A Theory of Neutrality Regimes in War: International Law and the Value of Non-Compliance

Scott Wolford
University of Texas
swolford@austin.utexas.edu

June 29, 2018

Abstract

I analyze a model of war expansion in the shadow of international law, where a neutrality regime serves as a collectively prearranged screening device that sorts belligerents according to their wartime ambitions (satiated or insatiated). The regime helps resolve a third party’s uncertainty over the desirability of intervention when only insatiably states violate the law, such that enforcement of international law can emerge from power calculations and not principled legal commitment. The model shows that neutrality regimes can be effective not despite but because of inconsistent compliance, that strong third parties are uniquely likely to regret not intervening in wars, that neutrality regimes are easiest to ratify when mutual suspicion is greatest, and more generally that neutrality regimes need not be epiphenomenal to power politics; rather, they can support the working of classical balance of power systems.

*Thanks to Phil Arena, Terry Chapman, Dave Clark, Jim Fearon, Nate Jensen, Leslie Johns, Amy Liu, Dan McCormack, Pat McDonald, Jim Morrow, Toby Rider, Emily Ritter, Greg Robinson, Bruce Russett, Patrick Shea, and audiences at the University of Maryland, University of California Merced, SUNY Binghamton, the 2014 Annual Conference of the Peace Science Society (International), and the 2016 Formal Models of International Relations conference at the University of Southern California for helpful comments and suggestions.
Neutrality regimes, which establish public commitments to the rights of states to remain aloof from others’ military affairs, have been part of international politics for centuries. Some states, like Great Britain during most 19th century European wars (Abbenhuis, 2014), declare neutrality, while others, like Belgium in 1839, have it declared (and guaranteed) for them (Partem, 1983, p. 24). Some violate others’ neutrality, as Germany did Belgium’s in 1914, while others studiously honor it. Japan, for example, nearly compromised its opening gambit in the Russo-Japanese War by refusing to attack enemy ships docked near those of neutral powers (Connaughton, 2004, pp. 37-40). Opinions also vary among decisionmakers on whether neutrality laws truly shape state behavior. In 1914, Germany’s chancellor accused England of going to war over the “scrap of paper” guaranteeing Belgian neutrality, yet it was Belgium’s peril—neither France’s nor also-neutral Luxembourg’s—that drew Britain into the war (Fromkin, 2004, p. 236; Hastings, 2013, pp. 85-102; Hull, 2014, pp. 41-43). Neutrality also appeared discredited after widespread disregard for it during the Second World War (Wylie, 2002, p. 2). Inconsistent compliance and beliefs about their efficacy notwithstanding, neutrality regimes both global, e.g. the Hague Conventions, and local, e.g. the 1957 Austrian State Treaty, are widely and often ratified by the same states that refuse to comply, punish violations inconsistently, or dismiss the law as irrelevant.  

A useful theory of neutrality regimes should account for each of these facts. Most work on international law, however, is either so skeptical as to render ratification, compliance, and

---

punishment puzzling or so concerned with explaining compliance that violations of a neutral’s rights are confused with failures of neutrality regimes themselves. Both perspectives elide the problem that neutrality regimes are designed to solve: undesirable war expansions (Bull 1977 pp. 36-38). States generally aim to save blood and treasure by staying out of ongoing wars, but they also wish to intervene against some belligerents in order to eliminate future threats (see Powell 1999 Ch. 5). One state that attacks another may or may not have plans for further postwar aggrandizement, but it has an incentive to claim benign intent regardless, if doing so can discourage third parties from balancing against it (see Kydd 2005). Third parties are thus uncertain over the value of balancing against states that attack others; intervening is useful if it weakens or eliminates a future enemy, but expanding the war is wasteful if the attacker would not have become a threat. States may signal and draw inferences about the scope of their foreign policy ambitions in a number of ways (see Chapman 2011; Ikenberry 2001; Voeten 2005), but if neutrality regimes are to help solve this problem, a useful theory must show how publicly ratified and accepted standards of behavior can facilitate judgments over the attractiveness of balancing.

I analyze a stylized model of war-fighting and war expansion under asymmetric information, where neutrality rights work as a collectively prearranged screening device, resolving a third party’s uncertainty over an attacker’s foreign policy ambitions. The theory rests on two main premises. First, publicly ratified international law establishes shared expectations that help states select from among many possible equilibria in response to violations (Morrow 2002, 2007). It defines the common conjecture, the broad agreement over how the game of international politics is to be played, that supports the strategies states choose in pursuit of their interests (Morrow 2014 Ch. 2). Second, states are concerned not about
compliance with the law *per se*, but with the scope of each other's foreign policy ambitions. Third parties would like to intervene against states that pose threats beyond today's war (insatiable types), but they prefer to stay out of the war when belligerents are not a long term-threat (satiable types), *regardless of observable respect for the law*. Neutrality regimes can clarify these intervention choices when third party states believe that only belligerents with insatiable aims violate the law, which also shapes initial choices over compliance with neutrality regimes. I show in an extension that this potential to avoid regretted interventions and provocations of balancing also encourages states to ratify neutrality regimes, particularly when they are mistrustful of each other's future intentions.

The theory accounts for variation in compliance with a particular class of international regimes, disagreements over their effectiveness, and their widespread ratification. By separating belligerents according to their ambitions, international law “works” not because it generates compliance but because it sorts belligerents according to their willingness to provoke an expanded war. Even if deterrence fails, the system of law itself does not. Perfect compliance, in fact, can lead to ex post regret as insatiable belligerents go unchecked. Thus, violations of the law can *make* it effective by resolving uncertainty over the desirability of intervention. Neutrality regimes are also easiest to ratify when mistrust among potential adopters is greatest—i.e., when states suspect each other of future insatiable ambitions. The model makes additional predictions about the conditions under which international law has an independent effect on state behavior, the characteristics of those states least able to take advantage of the informational value of neutrality regimes, and patterns of war expansion and localization. Contrary to claims that subordinate an epiphenomenal international law to the workings of *realpolitik* (e.g., [Mearsheimer, 1994/95], collectively prearranged screen-
ing devices like neutrality regimes can both exist and alter behavior not despite but because of the high stakes nature of war and the deep mistrust out of which it can emerge.

**The Puzzle of Neutrality Regimes**

Neutrality regimes come in both global and local flavors. First, states can invoke global legal standards like the Hague Convention to declare neutrality with respect to a specific conflict. Second, local regimes can emerge from active neutralization, or what Partem (1983, pp. 16-20) calls “multilateral declarations of neutrality,” by which neutral status for particular states is guaranteed by treaty, such as 1839’s Treaty of London, its 1867 extension to Luxembourg, or the 1955 Austrian State Treaty. These balance of power treaties remove issues of contention from relations between other powers that might be tempted to use the territory or resources of the state in question to menace each other. However declared, neutrality stipulates certain rights and obligations; neutrals aid neither side in a war, and belligerents honor neutrals’ independence and territorial integrity. Should a belligerent violate a neutral’s rights, the victim may retaliate, and other states that wish—or are obligated—to uphold its neutrality can punish violations legally by joining the war.

Neutrality regimes are designed in principle to limit the spread of war. Yet to the extent that international law is enforced, states (or other actors) must find it in their interest to punish those that violate the rules. This renders compliance puzzling. Why, in the absence of a centralized authority able to punish them, do states accept and honor legal obligations? One answer focuses on the endogenous nature of the terms of cooperation: states only consent to laws with which they expect to comply, rendering compliance spurious to actions
states would already take absent legal obligations (Carrubba, 2005; Chapman and Chaudoin, 2013; Downs, Rocke and Barsoom, 1996). Others invoke the realization of a trivially natural propensity to comply (Chayes and Chayes, 1993) or a lack of compelling reasons to violate (Valentino, Huth and Croco, 2006), the activation of domestic pressure groups (Conrad and Ritter, 2013; Dai, 2005), the spurious signaling value of ratification (Simmons, 2000; Simmons and Danner, 2010), or the role of commonly known rules in coordinating voluntary punishments against violators (Johns, 2012; Morrow, 2002; Reinhardt, 2001).

I focus here on the latter mechanism. By stipulating a clear set of neutrality rights, through global conventions or local treaties, states coordinate expectations on what types of state violate the law and the appropriate response. States often make decisions over violating international law long after acceding to the relevant protocols, and this is particularly true with respect to the laws of neutrality. The rights and obligations of neutrality constituted customary rules and practices long before the Hague Convention (Bull, 1977, Ch. 2), where states accepted protocols under “a veil of ignorance of which wars [would] be fought” (Morrow, 2002, pp. S54-S55) in the future. Unilateral declarations of neutrality can thus rely on long-established shared beliefs about neutrality rights. Spain remained neutral during both World Wars, while the United States attempted the same (Hastings, 2012; Knock, 1995); Great Britain and France declared neutrality in the American Civil War, angering the United States by recognizing Confederate belligerency (Jones, 2010, pp. 40, 45); the United States declared neutrality during the Anglo-French War of 1793 (ibid., p. 52); and the United Kingdom declared neutrality during nearly all 19th Century European wars, notably excepting the Crimean War (Abbenhuis, 2014). Even multilateral declarations of neutrality occur in the shadow of an uncertain future; Belgian neutrality was guaranteed in 1839, seventy-
five years before it was violated, and with the nearby powers—France, Prussia, and the United Kingdom—promising to punish violations with force (see Hull 2014 pp. 17, 31).

However states come by their neutrality, the question of whether and how the relevant laws shape state behavior remains an open question. Skepticism borne of its apparent failure in the Second World War lingers among both academics and national policymakers (Mearsheimer 1994/95; Reiter 1996), and obligations of collective security embodied in the United Nations charter—and the League of Nations convenant before it—would seem to place the rights of neutrals on questionable ground in the postwar era (Carr, 1964; Wylie 2002). Yet as the international system changes, emerging from a unilateral arrangement with rising and resurgent great powers mounting challenges to the liberal postwar order, we should not expect neutrality regimes to remain obscure—nor should we forego an improved understanding of historical institutions or regimes at the expense of gaining insight to modern-day or future analogues. Nonetheless, realist traditions adopt something akin to Germany’s 1914 view of international law (e.g. Mearsheimer, 1994/95, 2001), and neutrality has received little treatment in either more optimistic assessments of international law or studies of laws designed to limit violence between belligerents.

Recent research on laws that limit violence in war offers a useful framework in which to view the laws of neutrality, because it matches theories of law with theories war-fighting (Morrow 2007, 2014; Posner 2003), ensuring that law has the chance to affect specific choices over the behaviors and outcomes in question. States in a conflict consider the value of compliance with the relevant laws—the treatment of prisoners and civilians or prohibitions against certain weapon systems or tactics—by calculating the risks of reciprocal punishment or normative sanction that would follow violation (see also Wallace 2012b).
When it comes to laws governing the rights of neutrals, however, belligerents are typically concerned over the possibility of additional states, including those guaranteeing the neutrality of others, joining the war in response to a violation. Japan’s reason for restraint in 1904, for example, was tied directly to the fear that striking American, French, or British ships would bring about swift—and potentially disastrous—intervention in the war. Therefore, understanding neutrality regimes requires matching a theory of law with a theory of war expansion, in which codified neutrality rights can shape wartime decisions by clarifying what constitutes violations and how others may respond.

Codified, publicly ratified neutrality regimes address this problem; they can help potential violators and punishers agree on which of many possible equilibria will be played (Morrow, 2002), confirming for a belligerent what a third party will view as a violation and supporting an expectation that violation will result in punishment. Law thus informs the common conjecture, or the shared beliefs about actions, appropriateness, and social roles that help actors agree on which equilibrium to play (Morrow, 2014, Ch. 2). Coordinating punishment is especially relevant for compliance due to threats of reciprocity, which supports laws limiting violence or the treatment of prisoners of war (Morrow, 2001, 2002, 2007, 2014; Posner, 2003; Wallace, 2012a). Such laws govern the actions of belligerents with respect to other belligerents; however, neutrality regimes govern the behavior of belligerents with respect to non-belligerents. Theories concerned with direct reciprocity often assume that states have no incentive to punish if no violations occur (Carrubba, 2005; Downs, Rocke and Barsoom, 1996), and in many instances, e.g. trade agreements (Reinhardt, 2001) or the abuse of prisoners of war (Morrow, 2014; Wallace, 2012b), this is reasonable. But when states respond to violations of neutrality, the stuff of ostensible punishment for violating
the neutrality regime can be desirable for some potential interveners, for whom failing to join the war against a potential future threat might entail ex post regret.

To see how punishing a violator can be desirable, consider the anticipated consequence of the German and (prospective) Japanese violations of neutrality in 1914 and 1904: an expansion of the war by or on behalf of aggrieved neutrals. In contrast to the treatment of prisoners, which states threaten to wield reciprocally, third parties may view joining and expanding the war as attractive whether or not laws were violated. When states join ongoing wars, they do so mindful of the costs of fighting, their ability to affect the outcome, their valuation of the issues at stake, and future threats posed by the belligerents (Altfeld and Bueno de Mesquita, 1979; Powell, 1999; Waltz, 1979). Potential interveners are potential balancers, trying to gauge whether a belligerent is “generally aggressive” or simply “dissatisfied with the particular state it attacked” (Powell, 1999, pp. 193-194). Observers are thus concerned over the scope of a belligerent’s ambitions. Is it satiable, and therefore likely to be satisfied after victory, or is it insatiable, and thus unlikely to cease its aggrandizement after winning today’s war? Defeating an insatiable belligerent alongside others today is easier than facing it alone in the future, but doing so is costly and risky, and intervention is still more wasteful if the belligerent turns out after the fact to have been satiable. With full information, states might join the war against an insatiable belligerent, banding with its current target to balance against a future threat, regardless of its respect for the law.

Contrast this with the logic of reciprocal punishment that sustains limitations on violence in war, where in many cases (a) neither side has an incentive to punish the other—that is, violate laws it hopes to see upheld—unless the other has first violated and (b) there are strong incentives to return to cooperation rather than repeatedly punish one another in
a spiral of costly violations. The threat of punishment may be necessary to preserve deter-
rence, but carrying out the threat is undesirable in and of itself, unless a belligerent, like
Imperial Japan, the Soviet Union, or Germany (with respect to the Soviet Union) in the
Second World War, prefers an unrestrained battlefield (Morrow 2014 Ch. 5). Yet when join-
ing a war promises benefits that remaining aloof does not, such as forestalling an adverse
shift in the distribution of power that would follow an insatiable state’s victory, joining the
war is desirable. The challenge for non-belligerents is identifying desirable interventions,
of divining aggressive aims in the face of strong incentives to claim benign intent. As such,
understanding the laws of neutrality requires asking how codified neutrality rights might
help resolve uncertainty about the desirability of joining ongoing wars.

Connecting neutrality regimes to changes in state behavior requires a focus on war ex-
pansion, which depends on the coordinating role of international law in (a) clarifying ac-
ceptable and unacceptable behaviors and (b) establishing shared expectations over the con-
sequences of violations. When states agree on what constitutes a violation of neutrality, and
when those laws are believed to be broken only by states with insatiable ambitions, neu-
trality regimes work as a collectively prearranged screening device, a commonly accepted
mechanism by which states draw inferences about each other’s preferences and intentions.
Effective neutrality regimes draw lines in the sand and dare insatiable belligerents to cross
them. I show below that these lines not only facilitate accurate judgments about joining
wars but also imply that neutrality regimes work best when they are occasionally violated.
Further, states submit themselves to these potentially revelatory regimes not when trust
and cooperation are easy but when mistrust is greatest.
**The Model**

Suppose that a belligerent state (B) is engaged in a war in which it can increase its chances of victory by violating the rights of some neutral state. It might invade the neutral along the way to its main target, or it might launch an attack on its main target that puts at risk a neutral’s citizens or military forces; each case entails a violation of the laws of neutrality.

Next, suppose that some non-belligerent state (A), which may or may not be the state whose neutrality B can exploit, chooses whether or not to join the war against B. A can observe B’s actions, but it is uncertain over the desirability of intervention. If B is insatiable, placing a high value on today’s war for further aggrandizement, A would like to defeat B now, but if today’s victory will satiate the belligerent, A would rather save the costs of war.

---

2The target of B’s attack, if not A itself, is not a strategic player in this model.
The game tree in Figure 1 shows that Nature first chooses B’s type, insatiable with probability \( \phi \) and satiable with probability \( 1 - \phi \), and reveals it only to B.\(^3\) When insatiable, B values victory in today’s war more highly than when it is satiable, because its interests encompass not only today’s conflict but also the position in which it will be placed for future aggrandizement. A satiable B is interested only in defeating today’s opponent, and violating the laws of neutrality offers no additional benefits beyond an increased chance of victory, since B has no further aims after today’s war. An insatiable B is a long-term threat with extensive aims that impinge on A’s security, but a satiable type will be relatively content with simply winning today’s war, with no consequences for A’s future interests.

To formalize the scope of B’s ambitions, I define its type as its valuation \( \beta > 0 \) for winning today’s war. When \( \beta = \bar{\beta} \), B is insatiable, and when \( \beta = \hat{\beta} \), it is satiable, where \( 0 < \hat{\beta} < \bar{\beta} \). Belgium, for example, was attractive to Germany in 1914 not merely as a path to France and major producer of iron and steel (Stevenson, 1982, p. 504) but as a location from which to menace Great Britain and isolate it from the Continent. An insatiable Germany with designs on using Belgian territory for these ends would value defeating Belgium more highly than one that merely wished to fight a holding action against an aggressive France. Thus, the scope of B’s ambitions also shapes A’s incentives to join the war against B; it would prefer, all else equal, to remain neutral if B is satiable but to intervene if B is insatiable, because fighting alongside B’s current target is more attractive than waiting to face a vic-

\(^3\)I have also solved a variant of the model in which B can be one of a continuum of types, but the results are substantively identical and more difficult to discuss than the two-type model.
torious $B$ alone in the future. Next, $A$ knows the probability distribution from which $B$’s type is drawn, so it begins the game uncertain over the desirability of intervention, using the type distribution to inform its prior beliefs ($\phi_0$) over $B$’s type. $A$ believes initially that $B$ is insatiable with probability $\phi_0 = \phi$ and satiable with probability $1 - \phi_0$.

Next, $B$ chooses whether to violate ($v$) or honor ($h$) the rights of a neutral state, which can be $A$ or some other neutral whose neutrality $A$ guarantees. If $B$ violates the neutrality regime, then its chances of victory increase; honoring the regime does not change its chances of victory. $A$ then decides whether to join the war ($j$) against the belligerent, which is costly up front but reduces $B$’s chances of victory, or not join ($\neg j$). The game can end in four ways: (a) $B$ honors the neutrality regime and $A$ does not join the war ($h, \neg j$); (b) $B$ honors neutrality and $A$ joins the war ($h, j$); (c) $B$ violates neutrality and $A$ does not join the war ($v, \neg j$); and (d) $B$ violates neutrality and $A$ joins the war ($v, j$).

After $A$’s decision, each player receives a payoff reflecting the expected outcome of the war, less the costs of violating neutrality ($c_B > 0$) or joining the war ($c_A > 0$). If $B$ wins the war, it receives $\beta = \{\bar{\beta}, \bar{\beta}\}$, and it receives zero if it loses, though it pays upfront costs ($c_B > 0$) to divert resources to fight a second target if it violates neutrality. $A$ receives its best outcome of zero if $B$ loses the war. Should $B$ win, $A$ receives $u_A(\bar{\beta}) \leq 0$, which decreases in the scope of $B$’s ambitions, such that $u_A(\bar{\beta}) < u_A(\bar{\beta}) \leq 0$. $A$’s worst outcome is victory by an insatiable $B$, while victory by a satiable $B$ is at least as bad for $A$ as either type of $B$.

---

4Since $B$ has already decided to fight, all other costs of war are sunk.
losing. Next, $B$’s probability of winning the war, $p_B \in (0, 1)$, is

$$p_B \equiv (b \lambda^v - ja),$$

which increases in $B$’s fighting capacity ($b > 0$) and decreases in $A$’s ($a > 0$), where $b > a$ and $j = 1$ if $A$ joins the war against $B$ and $j = 0$ if $A$ does not join. Next, $\lambda > 1$ is a military boost that $B$ receives only if it violates the law (i.e., if $v = 1$). Finally, I assume that

$$\frac{c_B}{b(\lambda - 1)} < \beta < \frac{c_B}{b(\lambda - 1) - a},$$

which rules out uninteresting equilibria by ensuring that $B$ is willing to violate if it knows that it will not provoke intervention, but that the satiable type will not value victory so highly as to violate when intervention is certain.

The model ensures that $B$ is tempted to violate the law while $A$ is uncertain over the value of intervening, highlighting a key distinction from other treatments of international law: if $A$ knew $B$ to be insatiable, it would join the war whether or not $B$ honored the law. To the extent that $A$ “punishes” violations, the action that triggers punishment may be spurious to its effects on $A$’s beliefs over $B$’s ambitions. Thus, the law can affect both compliance with the law and the expansion of war (as neutrality regimes aim to do), even when states do not place any principled value on each other’s compliance. It is important to note, however, that the model (as well as the ratification extension below) assumes the existence of

---

5To ensure that $p_B < 1$, I impose $b < (1 - ja)/\lambda^v$ throughout the analysis.
a clear redline, a specific behavior agreed to constitute a violation. The precise location of that line is often contentious, as evidenced by extensive negotiations over contraband goods headed to Europe, as well as American president Wilson’s inclusive understanding of the rights of neutral citizens on belligerent ships, during the First World War \cite{Philpott2014, Tooze2014}. Nonetheless, this exploratory model \cite{Clarke2012} is a necessary step in (a) defining the value of agreeing to such a redline, as the Hague Conventions provide in questions of neutrals’ territorial integrity, and (b) setting up the extension to the ratification of neutrality regimes after the primary analysis.

**Analysis**

Accounting for patterns of compliance with, enforcement of, and beliefs about the value of neutrality regimes requires that we identify when a neutrality regime exists and when it has an independent effect on state behavior. A neutrality regime exists when (a) states believe that only insatiable types violate and (b) agree on the likely response: if $B$ violates, $A$ will believe it to be insatiable, which $B$ knows will be followed by $A$ joining the war against it. Any equilibrium in which a neutrality regime exists must be supported by these particular strategies and beliefs. But states can in principle settle on the same common conjecture absent codified neutrality rights. If law is not to be epiphenomenal, it must coordinate state expectations when there exist multiple equilibria from which to choose; if an equilibrium consistent with a neutrality regime is unique, then there are no alternative equilibria on which to coordinate. Therefore, the law affects state behavior only when states play an equilibrium consistent with the neutrality regime rather than an equilibrium that exists for
the same parameter values but is inconsistent with the regime.

There are several Perfect Bayesian Equilibria, each of which involves a unique combination of strategy profiles and beliefs, including B’s choice over honoring or violating and A’s beliefs and strategies in response. Figure 2 plots the equilibrium space as a function of two parameters: A’s prior belief that B is insatiable ($\phi_0 = \phi$) and the insatiable type’s valuation for victory ($\beta$). Moving up the vertical axis, A begins the game increasingly certain that B is insatiable; as $\phi$ approaches 1, A’s threat to join the war becomes more credible regardless of B’s behavior. The horizontal axis tracks the extremity of B’s aims ($\bar{\beta}$); near the origin, the difference between $\beta$ and $\bar{\beta}$ is negligible, but moving right along the axis, the difference between types increases, and an insatiable type poses an ever larger threat to A than a satiable type. This allows for a characterization of A’s ex ante assessment of the desirability of intervention, which depends on both (a) prior beliefs and the difference between types and (b) how B’s strategy affects those beliefs.

Figure 2 shows the parameter values for which neutrality regimes exist and for which they can shape state behavior. In every equilibrium marked by an asterisk (*), A believes that only insatiable types violate, which also implies that A is sure to join the war in response. Equilibria consistent with the neutrality regime exist throughout the space, but the neutrality regime can affect behavior only when there also exists an equilibrium in which satiable types can be tempted by a weak threat of intervention to violate opportunistically. When A’s threat of intervention is sufficiently strong—i.e., when $\phi > \bar{\phi}$ such that it begins the game relatively confident that B is insatiable—then a satiable B is deterred from violating in all equilibria (characterized in Proposition 3 in the appendix). There exists either a partial compliance equilibrium in which B probabilistically honors if insatiable (hoping
to mask its type and avoid intervention) or a separating equilibrium in which $B$ violates if insatiable (revealing its type to $A$ and provoking intervention). Since these equilibria are unique in their regions of the parameter space, coordination is moot, and the neutrality regime is epiphenomenal—i.e., there is no behavioral difference between worlds with and without a neutrality regime. But a neutrality regime can shape state behavior when $A$ begins the game less certain that $B$ is insatiable.

**Proposition 1.** When $\phi \leq \bar{\phi}$, PBE with the following strategies exist.

- For all $\bar{\beta}$, there exists a semi-separating **opportunistic violation** equilibrium in which $B$ violates if insatiable and violates probabilistically if satiable; $A$ does not join if $B$ honors but joins probabilistically if $B$ violates.

- When $\bar{\beta} \leq c_B/(b(\lambda - 1) - a)$, there exists a pooling **full compliance** equilibrium in which
\(B\) honors regardless of type; \(A\) joins iff \(B\) violates.

- When \(\bar{\beta} > c_B/(b(\lambda - 1) - a)\), there exists a **separating** equilibrium in which \(B\) violates iff insatiable; \(A\) joins iff \(B\) violates.

Therefore, a neutrality regime both exists and shapes state behavior.

Proposition 1 characterizes the multiple equilibria that exist when \(A\) is sufficiently doubtful of \(B\)'s insatiability that its threat to intervene is not inherently credible (\(\phi \leq \bar{\phi}\)), creating an opportunity for law to help players select one equilibrium over the other. Key to this part of the space is an opportunistic violation equilibrium in which \(B\) violates when insatiable and occasionally when satiable, seizing a military advantage by violating the law not because of aggressive intent but because of the low risk that \(A\) intervenes in response. In this equilibrium, \(A\) probabilistic intervention produces regret over both wasted interventions (when \(B\) turns out to have been satiable) and failures to join the war (when an insatiable is allowed to go unchecked). Therefore, a neutrality regime can coordinate expectations on one of two equilibria in which \(B\) violates only when insatiable, rendering violations informative for \(A\) over the desirability of joining the war. In the **separating** equilibrium, \(B\) violates when insatiable, prompting \(A\) to join the war, and this threat deters a satiable \(B\) from violating. In the **full compliance** equilibrium, \(A\)'s threat to punish violations remains credible, but it is not called in; \(B\) honors the law regardless of its type.

The **full compliance** equilibrium exists when \(A\) does not believe \(B\) too likely to be insatiable and when the insatiable type's aims are not too extreme (though it is still worth
fighting), or when

\[ \bar{\beta} \leq \frac{c_B}{b(\lambda - 1) - a}. \]  

(1)

At this equilibrium, A wields a perfectly credible threat of intervention in the event of violation, which ensures that B honors the law whether satiable or insatiable. Should B violate, A believes it to be insatiable, such that A is happy to join the war in hopes of preventing B from posing a future threat. In this quadrant of the parameter space, full compliance with the law occurs simply by the coordination of expectations. The opportunistic violation equilibrium is a plausible way to play the game, but if a common conjecture can develop around a neutrality regime, a simple change in expectations can move states to an equilibrium in which the neutrality regime “works.” States agree on what constitutes a violation and the likely response, shifting them from an equilibrium in which both insatiable and satiable types violate to one in which both satiable and insatiable types comply with the law.

In standard accounts of international law, the full compliance equilibrium appears to be the best possible case for its effectiveness, ebcouse the law is never violated. A’s threat is sufficiently credible that it is never called in on the equilibrium path, but as a result it foregoes a potentially desirable intervention against an insatiable belligerent, which has masked its true intentions by honoring the law. Since insatiability is relatively unlikely and the difference \( (\bar{\beta} - \hat{\beta}) \) between the threat posed by satiable and insatiable types is small, A tolerates a low probability of regretted non-intervention in return for saving the costs of war associated with attacking a compliant belligerent that A believes likely to be satiable \( (\phi \leq \hat{\phi}) \). A engages in no wasteful interventions, but it foregoes desirable interventions that
might otherwise eliminate a future threat. In this sense, the neutrality regime supports a threat of intervention that may be too strong; a breach of the law would be informative if it occurred, but $B$ refuses to reveal itself as insatiable, mimicking a satiable type by complying on the equilibrium path and leaving $A$’s information problem unsolved. Compliance is achieved at the cost of $A$ regretfully allowing some future threats to go unchecked.

Next, a _separating_ equilibrium in which $B$ violates only when insatiable (and despite certain intervention) exists alongside violation when the difference between insatiable and satiable types is relatively large, or $\bar{\beta} > c_B / (b(\lambda - 1) - a)$. Like it does in the full compliance equilibrium, $A$ wields a credible threat of intervening after violation, and the satiable $B$ remains content to honor the law. Nonetheless, $B$ violates when insatiable, sufficiently bent on reaping the gains of violating neutrality that it cannot be deterred. The law is violated, but given $A$’s preference to intervene against insatiable belligerents, the regime itself is not a failure. $B$’s violation is informative, and in contrast to the full compliance equilibrium it occurs on the equilibrium path, touching off what is for $A$ a desired expansion of the war. Given the certainty of provoking intervention, $B$ only violates when doing so offers a substantial military boost, disregarding the neutrality regime despite its commonly known existence, just as Germany did in 1914. Both observers and violators might be skeptical or contemptuous of the neutrality regime—the law is violated and punishment is supported by self-interest—but failures of deterrence are not a failure of the regime itself.

While it cannot be said that international law “works” at the separating equilibrium in the sense of deterring violations as it does under full compliance, it does affect state behavior by forcing a separation of belligerent types that allows $A$ to identify _desirable_ interventions. By establishing a collectively agreed upon redline and daring insatiable states to cross it,
states that ratify neutrality regimes create a prearranged screening device that aligns desirable military interventions with punishment of states that violate the law. In the separating equilibrium, $A$ avoids both wasteful interventions and regretted non-interventions. If $A$ valued only $B$’s compliance—that is, if intervention were useful only as a punishment with the aim of deterrence—then we would rightly interpret the neutrality regime as doing a poor job as a deterrent when it guides states to this equilibrium. But when securing others’ compliance is less important than choosing desirable interventions, punishment is spurious to the knowledge $A$ gains from the fact that $B$ cannot be deterred: $B$ is insatiable and thus worth attacking in order to minimize a future threat.

Since $A$ would intervene if it knew $B$ to be insatiable, even if $B$ honored the law, this analysis presents a different view of the role of international law in shaping state behavior. In the full compliance equilibrium, threats of intervention work in the sense that moderately insatiable states honor neutrality rights. This comes at a cost to $A$, who forgoes intervention when it observes rights being honored, despite the fact that it might be allowing a future threat—one that it would wish to fight under complete information—to grow under its nose. When the insatiable $B$’s aims are so expansive as to render it undeterrable, we observe the separating equilibrium, where neutrality regimes “work” precisely because they are violated: only if the insatiable $B$ violates the law can $A$ can tailor its strategy to join the war against an insatiable belligerent. Only under the separating equilibrium, in which the law is occasionally violated, can $A$ avoid both wasted interventions and regretted non-interventions. As discussed further below, this recommends caution in choosing standards by which to judge the effectiveness of international law.
Extension: Ratification and Neutralization

Neutrality regimes can shape state behavior by coordinating expectations, but if they limit insatiable states’ military options and expose them to potential balancing interventions, why ratify them in the first place? International law is the product of voluntary ratification of documents like the Treaty of London, the Austrian State Treaty, or the Hague Conventions, and states are often wary of committing to regimes that restrict their freedom of action. However, states ratify local or global neutrality regimes under a veil of ignorance over the future wars in which they will be involved, what their roles in those wars may be, and thus whether they may benefit from prearranging a particular screening device. I show in this section that sufficient levels of mutual mistrust—of bilateral fears that the other side will one day harbor insatiable goals—can actually facilitate ratification.

Consider an extension in which each player $i = \{A, B\}$ chooses simultaneously whether or not to ratify a neutrality regime before playing the game in Figure 1. When both states ratify, they acknowledge that playing violate ($v$) is against the law, but unless both ratify, there is no such shared expectation and the neutrality regime does not exist. To ensure that law can meaningfully coordinate expectations, I restrict my focus to $\phi \leq \bar{\phi}$, where law can select among multiple equilibria. Failing to jointly ratify puts players into the opportunistic violations (i.e., “no law”) equilibrium, in which both types of $B$ violate, and $A$ experiences both wasted interventions and regretted non-interventions. Joint ratification, which is costless,

---

$6$Since I model ratification as costless, states would be indifferent over acceding to neutrality rights and rejecting them in this part of the equilibrium space.

21
creates a shared expectation that states will play either the full compliance or separating equilibrium consistent with a neutrality regime, depending on the value of $\beta^7$.

Ratification decisions occur under a veil of ignorance, such that neither state $(i, k)$ knows its identity ($A$ or $B$) in a future war—though it is sure to be one of the two—or its type (insatiable or satiable) of it turns out to be state $B$. In the first move, each state receives a private signal $\theta_{i,k}$ related to the true probability $\theta$ with which $i$ will be state $A$ in a future interaction. When $\theta_i$ is high, $i$ believes itself likely to be $A$; when $\theta_k$ is high, on the other hand, $k$ believes itself likely to be $B$. The sequence of play is: (1) Nature sends private signals of identity, (2) states choose simultaneously whether or not to ratify, (3) Nature chooses identities, and (4) the game then proceeds as it does in Figure 1. Thanks to its built-in veil of ignorance, this specification allows both states to believe the other likely to be an insatiable type. In other words, mistrust can be both two-sided and severe.

**Proposition 2.** Suppose that $\phi \leq \Phi$ and $\lambda > 1 + a/b$. When states $i, k$ do not know their future identities, ratification occurs when $\theta_k \leq \hat{\theta} \leq \theta_i$. The specific value of $\hat{\theta}$ depends on $\beta$.

Proposition 2 states that, as long each side states assess a sufficiently high probability of being state $A$ itself (and thus menaced by $B$’s possible insatiability), it ratifies the neutrality regime. Thus, states choose to coordinate their expectations on a set of laws protecting the rights of neutral states, collectively arranging a screening device that allows them to make informed decisions in the future over joining ongoing wars. The ratification constraint,

---

Costless ratification ensures that no exogenous costs of signing the agreement stop states from acceding to neutrality rights; any disincentive to ratification emerges from the consequences of doing so.
\[ \theta_k \leq \hat{\theta} \leq \theta_i, \] requires that \( i \) believe itself likely enough to be \( A \) in the future (\( \theta_i \geq \hat{\theta} \)) and that \( k \) believe itself likely enough to be \( A \) in the future (\( \theta_k \leq \hat{\theta} \)). In other words, each player expects to confront the other as an insatiable type at some point in the future. The ratification constraint is also easier to satisfy when insatiability is extreme (\( \bar{\beta} > c_B/(b(\lambda - 1) - a) \)); state \( i \) will be comparatively more menaced by the other’s insatiability if it turns out to be \( A \) itself, but if it turns out to be state \( B \) it is also more likely to tolerate threats of intervention in pursuit of military expediency. This is because more extreme levels of insatiability, i.e. \( \bar{\beta} > c_B/(b(\lambda - 1) - a) \), are associated with the separating equilibrium in which an insatiable \( B \) cannot be deterred from violating, revealing its type and provoking \( A \) to intervene.

We can equate the two-sided mistrust of this veil of ignorance with precisely those conditions under which states ratify global neutrality regimes, like the Hague Conventions, and multilateral declarations of neutrality in which signatories have strong reasons not to trust one another. Similar conditions of shared mistrust drove Great Britain, France, and Prussia into a collective guarantee of Belgian neutrality in 1839. If this runs counter to intuition, suppose that only one of this great power triad were expected to be the belligerent power in the future. If all states agreed that, say, France would be tempted to violate Belgian neutrality as part of a program to attack the German states or menace the United Kingdom, such a multilateral guarantee of Belgium’s neutrality would have been less viable; this hypothetically insatiable France would refuse to pre-commit to reveal its type and provoke intervention. Thus both local and global regimes are “acutely dependent on the existence and reasonable continuity of convergent (though not necessarily common) interests among the concerned governments” ([Black et al., 1968] p. vii).
Various small powers have occasionally found their independence guaranteed because their powerful neighbors were unable to agree on any other solution short of a war they preferred to avoid. (Rothstein 1968, p. 193, quoted in Partem 1983, pp.16-17)

When powerful states mistrust one another, they reasonably try to minimize the costs of that mistrust, and both local and global neutrality regimes can be an attractive means of doing so. International cooperation over neutrality regimes is possible not despite but because of mutual mistrust. Notably, this logic places buffer states in a rather different light than that cast by Fazal's (2007) treatment of state death; buffers may die at greater rates than non-buffer states, but while they live their neutralization can keep the peace between mistrustful enemies that might otherwise intervene in one another's wars (see also Wagner, 2007, Ch. 4, 5), and they might not exist at all if not for that underlying mistrust.

Summary and Discussion

Neutrality regimes see inconsistent patterns of compliance, enforcement, and apparent normative commitment to the law. Nonetheless, a global regime has existed formally since 1905—and informally before that—while local regimes remain a common guarantee of the neutrality of individual states. Inconsistent compliance despite ratification may tempt us to infer that neutrality regimes are epiphenomenal, but that reading cannot account for instances of costly compliance, like Japan's in 1905, and enforcement, like Britain's in 1914. Judging neutrality regimes by the standard of compliance elides the very reason that many states find such regimes valuable: violations that clarify the value of intervening in ongoing
wars. The present theory can account for these patterns, from compliance and enforcement to ratification, as well as conflicting views of the value of law, by viewing neutrality regimes not toothless seekers of compliance but as collectively prearranged screening devices.

If neutrality regimes identify desirable interventions, avoiding both wasteful fighting and regretted decisions not to intervene, then we should expect to see inconsistent compliance with and respect for the associated laws. In the separating equilibrium, an insatiable state violates the law, a third party infers the violator’s type and takes actions to balance against it—and the target of this “enforcement” can rightly note that the intervener’s claim to uphold international law is less than principled. And yet, given the third party’s goal of making accurate judgments about when to join wars, it would be incorrect to say that the neutrality regime is ineffective. Nor should it be surprising that those paying a price for running afoul of the regime dismiss it as a “scrap of paper” upheld by hypocritical, self-interested states no more committed to the spirit of the law than the violator. The third party’s enforcement may be spurious to the knowledge it gains from the belligerent’s willingness to violate, but absent this individual incentive to join the war, violations of neutral rights would be even more common—as implied by a comparison to the opportunistic violation equilibrium. In this way, neutrality regimes can actually support the operation of a classic balance of power system, sometimes identified as a realm in which international law should not be particularly effective (see, inter alia, Mearsheimer [1994/95]).

Viewing neutrality regimes as prearranged screening devices sheds light on several more patterns. First, neutrality regimes have their greatest impact when insatiability is believed to be rare. Coordination is trivially easy when insatiability is very likely, because $A$ has an inherently credible threat to join the war. Yet when $A$ begins the game less certain that it
faces an insatiable belligerent, there is room for law to alter behavior, because there also exists an equilibrium in which satiable types opportunistically violate. Law selects equilibria only when states do not already share expectations about intervention. If states can already coordinate on responses to violations, then there is little scope for international law to “matter.” Setting aside specific behavioral predictions, looking for the effects of neutrality regimes requires that we identify cases where insatiability is not believed likely ex ante. Sampling on cases involving both likely and unlikely insatiability risks over-attributing the effects of international law on patterns of compliance/violation and war expansion.

Second, thanks to their screening properties, overt violations of neutrality—e.g., Germany’s invasion of Belgium in 1914—may represent a failure of deterrence but not of the neutrality regime itself. Full compliance entails regretted non-intervention, where (a) A is sufficiently doubtful of B’s insatiability and (b) its threat to intervene following violation is so credible that insatiable states mask their type, pooling with satiable states and avoiding interventions that, all else equal, A would have wished to launch. The separating equilibrium, on the other hand, occurs when the threat of intervention cannot deter an insatiable type of B from violating but can deter a satiable type; here, the law solves the information problem with which A begins the game, allowing A to both avoid wasteful interventions and engage in those that are desirable. Full compliance is not the best-case scenario for a neutrality regime designed to screen out insatiable states; the separating equilibrium is.

Further, and counter to some specific realist critiques of international law and institutions (Mearsheimer, 1994/95, p. 29), punishment for violations of the law requires nothing more than A’s self-interested desire to identify worthwhile preventive wars.

Because neutrality regimes are screening devices, deterring some violations but encour-
aging desirable interventions when deterrence fails, compliance is a problematic standard by which to judge its effects on behavior. Rates of compliance are a useful explanatory variable when cooperation is sustained by threats of reciprocal punishment (see Morrow, 2002, p. S44), but when ostensible punishment is spurious to desirable interventions, noncompliance does not indicate that international law has no effect—much less that it does not have a desirable effect—on state behavior and outcomes. To assess the model's empirical implications, it is necessary, first, to identify cases where neutrality regimes can have an effect, such as a low ex ante probability of insatiability. Second, the analyst must identify differences between the opportunistic violations equilibrium and its alternatives. In this case, the most prominent difference is a reduced rate of wasteful interventions, because both violations and regretted non-interventions still occur in equilibrium. Identifying regretted interventions is difficult, particularly because war entails regret over its costs regardless of its underlying aims (Fearon, 1995), but other outcomes such as rates of compliance and violation may not be sufficient to distinguish laws with and without effect.

The model also suggests that more powerful third-party states should be more prone to regretted non-intervention than weaker states, despite their inherent military advantages in threatening war. At first blush, this runs counter to intuition: powerful states are better able to recover the costs of war, and as such they should have fewer excuses for failing to intervene or to convince their targets of that fact. But this line of reasoning misses the point that the very credibility of a powerful state's threat to intervene, and the extent to which it can affect $B$'s military fortunes, alters an insatiable $B$'s incentives to violate the law in the first place. We can see this in Figure 2, where the threshold dividing full compliance from the separating equilibrium is $\bar{\beta} = c_B/(b(\lambda - 1) - a)$. This threshold falls as $A$'s military
power \((a)\) increases; as a result, full compliance exists for values of \(\bar{\beta}\) where a separating equilibrium would otherwise exist for lower levels of \(A\)’s military power.

The very credibility of \(A\)’s threat to intervene forces the insatiable \(B\) to mimic the satiable type, leaving \(A\)’s information problem unsolved and forcing it, thanks to its prior doubts that \(B\) is insatiable (low \(\phi\)), not to intervene. \(A\) then runs the risk of allowing an insatiable state to go unchecked. Weaker third parties should thus appear better able than the strong to tailor their intervention decisions to their desire to balance against future insatiable threats. This is consistent with the violation of Belgian neutrality in the First World War; Germany, convinced that any British military contribution to the war would be marginal, was more than willing to risk intervention in order to realize its plans of sweeping into France through Belgium (Hastings, 2013, p. 101).8 Had the United Kingdom been able to meaningfully intervene in what Germany expected to be a short land campaign on the Western Front, then it might have ended up in a situation of regretted non-intervention had Germany adjusted its strategy for the invasion of France.

The model has further implications for the expansion of war, because some violations of neutrality are not deterred by the threat of intervention on the equilibrium path. Thus, 8

8Germany’s ambassador in London, Prince Lichnowsky, certainly expected that violating Belgian neutrality would influence the British decision: “The question as to whether we invade Belgian neutrality in our war with France may be of decisive importance in determining that of England’s neutrality” (quoted in Fromkin, 2004, p. 246). The issue is that Moltke and most of the general staff believed that British intervention simply would not matter on the ground, as France would be knocked out of the war before a sufficient British army could land on the Continent, if it could at all.

28
violations of neutrality increase the probability that wars expand to include new belligerents, a hypothesis thus far missing from studies of conflict expansion (see, *inter alia*, Altfeld and Bueno de Mesquita, 1979; Gartner and Siverson, 1996; Joyce, Ghosn and Bayer, 2014; Siverson and Starr, 1990). While attacking a neutral state accounts for expansion in a trivial way, some violations might on their own be small enough to make retaliation difficult; the United Kingdom was more than willing to disavow itself of its obligations to Luxembourg in 1914 (Fromkin, 2004, p. 244). Yet violating one state’s neutrality may bring other states, its guarantors in particular, into the conflict against the violator. Further, the states most likely to respond to violations of neutrality by joining wars are relatively weak, because their initial threats to intervene are not sufficiently credible to induce an insatiable B into full compliance. Precisely because powerful states would intervene, their threat is never called in, and insatiable states violate neutrality rights and provoke war expansion only when the expected intervention is not too significant. This explains Germany’s willingness to violate Belgian neutrality in 1914; the leadership knew that the British might intervene, but their military contribution was expected to be negligent.

Finally, the extended model shows that neutrality regimes are easiest to ratify when concerned states are most fearful of one another. The veil of ignorance under which treaties are ratified ensures that the law can have its proposed effects in observational data, but it also suggests that neutrality regimes emerge from the interactions of those states that trust each other the least. Only when each side believes itself likely to be the victim of the other’s insatiable aims in the future do we see neutrality regimes ratified in equilibrium. Even when neutrality regimes born of mutual mistrust break down, their violation signals to one side that its initial mistrust was justified, revealing that intervention in today’s war is
worthwhile to eliminate a future threat. This explains both why states ratify global regimes like the Hague Conventions, as well as why antagonistic great powers—such as France, the UK, and Prussia in 1839—might guarantee the neutrality of particular states that offer military advantages in hypothetical future wars.\footnote{Such mutual mistrust may also explain why the Ottoman Empire, the “Sick Man of Europe,” managed to survive for as long as it did, and why it ultimately died when it did: the breakdown of mutual mistrust among blocs of great powers during the First World War.} Multilateral neutralizations often create buffer states, whose rates of survival are lower than the average (Fazal, 2007), but the act of neutralization creates a screening device by which states can identify insatiable belligerents worthy of fighting before they pose still larger threats after victory.

**Conclusion**

Understanding the effects of neutrality regimes on patterns of war and peace requires that we combine a theory of international law with a theory of the problem it is designed to solve: the expansion and spread of war. Neutrality regimes, whether global or local, are ratified under a veil of ignorance about future wars, and they serve not as mechanisms of creating compliance but as collectively prearranged screening devices. When neutrality regimes “work,” they ensure that only insatiable states violate the law, helping third parties make informed decisions not about punishing violators per se but about joining wars when doing so can help eliminate a future threat. In contrast to other bodies of international law that focus on trade policy, human rights, or limits on violence in war, neutrality regimes are
most effective when they are occasionally (and informatively) violated. Compliance is thus a poor standard by which to judge neutrality regimes. When neutrality regimes coordinate expectations on equilibria in which third parties can avoid both wasted interventions and regrets over staying on the sidelines, they can be viewed as a working part of a balance of power system, where international law is often dismissed as epiphenomenal.

Combining a theory of war expansion with a theory of law allows this modeling exercise account for numerous, apparently incompatible, stylized facts about neutrality regimes: varying rates of compliance and punishment, divergent beliefs among states about their meaning, and why—in spite of an apparently poor record—even mistrustful great powers often ratify them. Further, neutrality regimes are not the only prearranged screening devices in international relations. Arms control agreements, for example, might work in a way similar to the model analyzed here. To the extent that pursuing a certain weapon system, from the battleships covered by 1922’s Washington Naval Treaty to the nuclear weapons prohibited by the Nuclear Nonproliferation Treaty, is associated with a violator’s future ambitions, enforcing the regime may be compatible with eliminating a future threat—just as it is under neutrality regimes. Punishment is spurious to the revealed desirability of acting against the violator, through preventive war or sanctions (see Coe and Vaynman [2015]; McCormack and Pascoe [n.d.]), and as such not all violations are “failures” of the regime if they clarify and help states address threats that would otherwise be difficult to identify. The analogy to arms control agreements suggests that they may be easy to ratify when mistrust is both deep and mutual, just as it is with neutrality regimes.

Their similarities with other screening institutions notwithstanding, it is worth asking what neutrality regimes mean for contemporary international relations. The global regime
supported by the Hague Conventions remains in force, yet in an age of total war and international institutions inspired (at least in part) by collective security, the “restraint and balance” on which neutrality regimes depend seem to carry less weight than in the past (Wylie, 2002, p. 10). Belgium may have given up its neutrality after its experience in the World Wars, but proposals for the neutralization of European states farther east—most notably Ukraine after Russia’s invasion and seizure of Crimea (Mearsheimer, 2014)—are not uncommon. Should nuclear weapons continue to limit the potential scale of great power conflict and unipolarity give way to wider-ranging multipolar competition, neutrality may return to the lexicon of international politics. In the South China Sea, where China disputes control over some of the world’s most valuable shipping lanes with multiple smaller states, declarations of neutrality that invoke the Hague Conventions in a potential future conflict could prove important in shaping third party decisions to intervene. Even Ukraine, dismembered after Russian intervention, may be a candidate for neutralization, a buffer between East and West whose agreed status can serve as a redline for those worried about curbing future attempts at aggrandizement in European security politics. Such an agreement would pay little heed to the preferences of those who live in Ukraine, but the state’s neutrality might serve as a common, prearranged means by which Russia on the one hand and NATO on the other can judge each side’s foreign policy ambitions.

Finally, the model makes a key assumption that players agree on the specific actions that constitute violations of neutrals’ rights. In some cases, such as the preservation of territorial integrity against physical encroachment, the line is longstanding and clear. In others, from the transfer of contraband material through neutral ports (an issue in the First World War) to the use and manipulation of nonbelligerent networks in cyberwar, the lines
are not so clear. Applying the present theory of neutrality regimes as screening devices to specific conflicts requires grappling with these very issues, but a theory of what redlines do once established is a necessary first step in developing a theory of why and how specific redlines are chosen over others. And if declarations of neutrality reemerge as part of the post-unipolar international order, an understanding of how neutrality regimes work in general will be crucial to anticipating their effects on a changing security environment.

Appendix

Proof of Proposition 1. Begin with the semi-separating opportunistic violation equilibrium. The satiable type $\beta$ is rendered indifferent by $A$’s strategy of randomizing after violation, such that $EU_B(h|\beta) = EU_B(v|\beta)$. This requires that the probability of $A$ joining following violation ($y$) satisfy $(b) \beta = y(b\lambda - a)\beta + (1 - y)(b\lambda)\beta - c_B$, yielding

$$y^* = \frac{b(\lambda - 1) - c_B/\beta}{a}.$$

To ensure that $y^* \in [0, 1]$, it must be the case that

$$\frac{c_B}{b(\lambda - 1)} \leq \beta \leq \frac{c_B}{b(\lambda - 1) - a}.$$

Next, we verify that the insatiable type is happy to violate, or $EU_B(v|\bar{\beta}) \geq EU_B(h|\bar{\beta})$. Given $A$’s strategy, this constraint is satisfied when $y^* (b\lambda - a)\bar{\beta} + (1 - y^*)(b\lambda)\bar{\beta} - c_B \geq (b) \bar{\beta}$. This is sure to be true as long as $\bar{\beta} > \beta$, which is true by assumption.

Next, we verify that $A$’s beliefs are consistent and strategies optimal given $B$’s strategy.
If $B$ honors, $A$ believes $\phi' = 0$, which implies that $EU_A(\neg j| h) \geq EU_A(j| h)$ can be represented by $(b) u_A(\bar{\beta}) \geq (b - a) u_A(\bar{\beta}) - c_A$, which is true when

$$c_A \geq -a u_A(\bar{\beta}). \tag{2}$$

If $B$ violates, $A$ is rendered indifferent between joining and remaining neutral by the probability with which the satiable type violates ($g^*$), such that $EU_A(j) = EU_A(n)$. By Bayes’ Rule, $A$’s posterior belief that $B$ is insatiable given violation is

$$\phi' = \frac{\phi}{\phi + (1 - \phi) g} \equiv \chi.$$ 

Therefore, the indifference condition is

$$\chi (b \lambda - a) u_A(\bar{\beta}) + (1 - \chi) (b \lambda - a) u_A(\bar{\beta}) - c_A = \chi (b \lambda) u_A(\bar{\beta}) + (1 - \chi) (b \lambda) u_A(\bar{\beta}),$$

which yields the satiable type’s probability of violation,

$$g^* = \frac{\phi \left( c_A + au_A(\bar{\beta}) \right)}{-(1 - \phi) \left( c_A + au_A(\bar{\beta}) \right)}.$$

Next, ensuring that $g^* \in [0, 1]$ requires, first, that $-au_A(\bar{\beta}) < c_A < -au_A(\bar{\beta})$, which binds over Inequality (2) since

$$\phi \leq \frac{c_A + au_A(\bar{\beta})}{au_A(\bar{\beta}) - au_A(\bar{\beta})}. \tag{3}$$
Therefore, opportunistic defection exists under the stipulated conditions.

Next, consider the full compliance equilibrium. Begin with $A$’s strategy, which is not to join if $B$ honors and join if $B$ violates. In the event that $B$ honors, $A$ chooses not to join when $EU_A(\neg j) \geq EU_A(j)$, or when

$$
\phi(b)u_A(\bar{\beta}) + (1-\phi)(b)u_A(\beta) \geq \phi(b-a)u_A(\beta) + (1-\phi)(b-a)u_A(\beta) - c_A,
$$

where its posterior belief is simply its prior, $\phi' = \phi$. This inequality can be satisfied, first, when $c_A \geq -au_A(\bar{\beta})$ or, second, when $-au_A(\beta) < c_A \leq -au_A(\bar{\beta})$ and

$$
\phi \leq \frac{c_A + au_A(\beta)}{au_A(\beta) - au_A(\bar{\beta})} \equiv \overline{\phi}.
$$

Next, $A$ joins if $B$ violates. Letting $\phi''$ denote its posterior (out of equilibrium) belief that $B$ is insatiable, it joins when $EU_A(j) \geq EU_A(\neg j)$, or

$$
\phi''(b\lambda-a)u_A(\bar{\beta}) + (1-\phi'')(b\lambda-a)u_A(\beta) - c_A > \phi''(b\lambda)u_A(\bar{\beta}) + (1-\phi'')b\lambda)u_A(\beta).
$$

This inequality is satisfied for any out of equilibrium beliefs $\phi''$ when $c_A < -au_A(\bar{\beta})$, and for

$$
\phi'' > \frac{c_A + au_A(\beta)}{au_A(\beta) - au_A(\bar{\beta})}
$$

when $-au_A(\beta) < c_A < -au_A(\bar{\beta})$. $A$ must believe that any type of $B$ that deviates from the equilibrium action is sufficiently likely to be insatiable, which satisfies the Intuitive Cri-
terion (Cho and Kreps, 1987), because A should be more likely to rule out a satiable type that can only do strictly worse by deviating from the equilibrium than an insatiable type. Now, consider $\beta$, who honors when $EU_B(h|\beta) \geq EU_B(v|\beta)$, or $(b)\beta \geq (b\lambda - a)\beta - c_B$. This inequality is satisfied when $\beta \leq c_B/(b(\lambda - 1) - a)$. The insatiable type honors when $EU_B(h|\beta) \geq EU_B(v|\beta)$, or $(b)\beta \geq (b\lambda - a)\beta - c_B$. This inequality is satisfied when $\beta \leq c_B/(b(\lambda - 1) - a)$. Since $\beta > \beta$, the binding constraint is $\beta \leq c_B/(b(\lambda - 1) - a)$.

Finally, to prove the existence of the separating equilibrium, begin with B’s incentive-compatibility constraints. The satiable type honors, such that $EU_B(h|\beta) \geq EU_B(h|\beta)$, or $(b)\beta \geq (b\lambda - a)\beta - c_B$, and the insatiable type violates, such that $EU_B(v|\beta) \geq EU_B(v|\beta)$, or $(b\lambda - a)\beta - c_B \geq (b)\beta$. Both constraints are satisfied when

$$\beta \leq \frac{c_B}{b(\lambda - 1) - a} \leq \bar{\beta} \quad \text{and} \quad \lambda > 1 + \frac{a}{b}.$$  

Since player-types of B perfectly separate, A’s posterior beliefs are $\phi' = 1$ if B violates and $\phi' = 0$ if B honors. As such, it intervenes following a violation when $EU_A(j) > EU_A(\neg j)$, or when $(b\lambda - a)u_A(\beta) - dc_A > (b\lambda)u_A(\beta)$, which is satisfied when $c_A < -au_A(\beta)$. A refuses to intervene if B honors when $EU_A(\neg j) \geq EU_A(j)$, or when $(b)u_A(\beta) \geq (b - a)u_A(\beta) - c_A$, which is satisfied when $c_A \geq -au_A(\beta)$. Combining these constraints, $-au_A(\beta) \leq c_A < -au_A(\beta)$ ensures that A’s strategy is a best response and consistent with its beliefs. $\square$

**Proposition 3.** When $\phi > \bar{\phi}$, A believes that only the insatiable B violates in all PBE, which are also unique for their parameter values. Therefore, a neutrality regime exists but does not shape state behavior.

**Proof of Proposition 3.** The proof has three steps. First, the existence of the separating equi-
librium when $\bar{\beta} > c_B/(b(\lambda - 1) - a)$ is established by Proposition 1, so it remains to prove the existence of a partial compliance equilibrium that exists for $\phi > \bar{\phi}$ and $\bar{\beta} \leq c_B/(b(\lambda - 1) - a)$.

Second, I show that no other equilibria exist in these regions of the parameter space.

The following strategies and beliefs constitute the PBE with partial compliance. If satiable, $B$ honors. If insatiable, $B$ honors with probability $h^*$ and violates with probability $1 - h^*$. If $B$ honors, $A$ believes $\phi' = \sigma$ and intervenes with probability $r^*$, choosing not to intervene with probability $1 - r^*$. If $B$ violates, $A$ believes $\phi' = 1$ and intervenes.

Begin with $\bar{\beta}$, rendered indifferent by $A$’s strategy of mixing after $B$ honors, such that $EU_B(\mathbf{v}|\bar{\beta}) = EU_B(\mathbf{h}|\bar{\beta})$. This requires that $r$ satisfy $(b\lambda - a)\bar{\beta} - c_B = r(b - a)\bar{\beta} + (1 - r)(b)\bar{\beta}$, yielding

$$r^* = \frac{(b(\lambda - 1) + a)\bar{\beta} + c_B}{a\bar{\beta}}$$

as the probability with which $A$ intervenes if $B$ honors. To ensure that $r^* \in [0, 1]$, it must be the case that

$$\frac{c_B}{b(\lambda - 1)} \leq \bar{\beta} \leq \frac{c_B}{b(\lambda - 1) - a}.$$

Next, we verify that the satiable type is happy to honor, or $EU_B(\mathbf{h}|\beta) \geq EU_B(\mathbf{v}|\beta)$. Given $A$’s strategy, this constraint is satisfied when $r^*(b - a)\bar{\beta} + (1 - r^*)(b)\bar{\beta} \geq (b\lambda - a)\bar{\beta} - c_B$, or when $\bar{\beta} > \beta$, which is true by assumption.

Next, we verify that $A$’s beliefs are consistent and strategies optimal given $B$’s strategy. If $B$ violates, $A$ believes $\phi' = 1$, which implies that $EU_A(j|\mathbf{v}) > EU_A(\neg j|\mathbf{v})$ can be represented
by \((b\lambda - a)u_A(\bar{\beta}) - dc_A > (b\lambda)u_A(\bar{\beta})\), which is true when \(c_A < -au_A(\bar{\beta})\). If \(B\) honors, \(A\) is rendered indifferent over joining and remaining neutral by the probability with which \(\bar{\beta}\) honors \((h^*\)) such that \(EU_A(j|h) = EU_A(\neg j|h)\). By Bayes’ Rule, \(A\)’s posterior belief that \(B\) is insatiable given that it honored is

\[
\phi' = \frac{\phi h}{\phi h + (1 - \phi)} \equiv \sigma.
\]

Therefore, the indifference condition is

\[
\sigma(b - a)u_A(\bar{\beta}) + (1 - \sigma)(b - a)u_A(\bar{\beta}) - c_A = \sigma(b)u_A(\bar{\beta}) + (1 - \sigma)(b)u_A(\bar{\beta}),
\]

which yields the insatiable type’s probability of honoring,

\[
h^* = \frac{-(1 - \phi)\left(c_A + au_A(\bar{\beta})\right)}{\phi\left(c_A + au_A(\bar{\beta})\right)}.
\]

Ensuring that \(h^* \in [0, 1]\) requires, first, that

\[
\phi \geq \frac{c_A + au_A(\bar{\beta})}{au_A(\bar{\beta}) - au_A(\bar{\beta})} \equiv \bar{\phi}
\]

and, second, \(-au_A(\bar{\beta}) \leq c_A \leq -au_A(\bar{\beta})\). Therefore, this PBE exists in the stipulated part of the parameter space.

Next, I show that these equilibria are unique in their parts of the space, by ruling out alternative equilibria. By Proposition 1, opportunistic violations and full compliance PBE do not exist when \(\phi > \bar{\phi}\). Remaining candidate equilibria entail pooling, such as full compliance.
or full violation, or strategies in which $A$ does not condition its choices on $B$’s strategy. First, we can rule out any pooling equilibrium in which $A$ does not intervene on the equilibrium path, because its priors are so high as to undermine any commitment not to intervene, as it does in the full compliance equilibrium. Therefore, since $\phi > \bar{\phi}$, there can exist no equilibrium in which $B$ pools (on either action) and $A$ does not join. Second, there exists a PBE in which $B$ pools on violation and $A$ joins only if $B$ honors (off the equilibrium path), but this requires implausible out-of-equilibrium beliefs that if $B$ honors it must be an insatiable type. Third, there can be no equilibrium in which $B$ pools on violation and $A$ only joins on the equilibrium path, because $\beta < c_B/(b(\lambda - 1) - a)$ (which is true by construction) ensures that $\beta$ has a profitable deviation to honoring. Fourth, there exists a PBE in which $B$ pools on compliance and $A$ joins only on the equilibrium path, but it requires $A$, off the equilibrium path, to believe implausibly that $B$ is likely to be satiable if it violates. Fifth, there exists an equilibrium in which $B$ pools on compliance and $A$ joins both in and out of equilibrium, but it requires $\bar{\beta} < c_B/(b(\lambda - 1) - a)$, which is untrue by construction. Finally, there is no separating equilibrium in which $B$ honors only if insatiable, because the insatiable type has a profitable deviation to violating. Therefore, separating and partial compliance PBE are unique in their respective areas of the parameter space when $\phi > \bar{\phi}$.

Proof of Proposition 2: Suppose that $\phi \leq \bar{\phi}$ and $\lambda > 1 + a/b$. To prove the claim, I first establish that ratification is a best reply for each player under full compliance and separating regimes, then specify what beliefs are weakly consistent with Bayes’ Rule along the equilibrium path.

When $\bar{\beta} \leq c_B/(b(\lambda - 1) - a)$, let $u_i(f, \theta_i)$ denote state $i$’s payoff for the implied full compli-
ance regime, in which it receives A’s payoff,

$$\varphi(b)u_A(\beta) + (1-\varphi)(b)u_A(\beta),$$

with probability $\theta_i$ and B’s payoff,

$$\varphi(b)\overline{\beta} + (1-\varphi)(b)\beta \geq \varphi \left( y^*(b\lambda - a)\overline{\beta} + (1-y^*)(b\lambda)\overline{\beta} - c_B \right) + (1-\varphi)(b)\beta,$$

with probability $1-\theta_i$. State i’s payoff for the opportunistic violations equilibrium is given by $u_i\text{nl,}\theta_i$, and it equals

$$\varphi \left( -\frac{c_A b\lambda}{a} \right) + (1-\varphi) \left( g^* \left( -\frac{c_A b\lambda}{a} \right) + (1-g^*)(b)u_A(\beta) \right)$$

(4)

For state i to ratify, it must be the case that $u_i\text{fc,}\theta_i \geq u_i\text{nl,}\theta_i$, which is true when

$$\theta_i \geq \frac{c_B \left( c_A + au_A(\beta) \right) (\overline{\beta} - \beta)}{\overline{\beta} c_B \left( c_A + au_A(\beta) \right) - \overline{\beta} \left( c_A \left( c_B + b(\lambda - 1) \left( u_A(\beta) - u_A(\overline{\beta}) \right) \right) \right) + ac_B u_A(\beta)} \equiv \hat{\theta}.$$

Next, since $\theta_k$ is k’s estimate that $i = A$, it follows that k will ratify when $\theta_k \leq \hat{\theta}$, which establishes the final condition for ratification under full compliance, $\theta_k \leq \hat{\theta} \leq \theta_i$.

Finally, let $\overline{\beta} > c_B/(b(\lambda - 1) - a)$ such that ratification entails a separating equilibrium. Let $u_i\text{se,}\theta_i$ denote i’s payoff for ratifying a separating equilibrium, which yields the left side of Inequality (4) with probability $\theta_i$ and $\varphi \left( (b\lambda - a)\overline{\beta} - c_B \right) + (1-\varphi)(b)\beta$ with probability $1-\theta_i$. As before, i compares this payoff to $u_i\text{nl,}\theta_i$, such that it ratifies when $u_i\text{se,}\theta_i >
\( u_i(nl, \theta_i) \), which is true when

\[
\theta_i \geq \frac{\bar{\beta} \left[ au_A (\bar{\beta}) + c_A \right] \left[ \beta(a - b \lambda + b) + c_B \right]}{\bar{\beta} \left[ au_A (\bar{\beta}) + c_A \right] \left[ \beta(a - b \lambda + b) + c_B \right] - \bar{\beta} \left[ au_A (\bar{\beta}) + c_A \right] \left[ u_A (\bar{\beta}) (a - b \lambda + b) + c_A \right]} \equiv \hat{\theta}.
\]

Next, since \( \theta_k \) is \( k \)'s estimate that \( i = A \), it follows that \( k \) will ratify when \( \theta_k \leq \hat{\theta} \), which establishes the final condition for ratification under full compliance, \( \theta_k \leq \hat{\theta} \leq \theta_i \).

Perfect Bayesian Equilibrium requires that beliefs are weakly consistent with Bayes’ Rule on the equilibrium path. Proposition 1 establishes weak consistency for each post-ratification subgame (post-ratification play is effectively a subgame, since beliefs acquired in the ratification stage are irrelevant once Nature draws identities and types), but it remains to clarify how players’ beliefs change along the equilibrium path in the ratification stage. First, let each state’s private signal over \( \theta \) (defining its type, or its mistrust of \( k \)) be distributed according to some continuous function \( f(\theta_{i,k}) \), where \( \theta_{i,k} \in (0,1) \). Second, since \( i \) ratifies when its private signal is \( \theta_i \geq \hat{\theta} \), \( k \)'s posterior belief over \( i \)'s type simply truncates the distribution to \( \theta_i \in [\hat{\theta},1) \). Third, by a similar logic, \( i \)'s posterior following \( k \)'s ratification is \( \theta_k \in (0,\hat{\theta}) \). Should joint ratification fail, \( i \)'s posteriors are truncated at \( \theta_k \in (0,\hat{\theta}) \) if \( k \) ratifies and \( \theta_k \in [\hat{\theta},1) \) if \( k \) does not ratify; \( k \)'s posteriors are \( \theta_i \in [\hat{\theta},1) \) if \( i \) ratifies and \( \theta_i \in (0,\hat{\theta}) \) if \( i \) does not ratify. Therefore, in the ratification equilibrium but before identities can be revealed in the final subgame, each state knows that the other is relatively mistrustful.
References


