.C. C selects into gray zone conflict and is not constrained by D's deterrent threat (*$w_C^* = 0$ *and* $g_C^* = \frac{\theta}{2\beta_C}$*) otherwise.*

Proof: See Appendix.

For Peer Review

We discuss Proposition 1 in two parts. In Section 5, we discuss what drives variation in gray zone activity. Then, in Section 7, we discuss why defenders may not always want to be effective at countering a challenger's gray zone activity.

5 What Drives Variation in Gray Zone Activity?

5.1 On the Defender's Deterrent Threat and Conflict Intensity

As the defender becomes more willing to go to war (κ_D decrease), the challenger selects a weakly less aggressive limited challenges (g_C^*) or avoids gray zone conflict altogether. For example, if NATO leaders are more willing to escalate to war over NATO's core states relative to its periphery states or non-members, then the model expects less aggressive gray zone action on Russia's part against core states. The defender's willingness to go to war thus creates an upper bound on the tolerated level of limited challenges. If the defender becomes more willing to go to war, the challenger must either scale back the intensity of their limited challenge in order to avoid war, or the challenger must forgo limited challenges and instead either accept the status quo or go to war. If the defender's deterrent threat becomes less credible, the challenger has more freedom to choose whatever amount of force is needed to get the job done as the challenger sees fit.

Observation 1*: If the defender becomes more willing to go to war (lower κ_D's), then the challenger selects weakly less intense limited challenges, or the challenger may no longer engage in limited challenges and instead accepts the status quo or goes to war.*

We provide a formal discussion of Observation 1 in the context of Proposition 1 in the Appendix. Figure 1 illustrates Observation 1. On the x-axis, moving left-to-right, D's deterrent threat from war is decreasing, as operationalized in D's costs of war κ_D increasing. On the y-axis, moving low-to-high, C's equilibrium challenge g_C^* is increasing. For the lowest costs of war—the region where the equilibrium is described in Case 1.A in Proposition 1—D is very willing to go to war, C is tightly constrained by D's external deterrent threat, and therefore gray zone conflict is not particularly productive for C. Under these parameters, C will select into the status quo.[12]

Moving to the right—to Case 1.C.—because D is less willing to go to war, C can engage in enough limited challenges to make gray zone conflict worthwhile. Here the intensity of C's limited challenge is constrained by D's deterrent threat, so C can select more aggressive limited challenges as D's deterrent threat decreases.

[12] If $\kappa_C \leq 0.41$, C prefers going to war in this region rather than selecting the status quo.

For Peer Review

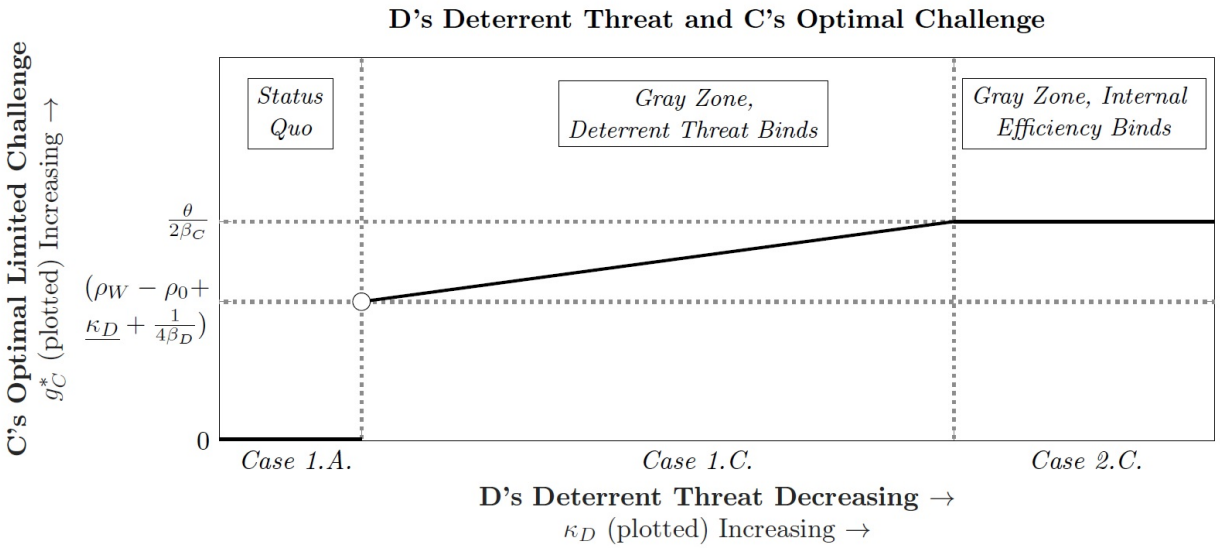

Figure 1: Optimal Limited Challenge as D's Deterrent Threat Decreases. C's selected limited challenge intensity under a range of κ_D's are plotted. We define $\underline{\kappa_D}$ implicitly as the value of D's war costs that satisfies $\theta = \frac{\beta_C\left(\rho_W - \rho_0 + \underline{\kappa_D} + \frac{1}{4\beta_D}\right)^2}{\left(\rho_W - \rho_0 + \underline{\kappa_D} - \frac{1}{4\beta_D}\right)}$. All equilibrium cases are described in Proposition 1. The parameters are $\rho_0 = 0.1$, $\rho_W = 0.7$, $\beta_C = 0.34$, $\kappa_C = 0.5$, $\beta_D = 1$, and $\theta = 0.69$. Values of κ_D fall between 0.01 and 0.24. Y-axis not drawn to scale.

In Case 2.C, because D's deterrent threat has sufficiently declined, D is willing to tolerate aggressive limited challenges without escalating to war. Here, D's deterrent threat no longer binds, and instead C's limited challenge is constrained only by C's internal efficiency constraint.

Altogether, Figure 1 illustrates a natural intuition. As D's deterrent threat from war decreases, C can select more aggressive limited challenges without provoking D to war. This expands the possible set of limited challenges that will not trigger an escalation, thus resulting in a greater intensity of gray zone conflict, until C no longer finds more gray zone conflict worthwhile.

5.2 On the Challenger's Resolve and Conflict Intensity

As the challenger becomes more resolved (θ increases), the challenger's chosen level of limited challenge (g_C^*) weakly increases, unless its resolve is so high that it forgoes gray zone conflict and resorts to war. For example, during the Syrian Civil War, while both Moscow and the Syrian Ba'athist party clearly preferred retaining Bashar al-Assad's power, the Ba'athist party was presumably more resolved and fought harder than Russia.

Consider a setting where the challenger selects a limited challenge based on their internal efficiency constraint. As the challenger's resolve increases, the challenger benefits more from shifting the status quo and is willing

For Peer Review

to select more aggressive limited challenges. However, a highly resolved challenger could be willing to select a limited challenge intensity that exceeds the defender's tolerance within gray zone conflict. For these high levels of resolve, the challenger must either choose a non-internally-optimal intensity of its limited challenges to keep the defender from going to war—reflecting the external constraint of the defender's deterrent threat—or accept escalation to war.

Observation 2: *As the challenger's resolve increases, the challenger selects weakly more intense limited challenges, or may forgo gray zone conflict for war.*

Observation 2 follows naturally from Proposition 1 and is illustrated in Figure 2. Moving from left-to-right, C's resolve θ is increasing. For the lowest resolve considered—equilibrium Case 2.A.—C does not benefit sufficiently from altering the status quo and prefers to make due with the status quo.

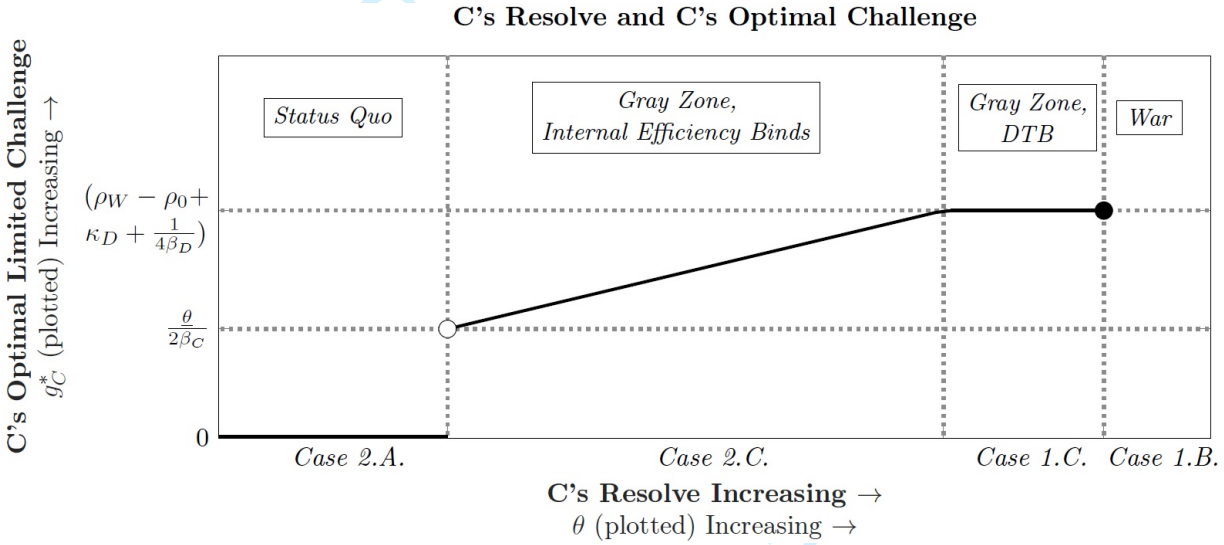

Figure 2: Optimal limited challenge as C's Resolve Increases. C's selected limited challenge intensity under a range of θ's are plotted. We define $\underline{\theta} = \frac{2\beta_C}{\beta_D}$. "DTB" is an abbreviation for "Deterrent Threat Binds." All equilibrium cases are described in Proposition 1. The parameters are $\rho_0 = 0.1$, $\rho_W = 0.8$, $\beta_C = 0.495$, $\kappa_C = 0.75$, $\beta_D = 1$, and $\kappa_D = 0.15$. Values of θ fall between 0.8 and 1.8 To simplify labeling, axes are not drawn to scale. We illustrate C selecting $w_A^* = 1$ as C selecting $g_C^* = 0$.

Moving to the right, in Case 2.C, because C is more resolved to alter the status quo, C does best by engaging in limited challenges. Here C's internal efficiency constraint binds, meaning C's selected limited challenge is increasing in C's resolve. In Case 1.C, C's resolve is even higher, but D's deterrent threat binds. C is resolved enough to engage in aggressive limited challenges beyond what D would tolerate (a level that would provoke D to escalate to war),[13] but C is not resolved enough to go to war. As a result, C selects a level of limited challenge bound by D's indifference threshold between war and gray zone conflict, which is unchanging in

[13] C's willingness to challenge grows along the trend line from the Case 2.C.

C's resolve. Finally, in Case 1.B, C's resolve has increased to the point where the level of limited challenges within gray zone conflict tolerated by D is not productive enough for C, and C optimally goes to war.

5.3 On the Challenger's Gray Zone Costs and Conflict Intensity

As the challenger's costs from gray zone conflict decrease (β_C decreases), the challenger will select weakly more aggressive limited challenges (g_C^*). When the internal efficiency constraint binds, the challenger's selection of their limited challenges will increase as their costs decrease. When the external deterrent threat binds, the challenger's selection of their limited challenge does not vary with their costs, but rather is dictated by the defender's willingness to go to war, which is static.

Observation 3: *As C's gray zone costs decrease, C selects weakly more intense limited challenges, and may forgo accepting the status quo or war for gray zone conflict.*

Figure 3 illustrates one example of Observation 3. Given the self-evident nature of this finding, we do not discuss these conditions at length.

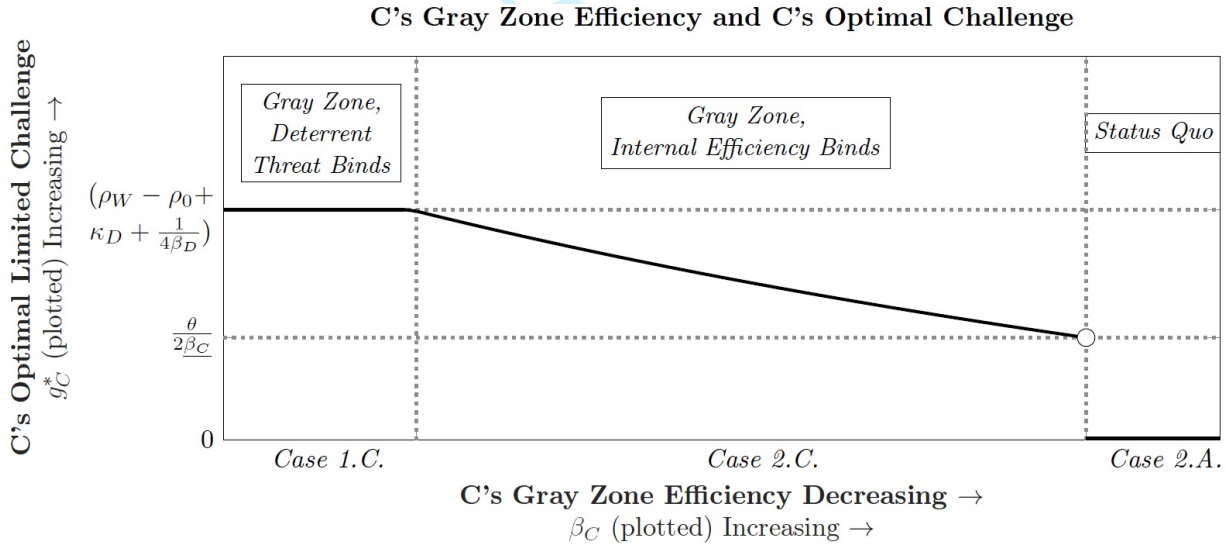

Figure 3: Optimal limited challenge as C's costs of limited challenges vary. C's selected limited challenge intensity under a range of β_C's are plotted. We define $\underline{\beta_C}$ as $\frac{\theta}{2\underline{\beta_C}} = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. All equilibrium cases are described in Proposition 1. The parameters are are $\rho_0 = 0.1$, $\rho_W = 0.8$, $\kappa_D = 0.15$, $\kappa_C = 0.2$, $\beta_D = 0.8$, and $\theta = 0.2$. Values of β_C fall between 0.091 and 0.13. To simplify labeling, y-axis not drawn to scale.

6 Empirical Application: Russian Efforts in the Gray Zone

We empirically assess our argument by analyzing data on the scope and intensity of Russian foreign interventions over the past two decades. Quantitative analysis supports the hypothesis discussed in Observation 1: that Russia chooses its level of provocation in response to NATO's implicit deterrent threat. Analysis also supports Observations 2 and 3: that Russian activity is inversely associated with decreased resolve and increased costs, as approximated by a geographical "loss of strength gradient" (Posen 2003).

We focus on Russia because its interventions are extensively referenced as paradigmatic examples of gray zone conflict (Marten 2015; Driscoll and Maliniak 2016; Jasper 2020). From 1994 to 2018, Russia has been involved in election interference in the United Kingdom and Moldova, cyberattacks in Estonia and Georgia, and special operations in Ukraine and Yugoslavia. Russia's military adventures are the "most likely" cases for the argument that gray zone conflict is an effective and low-cost innovation for revising the status quo. The diversity of Russian targets and means employed provides an opportunity to conduct a controlled comparison of Russian choices under different deterrent circumstances. We present a simple statistical analysis to test the implications of the model on Russian decision making while controlling for a number of relevant factors. By no means is this analysis causal, as our key independent variables—NATO membership, for example—are not exogenous treatments. Our results should be viewed as suggestive and taken with the necessary caveats. We then undertake a brief qualitative analysis of three cases: Estonia, Ukraine, and Georgia. We find that the possibility of a NATO intervention is associated with more limited gray zone operations, suggesting that Russian gray zone behavior occurs not despite NATO's deterrent threat, but because of it.

6.1 Data

Admittedly, data on any gray zone interventions are themselves ambiguous. Most quantitative studies of gray zone operations focus on a particular type, like cyber in the case of the Dyadic Cyber Incident and Dispute (DCID) data or electoral interference in the case of the Russian Electoral Interventions (REI) data (Valeriano and Maness 2014; Casey and Way 2017). Consequently, they cover almost entirely distinct samples with significant differences concerning the severity of Russian attacks.[14]

To address these discrepancies, we construct a new, expanded dataset of 82 cases of Russian intervention from 1994 to 2018. DCID and REI together describe 71 unique cases of Russian aggression that have either included some degree of cyber intervention or were cases of electoral interference. We identify 10 additional instances of Russian cyberattacks in the coding period that are not listed in previous datasets. Most of these

[14]Indeed, the only country-year that appears in both datasets is Ukraine 2014.

For Peer Review

new cases involve cyber conflict after 2011 (the latest year in DCID) that were non-electoral (the universe of cases in REI). We also include the 3 cases of non-cyber Russian action from the International Crisis Behavior (ICB) dataset (Brecher and Wilkenfeld 1997). For each incident, we create a new coding of the intensity of Russian attacks by coding whether Russia used five different types of military force in ascending order of intensity: (1) information operations (social media and disinformation), (2) cyber operations that result in disruption of infrastructure (service denial or industrial control system attacks), (3) overt use of special operations or unattributed military forces, (4) conventional air or sea forces, and (5) conventional ground forces. This data is then aggregated to the country-year level with a coding for the highest level of intensity of Russian intervention against each country in each year. Our dependent variable is thus an ordinal variable coded 5 for the highest level of intensity down to 0 for country-years experiencing no Russian attack.

Figure 4 plots the count and average intensity of Russian gray zone operations since 1994. Contrary to descriptions of gray zone conflict as the product of an expansive technological portfolio, there does not appear to be a clear temporal pattern in the intensity or frequency of activity. Instead, 2004 represents the most intense overall Russian interventions and 2014 experienced the highest number of interventions (most of which were associated with Ukraine).

Figure 5 depicts a pattern of the geographical coverage of Russian conflict events in Europe. Russia appears to be willing to use more force in countries in its "near abroad," relative to countries further away. Interpreting this geographic pattern on its own is difficult, as distance from Russia is plausibly related to Russian interest or resolve, ease of conducting operations, or the impact, or even the determinants of NATO membership. This pattern thus highlights the need for more sophisticated analysis.

Now we discuss our independent variables. Consistent with the discussion on the external deterrent threat (Observation 1), we propose that the external deterrent threat from war is a key driver of Russian gray zone behavior. We operationalize this concept through a dummy variable for NATO membership. NATO members plausibly possess lower costs for fighting—since they can rely on collective security. They may thus have a reduced willingness to tolerate aggressive low-level behavior from Russia. If Russia is responding to NATO's deterrent threat, we expect NATO states to experience less intense Russian activity.

Consistent with the discussion of the internal-efficiency constraint (Observations 2 and 3), insofar as military power is affected by a loss of strength gradient (Posen 2003), we propose the intensity of Russian gray zone operations could decrease as Russia has less resolve over the issue, or faces greater costs for conducting operations.[15] We operationalize Russian resolve and gray zone efficacy jointly through a variable of the

[15]We only analyze this variable in models that also include a NATO membership covariate.

For Peer Review

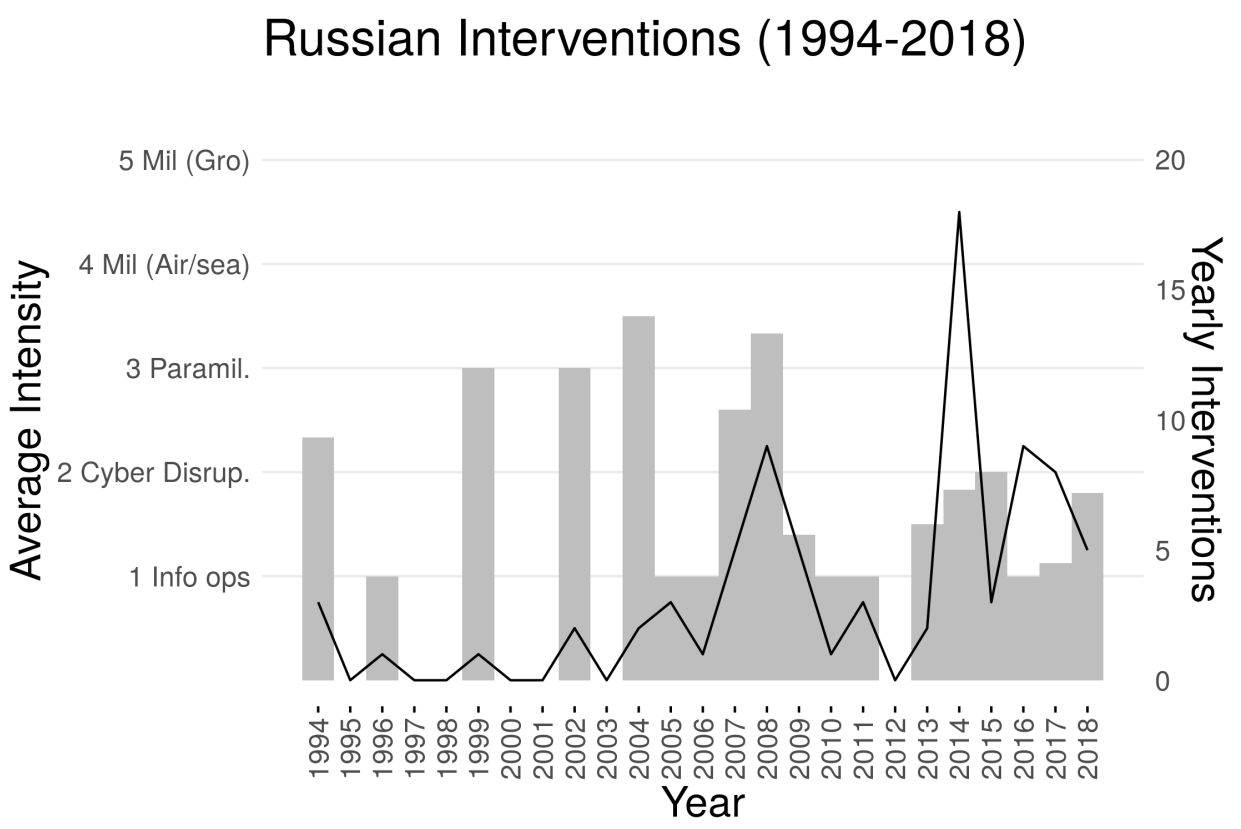

Figure 4: Intensity of Russian intervention across time. The line represents annual average intensity. Bars denote the number of interventions annually.

logged minimum distance between Russia and each potential-target state, as we expect Russia to care more about states on its periphery and to more easily conduct operations in its proximity (Weidmann, Kuse, and Gleditsch 2010).

We also include a series of control variables motivated by existing theories of relationships between the control variables and our dependent and independent variables of interest. We include a democracy dummy for states with a Polity V score greater or equal to 6 to control for potential Russian eagerness to target democracies (Early and Asal 2018). A state's possession of nuclear weapons may also alter Russia's calculus about how to pursue aggressive actions, so we include a dummy variable for nuclear states (Gartzke and Kroenig 2009). We include GDP per capita and the log of population as larger, richer states could afford more opportunities for Russian interventions, especially cyber interventions (Beckley 2010). Finally, we include military expenditure since it influences both alliance decisions and the cost of undertaking aggression against an adversary (Omitoogun and Skons 2006). Summary statistics are included in the Appendix.

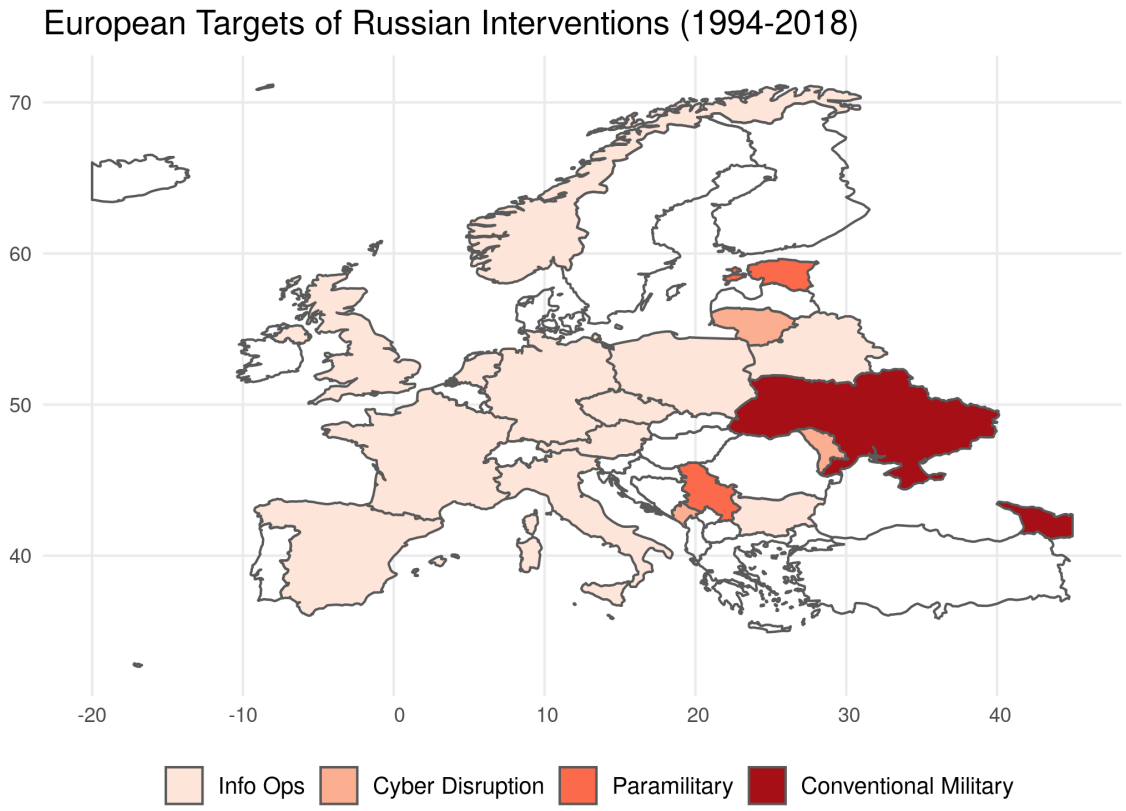

Figure 5: Geographic representation of Russia intervention. Shading represents the highest intensity of Russian intervention in each European state.

6.2 Model and Results

Because our outcome is an ordinal value, we estimate a series of ordered probit models with year fixed-effects and standard errors clustered by country. Our unit of analysis is the country-year. We run three empirical models on two samples. On the models, first we estimate the relationship between the intensity of Russian intervention and NATO membership and minimum distance from Russia without any control variables. Second, we re-run the first model while also controlling for the range of variables indicated above. We exclude military spending from the second model because it is missing across the entire panel for countries like Yugoslavia and Bosnia & Herzegovina. In model 3, we include the military spending variable, thus operationalizing military power using population and military spending. On the samples, our first sample (models (1)-(3)) includes all European states. We define state membership using the Gleditsch and Ward state list and continent location using the World Bank Development Indicator (Gleditsch and Ward 1999). This excludes micro-states with less than 250,000 people like Liechtenstein and San Marino. The downside is this sample may include states that are not of interest to Russia and may fall outside the scope of our model (for example, Luxembourg). To address this, our second sample (models (4)-(6)) represents "relevant

European states," includes only European states that meet any of the following three criteria: a) targets of a Russian attack from 1945-1993 as identified in the Militarized Interstate Dispute (MID) or International Crisis Behavior (ICB) datasets, b) Former Soviet Union or Warsaw Pact states, or c) states that are contiguous with Russia.

	Full sample			Relevant states sample		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Independent Variables						
NATO member	−0.28	−0.46**	−0.60***	−0.47*	−0.58**	−0.68***
	(0.22)	(0.20)	(0.22)	(0.26)	(0.26)	(0.25)
Russia distance	−0.10***	−0.11***	−0.12***	−0.05	−0.09**	−0.09**
	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Controls						
Democracy		0.16	0.46		0.12	0.43
		(0.43)	(0.42)		(0.45)	(0.43)
Nuclear power		0.93**	0.44		0.92*	1.06
		(0.42)	(0.44)		(0.48)	(0.82)
Population		0.19**	0.14		0.16	0.18
		(0.09)	(0.12)		(0.10)	(0.13)
GDP per cap		−0.01**	−0.02**		−0.01	−0.01
		(0.01)	(0.01)		(0.01)	(0.01)
Mil. spending			0.02			−0.00
			(0.01)			(0.02)
Observations	1,000	921	891	376	373	346

All models include year-fixed effects with country-clustered standard errors in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2: Intensity of Russian Intervention: Ordered Probit Results

Table 2 presents the coefficient estimates from the ordered probit regressions run on both samples. The results show that both NATO membership and distance from Russia decrease the intensity of Russian intervention against European states. Every models that utilizes control variables (models (2), (3), (5), and (6)) suggests the relationship is statistically significant at least at the 0.05 level. Similarly, the coefficient for distance from Russia is in the expected direction in all models and statistically significant at the 0.05 level in every model except the relevant state sample with no control (model (4)).

These results provide evidence of Observation 1: as the defender's deterrent threat increases, the challenger scales back the intensity of its limited challenges. For example, model (6) reports a proportional odds ratio coefficient of 0.44 on the NATO dummy.[16] This value means that for relevant NATO states, the odds of a non-cyber, non-information attack (categories 3, 4, or 5) are 49% lower than the odds of experiencing a cyber attack, an information attack, or no attack. Our findings are also consistent with Observations 2 and 3, insofar as Russian valuation for the stakes and ease of operation arguably increases in regions deeper

[16] A complete table of all odds ratios is provided in the Online Appendix.

For Peer Review

within its "Near Abroad" than for areas well beyond it. Together, this suggests that Russian behavior is shaped by the external deterrent threat in some cases and its own internal efficiency constraint in other cases, though the latter relationship is statistically weaker. Importantly, it is not just that NATO membership deters Russian aggression, but that NATO membership alters the intensity of Russian aggression in a way that explains its type. NATO members are not increasingly the victims of gray zone conflict because Russia is using a novel form of effective conflict they could not before. Rather, our results suggest that Russia is using gray zone forms of conflict because Western deterrence has taken its best options off the table.

Our findings are consistent across a range of alternate samples and modeling specifications detailed in the Online Appendix. As NATO membership is a somewhat coarse operationalization of external deterrent threat, we run additional models distinguishing between non-NATO members in the NATO accession membership process and those who are not.[17] The results from these models are consistent with the original model specifications. We also find consistent results when sampling on only country-years that experience an attack and also with OLS and ordered logit regression models. We also use multiple imputation with additive regression, bootstrapping, and predictive mean matching to replace missing values for control variables (Buuren 2012). These results provide consistent statistical evidence in support of our argument, thus mitigating concern that the initial results are an artifact of listwise deletion (Arel-Bundock and Pelc 2018). Additionally, we re-run the analysis including CINC ratios in place of population and the SIPRI military expenditure data (Singer, Bremer, and Stuckey 1972).[18] To addressing missingness—CINC ratios are not published for years 2012-2018—we similarly use multiple imputation and find similar support for our argument.

For Peer Review

6.3 Case Studies

The quantitative analysis can be thought of as a coarse technique for considering empirical trends. Some concerns about endogeneity and causal inference can be partially addressed by examining the logic of Russian interventions in detailed qualitative case studies. For a more fine-grained test of our hypotheses, we consider three major cyber campaigns attributed to Russia that feature prominently in the cybersecurity literature: Estonia, Georgia, and Ukraine. These cases are typically highlighted as examples of the increasing potential of gray zone conflict, making them a most likely case for the conventional efficiency logic. We follow a most similar case study design in selecting cases that feature cyber attacks by the same contiguous challenger

[17]The pre-NATO stages includes membership in Partnership for Peace (PfP), Intensified Dialogue, and Membership Action Plan (MAP). For details on what these entail, see Amara and Paskevics (2010).

[18]Population and military expenditure are two of the six components that comprise the CINC index.

(Russia) but differ in other military instruments employed (Bennett and Elman 2007).[19] Additionally, to the extent that Russia wants to influence its immediate neighbors, Russia has an interest in all states and could intervene with relative ease. There are many potential explanations for why Russia wanted what it wanted in each instance, but here we set aside Russia's foreign policy formulation (McFaul 2020). Instead, we highlight geo-strategic context and military effectiveness. A summary of the extent of Russian conflict behavior is provided in Table 3.

Russian Response	Estonia (2007)	Ukraine (2014)	Georgia (2008)
Conventional Forces			X
Special Operations		X	X
Cyber Operations	X	X	X

Table 3: Case comparison of Russian gray zone conflicts

6.3.1 Estonia (2007)

Of the three states, Estonia experienced the most limited operations. Moscow coordinated a wave of DDoS attacks against Estonia following the relocation of a Soviet statue (Schmidt 2013). The gap in time between Estonia's 2004 ascension to NATO and the 2007 Russian cyber campaign is telling. In Georgia and Ukraine, the prospect of NATO ascension (announced in the April 2008 Bucharest Summit Declaration) provoked a Russian response. The Estonian attacks, by contrast, were muted and opportunistic, not a determined bid to change conditions on the ground. No one issued any clear demands or claimed responsibility, and Estonia did not replace the statue. The DDoS attacks were an ambiguous symbolic gesture calibrated to fall well below the threshold that might trigger a NATO response. The ambiguous legal status of a cyberattack in 2007 both enabled and constrained Russia in this respect (Joubert 2012). NATO was highly unlikely to escalate so long as Russia did not inflict serious harm. Estonia's defense minister considered but ultimately rejected invoking Article V, the collective defense clause of the NATO treaty, instead treating the episode as a domestic law enforcement matter (Traynor 2007). Overall, Russian moves seemed aimed at avoiding a greater escalation.[20]

6.3.2 Ukraine (2014)

Consistent with the logic of NATO's deterrent threat, Russian actions in Ukraine have been more extensive than those in Estonia, but less than what occurred in Georgia. Despite six years of protracted war there has occurred neither large-scale combined arms warfare, as in Georgia, nor unrestrained ethnic cleansing

[19] We include a fourth case study of Russian intervention in the 2016 US election in the appendix.

[20] In the context of our model, the decision to move the statue represented a "status quo," where Estonia could engage in independent, nationalist, possibly anti-Russian policy choices. The Russian cyberattack undermined the Estonian government's ability to proceed behaving in this manner.

For Peer Review

(Driscoll and Steinert-Threlkeld 2020). The fact that Russia could have exerted more effort, together with the actions made to allow both sides to save face, suggest Russian restraint.[21] Even though NATO has no formal commitment to Ukraine, conflict in a country that borders NATO allies like Poland and Hungary is implicitly shaped by the possibility of Western intervention, risking nuclear escalation in the process. As a result, we conjecture that Russia acts circumspectly. Endemic Russian cyberattacks and information operations have had little impact on battlefield events (Kostyuk and Zhukov 2019). Even as social media manipulation is supposedly a Russian specialty, pro-Kremlin narratives have not taken hold in Western Ukraine (Driscoll and Steinert-Threlkeld 2020).

6.3.3 Georgia (2008)

While Georgia was hit by DDoS service attacks (similar to Estonia) (Deibert, Rohozinski, and Crete-Nishihata 2012), Russia also intervened militarily in South Ossetia and Abkhazia, an early example of cross-domain operations leveraging cyberspace. Russia's intervention choices in this conflict, situated at the far end of the Western deterrence gradient and deep in Russia's traditional sphere of influence, were relatively unconstrained. The same month that NATO announced a pathway to membership for Georgia, Russia announced that it would unilaterally increase peacekeepers in Abkhazia. Russia then used whatever mix of tools it needed to accomplish its objectives and did not pull its punches given that Western counteraction was unlikely (Binnendijk 2020). As Driscoll and Maliniak (2016, 590) point out, Georgia's location makes it a security liability in the eyes of many Western leaders. The Russian intervention served to clarify the stakes of Western interference in its near abroad. While Russia's tactical performance left much to be desired, the mission was a strategic success that reinforced the status quo ante and ended the conversation about Georgia joining NATO. The forceful nature of the Russian intervention is notable, with long columns of conventional armor, something not considered elsewhere, despite the military imperatives of mass and firepower.

6.3.4 Discussion of Cases

The overall pattern of recent Russian intervention is consistent with our hypothesis that deterrence encourages capable actors to engage in calculated restraint. Moving from Estonia to Ukraine and finally to Georgia, as the deterrent threat from NATO becomes less salient, Russia pursues its international objectives with greater intensity. One might argue that Russia has different levels of resolve across these cases. For example, one might argue that Russia places a very different value on the outcome in Ukraine than Estonia. Indeed, Russia

[21] Mixed messages of resolve and restraint are common in covert action (Carnegie and Carson 2018; Carson 2018).

For Peer Review

let Estonia join NATO without a fight in 2004. By contrast, Russia had supported Georgian separatists since the early 1990s and was highly resolved to ward off Western encroachment. The Ukraine case, however, finds this alternative account wanting. The seat of the medieval Kievan Rus empire is arguably more salient in Russian nationalist mythology than Georgia, a peripheral outpost in the Caucuses far from Moscow, and the Black Sea port of Sevastopol also makes Crimea strategically important. If Russian moves were motivated by resolve rather than external deterrence, then we would expect more robust, and more overt, Russian military efforts in Ukraine. Yet, despite Russia's undoubted higher valuation for the stakes in Ukraine, one observes considerable restraint.

7 Implications Concerning Escalation Dynamics

How can a defender best react to the prospect of gray zone conflict? It might seem intuitive that improvements in a defender capacity to counter gray zone aggression should reinforce the strength of deterrence, but this is not necessarily the case. If a defender has low costs for conducting gray zone conflict, then the defender can operate effectively and aggressively within gray zone competition; being faced with such an opponent, a challenger may forgo gray zone operations in favor either of the status quo or going to war. In other words, developing the tools to effectively thwart gray zone conflict could lead to peace or to greater escalation and more war.

Figure 6 details this logic. On the x-axis, we plot C's resolve. On the y-axis, we plot D's costs of gray zone conflict. Each region of the graph is an equilibrium type. For example, "C Selects Gray Zone, No Deterrent Threat" is described in Case 2.C. in Proposition 1. The dotted line represents the cut-point where, to the left, C prefers the status quo to war, and to the right, C prefers war to the status quo. As can be seen, D does not always benefit by increasing its ability to resist gray zone conflict. Consider a point in the "C Selects Gray Zone, Internal Efficiency Binds" region that is to the left of the dotted line. Here if D's gray zone efficiency increases, then after some point, C will forgo gray zone conflict and accept the status quo. In this circumstance, C has a low-enough resolve that, with sufficient pressure, C can be deterred from all forms of conflict. However, now consider a point in the same equilibrium region, but to the right of the dotted line. If β_D increases above some point, C will choose war. In this circumstance, C is resolved enough that as D increases their gray zone efficiency, C will opt into war.

Observation 4: *Decreases in D's gray zone costs (β_D) can lead to less or more conflict depending on C's level of resolve.*

For Peer Review

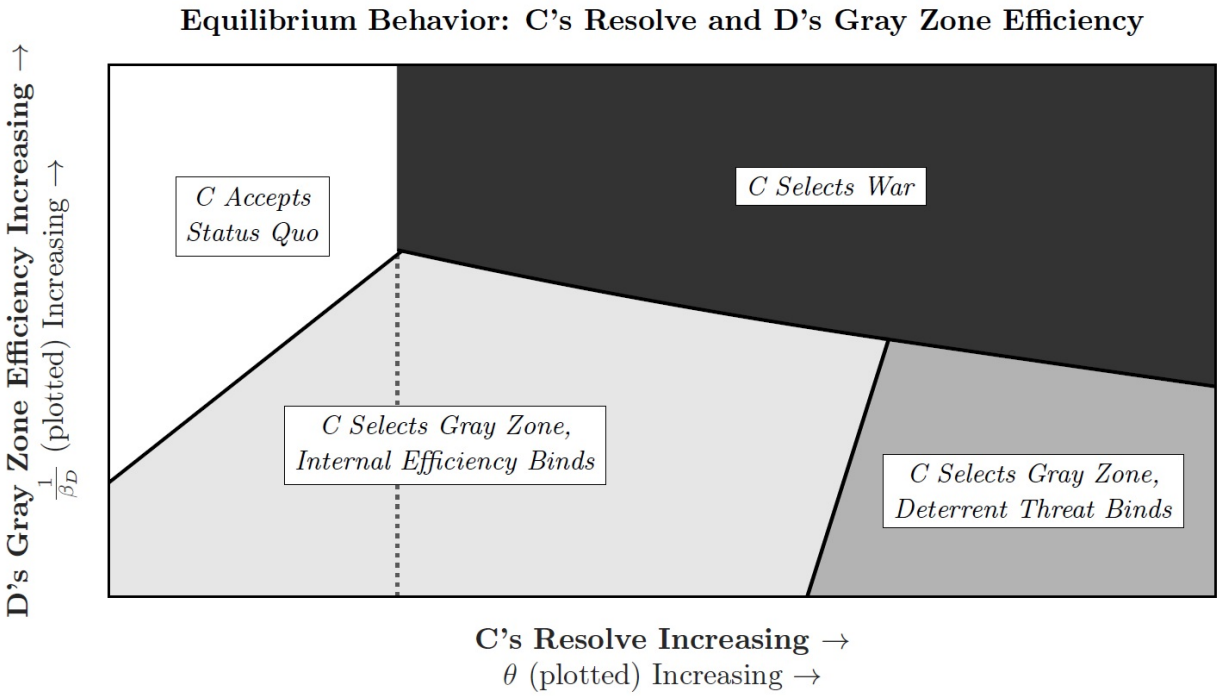

Figure 6: Equilibrium behavior as C's resolve and D's gray zone efficiency varies. C's resolve θ and the inverse D's gray zone efficiency $\frac{1}{\beta_D}$ are plotted. The dashed line is on the θ value where $\theta = \frac{\kappa_C}{\rho_W - \rho_0}$, or where C is indifferent between initially accepting the status quo and going to war. The parameters are $\rho_0 = 0$, $\rho_W = 0.5$, $\beta_C = 1$, $\kappa_C = 0.53$, and $\kappa_D = 0.1$.

There is an alternate interpretation; β_D could be influenced by the defender's prior, un-modeled moves in the game that set up current gray zone operations. For example, if the defender aggressively pursued counterinsurgency operations against foreign-backed rebels before this game began, the defender could be in a position to pursue aggressive gray zone conflict within the game. This interpretation illustrates how it is not just latent, exogenous costs that influence the challenger's activity, but rather war can result from the defender behaving too aggressively in prior actions against a highly resolved challenger.

This final interpretation has important policy consequences. Even if most actors are assumed to harbor challenger ambitions (Schweller 1996), would-be defenders still face a security-like dilemma in shaping how aggression is expressed. In the security dilemma, the outcome of a "threat" from the target is a function of whether or not a challenger is resolved. Conflict short of war complicates this picture because gray zone behavior by a highly resolved challenger and by a marginally-resolved challenger may be observationally indistinguishable. And yet, the consequence of the defender threatening to escalate within gray zone conflict vastly differs. The new U.S. Cyber Command doctrine of "persistent engagement" aims to establish dominance in strategic competition short of war through proactively "defending forward," but its very success could become the trigger for inadvertent escalation (Healey and Jervis 2020).

Peer Review

Of course, this is not to say that any improvement in the defender's gray zone efficiency will always result in war. One could reasonably believe that improved U.S. election security could counter Russian election interference without leading to war, and our model captures this. However, this model also suggests that other issues might occupy a more precarious space. A defender's failure to understand a challenger's resolve might lead that defender to over-invest in gray zone capabilities thus leading to war, or to under-invest in gray zone capabilities thus allowing low-level conflict to simmer. What issues occupy what spaces is beyond the scope of this (already extensive) paper, and is left as a topic for future research.

8 Every Silver Lining's Got a Touch of Gray

Gray zone conflict occurs when capable actors intentionally limit the intensity or capacity of aggression and refrain from escalation. Deterrence shapes the way that conflict emerges, but it may not suppress conflict altogether. The good news is that gray zone conflict is symptomatic of deterrence success. Adversaries are "designing around" deterrence in an effort to avoid the anticipated retaliation of the defender (Lieberman 2012). The bad news is that gray zone conflict probes the threshold of deterrence effectiveness. We expect conflict severity to be greater wherever there are questions about the willingness or ability of defenders to respond forcefully. An adversary is seldom passive. There will always be attempts at end-runs or push-back, even when deterrence is credible. It is thus important to think carefully before overextending commitments where credibility is in doubt.

Just as there is a gray zone between war and peace, the distinction between effective and ineffective deterrence is also fuzzy. We have introduced the notion of a deterrence gradient, a straightforward extrapolation from the military loss of strength gradient, to describe credible deterrence as a continuous variable. Wherever deterrence is credible, or the challenger's resolve is low, challengers can be expected to exercise restraint as they probe to see what they can get away with. Wherever deterrence is not credible or challengers are highly resolved, however, a challenger must be more emboldened to use whatever means they have at their disposal to meet their objectives, limited only by internal efficiency constraints. The challenge lies between these extremes, where the variable threshold of credibility creates a policy arena for limited conflict, and where it can be difficult to distinguish efficiency motivations from risk sensitivity. Doubling down on deterrence can mitigate conflict in the latter case but provoke escalation in the former.

We have used the same cases that have raised alarms about the dangers of gray zone conflict to suggest the validity of an alternative explanation. The evidence suggests that Russia systematically reduces the intensity of its interventions along the deterrence gradient, employing a greater variety of means with more lethal

For Peer Review

intensity where deterrence is weakest but conducting only ambiguous information operations where deterrence is most robust. The conventional wisdom is right that Russian interventions are paradigmatic exemplars of gray zone conflict, but it is wrong about Russian motivations and the effectiveness of these operations. Revisionist powers have not discovered a secret formula or novel tools destined to destabilize Western democracies or undermine NATO's deterrence posture. Rather it acts opportunistically as circumstances enable it to hassle adversaries and their clients without, however, risking a costly military confrontation. The flip side of this logic, however, is that Russia is willing to call NATO's bluffs in cases where it can reasonably expect that NATO is unwilling to intervene. In Georgia, and even more so in Chechnya, Russian willingness to prioritize effectiveness at the price of efficiency is clear.

This argument has implications for the debate over NATO expansion after the Cold War (Shifrinson 2016; Lanoszka 2020). When expansion is posed in starkly binary terms, it can be seen as either a stabilizing force for Europe or an irresponsible provocation of legitimate Russian security interests fueled by liberal expansionism (McFaul, Sestanovich, and Mearsheimer 2014; Mearsheimer 2014). If deterrence and conflict are continuous variables, however, then the real question is not simply whether NATO should or should not have expanded its security guarantees, but how far. One might thus argue that the first round of expansion to include the Eastern-Central countries (Poland, Hungary, Czech Republic) under the NATO umbrella helped to stabilize an historically conflict-prone portion of Europe. After the fall of the Soviet Union and during a period of military and economic weakness, moreover, Russia was grudgingly willing to accept a reduction in its European influence. One might also debate whether later rounds which brought in Baltic and Balkan countries made sense in whole or part. This is not the place to debate this history. We merely wish to point out that the alternative perspectives of NATO provocation and Russian aggression are better conceived of as context specific variables rather than absolute qualities of either actor. The right question is not whether NATO should have expanded, but how far.

Just as deterrence varies along the gradient, the contours of the gradient can shift over time. When NATO's relative power was increasing, expansion was defensible. If NATO's relative power decreases for whatever reason, then retrenchment makes more sense. Conversely, declining Russian relative power may enable NATO to bolster the line, rendering today's gray zone provocations more prohibitive tomorrow. Gray zone conflict allows keen observers to map the contours of the deterrence gradient, especially in areas where the "defender" has overreached its ability or will to respond. As information about the gradient is revealed, actors can take steps to shore up defenses and reassess priorities. Russia has advertised its willingness to interfere in elections, distort public debate, mobilize nationalist movements, and engage in other provocations. This in turn has led Western governments and publics to heighten awareness, increased vigilance, renewed defenses,

For Peer Review

and deterrence postures. Much as the shooting down of the Malaysian Airlines flight over Donetsk led both to renewed debate within NATO about intervention and to greater restraint in Moscow, so too the lowering of credible escalation thresholds can help to contain risk-averse opportunists. Just as gray zone conflict is symptomatic of deterrence success, the increasing incidence of Russian provocation may be symptomatic of a closing window for its effectiveness, such as it is.

The very fact that an adversary opts to engage in limited conflict suggests both vulnerabilities and opportunities. Instead of worrying about Western paralysis in response to Russian cunning, we can acknowledge that NATO has already blocked Russia from wielding even greater influence. NATO's implicit general deterrence posture has arguably succeeded in keeping more extreme forms of Russian aggression in check. The unfortunate fact remains, however, that a simple remedy for gray zone conflict does not exist, if only because there is always some uncertainty about the precise contours of the deterrence gradient. We may be able to choose *how* our adversary confronts us, but not *whether*. Deterrence is not an on/off switch, but a rheostat, providing a range of causes and variable effects. Gray zone conflict is, then, not simply deterrence failure but also a modest kind of success.

For Peer Review

References

Altman, Dan. 2018. "Advancing Without Attacking: The Strategic Game Around the Use of Force." *Security Studies* 27 (1): 58–88. https://doi.org/10.1080/09636412.2017.1360074.

Amara, Jomana, and Martins Paskevics. 2010. "Unfulfilled Promises: The Impact of Accession on Military Expenditure Trends for New NATO Members." *Comparative Strategy* 29 (5): 432–49. https://doi.org/10.1080/01495933.2010.520988.

Arel-Bundock, Vincent, and Krzysztof J. Pelc. 2018. "When Can Multiple Imputation Improve Regression Estimates?" *Political Analysis* 26 (2): 240–45. https://doi.org/10.1017/pan.2017.43.

Baliga, Sandeep, Ethan Bueno De Mesquita, and Alexander Wolitzky. 2020. "Deterrence with Imperfect Attribution." *American Political Science Review* 114 (4): 1155–78. https://doi.org/10.1017/S0003055420000362.

Beckley, Michael. 2010. "Economic Development and Military Effectiveness." *Journal of Strategic Studies* 33 (1): 43–79. https://doi.org/10.1080/01402391003603581.

Bennett, Andrew, and Colin Elman. 2007. "Case Study Methods in the International Relations Subfield." *Comparative Political Studies* 40 (2): 170–95. https://doi.org/10.1177/0010414006296346.

Binnendijk, Anika. 2020. "Understanding Russian Black Sea Power Dynamics Through National Security Gaming." Santa Monica, CA: RAND Corporation.

Bragg, Belinda. 2017. "Integration Report: Gray Zone Conflicts, Challenges, and Opportunities." Arlington, VA.

Brands, Hal. 2016. "Paradoxes of the Gray Zone." SSRN Scholarly Paper ID 2737593. Rochester, NY: Social Science Research Network.

Brecher, Michael, and Jonathan Wilkenfeld. 1997. *A Study of Crisis*. University of Michigan Press.

Brodie, Bernard. 1957. "More About Limited War." Edited by RN Rear Admiral Sir Anthony W. Buzzard, Robert E. Osgood, and P. M. S. Blackett. *World Politics* 10 (1): 112–22. https://doi.org/10.2307/2009228.

Buuren, Stef van. 2012. *Flexible Imputation of Missing Data*. CRC Press.

Carcelli, Shannon, and Erik A. Gartzke. 2017. "The Diversification of Deterrence: New Data and Novel Realities." In *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.745.

For Peer Review

Carnegie, Allison, and Austin Carson. 2018. "The Spotlight's Harsh Glare: Rethinking Publicity and International Order." *International Organization* 72 (3): 627–57. https://doi.org/10.1017/S0020818318000176.

Carson, Austin. 2016. "Facing Off and Saving Face: Covert Intervention and Escalation Management in the Korean War." *International Organization* 70 (1): 103–31. https://doi.org/10.1017/S0020818315000284.

———. 2018. *Secret Wars: Covert Conflict in International Politics*. Princeton Studies in International History and Politics. Princeton, NJ: Princeton University Press.

Casey, Adam, and Lucan Ahmad Way. 2017. "Russian Electoral Interventions, 1991-2017." Scholars Portal Dataverse. https://doi.org/10.5683/SP/BYRQQS.

Coe, Andrew J. 2018. "Containing Rogues: A Theory of Asymmetric Arming." *The Journal of Politics* 80 (4): 1197–1210. https://doi.org/10.1086/698845.

Danilovic, Vesna. 2001. "The Sources of Threat Credibility in Extended Deterrence." *Journal of Conflict Resolution* 45 (3): 341–69. https://doi.org/10.1177/0022002701045003005.

Danilovic, Vesna, and Joe Clare. 2010. "Deterrence and Crisis Bargaining." In *Oxford Research Encyclopedia of International Studies*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190846626.013.78.

Debs, Alexandre, and Nuno P. Monteiro. 2014. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68 (1): 1–31. https://doi.org/10.1017/S0020818313000192.

Deibert, Ronald, Rafal Rohozinski, and Masashi Crete-Nishihata. 2012. "Cyclones in Cyberspace: Information Shaping and Denial in the 2008 RussiaGeorgia War." *Security Dialogue* 43 (1): 3–24. https://doi.org/10.1177/0967010611431079.

Driscoll, Jesse, and Daniel Maliniak. 2016. "With Friends Like These: Brinkmanship and Chain-Ganging in Russia's Near Abroad." *Security Studies* 25 (4): 585–607. https://doi.org/10.1080/09636412.2016.1220208.

Driscoll, Jesse, and Zachary Steinert-Threlkeld. 2020. "Social Media and Russian Territorial Irredentism: Some Facts and a Conjecture." *Post-Soviet Affairs* 36 (2): 101–21.

Dunford, Joseph. 2016. "Gen. Dunford's Remarks and Q&A." Center for Strategic and International Studies.

Early, Bryan, and Victor Asal. 2018. "Nuclear Weapons, Existential Threats, and the StabilityInstability Paradox." *The Nonproliferation Review* 25 (3-4): 223–47. https://doi.org/10.1080/10736700.2018.1518757.

Fallon, Michael. 2017. "Speech Delivered by Secretary of State for Defence Sir Michael Fallon at the RUSI Landwarfare Conference." Speech.

For Peer Review

Farmanfarmaian, Roxane. 2021. "Strategies and Ethics of Hybrid Warfare in the Gulf: Cyberwar and AI in the High-Stakes Struggle Between Iran and Saudi Arabia." London, England: Cambridge Middle East and North Africa Forum.

Fearon, James D. 1997. "Signaling Foreign Policy Interests Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41 (1): 68–90. https://doi.org/10.1177/0022002797041001004.

———. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414. https://doi.org/10.1017/S0020818300033324.

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): 149–69. https://doi.org/10.1111/j.1540-5907.2010.00486.x.

Filson, Darren, and Suzanne Werner. 2002. "A Bargaining Model of War and Peace: Anticipating the Onset, Duration, and Outcome of War." *American Journal of Political Science* 46 (4): 819–37. https://doi.org/10.2307/3088436.

Freedman, Lawrence. 2004. *Deterrence*. United Kingdom: Wiley.

Galula, David. 1964. *Counterinsurgency Warfare: Theory and Practice*. Hailer Publishing.

Gartzke, Erik, and Matthew Kroenig. 2009. "A Strategic Approach to Nuclear Proliferation." *Journal of Conflict Resolution* 53 (2): 151–60. https://doi.org/10.1177/0022002708330039.

George, Alexander, and Richard Smoke. 1989. "Deterrence and Foreign Policy." *World Politics* 41 (2): 170–82. https://doi.org/10.2307/2010406.

Gleditsch, Kristian S., and Michael D. Ward. 1999. "A Revised List of Independent States Since the Congress of Vienna." *International Interactions* 25 (4): 393–413. https://doi.org/10.1080/03050629908434958.

Goldstein, Lyle. 2017. "The USChina Naval Balance in the Asia-Pacific: An Overview." *The China Quarterly* 232: 904–31. https://doi.org/10.1017/S030574101700131X.

Green, Michael, Kathleen Hicks, Zack Cooper, John Schaus, and Jake Douglas. 2017. *Countering Coercion in Maritime Asia: The Theory and Practice of Gray Zone Deterrence*. Rowman & Littlefield.

Gurantz, Ron, and Alexander V. Hirsch. 2017. "Fear, Appeasement, and the Effectiveness of Deterrence." *The Journal of Politics* 79 (3): 1041–56. https://doi.org/10.1086/691054.

Hart, Sir Basil Henry Liddell. 1954. *Strategy: The Indirect Approach*. Faber & Faber.

For Peer Review

Healey, Jason, and Robert Jervis. 2020. "The Escalation Inversion and Other Oddities of Situational Cyber Stability." *Texas National Security Review* 3 (4): 30–53. https://doi.org/http://dx.doi.org/10.26153/tsw/10962.

Hicks, Kathleen H., and Alice Hunt Friend. 2019. "By Other Means Part I: Campaigning in the Gray Zone." Lanham: Center for Strategic & International Studies.

Hoffman, Frank G. 2007. "Conflict in the 21st Century: The Rise of Hybrid Wars." Arlington, VA: Potomac Institute for Policy Studies.

Holmes, James R., and Toshi Yoshihara. 2017. "Deterring China in the "Gray Zone": Lessons of the South China Sea for U.S. Alliances." *Orbis* 61 (3): 322–39. https://doi.org/10.1016/j.orbis.2017.05.002.

Hughes, Geraint. 2020. "War in the Grey Zone: Historical Reflections and Contemporary Implications." *Survival* 62 (3): 131–58. https://doi.org/10.1080/00396338.2020.1763618.

Huth, Paul K. 1999. "Deterrence and International Conflict: Empirical Findings and Theoretical Debates." *Annual Review of Political Science* 2 (1): 25–48. https://doi.org/10.1146/annurev.polisci.2.1.25.

Jackson, Colin F. 2016. "Information Is Not a Weapons System." *Journal of Strategic Studies* 39 (5-6): 820–46. https://doi.org/10.1080/01402390.2016.1139496.

Jackson, Van. 2017. "Tactics of Strategic Competition: Gray Zones, Redlines, and Conflict Before War." *Naval War College Review* 70 (3): 39–61.

Janičatová, Silvie, and Petra Mlejnková. 2021. "The Ambiguity of Hybrid Warfare: A Qualitative Content Analysis of the United Kingdom's PoliticalMilitary Discourse on Russia's Hostile Activities." *Contemporary Security Policy*, February. https://doi.org/10.1080/13523260.2021.1885921.

Jasper, Scott. 2020. *Russian Cyber Operations: Coding the Boundaries of Conflict.* Georgetown University Press.

Jensen, Benjamin, Brandon Valeriano, and Ryan Maness. 2019. "Fancy Bears and Digital Trolls: Cyber Strategy with a Russian Twist." *Journal of Strategic Studies* 42 (2): 212–34. https://doi.org/10.1080/01402390.2018.1559152.

Jervis, Robert. 1984. *The Illogic of American Nuclear Strategy.* Cornell University Press.

Joseph, Michael F. 2021. "A Little Bit of Cheap-Talk Is a Dangerous Thing: States Can Communicate Intentions Persuasively and Raise the Risk of War in the Processes." *The Journal of Politics* 83 (1): 166–81. https://doi.org/10.1086/709145.

For Peer Review

Joubert, Vincent. 2012. "Five Years After Estonia's Cyber Attacks: Lessons Learned for NATO?" 76. Rome, Italy: NATO Defense College.

Kilcullen, David. 2010. *Counterinsurgency.* Hurst.

Kissinger, Henry. 1955. "Military Policy and Defense of the "Grey Areas"." *Foreign Affairs* 33 (3): 416–28. https://doi.org/10.2307/20031108.

———. 1957. "Strategy and Organization." *Foreign Affairs* 35 (3): 379–94. https://doi.org/10.2307/20031235.

Kostyuk, Nadiya, and Yuri Zhukov. 2019. "Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events?" *Journal of Conflict Resolution* 63 (2): 317–47. https://doi.org/10.1177/0022002717737138.

Lanoszka, Alexander. 2020. "Thank Goodness for NATO Enlargement." *International Politics* 57: 451–70. https://doi.org/10.1057/s41311-020-00234-8.

———. 2016. "Russian Hybrid Warfare and Extended Deterrence in Eastern Europe." *International Affairs* 92 (1): 175–95. https://doi.org/10.1111/1468-2346.12509.

Lebow, Richard Ned. 2010. "The Past and Future of War." *International Relations* 24 (3): 243–70. https://doi.org/10.1177/0047117810377277.

Lieberman, Elli. 2012. *Reconceptualizing Deterrence: Nudging Toward Rationality in Middle Eastern Rivalries.* Routledge.

Lindsay, Jon R., and Erik A. Gartzke. 2018. "Coercion Through Cyberspace: The Stability-Instability Paradox Revisited." In *Coercion: The Power to Hurt in International Politics*, edited by Kelly M. Greenhill and Peter Krause. New York, NY: Oxford University Press.

Marten, Kimberly. 2015. "Putin's Choices: Explaining Russian Foreign Policy and Intervention in Ukraine." *The Washington Quarterly* 38 (2): 189–204. https://doi.org/10.1080/0163660X.2015.1064717.

Matisek, Jahara W. 2017. "Shades of Gray Deterrence: Issues of Fighting in the Gray Zone." *Journal of Strategic Security* 10 (3): 1–26.

Mazarr, Michael. 2015a. "Mastering the Gray Zone: Understanding a Changing Era of Conflict." Research Report. Strategic Studies Institute: US Army War College.

———. 2015b. "Struggle in the Gray Zone and World Order." *War on the Rocks.* http://warontherocks.com/2015/12/struggle-in-the-gray-zone-and-world-order/.

McCarthy, Michael C., Matthew A. Moyer, and Brett H. Venable. 2019. "Deterring Russia in the Gray Zone."

For Peer Review

Carlisle Barracks, PA: Strategic Studies Institute.

McCormack, Daniel, and Henry Pascoe. 2017. "Sanctions and Preventive War." *Journal of Conflict Resolution* 61 (8): 1711–39. https://doi.org/10.1177/0022002715620471.

McFaul, Michael. 2020. "Putin, Putinism, and the Domestic Determinants of Russian Foreign Policy." *International Security* 45 (2): 95–139. https://doi.org/10.1162/isec_a_00390.

McFaul, Michael, Stephen Sestanovich, and John Mearsheimer. 2014. "Faulty Powers." *Foreign Affairs*.

Mearsheimer, John. 2014. "Why the Ukraine Crisis Is the West's Fault." *Foreign Affairs*.

Nagl, John. 2005. *Learning to Eat Soup with a Knife: Counterinsurgency Lessons from Malaya and Vietnam*. University of Chicago Press.

Olson, Erik. 2015. "America's Not Ready for Today's Gray Wars." *Defense One*. https://www.defenseone.com/ideas/2015/12/a not-ready-todays-gray-wars/124381/.

Omitoogun, Wuyi, and Elisabeth Skons. 2006. "Military Expenditure Data: A 40-Year Overview." In *SIPRI Yearbook 2006: Armaments, Disarmament and International Security*, 269–94. Stockholm International Peace Research Institute.

O'Rourke, Lindsey. 2018. *Covert Regime Change: America's Secret Cold War*. Cornell Studies in Security Affairs. Ithaca, NY: Cornell University Press.

Osgood, Robert. 1969. "The Reappraisal of Limited War." *The Adelphi Papers* 9 (54): 41–54. https: //doi.org/10.1080/05679326908448127.

Pettyjohn, Stacie L., and Becca Wasser. 2019. "Competing in the Gray Zone: Russian Tactics and Western Responses." Santa Monica, CA: RAND Corporation.

Plana, Sara. 2020. "Proxy War: The 'Least Bad Option' or the 'Second-to-Last Resort'?" *Texas National Security Review*, March.

Posen, Barry. 2003. "Command of the Commons: The Military Foundation of U.S. Hegemony." *International Security* 28 (1): 5–46. https://doi.org/10.1162/016228803322427965.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169–203. https://doi.org/10.1017/S0020818306060061.

———. 2015. "Nuclear Brinkmanship, Limited War, and Military Power." *International Organization* 69 (3): 589–626. https://doi.org/10.1017/S0020818315000028.

For Peer Review

Quackenbush, Stephen L. 2006. "Not Only Whether but Whom: Three-Party Extended Deterrence." *Journal of Conflict Resolution* 50 (4): 562–83. https://doi.org/10.1177/0022002706290431.

———. 2011. "Deterrence Theory: Where Do We Stand?" *Review of International Studies* 37 (2): 741–62. https://doi.org/10.1017/S0260210510000896.

Ramsay, Kristopher W. 2017. "Information, Uncertainty, and War." *Annual Review of Political Science* 20 (1): 505–27. https://doi.org/10.1146/annurev-polisci-051215-022729.

Rid, Thomas. 2013. "Cyberwar and Peace." *Foreign Affairs*.

Sagan, Scott, and Kenneth Waltz. 2003. *The Spread of Nuclear Weapons: A Debate Renewed*. Norton.

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

Schmidt, Andreas. 2013. "The Estonian Cyberattacks." In *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*, edited by Jason Healey, 174–93. Cyber Conflict Studies Association.

Schram, Peter. 2021. "Hassling: How States Prevent a Preventive War." *American Journal of Political Science* 62 (2): 294–308. https://doi.org/10.1111/ajps.12538.

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

Schweller, Randall. 1996. "Neorealism's Status-Quo Bias: What Security Dilemma?" *Security Studies* 5 (3): 90–121. https://doi.org/10.1080/09636419608429277.

Shifrinson, Joshua R. Itzkowitz. 2016. "Deal or No Deal? The End of the Cold War and the U.S. Offer to Limit NATO Expansion." *International Security* 40 (4): 7–44. https://doi.org/10.1162/ISEC_a_00236.

Shultz, George. 1986. "Low-Intensity Warfare: The Challenge of Ambiguity." Conference Address. National Defense University, Washington, DC.

Singer, David, Stuart Bremer, and John Stuckey. 1972. "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." In *Peace, War, and Numbers*, 19–48. Sage Publications.

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

Smith, Alastair, and Allan C. Stam. 2004. "Bargaining and the Nature of War." *Journal of Conflict Resolution* 48 (6): 783–813. https://doi.org/10.1177/0022002704268026.

For Peer Review

Snyder, Glenn. 1965. "The Balance of Power and the Balance of Terror." In *World in Crisis: Readings in International Relations*, edited by Frederick Hartmann, 180–91. New York: The Macmillan Company.

Sobek, David, and Joe Clare. 2013. "Me, Myself, and Allies: Understanding the External Sources of Power." *Journal of Peace Research* 50 (4): 469–78. https://doi.org/10.1177/0022343313484047.

Spaniel, William. 2019. *Bargaining over the Bomb: The Successes and Failures of Nuclear Negotiations*. Cambridge University Press.

Taber, Robert. 1965. *War of the Flea: The Classic Study of Guerrilla Warfare*. L. Stewart.

Thornton, Rod. 2015. "The Changing Nature of Modern Warfare." *The RUSI Journal* 160 (4): 40–48. https://doi.org/10.1080/03071847.2015.1079047.

Tor, Uri. 2015. "'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence." *Journal of Strategic Studies* 40 (1-2): 92–117. https://doi.org/10.1080/01402390.2015.1115975.

Traynor, Ian. 2007. "Russia Accused of Unleashing Cyberwar to Disable Estonia." *The Guardian*, May.

Turbiville, Graham. 2002. "Preface: Future Trends in Low Intensity Conflict." *Low Intensity Conflict & Law Enforcement* 11 (2-3): 155–63. https://doi.org/10.1080/0966284042000279957.

Valeriano, Brandon, and Ryan C Maness. 2014. "The Dynamics of Cyber Conflict Between Rival Antagonists, 200111." *Journal of Peace Research* 51 (3): 347–60. https://doi.org/10.1177/0022343313518940.

Wagner, R. Harrison. 2000. "Bargaining and War." *American Journal of Political Science* 44 (3): 469–84. https://doi.org/10.2307/2669259.

Weidmann, Nils, Doreen Kuse, and Kristian Skrede Gleditsch. 2010. "The Geography of the International System: The CShapes Dataset." *International Interactions* 36 (1): 86–106. https://doi.org/10.1080/03050620903554614.

Wirtz, James J. 2017. "Life in the "Gray Zone": Observations for Contemporary Strategists." *Defense & Security Analysis* 33 (2): 106–14. https://doi.org/10.1080/14751798.2017.1310702.

Zagare, Frank C. 1996. "Classical Deterrence Theory: A Critical Assessment." *International Interactions* 21 (4): 365–87. https://doi.org/10.1080/03050629608434873.

Zagare, Frank C., and D. Marc Kilgour. 1998. "Deterrence Theory and the Spiral Model Revisited." *Journal of Theoretical Politics* 10 (1): 59–87. https://doi.org/10.1177/0951692898010001003.

For Peer Review

Online Appendix: Supporting Information for *The Shadow of Deterrence: Why capable actors engage in contests short of war*

Author names redacted

2021-05-24

Contents

This appendix accompanies the paper "The Shadow of Deterrence: Why capable actors engage in contests short of war". It provides supplemental information concerning proofs for the formal model, the data set of Russian gray zone campaigns introduced in the paper, and robustness checks and alternate specifications for the statistical model.

1 Formal Model

1.1 Formal statement of assumptions

We formally express the assumption that the kinks in the P function are never activated in equilibrium. Letting $\tilde{g_C}$ and $\tilde{g_D}$ denote the optimal levels selected by C and D conditional on the actors selecting into

For Peer Review

gray zone conflict (these are defined below), when Assumption 1 holds, the "min-max" statements in the P function will never be relevant to analysis.

Assumption 1*: In equilibrium,* $\rho_0 < P(\tilde{g_C}, \tilde{g_D}) < \rho_W$.

Based on the optimal $\tilde{g_C}$ and $\tilde{g_D}$ (solved below), this condition amounts to $\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D} > 0$ and $0 < \rho_W - \rho_0 - \frac{\theta}{2\beta_C} + \frac{1}{2\beta_D}$ if $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, and $\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D} > 0$ and $\kappa_D - \frac{1}{4\beta_D} < 0$ if $\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D} \leq \frac{\theta}{2\beta_C}$.

1.2 Proving Proposition 1

1.2.1 Equilibrium Intuition

Outside of gray zone conflict, C will prefer the status quo to initially going to war when

$$\theta\rho_0 \geq \theta\rho_W - \kappa_C$$

or

$$\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}.$$

Now we discuss the intuition of the equilibrium in the paper. Assume that C is optimally selecting a g_C^* such that the game ends in gray zone conflict (in other words assume that $w_C^* = 0$ and $g_C^* \geq 0$). Also assume that D selects an optimal g_D^* such that $g_D^* \leq g_C^*$ (this will be borne out by Assumption 1). D selects g_D^* characterized by

$$g_D^* \in argmax_{g_D \geq 0} \left\{1 - \rho_0 - g_C + g_D - \beta_D g_D^2\right\}.$$

We take first-order conditions with respect to g_D and solve the expression above to identify the optimal level of D's gray zone response g_D^*. This unique value is

$$g_D^* = \frac{1}{2\beta_D}.$$

Using the expression for g_D^*, D's utility in terms of the selected g_C^* is $U_D = 1 - \rho_0 - g_C^* + \frac{1}{4\beta_D}$.

We can then begin considering C's utility. There are two matters to consider. First, it could be that C will select an optimal g_C^* that is constrained by D's willingness to go to war. Essentially, if $g_C > \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, then D's utility from war is greater than D's utility from gray zone conflict; thus, if C wants to remain in gray zone conflict and will be constrained by D's deterrent threat, C will select $\hat{g_C}$, where $\hat{g_C}$ is the greatest g_C that would make D indifferent between gray zone conflict and war, or

$$\hat{g_C} = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}.$$

Second, C may select an optimal g_C^* that is constrained by their own internal costs. When this is the case, C will select $\check{g_C}$, defined by the optimization

$$\check{g_C} \in argmax_{g_C \geq 0} \left\{\theta\left(\rho_0 + g_C - \frac{1}{2\beta_D}\right) - \beta_C g_C^2\right\},$$

which yields

$$\check{g_C} = \frac{\theta}{2\beta_C}.$$

Before discussing the true behavior, we highlight two things that do not happen. First, note that C will never select an g_C that provokes D to go to war in the final stage, because this is strictly worse than initially going

For Peer Review

to war. Second, note that C will never select into gray zone conflict (i.e. set $w_R = 0$ and $g_C^* > 0$) if g_D^* as defined above is greater than g_C^* because C could do strictly better not paying the costs of war and selecting into the status quo ($g_C^* = 0$).

With this is place, if C optimally selects into gray zone conflict, C will select $g_C^* = \tilde{g_C}$, where

$$\tilde{g_C} = min\left\{\hat{g_C}, \check{g_C}\right\}.$$

We have now characterized what happens within gray zone conflict. We now need to describe how the game optimally plays out across the possibility of selecting into the status quo, war (at the onset; $w_A = 1$), or gray zone conflict. Because C moves first, this is ultimately C's choice. We can calculate C's decision within the two cases of gray zone conflict.

First, we consider the case when $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. This condition implies that the selected gray zone conflict will be constrained by D's deterrent threat and not C's internal costs. So, if C selects into gray zone conflict, C will select $g_C^* = \hat{g_C} = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. We can then express C's behavior in terms of θ. C prefers the status quo to gray zone conflict when

$$\theta\rho_0 \geq \theta\left(\rho_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$$

or

$$\theta \leq \frac{\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)}.$$

Note that the above derivation relies on $\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D} > 0$, lest the inequality sign would flip. This holds by Assumption 1.

Next, C prefers war to gray zone conflict when

$$\theta\rho_W - \kappa_C > \theta\left(\rho_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$$

or

$$\theta > \frac{\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}.$$

Note that the above derivation relies on $\frac{1}{4\beta_D} - \kappa_D > 0$, lest the inequality sign would flip. This holds by Assumption 1.

Next, we assume $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. This condition implies that the selected gray zone conflict will be constrained by C's internal costs and not D's deterrent threat. So, if C selects into gray zone conflict, C will select $g_C^* = \check{g_C} = \frac{\theta}{2\beta_C}$. We can then express C's behavior in terms of θ. C prefers the status quo to gray zone conflict when

$$\theta\rho_0 \geq \theta\rho_0 + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}$$

or

$$0 \geq \theta\left(\frac{\theta}{4\beta_C} - \frac{1}{2\beta_D}\right).$$

Next, C prefers war to gray zone conflict when

$$\theta\rho_W - \kappa_C > \theta\rho_0 + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}$$

For Peer Review

or

$$\theta > \frac{\kappa_C}{\rho_W - \rho_0 - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}.$$

Note that the above derivation relies on $\rho_W - \rho_0 - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D} > 0$, lest the inequality sign would flip. This holds by Assumption 1.

When $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, it is straightforward to see that, for a great enough θ, C's will declare war. We now demonstrate this for $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. We calculate

$$\frac{d}{d\theta}\left(U_C(war) - U_C(grayzone)\right) = \frac{d}{d\theta}\left(\theta\rho_W - \kappa_C - \left(\theta\rho_0 + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}\right)\right) = \quad \rho_W - \rho_0 - \frac{\theta}{2\beta_C} + \frac{1}{2\beta_D}$$

By Assumption 1, the right hand side is positive. Therefore, as θ increases, the war payoffs are rising faster than the gray zone payoffs, and for a great enough θ C will prefer war to gray zone conflict.

With all of this defined, we can characterize C's strategy in terms of θ; as θ increases, C prefers more degrees of conflict (i.e. larger g_C^*'s or war) to get what they want.

1.2.2 Equilibrium Behavior

Proposition 1A and the text below contains a more complete discussion of the equilibrium behavior characterized in Proposition 1.

Proposition 1A: *In equilibrium, the game will play out in the following manner.*

Case 1, $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*:*

- *1.A. If* $\theta \leq \frac{\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)}$ *and* $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$*, then C accepts the status quo. C selects* $w_C^* = 0$ *and* $g_C^* = 0$*, and D selects* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \rho_0$ *and* $U_C = \theta\rho_0$*.*
- *1.B. If* $\theta > \frac{\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$ *and* $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$*, then C declares war. C selects* $w_C^* = 1$*, and payoffs are* $U_D = 1 - \rho_W - \kappa_D$ *and* $U_C = \theta\rho_W - \kappa_A$*.*
- *1.C. Otherwise, the game end in gray zone conflict where* C's *limited challenge is constrained by D's deterrent threat. C selects* $w_C^* = 0$ *and* $g_C^* = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*, and D selects* $w_D^* = 0$ *and* $g_D^* = \frac{1}{2\beta_D}$*. Payoffs are* $U_D = 1 - \rho_W - \kappa_D$ *and* $U_C = \theta\left(\rho_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$*.*

Case 2, $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$*:*

- *2.A. If* $\theta \leq \frac{2\beta_C}{\beta_D}$ *and* $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$*, then C accepts the status quo. C selects* $w_C^* = 0$ *and* $g_C^* = 0$*, and D selects* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \rho_0$ *and* $U_C = \theta\rho_0$*.*
- *2.B. If* $\theta > \frac{\kappa_C}{\rho_W - \rho_0 - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$ *and* $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$*, then C declares war. C sets* $w_C^* = 1$*. Payoffs are* $U_D = 1 - \rho_W - \kappa_D$ *and* $U_C = \theta\rho_W - \kappa_A$*.*
- *2.C. Otherwise, the game will end in gray zone conflict where* C's *limited challenge is constrained by* C's *internal efficiency. C selects* $w_C^* = 0$ *and* $g_C^* = \frac{\theta}{2\beta_C}$*, and D selects* $w_D^* = 0$ *and* $g_D^* = \frac{1}{2\beta_D}$*. Payoffs are* $U_D = 1 - \rho_0 - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D}$*, and* $U_C = \theta\rho_0 + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}$*.*

Working backwards, D will declare war for all $g_C > \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. If $g_C \leq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, D will select $g_D = min\left\{\frac{1}{2\beta_D}, g_C\right\}$. When $g_D = \frac{1}{2\beta_D}$, D is selecting their optimal level of gray zone response based on their internal optimization. When $g_D = g_C$, it implies that D would be willing to select a greater gray zone response, but does not need to, essentially driving the political impact of C's limited challenges back to zero (at cost).

For Peer Review

1.3 Observation 1 Discussion

Assume for now the parameters are such that the Case 1.C. conditions hold, and consider what happens when κ_D decreases. Because here C selects the greatest level of limited challenges that will not provoke D to war, C's selected g_C^* is a decreasing function of κ_D; therefore, because g_D^* is fixed, the final extent of gray zone conflict will be less. Of course, the analysis does not stop there. Improvements in D's willingness to go to war constrain how useful gray zone conflict is to R, and, within Case 1.C., C's utility is decreasing in $-\kappa_D$.[1] Thus, if κ_D becomes small enough, C will leave gray zone conflict and instead select into either accepting the status quo (entering into Case 1A) or going to war (entering into Case 1B). Additionally, it is worthwhile noting that as κ_D decreases, the condition that selects into Case 1 (over Case 2) has more slack, implying that improvements in D's willingness to go to war will keep D within Case 1.

Now assume the parameters are such that the Case 2.C. conditions hold, and consider what happens when κ_D decreases. Note that this will not change the selected g_C^* here, but it could break the inequality $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$ that determines whether the equilibrium is defined in Case 1 or Case 2. Thus, for a small enough κ_D, the conditions for Case 2 will break and the conditions for Case 1 will hold. When this happens, either the selected g_C^* is increasing in κ_D (Case 1.C.) or gray zone conflict is not selected (Case 1.A. or 1.B.).

1.4 Extension 1: Endogenous β_D

In the model in the paper, we treated D's gray zone efficiency β_D as exogenous. In some special cases or under some conditions, this may be too strong an assumption. In this section, we characterize an equilibrium for the game when D can have complete flexibility in selecting some $\beta_D \geq \underline{\beta_D} > 0$, where β_D cannot equal zero because D's costs from their gray zone response will then be undefined.[2] The key take away from this extension is that if β_D is endogenous (and its selection cost-less), then D's selection of β_D^* will be arbitrated by two properties. As the first property, it matters whether C prefers war to the status quo (formally, if C is type $\theta > \frac{\kappa_D}{\rho_W - \rho_0}$), or C prefers the status quo to war ($\theta \leq \frac{\kappa_D}{\rho_W - \rho_0}$). When C prefers the status quo to war, then D is in a position where D can, by selecting a low enough β_D, influence C to stop undertaking limited challenges and select into the status quo. Intuitively, when D is very good at gray zone conflict, D would select a high g_D^*, which makes gray zone conflict less productive for C. But, when C prefers war to the status quo, then D could pressure C to stop undertaking limited challenges, but this will result in C going to war with D.

As the second property, D's decision will also be arbitrated by whether D can select a gray zone efficiency β_D^* that pushes C into a level of gray zone conflict where the deterrent threat does not bind. Recall that if C optimally conducts gray zone conflict, C selects $g_C^* = min\{\hat{g_C}, \check{g_C}\}$, implying that C will either select an optimal $g_C^* = \check{g_C} = \frac{\theta}{2\beta_C}$ based on their own internal cost-benefit analysis, or select an optimal $g_C^* = \hat{g_C} = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$ tailored to make D indifferent between war and gray zone conflict (where the deterrent threat binds), with C ultimately choosing the smaller of the two. This means that if D can select a small enough β_D so that $\check{g_C} < \hat{g_C}$, then C will selecting a level of limited challenge that is below the point that would make D indifferent between war and gray zone conflict, thus granting D some surplus.

The above two properties interact. D will always prefer the status quo to gray zone conflict where the deterrent threat doesn't bind, and gray zone conflict where the deterrent threat doesn't bind to gray zone conflict where the deterrent threat does bind or war. Proposition A identifies how D selects β_D^* in one possible equilibrium. Note that this is not the only possible equilibrium.[3]

Proposition A. *As one equilibrium, in the game with endogenous β_D, D will select the following levels of β_D^*:*

[1] This follows from $\frac{d}{d\kappa_D} U_D = \theta - 2\beta_C \left[\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right] > 0$, as determined by the conditions for Case 1 to hold.

[2] For ease, we will assume that all parameters imply that the selected equilibrium is such that the selected β_D^* is strictly greater than $\underline{\beta_D}$.

[3] Consider the equilibrium space for the range of θ where the selected β_D will either push C into war or gray zone conflict where the deterrent threat binds. In the figure below, this is the far right region of the graph. Here D can select any β_D and it will grant D the same final expected utility of their wartime utility.

For Peer Review

Case 1: $\theta \leq \frac{\kappa_D}{\rho_W - \rho_0}$*:*

- *1.A. We define* $\tilde{\beta_D}$ *as* $\theta = \frac{2\beta_C}{\tilde{\beta_D}}$ *. So long that* $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\tilde{\beta_D}}$*, then D selects* $\beta_D^* = \tilde{\beta_D}$*. The game will proceed as defined in Proposition 1, Case 2.A., where the final outcome is the status quo.*
- *1.B. Otherwise, D selects* $\beta_D^* = \hat{\beta_D}$*, here* $\hat{\beta_D}$ *is defined implicitly as* $\theta = \frac{\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\hat{\beta_D}}\right)^2}{\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\hat{\beta_D}}\right)}$ *(also note from earlier assumptions* $\hat{\beta}_D > 0$*). The game will proceed as defined in Proposition 1, Case 1.A., where the final out come is the status quo.*

Case 2: $\theta > \frac{\kappa_D}{\rho_W - \rho_0}$

- *2.A. We define* $\check{\beta_D}$ *implicitly as* $\theta = \frac{\kappa_C}{\left(\rho_W - \rho_0 - \frac{\theta}{4\check{\beta_C}} + \frac{1}{2\check{\beta_D}}\right)}$*. As long as* $\frac{\theta}{2\check{\beta_C}} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\check{\beta_D}}$*, then D selects* $\beta_D^* = \check{\beta_D}$*. The game will proceed as defined in Proposition 1, Case 2.C., where the final outcome is gray zone conflict where C is not bound by D's deterrent threat.*
- *2.B. Otherwise, D selects* $\beta_D^* = \dot{\beta_D}$*, here* $\dot{\beta_D}$ *is defined implicitly as* $\theta = \frac{\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\dot{\beta_D}}\right)^2}{-\kappa_D + \frac{1}{4\dot{\beta_D}}}$*. The game will proceed as defined in Proposition 1, Case 1.C., where the final outcome is gray zone conflict where C is not bound by D's deterrent threat.*

As one example of how this one equilibrium plays out, we adapt Figure 4 in the text. Now the solid black lines denote the selected levels of β_D^* (with $1/\beta_D$ plotted so that greater y-axis values represent greater gray zone efficiencies for D), and the dotted lines separate equilibrium spaces.

Moving left to right, for θ between 1.285 and $\frac{\kappa_C}{\rho_W - \rho_0}$, D's optimal β_D^* is described in Proposition A Case 1.A. As the outcome, C will optimally select into the status quo. For this selected β_D^*, C knows that C would face enough of a challenge in gray zone conflict to make competing there too costly. Thus within this region, D could select a low enough β_D^* to compel C to forgo limited challenges and conflict, and stick to the status quo.

Moving right, for θ between $\frac{\kappa_C}{\rho_W - \rho_0}$ and $2\beta_C(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\check{\beta_D}})$, D's optimal β_D^* is described in Proposition A Case 2.A. As the outcome, C will optimally select into gray zone conflict, but will be constrained by C's internal costs. For this selected β_D^*, D wants to challenge C in gray zone conflict (which a lower β_D^* accomplishes), but does not want to push C into forgoing gray zone conflict, because within this region C prefers war to accepting the status quo. Thus here, D selects the β_D^* where C selects into gray zone conflict and is not bound by the deterrent threat, because this gives D some surplus beyond what war or C selecting gray zone conflict and being bound by the deterrent threat produces.

Finally, for θ between $2\beta_C(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\check{\beta_D}})$ and 1.4, D's optimal β_D^* is described in Case 2.B. As the outcome, C will optimally select into gray zone conflict, and will be constrained by D's deterrent threat. This situation is problematic for D. If D modifies β_D^*, either C will adapt by selecting the new g_C^* that makes D indifferent between war and gray zone conflict, or will go to war over the issue. Within this region, it does not matter what β_D^* is selected, because C will always select an action that gives D their wartime utility.

1.5 Extension 2: Probabilistic Escalation to War

A useful feature of the model above is that everything that occurs is deterministic. Only if a state wants to go to war or wants to enter gray zone conflict does it actually happen. However, this represents a simplification. Perhaps in some cases, one state behaving aggressively in lower-levels of conflict can create an incident that necessitates an escalation to higher levels of conflict. To speak to this issue, we introduce the possibility of probabilistic escalation out of gray zone conflict. Our results are substantively similar, but this change shifts some equilibrium properties. Intuitively, now gray zone conflict can probabilistically lead to C's worst outcome: where C invests in limited challenges, war happens, and C must pay the costs of limited challenges with the costs of war. Strategically, because here gray zone conflict is overall worse for C, C will be more willing to accept the status quo or go to war.

For Peer Review

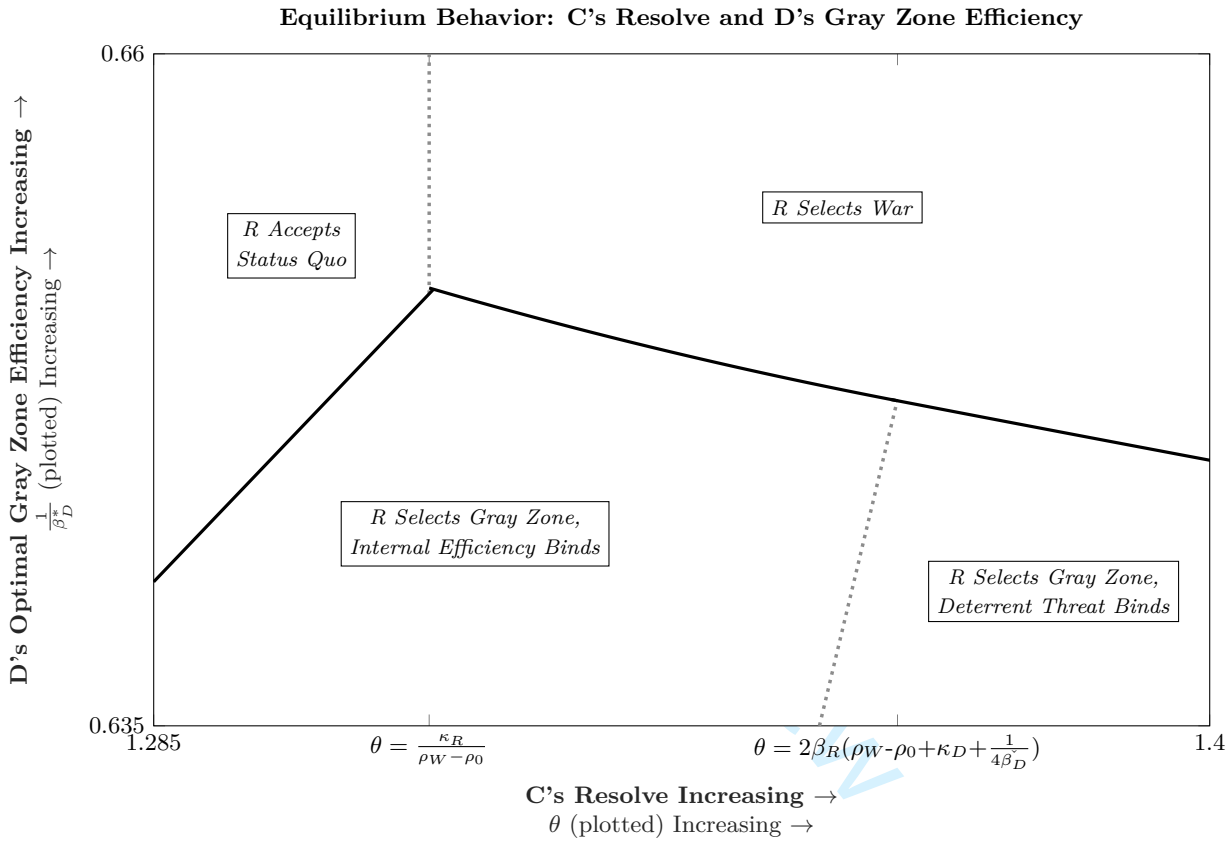

Figure A1: Extension 1: D's Optimal d^*. C's resolve θ and the inverse D's gray zone efficiency $\frac{1}{\beta_D}$ are plotted. The dotted lines separate different kinds of equilibrium play, and the dark black lines denote D's optimal selected β_D. The parameters are $\rho_0 = 0$, $\rho_W = 0.5$, $\beta_C = 1$, $\kappa_C = 0.53$, and $\kappa_D = 0.1$.

There are many possible ways to model this. For ease, we choose (in our opinion) the simplest way, which is that selecting $g_C > 0$ introduces a $1-\zeta \in (0,1)$ likelihood of an escalation to war. Thus, when C selects $g_C > 0$, C's new expected utility is

$$U_C = \theta\left(\zeta P(g_C, g_D) + (1-\zeta)\rho_W\right) - (1-\zeta)\kappa_C - \beta_C g_C.$$

To offer some intuition, g_D^*, $\hat{g_C}$, $\check{g_C}$, and $\tilde{g_C}$ remain the same as it was in the model in the text (as defined in Proposition 1). However, the cut-points that distinguish C's decision to enter into the status quo, gray zone conflict, or war change slightly; overall, the key take-away is that considering probabilistic escalation makes gray zone conflict less appealing relative to the status quo and war.

We express equilibrium behavior in Proposition B. Then below, we derive the new cut-points, Additionally in the derivations, we discuss how the new cut-points imply that gray zone conflict is less appealing and fewer types θ will select into it relative to the game without a probabilistic likelihood of escalation to war from gray zone conflict.

Proposition B: *In equilibrium, the game with a $1-\zeta$ chance of escalation out of gray zone conflict to war will play out in the following manner.*

Case 1, $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$:

- *1.A. If $\theta \leq \frac{(1-\zeta)\kappa_C + \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{(1-\zeta)(\rho_W - \rho_0) + \zeta\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)}$ and $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$, then C accepts the status quo. C selects $w_C^* = 0$ and $g_C^* = 0$, and D selects $w_D^* = 0$ and $g_D^* = 0$.*
- *1.B. If $\theta > \frac{\zeta\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\zeta\left(\frac{1}{4\beta_D} - \kappa_D\right)}$ and $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$, then C declares war. C selects $w_C^* = 1$.*
- *1.C. Otherwise, the game end in gray zone conflict where* C's *limited challenge is constrained by D's deterrent threat. C selects $w_C^* = 0$ and $g_C^* = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$, and (assuming the game does not probabilistically escalate to war) D selects $w_D^* = 0$ and $g_D^* = \frac{1}{2\beta_D}$.*

Case 2, $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$:

- *2.A. If $(1-\zeta)\kappa_C \geq \theta\left((1-\zeta)(\rho_W - \rho_0) + \zeta\left(\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D}\right) - \frac{\theta}{4\beta_C}\right)$ and $\theta \leq \frac{\kappa_C}{\rho_W - \rho_0}$, then C accepts the status quo. C selects $w_C^* = 0$ and $g_C^* = 0$, and D selects $w_D^* = 0$ and $g_D^* = 0$.*
- *2.B. If $\theta > \frac{\zeta\kappa_C}{\left(\zeta\left(\rho_W - \rho_0 - \frac{\theta}{2\beta_C} + \frac{1}{2\beta_D}\right) + \frac{\theta}{4\beta_C}\right)}$ and $\theta > \frac{\kappa_C}{\rho_W - \rho_0}$, then C declares war. C sets $w_C^* = 1$.*[4]
- *2.C. Otherwise, the game will end in gray zone conflict where* C's *limited challenge is constrained by* C's *internal efficiency. C selects $w_C^* = 0$ and $g_C^* = \frac{\theta}{2\beta_C}$, and (assuming the game does not probabilistically escalate to war) D selects $w_D^* = 0$ and $g_D^* = \frac{1}{2\beta_D}$.*

1.5.1 Equilibrium Intuition

First, we consider the case when $\frac{\theta}{2\beta_C} \geq \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. This implies that C will select $g_C^* = \hat{g_C} = \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. We can then express C's behavior in terms of θ. C prefers the status quo to gray zone conflict when

$$\theta\rho_0 \geq \theta\left(\zeta\left(\rho_W + \kappa_D - \frac{1}{4\beta_D}\right) + (1-\zeta)\rho_W\right) - (1-\zeta)\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$$

or

$$\frac{\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\zeta\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)} + \frac{(1-\zeta)(\theta\rho_0 - \theta\rho_W + \kappa_C)}{\zeta\left(\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D}\right)} \geq \theta.$$

[4] *While the right-hand-side of this condition is also increasing in θ, the left-hand-side increases faster with increases in θ.*

For Peer Review

Note that the inequality sign does not flip because, by Assumption 1, $\rho_W - \rho_0 + \kappa_D - \frac{1}{4\beta_D} > 0$. We are able to say that $\frac{\beta_C\left(\rho_W-\rho_0+\kappa_D+\frac{1}{4\beta_D}\right)^2}{\zeta\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)} > \frac{\beta_C\left(\rho_W-\rho_0+\kappa_D+\frac{1}{4\beta_D}\right)^2}{\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)}$ because $\zeta \in (0,1)$. Furthermore, this constraint (on when the status quo is preferred to gray zone conflict) matters only when C prefers the status quo to war, or when $\theta\rho_0 - \theta\rho_W + \kappa_C \geq 0$; this condition implies $\frac{(1-\zeta)(\theta\rho_0-\theta\rho_W+\kappa_C)}{\zeta\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)} \geq 0$, which means $\frac{\beta_C\left(\rho_W-\rho_0+\kappa_D+\frac{1}{4\beta_D}\right)^2}{\zeta\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)} + \frac{(1-\zeta)(\theta\rho_0-\theta\rho_W+\kappa_C)}{\zeta\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)} > \frac{\beta_C\left(\rho_W-\rho_0+\kappa_D+\frac{1}{4\beta_D}\right)^2}{\left(\rho_W-\rho_0+\kappa_D-\frac{1}{4\beta_D}\right)}$, which in turn implies that there are more C's with some resolve θ that will select into the status quo in the game here relative to the game in the text without probabilistic escalation.

Next, C prefers war to gray zone conflict when

$$\theta\rho_W - \kappa_C > \theta\left(\zeta\left(\rho_W + \kappa_D - \frac{1}{4\beta_D}\right) + (1-\zeta)\rho_W\right) - (1-\zeta)\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$$

or

$$\theta > \frac{\zeta\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\zeta\left(\frac{1}{4\beta_D} - \kappa_D\right)}.$$

Note that based on Assumption 1, the above sign does not flip. We can say that $\zeta\kappa_C - \zeta\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2 > \zeta\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2$. This implies that

$$\frac{\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D} = \frac{\zeta\kappa_C - \zeta\beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\zeta\left(\frac{1}{4\beta_D} - \kappa_D\right)} > \frac{\zeta\kappa_C - \beta_C\left(\rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\zeta\left(\frac{1}{4\beta_D} - \kappa_D\right)}.$$

In other words, there are more C's with some resolve θ that will select into war in the game here relative to the game without probabilistic escalation.

Next, we assume $\frac{\theta}{2\beta_C} < \rho_W - \rho_0 + \kappa_D + \frac{1}{4\beta_D}$. This condition implies that the selected gray zone conflict will be constrained by C's internal costs and not D's deterrent threat. So, if C selects into gray zone conflict, C will select $g_C^* = \breve{g_C} = \frac{\theta}{2\beta_C}$. We can then express C's behavior in terms of θ. C prefers the status quo to gray zone conflict when

$$\theta\rho_0 \geq \theta\left(\zeta\left(\rho_0 + \frac{\theta}{2\beta_C} - \frac{1}{2\beta_D}\right) + (1-\zeta)\left(\rho_W\right)\right) - (1-\zeta)\kappa_C - \frac{\theta^2}{4\beta_C}$$

or

$$(1-\zeta)\kappa_C \geq \theta\left((1-\zeta)(\rho_W - \rho_0) + \zeta\left(\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D}\right) - \frac{\theta}{4\beta_C}\right).$$

To speak to this inequality, we will need to consider a few different cases here.

First, it could be possible that $\left((1-\zeta)(\rho_W - \rho_0) + \zeta\left(\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D}\right) - \frac{\theta}{4\beta_C}\right) \leq 0$. When this is the case, then C would never want to select into gray zone conflict as doing so would always be strictly worse for C.

Next, consider when $\left((1-\zeta)(\rho_W - \rho_0) + \zeta\left(\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D}\right) - \frac{\theta}{4\beta_C}\right) > 0$ and $(1-\zeta)(\theta\rho_W - \theta\rho_0 - \kappa_C) > 0$. In this case, C's wartime payoff $\theta\rho_W - \kappa_C$ is greater than C's status quo payoff, meaning that C would never select into the status quo over selecting into war, meaning this constraint would never be activated.

For Peer Review

Finally, consider when $\left((1-\zeta)(\rho_W-\rho_0)+\zeta\left(\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}\right)-\frac{\theta}{4\beta_C}\right)>0$ and $(1-\zeta)(\theta\rho_W-\theta\rho_0-\kappa_C)<0$. We can re-write the above as

$$0 \geq \theta\left(\zeta\left(\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}\right)-\frac{\theta}{4\beta_C}\right)+(1-\zeta)(\theta\rho_W-\theta\rho_0-\kappa_C)$$

Note that $\frac{\theta}{4\beta_C}-\frac{1}{2\beta_D}=\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}-\frac{\theta}{4\beta_C}>\zeta\left(\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}\right)-\frac{\theta}{4\beta_C}$, where the inequality holds by Assumption 1. Altogether, this means that $\theta\left(\frac{\theta}{4\beta_C}-\frac{1}{2\beta_D}\right)>\theta\left(\zeta\left(\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}\right)-\frac{\theta}{4\beta_C}\right)+(1-\zeta)(\theta\rho_W-\theta\rho_0-\kappa_C)$. This implies that there are more C's with some resolve θ that will select into the status quo in the game here relative to the game without probabilistic escalation.

Finally, assuming $\frac{\theta}{2\beta_C}<\rho_W-\rho_0+\kappa_D+\frac{1}{4\beta_D}$, C prefers war to gray zone conflict when

$$\theta\rho_W-\kappa_C>\theta\left(\zeta\left(\rho_0+\frac{\theta}{2\beta_C}-\frac{1}{2\beta_D}\right)+(1-\zeta)\left(\rho_W\right)\right)-(1-\zeta)\kappa_C-\frac{\theta^2}{4\beta_C}$$

or

$$\theta>\frac{\zeta\kappa_C}{\left(\zeta\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)+\frac{\theta}{4\beta_C}\right)}.$$

Note the inequality sign does not flip because $\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)>0$. Furthermore, by that condition, $\zeta\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)+\frac{\theta}{4\beta_C}>\zeta\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)+\zeta\frac{\theta}{4\beta_C}$. Therefore $\frac{\kappa_C}{\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)+\frac{\theta}{4\beta_C}}>\frac{\zeta\kappa_C}{\zeta\left(\rho_W-\rho_0-\frac{\theta}{2\beta_C}+\frac{1}{2\beta_D}\right)+\frac{\theta}{4\beta_C}}$. This implies that there are more C's with some resolve θ that will select into war in the game here relative to the game without a random chance of escalation.

Finally, note that D's strategies in this game are unchanged from the game without probabilistic escalation.

1.6 Extension 3: Endogenous Bargaining and Information Asymmetry

Here we offer one possible microfoundation for a key assumption in the game: that the game begins with C being potentially dissatisfied with the status quo. We do this by keeping the game structure we introduced in the main paper and by adding new initial moves to the game. We grant D the option to establish the "status quo" though an ultimatum offer, we assume D has private information, and we add some additional parameter assumptions. Importantly, reasonable readers may take issue with certain facets of the game form below; for that reason, we wish to highlight that what we present is not the only way for our conflict selection subgame (i.e. C selecting whether to accept, go to war, or engage in a low level challenge, and then D doing the same) to start with a potentially dissatisfied C. Past examinations of the causes for war—like private information (Fearon 1995), commitment problems (Powell 2006), costly peace (Coe 2011), political bias (Jackson and Morelli 2007), and the possibility of behavioral types (Acharya and Grillo 2015)—have all demonstrated that natural circumstances can lead to settings where one state is willing to make an offer to another state, knowing that the offer could lead to some form of conflict.

The take-away from the following analysis is that with endogenous bargaining, C's gray zone actions are still driven by the external deterrent constraint and the internal efficiency constraint, though other factors can also play a role. While it is well established that including information asymmetry can make an empirical analysis of conflict onset less pinned-down (Gartzke 1999), the model does predict that whenever gray zone conflict is observed, C will select a level of gray zone conflict benchmarked either to the external deterrent constraint of D, or based on their own internal efficiency constraint. Based on this prediction, we also include an empirical analysis in A8 below where we re-run our models on the sample of observations where some form of conflict occurred. We find that our key variables approximating the deterrent threat (NATO membership) and internal efficiency (distance from Russia) still predict the intensity of conflict. While these results should

For Peer Review

be taken with precautions—the sample is smaller and (due to the high number of recent attacks) a larger proportion of control variables are missing—this does still suggest our formal model is plausibly describing key drivers for gray zone conflict intensity.

For this new formal model, we assume the following game form.

1. Nature moves first and sets $\rho_W \in \{\underline{\rho}_W, \bar{\rho}_W\}$, with $\underline{\rho}_W < \bar{\rho}_W$. Nature fixes $\bar{\rho}_W$ with probability $\epsilon \in (0,1)$ and $\underline{\rho}_W$ with probability $1-\epsilon$. D observes the selected ρ_W, but C does not. The selected ρ_W can be thought of as D's "type," where D is stronger if they are type $\underline{\rho}_W$.
2. D makes C some offer $x \in \{\underline{x}, \bar{x}\}$, with $\underline{x} < \bar{x}$. It is worthwhile mentioning that issue indivisibility is not a driver of conflict in this game, and if we assumed complete information, there will be no conflict.[5]
3. C either goes to war by setting $w_C = 1$ or sets $w_C = 0$. If C goes to war, the game terminates and C and D receive payoffs $\theta\rho_W - \kappa_C$ and $1 - \rho_W - \kappa_D$ (with $\rho_W \in \{\underline{\rho}_W, \bar{\rho}_W\}$), respectively. If C sets $w_C = 0$, C also selects $g_C \in \mathcal{G}_C = \mathbb{R}_{\geq 0}$, where $g_C = 0$ is walking away from the crisis and accepting the offer, and $g_C > 0$ is conducting some limited, costly military action that shifts the offer in favor of the challenger.
4. As long as C did not previously go to war, D can either escalate to war by setting $w_D = 1$, or not by setting $w_D = 0$ and selecting some gray zone response $g_D \in \mathcal{G}_D = \mathbb{R}_{\geq 0}$, with $g_D = 0$ implying that D does not respond to the limited challenge. When war occurs, C and D receive payoffs $\theta\rho_W - \kappa_C$ and $1 - \rho_W - \kappa_D$. When D selects $g_D \geq 0$, C and D receive payoffs $\theta P(x, g_C, g_D) - \beta_C g_C^2$ and $1 - P(x, g_C, g_D) - \beta_D g_D^2$, with $P(x, g_C, g_D) = max\{min\{\bar{\rho}_W, x + g_C - g_D\}, x\}$.

The payoffs to the game are summarized in the table below. There are two key changes to highlight here.

First, we modify C's utility should D escalate to war after C selects some level of gray zone conflict. Here C faces no costs from this gray zone challenge. This assumption is made primarily for analytic ease. In the mixing equilibrium examined below, under some parameters, C may select a limited challenge that provokes strong-type D's to go to war. When gray zone challenge costs do not "carry-over" (as is assumed in this model) C selects their optimal gray zone challenge purely based on the weak-type D's parameters. If C faces carry-over costs and the carry-over costs of gray zone challenges when D declares war are too high, C may reduce their optimal gray zone challenge in order to mitigate some carry-over costs when C is paired with a strong-type D.[6] So long that the carry-over costs are sufficiently low, sufficiently dampened, or that D is a strong-type with a low-enough probability, this assumption makes little difference.[7]

Second, the P function now falls between the offer x and the expected political war outcome for the weak-type D. This still embraces that gray zone conflict is a limited challenge.

Scenario	**C's utility facing $\underline{\rho}_W$**	**C's utility facing $\bar{\rho}_W$**	**D's utility ($\rho_W \in \{\underline{\rho}_W, \bar{\rho}_W\}$)**
C initially initiates war ($w_C = 0$)	$\theta\underline{\rho}_W - \kappa_C$	$\theta\bar{\rho}_W - \kappa_C$	$1 - \rho_W - \kappa_D$
C and D select gray zone/accept status quo ($w_C = 0$, $g_C \geq 0$, $w_D = 0$, $g_D \geq 0$)	$\theta P(x, g_C, g_D) - \beta_C g_C^2$	$\theta P(x, g_C, g_D) - \beta_C g_C^2$	$1 - P(x, g_C, g_D) - \beta_D g_D^2$
D escalates to war after C acts ($w_C = 0$, $g_C \geq 0$, $w_D = 1$)	$\theta\underline{\rho}_W - \kappa_C$	$\theta\bar{\rho}_W - \kappa_C$	$1 - \rho_W - \kappa_D$

Table A1: Summarized payoffs for actors

We now examine perfect Bayesian Nash equilibria. To accommodate additional types, we slightly adapt Assumptions 1 from above. This is now the following.

***Assumption 1C**: In equilibrium, $x < P(\tilde{g_C}, \tilde{g_D}) < \bar{\rho}_W$, where $x \in \{\underline{x}, \bar{x}\}$ is the selected offer.*

[5] This model can still function with a continuum of offers, though it becomes much more complicated.

[6] If this assumption were not in place, this would change C's selected gray zone challenge cases 1D and 2D within Proposition C, and it would alter the decisions over which case to enter; this is what was observed in the second extension.

[7] As the most reasonable way to accommodate the assumptions in the paper and here, assume that the carry-over costs are non-zero but very small.

For Peer Review

Based on the optimal $\tilde{g_C}$ and $\tilde{g_D}$ (solved below), for a given x, this condition amounts to $\frac{\theta}{2\beta_C} - \frac{1}{2\beta_D} > 0$ and $0 < \bar{\rho}_W - x - \frac{\theta}{2\beta_C} + \frac{1}{2\beta_D}$ if $\frac{\theta}{2\beta_C} < \bar{\rho}_W - x + \kappa_D + \frac{1}{4\beta_D}$, and $\bar{\rho}_W - x + \kappa_D - \frac{1}{4\beta_D} > 0$ and $\kappa_D - \frac{1}{4\beta_D} < 0$ if $\bar{\rho}_W - x + \kappa_D + \frac{1}{4\beta_D} \leq \frac{\theta}{2\beta_C}$.

We also make several assumptions on $\underline{x}$ and $\bar{x}$. First, we assume that if C selects any kind of gray zone challenge after receiving an offer of $\underline{x}$, a strong-type D ($\underline{\rho}_W$) will go to war. This is the following, but notably, there are other ways for this assumption to hold as well (Abreu and Gul 2000; Acharya and Grillo 2015).[8]

Assumption 2C: $\underline{x} = \underline{\rho}_W + \kappa_D$.

We also assume that weak-type D's ($\bar{\rho}_W$) prefer attaining a final payoff of $1 - \bar{x}$ (i.e. making a high-offer to C) rather than setting $x = \underline{x}$ and experiencing gray zone conflict or war. This produces:

Assumption 3C: $1 - \bar{x} > 1 - \bar{\rho}_W - \kappa_D$, and if $\frac{\theta}{2\beta_C} < \bar{\rho}_W - \bar{x} + \kappa_D + \frac{1}{4\beta_D}$ then $1 - \bar{x} > 1 - \underline{x} - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D}$.

We also assume that if C receives a high offer ($\bar{x}$) from a weak-type D, C will accept the offer rather than escalate to war or implement gray zone conflict. This produces:

Assumption 4C: $\theta \leq \frac{\kappa_C}{\bar{\rho}_W - \bar{x}}$ and if $\frac{\theta}{2\beta_C} \geq \bar{\rho}_W - \bar{x} + \kappa_D + \frac{1}{4\beta_D}$ then $\theta \leq \frac{\beta_C\left(\bar{\rho}_W - \bar{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\bar{\rho}_W - \bar{x} + \kappa_D - \frac{1}{4\beta_D}\right)}$, or if $\frac{\theta}{2\beta_C} < \bar{\rho}_W - \bar{x} + \kappa_D + \frac{1}{4\beta_D}$ then $0 \geq \theta\left(\frac{\theta}{4\beta_C} - \frac{1}{2\beta_D}\right)$.

And finally, we assume that if C receives a low offer ($\underline{x}$) from a strong type D, then C prefers accepting the offer rather than going to war.

Assumption 5C: $\underline{x} > \underline{\rho}_W - \kappa_C$.

With these assumptions in place, the complete information game would play out as follows. Strong-type D's (when nature sets $\underline{\rho}_W$) would always make low offer $\underline{x}$ to C, and C would always accept (based on Assumptions 2C and 5C). Weak-type D's (when nature sets $\bar{\rho}_W$) choose one of two offers. First, suppose that in response to a low offer of $\underline{x}$ by a type $\bar{\rho}_W$, C's best response is to accept. Formally, this would imply that, when facing a weak-type D, C prefers accepting to war ($\underline{x} > \theta\bar{\rho}_W - \kappa_C$) and accepting to gray zone conflict (if $\frac{\theta}{2\beta_C} \geq \bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}$ then $\theta \leq \frac{\beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\bar{\rho}_W - \underline{x} + \kappa_D - \frac{1}{4\beta_D}\right)}$, and if $\frac{\theta}{2\beta_C} \geq \bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}$ then $0 \geq \theta\left(\frac{\theta}{4\beta_C} - \frac{1}{2\beta_D}\right)$). If this is the case, then weak-type D's will pool with strong-type D's, also set $x = \underline{x}$, and C will always accept the low offer. Second, suppose that in in response to a low-offer of $\underline{x}$ by a type $\bar{\rho}_W$, C's best response is to engage in a limited challenge or war. If this is the case, then weak-type D will always make a high offer to C rather than face any kind of conflict (based on Assumptions 3C and 4C). Thus, in the complete information version of the game, under the above assumptions, there will be no war or gray zone conflict.

Next, consider what happens when information asymmetry is introduced. Strong-type D's ($\underline{\rho}_W$) will always make the low offer $\underline{x}$ (based on Assumptions 2C and 5C). A fully separating equilibrium is not possible,[9] and type $\bar{\rho}_W$ D's will pool on $\underline{x}$ or semi-pool by mixing between the $\underline{x}$ and $\bar{x}$ offers. In response to a high offer of $\bar{x}$, C will accept ($g_C = 0$) by virtue of Assumption 4C. In response to a low offer of $\underline{x}$, C may mix, sometimes accepting ($g_C = 0$) and sometimes selecting some optimal response similar to what is outlined in Proposition 1 in the paper (with the new full equilibrium outlined below in Proposition C). Essentially, when a weak-type D makes the low offer ($\underline{x}$) to C, C's behavior is very close to the "dissatisfied state" discussed in the paper, who has been presented with a political status quo (the offer) that could tempt them to go to war.

We derive the equilibria here. For a semi-separating equilibrium to exist, weak-type D's must mix between making a high offer ($\bar{x}$) to C that will result in peace, and a low offer $\underline{x}$ to C, knowing that C will, with

[8] The simplest way to implement this would be to assume that the strong-type D is a behavioral type that will go to war if facing any challenge.

[9] Suppose a separating equilibrium exists where strong type D's select $x = \underline{x}$ and weak type D's select $x = \bar{x}$. C does not want to go to war with type $\underline{\rho}_W$ D's and would therefore never challenge or go to war when facing low offer $x = \underline{x}$. However, this would incentivize weak-type D's to deviate to $x = \underline{x}$.

For Peer Review

probability $\alpha \in (0,1)$, accept the low offer. What C does otherwise (with probability $1-\alpha$) is solved for below. For now, call what C does otherwise (with probability $1-\alpha$) C's "conflict response." Assume for now that a type $\bar{\rho}_W$ facing the conflict response after making the low offer $\underline{x}$ does worse than the type $\bar{\rho}_W$ D would have done by making the high offer $\bar{x}$.[10] Formally, letting $U_D(\bar{\rho}_W, CR)$ denote type $\bar{\rho}_W$ D's utility from C's conflict response (CR), this implies $U_D(\bar{\rho}_W, CR) < 1-\bar{x}$.

We solve for α below. For weak type $\bar{\rho}_W$ D's to be indifferent between making the high and low offer, the following condition must hold:

$$1-\bar{x} = \alpha(1-\underline{x}) + (1-\alpha) * U_D(\bar{\rho}_W, CR),$$

or, in terms of α,

$$\alpha = \frac{1-\bar{x}-U_D(\bar{\rho}_W, CR)}{1-\underline{x}-U_D(\bar{\rho}_W, CR)}.$$

We will reference this term again in the statement of the equilibria. Note that because $\bar{x} > \underline{x}$ and Assumption 3C, α always falls within 0 and 1 (non-inclusive).

Next, we assume that with probability $\gamma \in (0,1)$ that type $\bar{\rho}_W$ D's will make low offer $\underline{x}$. We use Bayes' rule to calculate probabilities of type conditional on offer. These are

$$Pr(\bar{\rho}_W|\underline{x}) = \frac{Pr(\underline{x}|\bar{\rho}_W) * Pr(\bar{\rho}_W)}{Pr(\underline{x})}$$

or

$$Pr(\bar{\rho}_W|\underline{x}) = \frac{(1-\epsilon)\gamma}{\epsilon + (1-\epsilon)\gamma},$$

and

$$Pr(\underline{\rho}_W|\underline{x}) = \frac{\epsilon}{\epsilon + (1-\epsilon)\gamma}.$$

This next derivation also depends on C's conflict response. We assume C attains utility $U_C(\bar{\rho}_W, CR)$ when selecting their optimal conflict response and facing a type $\bar{\rho}_W$ D. C is indifferent between their conflict response and accepting low offer $\underline{x}$ when

$$\underline{x} = \frac{(1-\epsilon)\gamma}{\epsilon + (1-\epsilon)\gamma} U_C(\bar{\rho}_W, CR) + \frac{\epsilon}{\epsilon + (1-\epsilon)\gamma}\left(\underline{\rho}_W - \kappa_C\right)$$

or

$$\gamma = \frac{\epsilon\left(\underline{x} - \left(\underline{\rho}_W - \kappa_C\right)\right)}{(1-\epsilon)\left(U_C(\bar{\rho}_W, CR) - \underline{x}\right)}.$$

The set of derivations above are all for semi-separating equilibria. In Proposition C below, Cases 1B, 1D, 2B, and 2D are all semi-separating equilibria following the structure described above.

Pooling equilibria could also exist. For example, suppose that C's optimal conflict response to a type $\bar{\rho}_W$ D offering $\underline{x}$ is to accept the bargained offer ($w_C = 0$ and $g_C = 0$). If this is the case, then type $\bar{\rho}_W$ D's does best by always fixing $x = \underline{x}$ because C cannot credibly commit to any form of conflict. In Proposition C below, Cases 1A and 2A take this form. These equilibria can be thought of as some combination of D's external deterrent threat and C's internal efficiency constraints binding. Essentially C is so ineffective at fighting war and gray zone conflict that C does best walking away with a low-offer rather than fighting.

[10] This is partly justified by Assumption 3C; however, sometimes C's optimal conflict response is accepting the status quo.

For Peer Review

Another type of pooling equilibrium can also exist. Sometimes C's war outcome against strong-type D's that C would never risk challenging a low-offer ($\underline{x}$) even if all type $\bar{\rho}_W$ D's (with probability $\gamma = 1$) are making the low-offer. Cases 1C, 1E, 2C, and 2E take this form. In these cases, C's optimal conflict response to a type $\bar{\rho}_W$ D offering $\underline{x}$ is some gray zone challenge or war, but C's expected utility from this optimal conflict response is too low for C to consider challenging due to the presence of strong-type D's. These equilibria are certainly shaped by D's external deterrent threat (as the deterrent threat from war with type $\underline{\rho}_W$ D's is central), and they may also be shaped by C's internal efficiency constraints; if C is ineffective at gray zone conflict, then challenging type $\bar{\rho}_W$ D's is less appealing.

The behavior that is described in Proposition 1 (in the text) is related to cases 1B, 1D, 2B, and 2D in Proposition C whenever a weak-type D makes a low-offer and C selects their optimal conflict response. Similarly, in cases 1A and 2A, the equilibria behavior matches what is in Proposition 1. In all of these cases, C is tailoring their gray zone challenge (or no challenge) based on weak-type D's deterrent threat or their own internal efficiency constraint.

Admittedly, there are several differences. For one, even within cases 1B, 1D, 2B, and 2D in Proposition C, the presences of the strong-type D's and the mixed strategies makes these equilibria play out differently. Additionally, the conditions for selection into the various equilibria have been refined further. What the proposition below suggests is that there is further nuance to D's deterrent threat and C's internal efficiency constraint when considering environments with information asymmetry. We can now write out the full equilibria conditions.

However, a key take-away from the equilibrium below is that, conditional on C selecting into gray zone conflict, C will make this selection based on its own internal efficiency constraint or on D's external deterrent constraint, as demonstrated in 1D and 2D.

Proposition C: *In equilibrium, the game will play out in the following manner.*

Case 1, $\frac{\theta}{2\beta_C} \geq \bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}$*:*

- *1.A. If* $\theta \leq \frac{\beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\bar{\rho}_W - \underline{x} + \kappa_D - \frac{1}{4\beta_D}\right)}$ *and* $\theta \leq \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, then both types of D always offer* $x^* = \underline{x}$ *and C always accepts the offer. C selects* $w_R^* = 0$ *and* $g_C^* = 0$*, and both types of D select* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*.*

- *1.B. If* $\theta > \frac{\kappa_C - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$*,* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, and* $\theta\underline{x} < \epsilon\left(\theta\bar{\rho}_W - \kappa_C\right) + (1-\epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$ *then type* $\underline{\rho}_W$ *D always offers* $x^* = \underline{x}$ *and type* $\bar{\rho}_W$ *D offers* $x^* = \underline{x}$ *with probability* γ *and offers* $x^* = \bar{x}$ *with probability* $1 - \gamma$*. When the offer is* $x^* = \bar{x}$*, C always accepts the offer by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, the type* $\bar{\rho}_W$ *D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \bar{x}$ *and* $U_C = \theta\bar{x}$*. When the offer is* $x^* = \underline{x}$*, C accepts the offer with probability* α *by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, in response the both type D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*. When the offer is* $x^* = \underline{x}$*, with probability* $1 - \alpha$ *C declares war; C selects* $w_R^* = 1$*, and payoffs are* $U_D = 1 - \bar{\rho}_W - \kappa_D$ *and* $U_C = \theta\bar{\rho}_W - \kappa_C$ *when D is type* $\bar{\rho}_W$*, and* $U_D = 1 - \underline{\rho}_W - \kappa_D$ *and* $U_C = \theta\underline{\rho}_W - \kappa_C$ *when D is type* $\underline{\rho}_W$*. The values* γ *and* α *are derived using values* $U_D(\bar{\rho}_W, CR) = 1 - \bar{\rho}_W - \kappa_D$ *and* $U_C(\bar{\rho}_W, CR) = \theta\bar{\rho}_W - \kappa_C$*.*

- 1.C. *If* $\theta > \frac{\kappa_C - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$*,* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, and* $\theta\underline{x} \geq \epsilon\left(\theta\bar{\rho}_W - \kappa_C\right) + (1-\epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then both types of D always offer* $x^* = \underline{x}$ *and C always accepts the offer. C selects* $w_R^* = 0$ *and* $g_C^* = 0$*, and both types of D select* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*.*

- *1.D. If* $\theta > \frac{\beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\bar{\rho}_W - \underline{x} + \kappa_D - \frac{1}{4\beta_D}\right)}$*,* $\theta \leq \frac{\kappa_C - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$*, and*

  $\theta\underline{x} < \epsilon\left(\theta\left(\bar{\rho}_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2\right) + (1-\epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then type* $\underline{\rho}_W$ *D always offers* $x^* = \underline{x}$ *and type* $\bar{\rho}_W$ *D offers* $x^* = \underline{x}$ *with probability* γ *and offers* $x^* = \bar{x}$ *with probability* $1 - \gamma$*. When the offer is* $x^* = \bar{x}$*, C always accepts the offer by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, the type* $\bar{\rho}_W$ *D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \bar{x}$ *and* $U_C = \theta\bar{x}$*. When the offer*

For Peer Review

is $x^* = \underline{x}$*, C accepts the offer with probability* α *by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, in response the both type D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*. When the offer is* $x^* = \underline{x}$*, with probability* $1 - \alpha$*, C conducts a limited challenge that is constrained by type* $\bar{\rho}_W$ *D's deterrence threat. C selects* $w_R^* = 0$ *and* $g_C^* = \bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}$*, and in response type* $\underline{\rho}_W$ *D's declare war setting* $w_D^* = 1$*and type* $\bar{\rho}_W$ *D's select* $w_D^* = 0$ *and* $g_D^* = \frac{1}{2\beta_D}$*. The payoffs are* $U_C = \theta\left(\bar{\rho}_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2$ *and* $U_D = 1 - \bar{\rho}_W - \kappa_D$ *when D is type* $\bar{\rho}_W$*, and* $U_C = \theta\underline{\rho}_W - \kappa_C$ *and* $U_D = 1 - \underline{\rho}_W - \kappa_D$ *when D is type* $\underline{\rho}_W$*. The values* γ *and* α *are derived using values* $U_D(\bar{\rho}_W, CR) = 1 - \bar{\rho}_W - \kappa_D$ *and* $U_C(\bar{\rho}_W, CR) = \theta\left(\bar{\rho}_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2$*.*

- *1.E. If* $\theta > \frac{\beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\left(\bar{\rho}_W - \underline{x} + \kappa_D - \frac{1}{4\beta_D}\right)}$*,* $\theta \leq \frac{\kappa_C - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2}{\frac{1}{4\beta_D} - \kappa_D}$*, and*

  $\theta\underline{x} \geq \epsilon\left(\theta\left(\bar{\rho}_W + \kappa_D - \frac{1}{4\beta_D}\right) - \beta_C\left(\bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}\right)^2\right) + (1 - \epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then both types of D always offer* $x^* = \underline{x}$ *and C always accepts the offer. C selects* $w_R^* = 0$ *and* $g_C^* = 0$*, and both types of D select* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*.*

Case 2, $\frac{\theta}{2\beta_C} < \bar{\rho}_W - \underline{x} + \kappa_D + \frac{1}{4\beta_D}$*:*

- *2.A. If* $\theta \leq \frac{2\beta_C}{\beta_D}$ *and* $\theta \leq \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, then both types of D always offer* $x^* = \underline{x}$ *and C always accepts the status quo. C selects* $w_R^* = 0$ *and* $g_C^* = 0$*, and both types of D select* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*.*

- *2.B. If* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x} - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$*,* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, and* $\theta\underline{x} < \epsilon\left(\theta\bar{\rho}_W - \kappa_C\right) + (1 - \epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then type* $\underline{\rho}_W$ *D always offers* $x^* = \underline{x}$ *and type* $\bar{\rho}_W$ *D offers* $x^* = \underline{x}$ *with probability* γ *and offer* $x^* = \bar{x}$ *with probability* $1 - \gamma$*. When the offer is* $x^* = \bar{x}$*, C always accepts the offer by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, the type* $\bar{\rho}_W$ *D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \bar{x}$ *and* $U_C = \theta\bar{x}$*. When the offer is* $x^* = \underline{x}$*, with probability* α *C accepts the offer by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, in response the both type D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*. When the offer is* $x^* = \underline{x}$*, with probability* $1 - \alpha$ *C declares war setting* $w_R^* = 1$*, and payoffs are* $U_D = 1 - \bar{\rho}_W - \kappa_D$ *and* $U_C = \theta\bar{\rho}_W - \kappa_C$ *when D is type* $\bar{\rho}_W$*, and* $U_D = 1 - \underline{\rho}_W - \kappa_D$ *and* $U_C = \theta\underline{\rho}_W - \kappa_C$ *when D is type* $\underline{\rho}_W$*. The values* γ *and* α *are derived using values* $U_D(\bar{\rho}_W, CR) = 1 - \bar{\rho}_W - \kappa_D$ *and* $U_C(\bar{\rho}_W, CR) = \theta\bar{\rho}_W - \kappa_C$*.*

- *2.C. If* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x} - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$*,* $\theta > \frac{\kappa_C}{\bar{\rho}_W - \underline{x}}$*, and* $\theta\underline{x} \geq \epsilon\left(\theta\bar{\rho}_W - \kappa_C\right) + (1 - \epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then both types of D always offer* $x^* = \underline{x}$ *and C always accepts the status quo. C selects* $w_R^* = 0$ *and* $g_C^* = 0$*, and both types of D select* $w_D^* = 0$ *and* $g_D^* = 0$*. Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*.*

- *2.D. If* $\theta > \frac{2\beta_C}{\beta_D}$*,* $\theta \leq \frac{\kappa_C}{\bar{\rho}_W - \underline{x} - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$*,* $1 - \underline{x} - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D} < 1 - \bar{x}$*, and* $\theta\underline{x} < \epsilon\left(\theta\underline{x} + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}\right) + (1 - \epsilon)\left(\theta\underline{\rho}_W - \kappa_C\right)$*, then type* $\underline{\rho}_W$ *D always offers* $x = \underline{x}$ *and type* $\bar{\rho}_W$ *D offers* $x = \underline{x}$ *with probability* γ *and offer* $x = \bar{x}$ *with probability* $1 - \gamma$*. When the offer is* $x = \bar{x}$*, C always accepts the offer by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, the type* $\bar{\rho}_W$ *D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \bar{x}$ *and* $U_C = \theta\bar{x}$*. When the offer is* $x = \underline{x}$*, C accepts the offer with probability* α *by setting* $w_R^* = 0$ *and* $g_C^* = 0$*, in response the both type D's set* $w_D^* = 0$ *and* $g_D^* = 0$*, and payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta\underline{x}$*. When the offer is* $x = \underline{x}$*, with probability* $1 - \alpha$ *C conducts a limited challenge that is constrained by C's own internal cost constraints. C selects* $w_R^* = 0$ *and* $g_C^* = \frac{\theta}{2\beta_C}$*, and in response type* $\underline{\rho}_W$ *D's declare war* $(w_D = 1)$ *and type* $\bar{\rho}_W$ *D's select* $w_D^* = 0$ *and* $g_D^* = \frac{1}{2\beta_D}$*. The payoffs are* $U_C = \theta\underline{x} + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}$ *and* $U_D = 1 - \rho_0 - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D}$ *when D is type* $\bar{\rho}_W$*, and* $U_C = \theta\underline{\rho}_W - \kappa_C$ *and* $U_D = 1 - \underline{\rho}_W - \kappa_D$ *when D is type* $\underline{\rho}_W$*. The values* γ *and* α *are derived using values* $U_D(\bar{\rho}_W, CR) = 1 - \underline{x} - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D}$*, and* $U_C(\bar{\rho}_W, CR) = \theta\underline{x} + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}$*.*

- *2.E. If* $\theta > \frac{2\beta_C}{\beta_D}$*,* $\theta \leq \frac{\kappa_C}{\bar{\rho}_W - \underline{x} - \frac{\theta}{4\beta_C} + \frac{1}{2\beta_D}}$*,* $1 - \underline{x} - \frac{\theta}{2\beta_C} + \frac{1}{4\beta_D} < 1 - \bar{x}$*, and* $\theta\underline{x} \geq \epsilon\left(\theta\underline{x} + \frac{\theta^2}{4\beta_C} - \frac{\theta}{2\beta_D}\right) + (1 -$

For Peer Review

$\epsilon) \left(\theta \underline{\rho}_W - \kappa_C\right)$, *then both types of* D *always offer* $x = \underline{x}$ *and* C *always accepts the status quo.* C *selects* $w_R^* = 0$ *and* $g_C^* = 0$, *and both types of* D *select* $w_D^* = 0$ *and* $g_D^* = 0$. *Payoffs are* $U_D = 1 - \underline{x}$ *and* $U_C = \theta \underline{x}$.

2 New data

The universe of cases was created by first identifying cases of Russian foreign interventions from 3 prior datasets; ICB (Brecher and Wilkenfeld 1997), DCID (Valeriano and Maness 2014), and REI (Casey and Way 2017). Code replicating those findings is provided in the appropriate RMarkdown files. These cases were then supplemented with additional cases of Russian interference the authors were able to identify.

2.1 Comparison of current datasets

A comparison of what cases were covered in each individual dataset is provided in Figure A2. Note that there are significant inconsistencies concerning the sample of post-1994 Russian interventions identified by the ICB, DCID, and REI datasets.

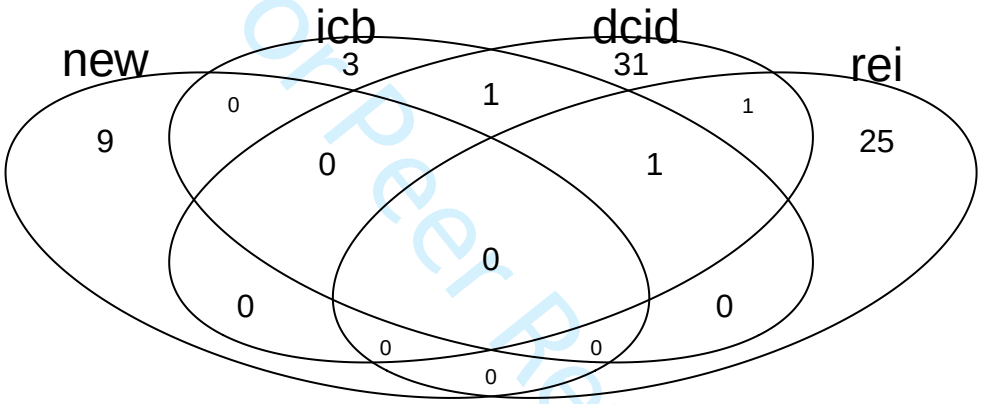

Figure A2: Venn diagram of case overlap among prior datasets

Aside from the cases covered, the intensity codings for current datasets are difficult to compare given their different scales. A more thorough analysis is provided in the appropriate R Markdown files, but a comparison of intensity codings in DCID (Valeriano and Maness 2014) and REI (Casey and Way 2017) is depicted in Figure A3. The DCID data identifies the United States, United Kingdom, Poland and Ukraine as targets of the most severe Russian cyber operations. In the cases documented by REI, the most severe Russian attacks occurred against France, Austria, and Ukraine. Part of this discrepancy is due to the respective foci of each dataset; DCID seeks out cases of cyber incidents and disputes while REI focuses on Russian electoral interference. While a majority of the REI cases include some form of Russian cyber activity, there are a few cases where only material support was provided (eg. Moldova 2014 and Belarus 1994).

This discrepancy exemplifies not only the challenges of relying on open source reporting for identifying cyber influence or disruption campaigns, but also differences in defining what counts as an attack. The only country-year that appears in both datasets is Ukraine 2014. We standardized codings across the two datasets using variable definitions from respective codebooks. A severity less than or equal to 2 in DCID's coding is synonymous in our recoding with REI's coding for disinformation, a severity between 3 and 7 equals REI's coding for cyberattack, and no cases in DCID have a severity greater than 7. We adopted Valeriano and Maness (2014)'s approach of sampling on intensity when there are multiple observations in a given time unit.

2.2 Variable codings

For each incident, we code whether Russia used conventional ground forces, conventional air or sea forces, paramilitary or covert forces, cyber disruption, and information operations. By distinguishing between these

For Peer Review

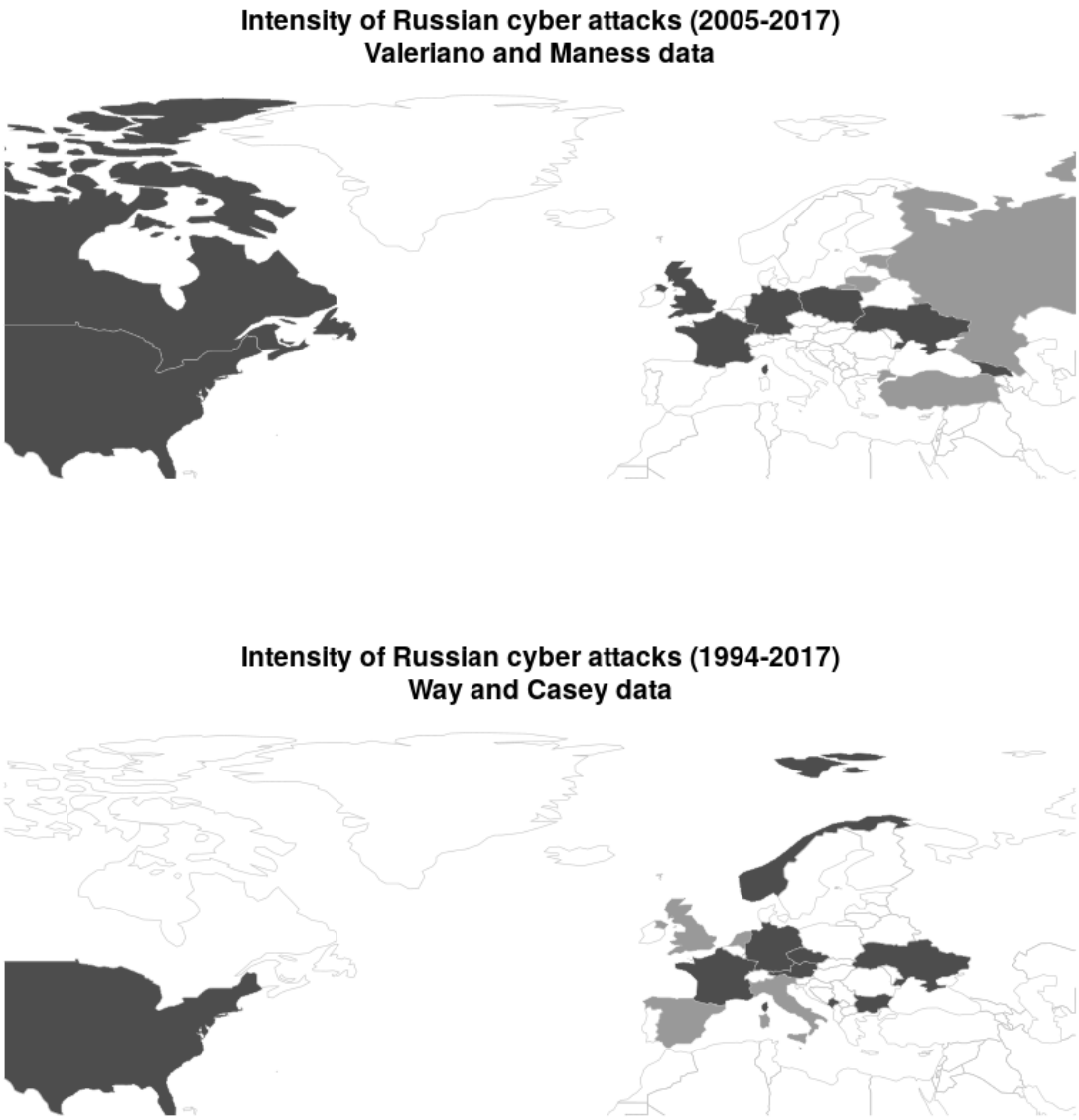

Figure A3: Comparison of coding for highest intensity Russian intervention in each target state

five types of aggression, we obtain a clearer picture of the intensity of each case of Russian intervention. The vast majority of cases include at least some type of cyber operations. In a few cases, data limitations preclude coding of non-kinetic activity by Russia or other actors. In Moldova 2005, for example, Russia provided material support for the Communist Party but there is no credible evidence of cyber activities.

The following binary coding criteria were used for each case:

- **resp_infoops** - Did Russia use information operations during this event? That includes propaganda, misinformation campaigns, theft of information, and other simple intrusions
- **resp_cyberdisrup** - Did Russia use cyber attacks during this operation? That includes hacking, phishing, cyber espionage, DDoS attacks, etc. that constitute a system shut down rather than simple intrusions
- **resp_paramil** - Did Russia use paramilitary troops during this event? Special forces, covert troops, speznatz, etc all count
- **resp_convmil_airsea** - Did Russia use conventional naval or air forces during this event?
- **resp_convmil_gro** - Did Russia use conventional ground troops like their army, artillery, tanks, etc during this event?

The complete dataset is provided in the appropriate .csv file. It includes sources used for the codings as well as justifications and explanations where needed.

2.3 Summary statistics

Although data was compiled on Russian intervention against all states from 1994-2018, the statistical analysis is limited to a sample from European states. In alignment with that, Table A2 present descriptive statistics of the sample used in the models provided in the main text.

Table A2: Covariate Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Intensity	1,000	0.1	0.4	0	0	0	5
NATO member	1,000	0.5	0.5	0	0	1	1
Dist. from Russia (minimum, log)	1,000	5.2	2.9	0.01	5.2	7.0	7.8
Democracy	926	0.9	0.3	0.0	1.0	1.0	1.0
Nuclear state	1,000	0.05	0.2	0	0	0	1
Population (log)	1,000	15.8	1.4	12.5	14.9	16.3	18.2
CINC ratio	754	0.1	0.1	0.0	0.01	0.1	0.4
GDP per capita	995	26.6	23.8	0.7	6.7	41.4	112.0
Military expenditure	962	7.3	13.3	0.0	0.3	5.7	59.8

Sample includes all European states (1994-2018). Binary variables converted to numeric.

The distribution of our dependent variable, intensity, is shown for the European sample in Figure A4. The figure only includes country-years with known attacks (omitting null cases) to allow an easier visual comparison of variation in attack intensity.

The bivariate correlations between the DV and the two EVs are shown in Figure A5. The intensity and NATO variables have been converted to numeric values to simplify visualizing the bivariate correlations.

3 Alternate model specifications

We run a set of alternate model specifications as robustness checks. Our results are consistent across alternate modeling specifications including different regression models, control variables, and imputation strategies. We choose the ordered probit results as the main results given the appropriateness of that model specification

For Peer Review

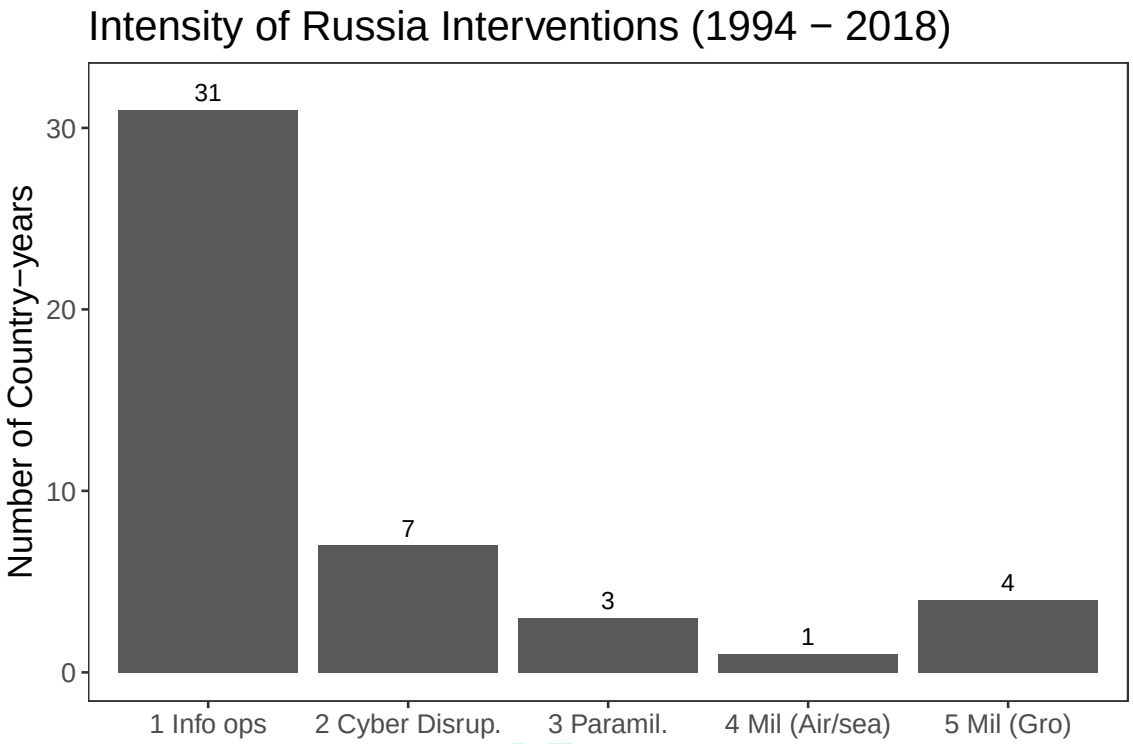

Figure A4: Intensity of Russian interventions

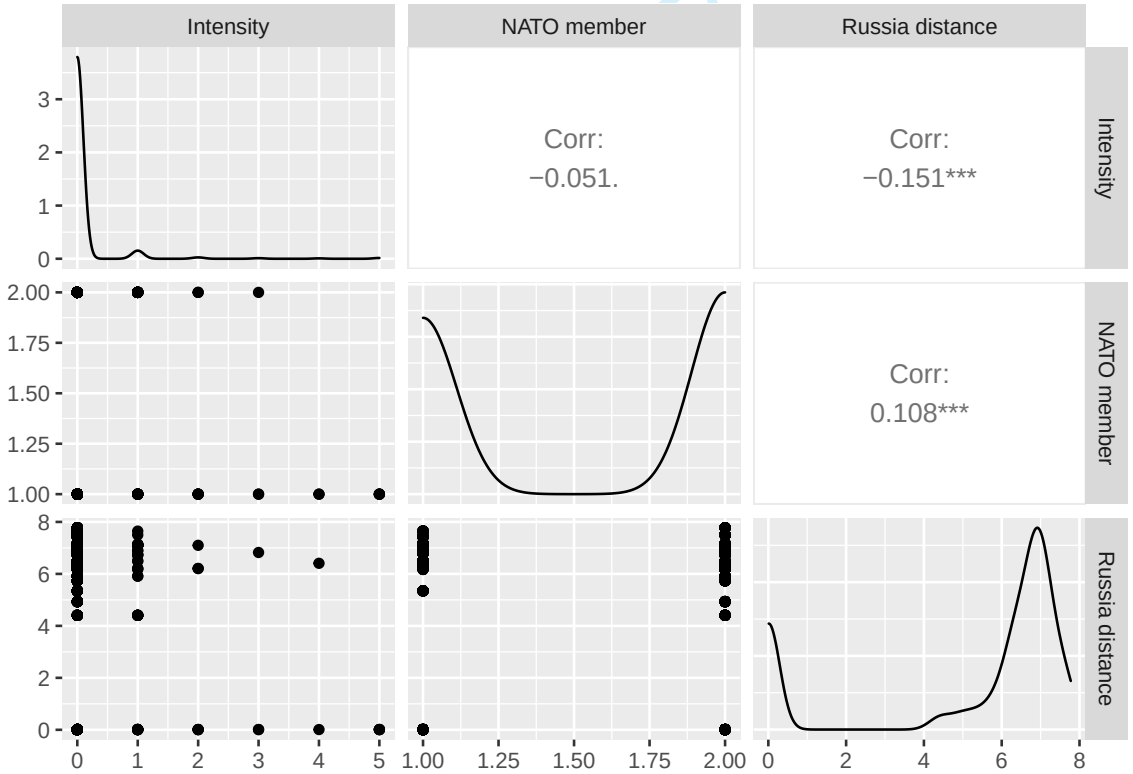

Figure A5: Bivariate correlation of dependent and independent variables

and to ensure our primary results are not simply an artifact of our imputation strategy. Those results are shown below.

3.1 Alternate alliance measure

Because the dichotomous NATO/Non-NATO independent variable could be viewed as too coarse, we also re-run our six empirical models with an additional dummy variable for if a state was in a NATO Membership Action Plan (MAP), part of the NATO Partnership for Peace (PfP), or if the state was engaged in Intensified Dialogue (ID). Figure A6 illustrates how states' NATO status and pre-status change over time. In line with our hypothesis that NATO membership could serve as a deterrent to Russian aggression, we would expect that entering into a partnership with NATO could also deter Russia actions. In order to be invited into NATO, states must be in good standing with existing members, suggesting that some NATO states might come to the aid of a state with (for example) a MAP.

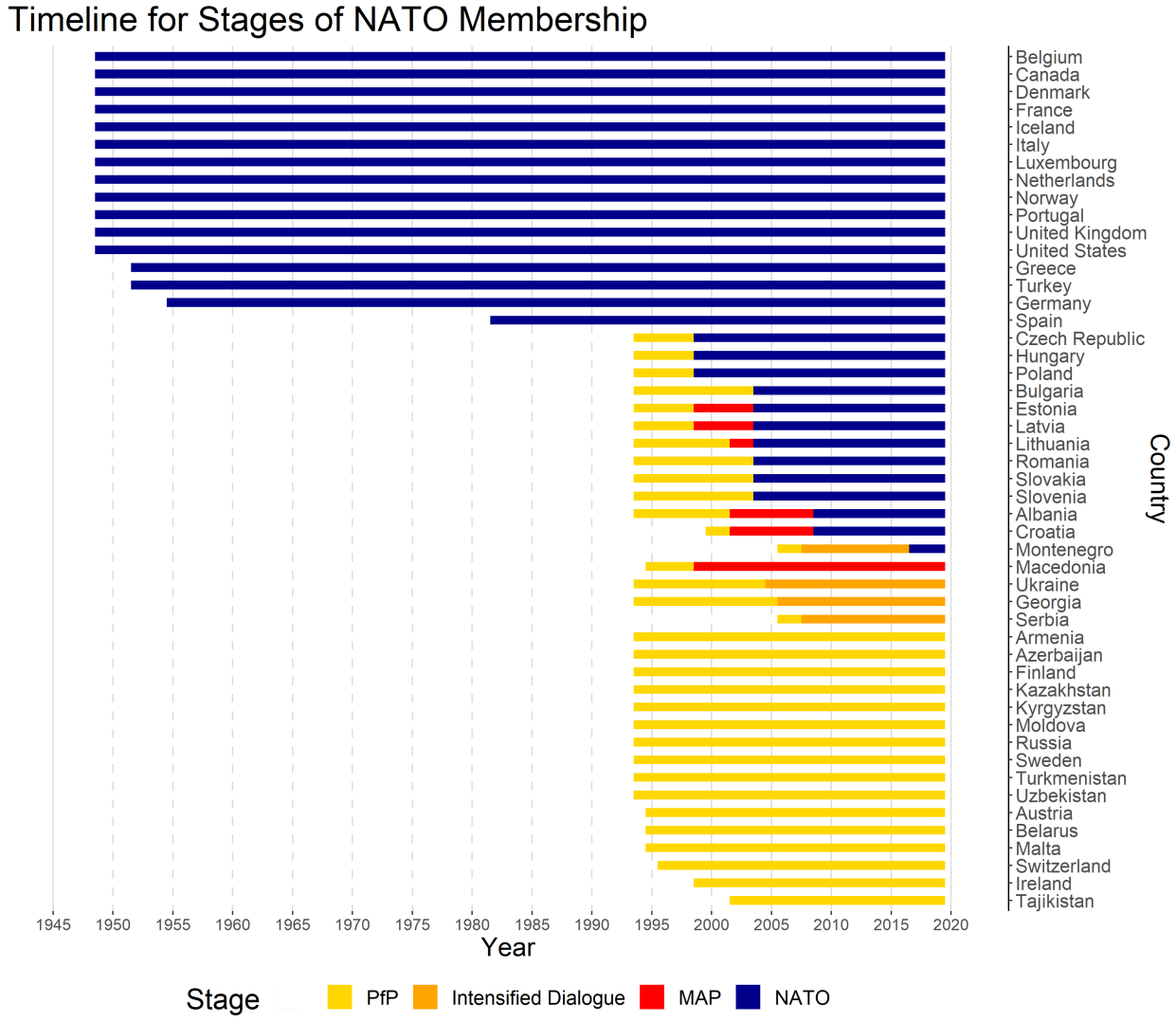

Figure A6: NATO membership timeline, including pre-NATO membership stages

As can be seen in Table A3, introducing the pre-NATO dummy variable makes little substantive changes to our existing results. We find that NATO membership is still associated with a decrease in the intensity of Russian attacks. We also find, consistent with the discussion above, taking the steps to become a NATO

member is also associated with a decrease in the intensity of Russian attacks. Once more, while we cannot claim our results are causal—just as NATO membership is not exogenous, NATO pre-membership is not exogenous—our results suggest that a state's warming relationship to NATO could decrease gray zone activity taken against it.

	Full sample			Relevant states sample		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Independent Variables						
NATO member	−0.37	−0.56***	−0.74***	−0.56**	−0.66**	−0.79***
	(0.23)	(0.21)	(0.23)	(0.26)	(0.26)	(0.25)
NATO pre-member	−0.85**	−0.91**	−1.04***	−4.74***	−4.46***	−4.60***
	(0.42)	(0.38)	(0.39)	(0.00)	(0.00)	(0.00)
Russia distance	−0.10***	−0.10***	−0.11***	−0.04	−0.08**	−0.08**
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)
Controls						
Democracy		0.25	0.53		0.22	0.51
		(0.43)	(0.42)		(0.43)	(0.43)
Nuclear power		0.97**	0.33		0.93*	0.89
		(0.42)	(0.43)		(0.48)	(0.73)
Population		0.15	0.08		0.14	0.15
		(0.09)	(0.13)		(0.09)	(0.11)
GDP per cap		−0.01**	−0.02**		−0.01	−0.01
		(0.01)	(0.01)		(0.01)	(0.01)
Mil. spending			0.02*			0.00
			(0.01)			(0.02)
Observations	1,000	921	891	376	373	346

All models include year-fixed effects with country-clustered standard errors in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table A3: Pre-NATO Variable Robustness Check

It is also worthwhile mentioning: in some cases, the coefficient on NATO pre-membership has a point estimate that is larger in magnitude than the coefficient on NATO membership. Taken as true, this would suggest that NATO pre-membership could be a more effective deterrent than NATO membership. While for the first three models this difference is not statistically significant at the 10% level, in the last three models this difference is statistically significant at the 1% level. Ultimately, it is difficult to know how to interpret this finding. While it is plausible that this relationship is actually borne out empirically—that NATO leadership may take gray zone attacks against its pre-member states more seriously than gray zone attacks against some of its existing states—this result could also be driven by biases in the data. It could be that, in order for a state to be included in our "relevant" sample and part of Intensified Dialogue, that state has been screened by NATO to be particularly unlikely to be attacked by Russian gray zone efforts; this would bias our results, and could create spurious results. While we cannot fully ascertain what is driving this quirk in our results, this robustness check has ultimately presented further positive (albeit suggestive) evidence that is in-line with our hypothesis: gray zone behavior is being shaped by NATO's deterrent threat. This does demonstrate that debates about both the causes and consequences of post-Cold War NATO expansion warrant important attention (Shifrinson 2016; Lanoszka 2020).

3.2 Odds ratios

Given the difficulty of interpreting ordered probit coefficients, Table A4 show the results as odds ratios with confidence intervals in parentheses when all other variables are held at their mean level. To use model 6 as an example for interpretation, for relevant NATO states, the odds of a non-cyber, non-information attack (categories 3, 4, or 5) versus a cyber attack, an information attack, or no attack is 49% lower.

For Peer Review

	Full sample			Relevant states sample		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Independent Variables						
NATO member	0.76	0.63*	0.55*	0.63	0.56*	0.51*
	[0.39; 1.12]	[0.29; 0.96]	[0.19; 0.91]	[0.21; 1.05]	[0.14; 0.98]	[0.09; 0.93]
Russia distance	0.90*	0.90*	0.88*	0.95	0.92*	0.91*
	[0.84; 0.96]	[0.85; 0.95]	[0.83; 0.94]	[0.89; 1.01]	[0.86; 0.98]	[0.85; 0.98]
Controls						
Democracy		1.17	1.58		1.13	1.54
		[0.46; 1.88]	[0.89; 2.26]		[0.39; 1.86]	[0.84; 2.25]
Nuclear power		2.53*	1.56		2.50*	2.88*
		[1.85; 3.21]	[0.84; 2.28]		[1.72; 3.29]	[1.53; 4.23]
Population		1.21*	1.15		1.17*	1.20
		[1.06; 1.36]	[0.95; 1.35]		[1.01; 1.34]	[1.00; 1.41]
GDP per cap		0.99*	0.98*		0.99	0.99
		[0.98; 1.00]	[0.97; 1.00]		[0.98; 1.00]	[0.97; 1.00]
Mil. spending			1.02			1.00
			[1.00; 1.03]			[0.97; 1.02]
Observations	1,000	921	891	376	373	346

All models are ordered probits and include year-fixed effects with country-clustered standard errors in parentheses.

Table A4: Odds Ratios

3.3 OLS regression

Although an ordered probit model is most appropriate given the dependent variable (intensity) is ordinal, we ensure that the sign on our coefficients are consistent with an OLS model that treats intensity as a continuous variable. See Table A5.

3.4 Ordered logit

We also run all models as ordered logits instead of ordered probits. Both are generalized linear models appropriate for an ordinal dependent variable that differ only in whether they use a logit link function as opposed to inverse normal link function (Johnston, McDonald, and Quist 2020). The results of the ordered logit in Table A6 are almost identical to those of the ordered probit, as expected.

3.5 Multiple imputation

Models 2, 3, 5, and 6 lose some observations due to missing values for control variables; primarily those not available after 2012. Variables with missing data are shown in Figure A7, with all but CINC being used in the models in the main text. We do not use the CINC ratio variable because listwise delection would lose 25% of our observations in a biased manner given the missingness is for all observations after 2012. Instead, the main model uses population and SIPRI military expenditure variables which adequately proxy for CINC given they are 2 of CINC's 6 components. When imputing missing values, we replace the population and military expenditure variables with CINC since it is more commonly used as an observable indicator for military power, the concept of interest.

Missing values are calculated using bootstrap re-sampling across 10 different imputations using predictive mean matching (Buuren et al. 2006; White, Royston, and Wood 2011). The imputation predictions account for the temporal nature of the data. The same ordinal probit for each model is run across all 10 imputations and the coefficient estimates and standard errors are pooled across the 10 imputations to account for uncertainty produced by variation across the imputations. Variables not included in the main regression like

For Peer Review

	Full sample			Relevant states sample		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Independent Variables						
NATO member	−0.06***	−0.07**	−0.08***	−0.14***	−0.14***	−0.16***
	(0.02)	(0.03)	(0.03)	(0.05)	(0.05)	(0.05)
Russia distance	−0.02***	−0.02***	−0.02***	−0.02**	−0.02**	−0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Controls						
Democracy		−0.10	−0.06		−0.08	−0.02
		(0.12)	(0.13)		(0.13)	(0.14)
Nuclear power		0.12***	0.09*		0.14**	0.11
		(0.04)	(0.05)		(0.06)	(0.10)
Population		−0.00**	−0.00**		−0.00**	−0.00**
		(0.00)	(0.00)		(0.00)	(0.00)
GDP per cap		0.02	0.01		0.04	0.04
		(0.01)	(0.02)		(0.03)	(0.04)
Mil. spending			0.00			0.00
			(0.00)			(0.00)
Observations	1,000	921	891	376	373	346

All models include year-fixed effects with country-clustered standard errors in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table A5: OLS Results

	Full sample			Relevant states sample		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Independent Variables						
NATO member	−0.43	−0.97**	−1.22**	−0.74	−1.12**	−1.29**
	(0.50)	(0.46)	(0.49)	(0.54)	(0.53)	(0.54)
Russia distance	−0.18**	−0.20***	−0.21***	−0.09	−0.17**	−0.15*
	(0.08)	(0.07)	(0.07)	(0.09)	(0.08)	(0.09)
Controls						
Democracy		0.79	1.27*		0.64	1.13
		(0.82)	(0.76)		(0.82)	(0.74)
Nuclear power		1.70**	0.97		1.67	2.41
		(0.82)	(0.88)		(1.03)	(1.88)
Population		0.48**	0.42		0.42*	0.50
		(0.22)	(0.28)		(0.24)	(0.30)
GDP per cap		−0.02*	−0.03*		−0.02	−0.02
		(0.01)	(0.02)		(0.02)	(0.02)
Mil. spending			0.02			−0.02
			(0.03)			(0.04)
Observations	1,000	921	891	376	373	346

All models include year-fixed effects with country-clustered standard errors in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table A6: Ordered Logit Results

ethno-linguistic fractionalization, and GDP per capita are included to increase the predictive performance of the imputation for variables that the literature suggests are correlated with the variables being imputed.

The results of the original models with imputed control variables are shown in Figure A8 and Table A7. We

For Peer Review

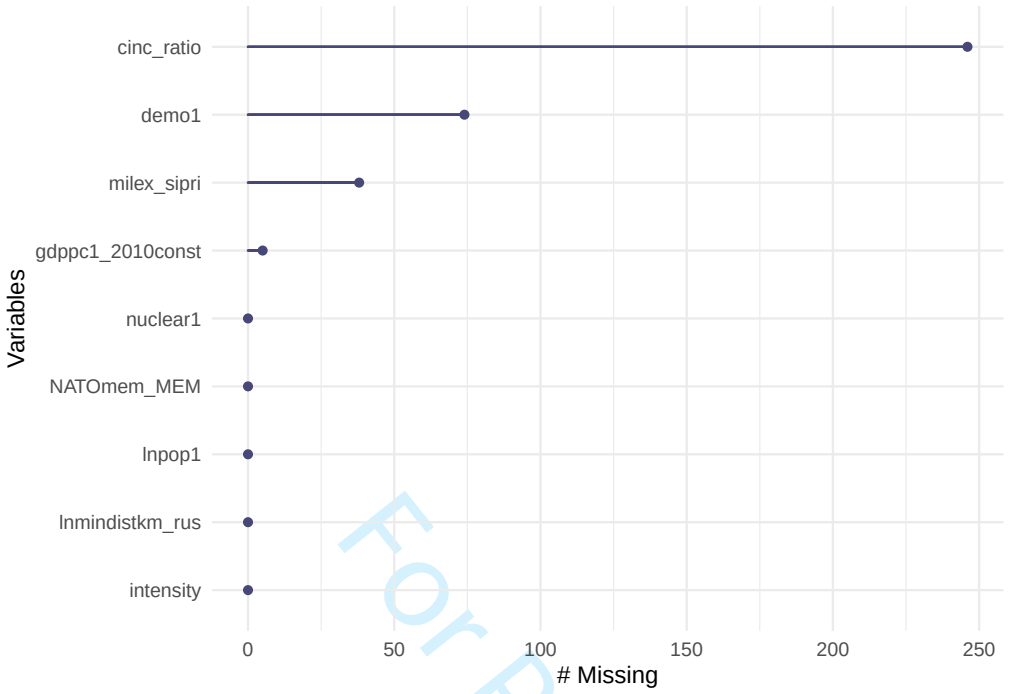

Figure A7: Number of missing observations for each variable in the dataset

show the results for models 1 and 4 separately to make clear that these models have no imputed values. The coefficients in A7 are the same as those reported in models 1 and 4 in the main text and are only reported here to enable comparison with the imputed models in Figure A8.

	Full sample	Relevant states sample
	Model 1	Model 4
NATO member	0.76	0.63
	[0.39; 1.12]	[0.21; 1.05]
Russia distance	0.90*	0.95
	[0.84; 0.96]	[0.89; 1.01]
Observations	1,000	376

All models are ordered probits and include year-fixed effects with country-clustered standard errors in parentheses.

Table A7: Odds Ratios (non-imputed models)

3.6 Targeted states sample

We run the same models on a third sample of just targeted states as empirical support for the endogeneous bargaining and information asymmetry model extension offered in section 1.6 in the appendix. This includes only country-years that were targets of Russian aggression, meaning the intensity variable is greater than 0. Those results are shown in Table A8 and are consistent with the results produced by the other models.

4 Case Study: US 2016

The main text presents case studies of Russian interventions in Estonia, Ukraine, and Georgia, which are all contiguous to Russia and former Soviet republics, and thus more comparable. Yet we expect the logic of the argument to apply more generally. Thus Russian intervention should be even more restrained against targets that are further away and more capable. Intervention in the 2016 U.S. election is consistent with this

For Peer Review

Models with Imputed Control Variables

	Model 2	Model 3	Model 5	Model 6
Predictors	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*
NATO member	0.39 **	0.37 **	0.28 **	0.29 **
	(0.16 – 0.94)	(0.16 – 0.87)	(0.09 – 0.85)	(0.11 – 0.79)
Russia distance	0.81 ***	0.81 ***	0.90 *	0.90 *
	(0.72 – 0.92)	(0.71 – 0.91)	(0.80 – 1.00)	(0.81 – 1.00)
Democracy	2.12	1.93	1.98	1.66
	(0.44 – 10.36)	(0.45 – 8.29)	(0.32 – 12.32)	(0.29 – 9.60)
Nuclear power	5.61 **	3.70 *	1.95	1.45
	(1.15 – 27.34)	(0.79 – 17.29)	(0.43 – 8.80)	(0.24 – 8.72)
GDP per cap	0.98	0.98 *	1.00	1.00
	(0.96 – 1.00)	(0.95 – 1.00)	(0.98 – 1.01)	(0.98 – 1.01)
Population	1.58 **		1.46 **	
	(1.04 – 2.42)		(1.02 – 2.10)	
CINC ratio		732.46 ***		120.07 *
		(12.60 – 42575.63)		(0.79 – 18277.02)
Observations	1000	1000	376	376
R^2	0.251	0.254	0.234	0.232

** p<0.1 ** p<0.05 *** p<0.01*

Figure A8: Intensity of Russian Intervention: Odds Ratios (Imputed models)

For Peer Review

	Model 1	Model 2	Model 3
Independent Variables			
NATO member	-2.16^{***}	-42.39^{***}	-43.40^{***}
	(0.64)	(1.86)	(1.72)
Russia distance	-0.19^{**}	-3.91^{***}	-2.19^{***}
	(0.08)	(0.14)	(0.08)
Controls			
Democracy		-83.35^{***}	-63.27^{***}
		(0.11)	(1.04)
Nuclear power		3.73	39.32^{***}
			(0.00)
Population		-19.79^{***}	-12.60^{***}
		(0.08)	(0.17)
GDP per cap		0.61^{***}	1.01^{***}
		(0.11)	(0.10)
Mil. spending			-1.67^{***}
			(0.36)

All models include year-fixed effects. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table A8: Targeted States Sample

expectation.

A U.S. intelligence assessment released soon after the 2016 election concluded with "high confidence" that "Russian President Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election. Russia's goals were to undermine public faith in the US democratic process, denigrate Secretary Clinton, and harm her electability and potential presidency. We further assess Putin and the Russian Government developed a clear preference for President-elect Trump" (Office of the Director of National Intelligence 2017). Moscow's influence operations might thus be described as unrestrained, even brazen, and thus motivated entirely by efficiency calculations. Yet the choice to pursue this course of action in the first place was very much constrained by the implicit deterrence posture of the United States. Russia could safely assume that the most powerful military in the world would retaliate for armed attacks against U.S. vital interests. While the United States had not designated its electoral process as "critical infrastructure" to explicitly signal that cyber interference was proscribed, Russia still had to consider the potential for American retaliation. Russia thus sought opportunities to impose costs and seek benefits while minimizing the risk of escalation. It found them through covert manipulation of democratic discourse. Indeed, Russia's electoral interference has gone essentially unpunished by the United States to date, aside from the expulsion of some Russian intelligence officers and the application of some additional sanctions to an already heavy regime put in place after Ukraine. Of course, if Trump's victory in 2016 or any of his administration's subsequent policies can ever be credited to active measures by the Russian Federation, even in part, it would amount to one of the most consequential intelligence coups in history. It is just as likely, however, that the Russian campaign simply added noise to one of the most chaotic campaigns in U.S. presidential history (Gelman and Azari 2017). Russian information operations appear to be a low-cost gamble to influence an over-determined outcome.

References

Abreu, Dilip, and Faruk Gul. 2000. "Bargaining and Reputation." *Econometrica* 68 (1): 85–117. https://doi.org/https://doi.org/10.1111/1468-0262.00094.

Acharya, Avidit, and Edoardo Grillo. 2015. "War with Crazy Types." *Political Science Research and Methods* 3 (2): 281–307. https://doi.org/10.1017/psrm.2014.23.

For Peer Review

Brecher, Michael, and Jonathan Wilkenfeld. 1997. *A Study of Crisis*. University of Michigan Press.

Buuren, S. Van, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76 (12): 1049–64. https://doi.org/10.1080/10629360600810434.

Casey, Adam, and Lucan Ahmad Way. 2017. "Russian Electoral Interventions, 1991-2017." Scholars Portal Dataverse. https://doi.org/10.5683/SP/BYRQQS.

Coe, Andrew J. 2011. "Costly Peace: A New Rationalist Explanation for War." Working Paper.

Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414. https://doi.org/10.1017/S0020818300033324.

Gartzke, Erik A. 1999. "War Is in the Error Term." *International Organization* 53 (3): 567–87. https://doi.org/10.1162/002081899550995.

Gelman, Andrew, and Julia Azari. 2017. "19 Things We Learned from the 2016 Election." *Statistics and Public Policy* 4 (1): 1–10. https://doi.org/10.1080/2330443X.2017.1356775.

Jackson, Matthew O., and Massimo Morelli. 2007. "Political Bias and War." *American Economic Review* 97 (4): 1353–73. https://doi.org/10.1257/aer.97.4.1353.

Johnston, Carla, James McDonald, and Kramer Quist. 2020. "A Generalized Ordered Probit Model." *Communications in Statistics - Theory and Methods* 49 (7): 1712–29. https://doi.org/10.1080/03610926.2019.1565780.

Lanoszka, Alexander. 2020. "Thank Goodness for NATO Enlargement." *International Politics* 57: 451–70. https://doi.org/10.1057/s41311-020-00234-8.

Office of the Director of National Intelligence. 2017. "Assessing Russian Activities and Intentions in Recent US Elections." Intelligence Community Assessment ICA 2017-01D. Washington, DC: National Intelligence Council. https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169–203. https://doi.org/10.1017/S0020818306060061.

Shifrinson, Joshua R. Itzkowitz. 2016. "Deal or No Deal? The End of the Cold War and the U.S. Offer to Limit NATO Expansion." *International Security* 40 (4): 7–44. https://doi.org/10.1162/ISEC_a_00236.

Valeriano, Brandon, and Ryan C Maness. 2014. "The Dynamics of Cyber Conflict Between Rival Antagonists, 2001–11." *Journal of Peace Research* 51 (3): 347–60. https://doi.org/10.1177/0022343313518940.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. https://doi.org/https://doi.org/10.1002/sim.4067.

For Peer Review