Title
Crisis Bargaining and War Initiation Before a Domestic Audience

Author
Gurantz, Ron Moti

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Crisis Bargaining and War Initiation Before a Domestic Audience

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Political Science

by

Ron Moti Gurantz

2014

ABSTRACT OF THE DISSERTATION

Crisis Bargaining and War Initiation

Before a Domestic Audience

by

Ron Moti Gurantz

Doctoral Candidate in Political Science

University of California, Los Angeles, 2014

Professor Arthur A. Stein, Chair

Studies of crisis bargaining have traditionally focused on the strategies for signaling resolve to other states, even when incorporating a domestic audience into the analysis. In this dissertation, I examine crisis bargaining strategies meant primarily to send signals to the domestic audience. Using game theoretic models, historical case studies and a survey experiment, I show that governments can successfully bait adversaries into minor incidents and deceive the public into authorizing war even though minor incidents can also be highly informative about an adversary's intentions under certain conditions. I also show that public opinion can lead governments to forgo preemptive strikes and preventive war, and that democracy can therefore reduce public welfare in some circumstances. I discuss the implications of these findings for the conventional wisdom on crisis behavior and on the value of democracy in foreign policy.

The dissertation of Ron Moti Gurantz is approved.

Barry O'Neill

Stergios Skaperdas

Marc Trachtenberg

Robert Trager

Arthur A. Stein, Committee Chair

University of California, Los Angeles

2014

Contents

List of Figures

ACKNOWLEDGEMENTS

I would like to thank my committee chair Arthur A. Stein and committee members Barry O'Neill, Stergios Skaperdas, Marc Trachtenberg and Robert Trager for their guidance. My appreciation goes to Tiberiu Dragu, Robert Powell and Mehdi Shadmehr for their comments on parts of the dissertation, and Albert Carnesale and Barbara Geddes for the opportunities they gave me in graduate school. I would also like to thank my friends and colleagues in graduate school who helped me in my work and my life, Liana Maris Epstein, Matthew Gottfried, Alexander V. Hirsch, Andrew MacDonald, Sarah Leary, Dov Levin, Chad Nelson, Steve Palley, Bradford Stapleton, Laura Weinstein and Lauren Wong. For their support, my gratitude goes to my family, Devorah, Itzhak and Maya Gurantz.

Chapter 3 is a version of *Fear, Appeasement and the Effectiveness of Deterrence*, co-written with Alexander V. Hirsch, that has been invited for resubmission at *International Organization*.

Ron Moti Gurantz

EDUCATION

University of California, Los Angeles

M.A., Political Science, October 2008

University of California, Berkeley

B.A., Political Economy, May 2004

WORK IN PROGRESS

"Fear, Appeasement and the Effectiveness of Deterrence" with Alexander V. Hirsch

Revise and resubmit at *International Organization*.

"War Initiation Before a Domestic Audience"

"Provoked Incidents as Pretexts for War"

"Mutual Optimism and the Vietnam War, 1964-1968"

"Commitment Problems and Neutralization in Laos and South Vietnam"

CONFERENCE PRESENTATIONS

Midwest Political Science Association, Chicago, Ill., April 2013

"Military Incidents as a Pretext for War"

American Political Science Association, Washington D.C., September 2010

"Deterrence with a Self-Enforcing Threat," with Alexander V. Hirsch

Midwest Political Science Association, Chicago, Ill., April 2010

"War as a Self-Enforcing Threat," with Alex V. Hirsch

RESEARCH EXPERIENCE

Prof. Barbara Geddes, September 2009-October 2010

Data collection and coding for NSF-funded project, Authoritarian Regimes and Time Horizons.

Prof. and Chancellor Emeritus Albert Carnesale, July 2007-December 2007

Research for projects on missile defense, nuclear forensics, homeland security spending and U.S. nuclear weapons posture.

FELLOWSHIPS

Charles F. Scott Fellowship, UCLA Graduate Division Endowed Fellowship, 2010-2011

Graduate Summer Research Mentorship, UCLA Graduate Division, 2009

Graduate Summer Research Mentorship, UCLA Graduate Division, 2008

Chapter 1

Introduction

Studies of crisis bargaining have understood crises primarily as contests of resolve between states, and have focused their attention on strategies used to signal resolve and influence the beliefs of adversaries. When they have examined the role of the domestic audience, as in the audience cost and opposition signaling literatures, the focus has still been on how these actors influence the ability of states to signal resolve to opponents (Fearon 1994, Schultz 1998).

While the study of crisis bargaining has centered on signaling between the participant states, crisis events often have multiple audiences and the signals have multiple targets. Governments may be interested in signaling resolve to their adversaries, but they may also be interested in signaling other qualities to other states or to domestic audiences. Signaling to third parties can even become the primary objective in a crisis, for example, if a government has decided that war is inevitable but still wants to convince the international community that diplomacy cannot resolve the dispute. Observers sometimes voice the suspicion that governments provoke crises to manipulate or divert the attention of a domestic audience, with the adversary being largely irrelevant to the objectives of the provoking state.

A number of behaviors observed in particular crises - or discovered after the fact - cannot be understood except as attempts to manipulate the beliefs of a third party. Governments have conducted "counterfeit diplomacy," engaging in negotiations they fully expect to fail or even negotiations they don't want to succeed, as the United States did prior to the Persian Gulf War (Montgomery 2013). They have allowed themselves to be attacked rather than launching preemptive strikes, as Israel did at the beginning of the Yom Kippur War (Reiter 1995). They have exaggerated or even provoked enemy attacks before adopting policies that were already decided upon, as the United States government did in the Gulf of Tonkin incident (Moise 1996). They have even conducted false flag operations, staging incidents as enemy attacks before initiating war, as Germany did before invading Poland (Allen 2005).

In none of these examples is one state attempting to convey its resolve to an opponent. Instead, the common thread is that one government is attempting to make another party view its opponent as belligerent. Rather than signaling their willingness to fight, they are seeking to convey to a third-party observer the enemy's eagerness to fight and their own willingness to settle the dispute peacefully. In some of these cases, the strategy had major political consequences, and in the Yom Kippur War case, the behavior had major military consequences as well.

The existence of these strategies has potentially far-reaching consequences for our understanding of crisis bargaining. They challenge the notion that crises should always be understood as contests of resolve between states and introduces the possibility that crises may be more important for the information they convey to third parties. They may also challenge the conventional wisdom about the value of democracy in foreign policy by raising the possibility that public opinion can distort foreign policy behavior.

This dissertation studies crisis bargaining strategies meant to influence audience beliefs and explores the implications of these strategies for our understanding of crisis bargaining. It focuses on three commonly observed strategies targeted at influencing the audience rather than the enemy: provoking incidents, exaggerating incidents and allowing enemy attacks. All three have similar objectives of winning public support for war by convincing the audience of the adversary's hostility. All three also raise similar questions about the role of crisis bargaining in international affairs and the value of democracy in foreign policy-making.

Crisis Bargaining and Domestic Politics

Research on interstate crisis bargaining has been central to the study of international relations, constituting a sizeable literature and providing some of the foundational texts in the field. Crisis behavior has been studied using just about every approach in the field, including rational choice (Schelling 1966), institutional analysis (Allison 1969), and psychology (Jervis 1976). In recent years, many of the important elements of international crises have been studied formally. This literature has examined uncertainty about preferences (Kydd 1997), uncertainty about capabilities (Bueno de Mesquita, Morrow & Zorick 1997), concerns about reputation (Treisman 2004), costly signaling through mobilization (Fearon 1998, Slantchev 2011), probabilistic threats (Nalebuff 1986, Powell 1987), and the credibility of verbal communication (Sartori 2002, Trager 2010).

Most of these studies have focused on the state as a unitary actor attempting to influence an adversary's beliefs, but some have extended their models to incorporate the role of domestic politics, mainly in the studies of audience costs and opposition signaling. The theory of audience costs proposes that government leaders are able increase the credibility of their

threats by announcing them publicly, because the domestic audience will impose extra punishment on the leadership for reneging on these commitments (Fearon 1994, Smith 1998). The opposition signaling literature has argued that support by a political opposition can improve a threat's credibility, while a lack of support can undermine it (Schultz 1998, Ramsay 2004).

By extending the study of crises to include domestic politics, scholars have enhanced our understanding of crisis dynamics and gained insight into the role played by regime type in interstate bargaining. Both theories imply that democratic countries are better able to credibly signal intentions, either because the opposition can reveal relevant information or because bluffing before a domestic audience is costly for the government. Empirical work has used these insights to explain why democracies appear to be more successful than autocracies in resolving crises in their favor (Partell & Palmer 1999, Gelpi & Griesdorf 2001).

While these literatures have produced insights about the influence of domestic politics in crises, they have done so within the conventional understanding of crises as contests of resolve, wherein two states attempt to convince each other of their willingness to fight for some objective. As a result, their insights have been about the impact of the domestic audience on the credibility of the state's signals to its opponent. They have not yet considered the implications for crisis dynamics if the domestic audience itself is a target of government signaling, even though government leaders certainly take actions aimed at influencing the beliefs of domestic observers.

Governments usually need some level of public support and public mobilization to successfully enact their foreign policies. As a result, they have to be concerned about multiple audiences in their conduct of foreign affairs. This can have a number of consequences, such as governments taking domestic actions to suppress criticism of foreign policies, or governments changing their foreign policies in response to domestic opposition. In crisis situations,

it can also mean taking foreign actions meant to influence the willingness of the domestic audience to support the policy.

That governments may take action in the foreign policy realm to influence the beliefs of domestic actors is not a novel idea in international relations. In the study of war, much has been written on how international actions send signals to an audience, for example in the literature on diversionary war (see Levy (1998) for review, also see Smith (1995)). Downs & Rocke (1994) argue that states may continue to fight wars, even when against the country's best interest, in the hope that a good outcome will reflect well on the leader and salvage his political career. Stasavage (2004) has argued that leaders may pander and posture in international economic negotiations to signal their political views to the public.

Even in the study of crises, some studies have looked at behavior meant primarily to influence third parties. Montgomery (2013) writes that governments may engage in "counterfeit diplomacy," where they take diplomatic initiatives that have little hope of succeeding, or which they don't want to succeed, to create the appearance of peaceful intentions. A series of papers has argued that governments may also subject their policies to review by international institutions, such as the United Nations Security Council, to receive an impartial endorsement of their actions (Chapman 2007, Fang 2008, Thompson 2006). Stein (2000) argues that even the very act of making a verbal case for one's policy indicates a concerns with influence the audience's beliefs.

These represent the small number of studies that have extended this perspective in a systematic way to the study of crisis behavior. Other than these, there is a surprising gap in the literature, especially since scholars in political science and elsewhere have identified a number of other crisis behaviors that don't make sense except as attempts to influence the beliefs of audiences. Many of these take the form of attempting to frame the enemy

for attacking first. Governments may exaggerate incidents that they suspect are isolated, as occurred with the Gulf of Tonkin (Siff 1999). They may forego the opportunity to launch a preemptive war, preferring to allow the enemy to attack them (Reiter 1995). They may attempt to provoke their enemies into committing acts of aggression, as Schuessler (2010) argues the United States did before WWII.[1]

I begin to fill this gap in the literature by focusing on these crisis behaviors. Specifically, I examine behaviors surrounding war initiation and the actions that lead to states opening fire on each other. When wars break out, the actor that attacks first is often labelled the "aggressor," and the victim of aggression can justify his response as necessary for self-defense. These labels can have a major impact on public opinion and public mobilization, as well as international law and allied support. Due to the importance of identifying the aggressor, states often adopt strategies during crises to ensure that their opponent strikes – or appears to strike – first.

I focus on a set of strategies used to achieve this goal: provoking incidents, exaggerating incidents and forgoing preemptive strikes. In the three main chapters of the dissertation, I address a series of questions regarding these behaviors. First, I show how the introduction of an audience can change a crisis from a contest of resolve into a contest of restraint and show how a state can lose in such a contest. I examine the logic behind provocations and explain how states are able to bait their adversaries into incidents and mislead their publics. Second, I ask whether this finding implies that all minor incidents should be viewed simply as pretexts for war by asking whether it is possible for minor incidents to spark major wars in the absence of an audience. I show that even minor incidents can reveal an opponent's

[1]Along the same lines, states may even fake incidents in "false flag" operations to frame their opponent for an act of aggression (Allen 2005). I don't study this topic in the dissertation, but I do address it in the conclusion.

belligerent intentions and lead to major conflict. Finally, I show that an audience can restrict a government's willingness to initiate war and that the public can therefore make itself worse off by creating the incentive for a government to forego preemptive strikes when anticipating an enemy attack.

An understanding of crisis strategies that are targeted at the public rather than the adversary has the potential to overturn some fundamental assumptions about crisis behavior. Crises may not be best understood at contests of resolve between states, but instead as performances meant to convince third parties of some quality other than resolve, such as one's benign intentions or aversion to war. This may lead us to reinterpret major historical events that we believed we understood.

An understanding of these strategies may also carry theoretical implications about the value of democracy in foreign policy that contradicts conventional wisdom. The main implication of the existing literature on crisis bargaining and domestic politics is that democracy increases the credibility of threats. However, the phenomena examined here raise the possibility that democracies may take actions in crises that are not in the public's interest but are instead meant to manipulate public beliefs. The possibility of a manipulated public undermines the idea that government actions before a domestic audience are informative to the adversary, or even that they are in the public's interest.

Finally, understanding these phenomena better may help guide the public in how they react to crisis events. Suspicion about government motivations are common during crisis events, and this research may help us both understand how and when the public can avoid being manipulated. In particular, this can help answer the question of whether it is always wise for the public to blame the side that fires the first shot. It can also be helpful to know when it may be in the public's interest to increase the accountability of its government or

to insulate the diplomatic process from democratic pressures.

Baiting Adversaries into Incidents

Almost by definition, wars begin with military attacks. In the study of crisis bargaining, these attacks usually signify the end of the crisis and the transition into a new and different phase of interstate interaction, the war. Most crisis bargaining models end with either settlement or war, with war resulting from a mutual failure of the states to convince each other of their resolve.

Often, however, the attacks are simply pretexts for implementing policies already decided upon. When faced with the possibility of war, states also may face the challenge of mobilizing their public and convincing them to support the war. Being attacked by an enemy is one of the most effective ways to achieve this. The extensive literature on the "rally around the flag" effect has shown that attacks and other acts of aggression, such as the Gulf of Tonkin incident, the Iranian Hostage Crisis or 9/11, can generate enormous support for the government (Mueller 1973, Callaghan & Virtanen 1993, Hetherington & Nelson 2003). Given the political utility a government can get out of being attacked, it is not surprising that states may want a war to begin this way, and may seize upon attacks and incidents to justify military action.

There are number of examples in American history of the government using an incident as a pretext for war. In 1846, President Polk was prepared to ask Congress for a declaration of war against Mexico. Instead, following an attack against U.S. military forces along the Rio Grande River, President Polk wrote to Congress that Mexico had invaded American soil and requested that Congress simply recognize that a state of war already existed (Stein 2000). In

1941, President Roosevelt seized upon a German attack against the USS Greer to announce his "shoot on sight" policy against German and Italian ships in the Atlantic, even though this policy had already been adopted (Kimball 2004, Schuessler 2010). In 1964, President Johnson seized upon the Gulf of Tonkin incident to get Congress to authorize military action in Southeast Asia, powers which Johnson had wanted in order to deepen U.S. involvement in the conflict in Vietnam (Siff 1999).

The different roles that attacks play in crises lead to different views of the ways crises begin, escalate and end. Attacks could be costly signals of resolve, or could be escalations after a failure to settle. But they could also be excuses, seized upon by states that have already decided that war is inevitable. In the second chapter, I develop a formal model that demonstrates how the introduction of an audience can generate a crisis in which attacks are seized upon as excuses rather than costly signals of resolve or acts that initiate fighting. In that way, a crisis can become a contest of restraint, in which states attempt not to provide their adversaries with an excuse for war, rather than a contest of resolve in which states attempt to demonstrate their willingness to fight.

I also examine how it is possible to lose in a contest of restraint. In a crisis where an adversary is attempting to frame you for an act of aggression, it should be easy to avoid being framed. The state that wants to avoid blame simply has to withhold action. Since attacking is detrimental to the attacking state, why would that state ever commit an act of aggression? Why would a country commit an incident that allows its adversary to place the blame for a war on it?

A possible answer is that governments provoke their adversaries into these incidents. In fact, the above examples all had some sort of provocation. President Polk sent the U.S. troops into territory that was claimed by Mexico and widely recognized as Mexican

territory, and the location of the incident and status of that territory became a major point of disagreement in the ensuing debates over the war (Schroeder 1973). The attack on the USS Greer, which Roosevelt explained as an unprovoked act of aggression, was actually a response to the Greer's pursuit of a German submarine (Schuessler 2010). The attack on the Gulf of Tonkin was likely a response to U.S. supported South Vietnamese raids into North Vietnam, something that was understood by administration officials at the time (Siff 1999).

These examples provide some clue to the answer, but in a sense they skirt the main issue. A provoked state still has the option of practicing restraint in the face of a provocation. The question then becomes, why would a country allow itself to be baited by responding to a provocation, when this serves the interest of the provoking country?

In the model, I demonstrate that a vigilant and war-averse adversary can be baited into an incident, and that a sophisticated legislature can be deceived into believing the incident is an attack. When both states are uncertain whether or not the provocative action will be revealed to the legislature, it can be rational for both the provoking state to attempt to provoke an incident and for the adversary to respond in the hope that the provocative actions will be exposed. This can be the case even if the adversary would prefer to have absorbed the incident in order to avoid fighting a war. While the attempt to provoke war will fail if the provocation is observed by the legislature, if they do not observe the provocation, it is rational for them to believe that the adversary initiated the attack for the purpose of starting a war.

I find that increasing the likelihood that the provocation will be exposed to the legislature decreases the likelihood that the legislature will be deceived. The implication is that, although having a legislature with veto powers can lead the government to engage in deceptive practices, increasing transparency can reduce the government's ability to deceive. I

also show that this logic can apply when the adversary wants either to reverse the cost the provocation imposes or to preempt an eventual attack.

To test the predictions of the model, I conduct an online survey experiment of 1,000 respondents to test public approval for war under different conditions. Respondents are asked to read a vignette about the outbreak of war with Iran, and randomly assigned to treatments in which Iran attacks or does not attack American ships in the Persian Gulf, and in which it is or is not revealed that the United States provoked the Iranian attack. The findings of the survey are consistent with the predictions of the model, with an incident increasing public approval for war but a U.S. provocation moderating this effect.

Finally, I examine executive decision-making in the face of provocations through a case study of German and Japanese government policy following provocative U.S. actions before WWII. I find that the states attempted to practice restraint in the face of provocations and that the visibility of the provocation played a crucial role when that restraint failed. I show that my argument helps to explain why the U.S. was unsuccessful in provoking an incident with Germany but successful with Japan.

Learning from Minor Attacks

Wars are typically major events, characterized by coordinated action among thousands, tens of thousands, or even millions of people. They can require enormous expense and result in the disruption of daily life, the destruction of property, the displacement of people and the loss of life, sometimes on a massive scale. Given the enormous cost and the major effort necessary to conduct wars, it is no surprise that the stakes in war are often correspondingly large, and the underlying disputes can be serious and long-standing.

In the second chapter, I assume that attacks may either be isolated incidents or the beginning of wars, and that it is difficult for the public to distinguish between the two. This is an assumption of the model, similar to assumptions made in bargaining models. However, in reality, most of the incidents I have discussed could easily have been ignored or treated as isolated incidents.

Sometimes, the initial attack is proportionate to the war that followed, such as Germany's invasion of Russia in 1941. More often, initial attacks are minor relative to the war that erupts in their wake, and are not part of a major campaign or offensive. In 1846, the United States declared war against Mexico following a Mexican attack against a small American unit that killed 16 soldiers (Stein 2000). In 1940, the Soviet Union invaded Finland following an incident in which seven shells, purportedly from Finland, landed in a Soviet village (Edwards 2006). In 1964, the U.S. Congress authorized military action in Southeast Asia following a pair of incidents in which North Vietnamese boats allegedly fired a few torpedoes at U.S. ships, and one of these incidents probably never occurred (Moise 1996).

Even in international crises where war didn't occur, there seemed to exist possibility of major war breaking out over a relatively minor attack. In the Offshore Islands Crisis in 1954, the United States was willing to launch a major war in response to a Chinese attack on the economically and militarily insignificant islands of Quemoy and Matsu (Chang 1988). Throughout the Cold War, the United States was similarly willing to go to war if the Soviet Union were to attack the relatively small and indefensible outpost of West Berlin (NSC 173 1953). In 1994, the United States took seriously North Korean threats to launch a full-scale war following a limited U.S. strike on their nuclear plant (Wit, Poneman & Gallucci 2004).

Given the findings in the previous chapter that minor incidents can be provoked, and the puzzle that major wars sometimes break out following seemingly minor incidents, it is

easy to assume that these incidents are always just pretexts. Governments exaggerate minor incidents as representing major acts of aggression, and respond to them in a disproportionate manner because these incidents represent the best opportunity to blame the enemy. In some of the above cases, this is very likely what occurred. If this is true for every case, then any minor incident leading to major war can be dismissed as a ruse. Must that always be the case, or is it possible that a major war could be a rational response to a minor attack?

A few mechanisms have been proposed for how a major war can be a rational response to a relatively minor transgression or attack in the literature on crisis bargaining and deterrence. The most commonly cited mechanism has been the need to maintain a reputation (Alt, Calvert & Humes 1988, Treisman 2004). This is a compelling explanation in some cases, though in cases where the cost of war is extremely high – nuclear war, for example – it may not be rational to be concerned with "the reputation of a country that would no longer exist."[2] Scholars have also focused on mechanisms that allow a country to commit itself to carrying out an otherwise irrational threat, such as placing "trip-wire" forces in the threatened area, making public threats to increase one's costs of backing down, or increasing the probability of war breaking out by accident or through mechanisms outside the government's control (Schelling 1966, Nalebuff 1986, Fearon 1994). Again, these mechanisms may have some utility, but for costly wars it is very difficult to make a commitment so costly that this mechanism is effective.[3] In addition, the notion that there exist mechanisms whereby war can occur without a conscious decision by at least one party to initiate war is also questionable.[4]

[2]Quoting O'Neill (1999, 86), who was paraphrasing Herman Kahn.

[3]It is telling that both Schelling and Fearon, in explaining why these mechanisms could lead to war, fell back on the idea that they engaged the country's honor. This is an essentially reputational argument, and it implicitly recognizes that commitment devices such as these cannot usually, in and of themselves, explain the decision to go to war.

[4]In his study of over 500 years of major conflict, Luard (1986) writes, "[t]hroughout the whole of the period we have been surveying it is impossible to identify a single case in which it can be said that a war started accidentally: in which it was not, at the time when war broke out, the deliberate intention of at

In the third chapter of this dissertation, co-written with Alexander V. Hirsch, we describe a mechanism through which a minor attack or incident can reveal a great deal of information about the attackers intentions, specifically that the attacker intends major war in the future. If a state commits a trangression while anticipating that the defender will respond with war, then the attacking state's willingness to commit the trangression reveals that he prefers war to the status quo. If that preference guarantees that the attacker will initiate such a war in the future, then it can be rational for the defender to respond with full-scale war immediately, which fulfills the attackers initial expectation.

This mechanism will exist under certain conditions on the trangression itself. The inherent value of the transgression must be small enough, and the military value for a future war large enough, that an attacker would not be appeased by being allowed to transgress. The main implication of this finding lies in the realm of deterrence. We demonstrate that the threat of major war is able to deter minor transgressions when this condition holds, since the mechanism can make the threat of major war credible, and that deterrence can be more likely the less valuable the transgression is to the attacker. This mechanism also implies that a defender can benefit from being uncertain about an attacker's intentions, since the fear of a belligerent attacker gives credibility to the defender's threat of major war.

We demonstrate that this mechanism can help to explain deterrence successes like the American deterrence of Soviet aggression in the 1946 Turkish Straits Crisis, the deterrence of a Soviet attack on Berlin, and the North Korean deterrence of an American airstrike in 1994. The focus of the chapter is on deterrence, but the implication is that a minor attack can indeed result to a major war. The implication is also that a minor attack can reveal

least one party that war should take place." Howard (1984) writes that "[i]f history shows any record of 'accidental' wars, I have yet to find them," and Trachtenberg (2000) expresses a similar skepticism.

the same aggressive intentions as a major campaign, thereby forcing the target of the attack to respond with major war in self-defense. This all occurs in a model where there is no audience or third party to influence. The implication is that minor incidents may not always be simply pretexts for war, but can legitimately be the beginning of a war.

Abstaining from Preemption

In the second and third chapter, I demonstrate that minor attacks can reveal hostile intentions and spark major war, but can also mislead even a vigilant public into supporting aggression. However, even attacks that were intended to be major, war-initiating attacks may not provide grounds for necessarily blaming the attacker as an aggressor. Preemptive attacks are by definition attacks by the side that was not intending to initiate war first. According to Reiter (1995), a preemptive war is one "in which one side attacks to forestall what it sees as an impending attack on itself." It is an attempt at gaining a tactical advantage in an inevitable war, rather than the implementation of a policy of choosing war over peace.

Perhaps the most well-known contemporary example is Israel's preemptive attack against Egypt in 1967 (Oren 2002). Reiter (1995) writes that the Chinese intervention in North Korea in 1950 and the German initiation of World War I should also be considered preemptive. The Continuation War in 1941 between the Soviet Union and Finland began with a preemptive attack by Russia in the belief that they faced an imminent offensive from Finnish and German units (Vehvilainen 2002).

A major deterrent to initiating war is the fear by the potential attacker of being blamed for the war. Reiter (1995) gives a number of examples of states not launching an attack for this reason. This can serve to deter one's own government from committing aggressive acts.

On the other hand, it can prevent one's own government from taking necessary preemptive and preventive measures. In 1973, for example, the Israeli government explicitly rejected a preemptive attack so that the international community would be able to clearly identify the aggressor, resulting in a successful attack by the Egyptian military (Druckman 2010).

The effect of public opinion in deterring war initiation is widely discussed in the debate over the value of democracy in foreign policy. Proponents argue that democracy leads to better foreign policies because it tempers the more war-like qualities of government. Authors from Kant to modern-day international relations scholars have argued that having what amounts to a public veto can prevent governments from initiating wars that are in the interests of elites but not in the interests of citizens. Proponents of more centralized, less democratic foreign policy decision-making argue that democratic publics are slow to recognize threats, and that an effective foreign policy requires a strong executive with a free hand to initiate war in a timely manner if necessary.

This same fundamental tension is reflected in the debate over the war powers of Congress in the United States. The logic behind the requirement in the U.S. Constitution that only Congress can declare war is that a body representing the interests of the people must be able to decide whether or not it is necessary for a country to enter war. However, the slow erosion of this power and the concentration of power in the executive has reflected the belief that, in today's more hostile and fast-moving security environment, decisions need to be made by those with a certain level of expertise and knowledge, and need to be made with a promptness that wouldn't exist if there was a public veto.

In the fourth chapter of the dissertation, I develop a formal model to examine whether blaming the initiator can leave the public worse off. I show that the public can prevent both unnecessary wars and necessary preemptive attacks by blaming the initiator and opposing

war initiation by its own government. I also examine the implications of this finding for the public's welfare. While it is possible that strengthening democracy can constrain a government from launching unnecessary aggressive wars and improve the public's welfare, I identify two scenarios in which weakening democracy can improve the public's welfare. I show that, first, under certain conditions reducing the public's ability to impose costs on its government can free all possible types of governments to launch preemptive war when it is warranted, at the price of freeing only very few types to launch aggressive war. I also show that it is possible that weakening democracy can free the government from being subject to a norm of non-aggression, which can persist even in scenarios when it is not in the public's interest to restrain its government.

The implication is that, first, blaming the initiator can backfire on the public in certain security environments. In a larger sense, this finding also implies that democracy's effect on foreign policy is contingent on the international environment and it may not necessarily be beneficial for the public. I also conclude that studies which examine democracy and foreign policy without recognizing that the effects of democracy will be different in different security environments, and that the adoption of democracy will probably vary depending on security environment, will come to misleading conclusions.

Crisis, Democracy and Manipulation

This dissertation modifies the conventional understanding of crises in multiple ways. First, it demonstrates that crises may not always be contests of resolve, but may be contests of restraint. When provoked, the challenge for a state that wants to avoid war is to avoid responding to the provocation, rather than demonstrating its resolve or willingness to fight.

Similarly, for a state that wants war, often their incentives are to wait to be attacked rather than preempting. This forces us to view crises in a different way, and to account not only for the resolve that states are trying to demonstrate to their opponents, but for the benign intentions that they try to demonstrate to their publics and to other states.

Second, it modifies our understanding of the relationship between democracy and crisis bargaining. The main argument in the crisis bargaining literature, which follows from works on audience costs and opposition signaling, is that democratic countries are better able to signal resolve because the opposition can convey information to the adversary and the public can punish concessions by its own government. There is also a line of argument that democracies perform better in foreign policy more broadly because of their need to cater to the public's preferences.

However, we also see that democracy can promote deceptive practices by the government in order to generate the consensus it needs to pursue certain policies. In the second chapter, I show that increasing transparency can moderate these effects, showing essentially that this problem can be defused with more democracy. In the fourth chapter, I show that democracy can be welfare enhancing. But I also show that the public can prevent necessary preemptive actions and that this can be welfare-reducing in security environments where the adversary is likely to be belligerent. The conclusion here is that the value of democracy is contingent on the security environment.

Finally, this dissertation has lessons for how the public should interpret crisis events and incidents given the possibility that they are facing manipulation. Certainly, the public should be suspicious of the events which transpire during a crisis, and I have found that they usually are. It is clear from this dissertation that the side that fires first is not always the true aggressor. On the other hand, even sophisticated publics and legislators can be de-

ceived because they lack crucial information about the intentions and actions of the relevant governments. They are placed in genuine dilemmas with very high stakes, and they must make the best decisions possible given their information.

The implication is that institutional changes may provide the best protection for the public. Increasing transparency should mitigate the ability of a government to use deception to provoke incidents and mislead a legislature. Lowering the probability that a government has biased preferences should also make deception and sub-optimal policy choices less likely. On the other hand, in particular security environments where the adversary is likely to be hostile, insulating the government from public pressures may allow it to make prudent foreign policy decisions without having to take actions to build public consensus that may be damaging for the public's well-being.

Chapter 2

Provoked Incidents as Pretexts for War

Public justifications for war often rely on blaming the enemy for starting the war, and this assignment of blame frequently hinges on narratives about which side attacked first. In the United States, the surprise attack on Pearl Harbor is widely recognized as both the beginning of the war and the main reason why the United States entered the war. Disagreements about the events surrounding the first battle of the Mexican-American War played a major role in the larger debates over the wisdom and morality of the war itself, as did revelations about the Gulf of Tonkin incident during the Vietnam War (Schroeder 1973, Siff 1999).

As later revelations about the Gulf of Tonkin incident make clear, these narratives can sometimes be misleading. The government's version of events can be false, or the whole incident fabricated. Attacks can be staged, as Nazi Germany did along the German-Polish border before invading Poland (Allen 2005). Even when attacks are real, they may not reflect the desire to fight a war. They may be acts of self-defense in anticipation of aggression, like

Israel's preemptive strike before the Six Day War. They may also be isolated incidents, either acts with limited objectives or responses to provocative actions.

Not infrequently, isolated incidents are intentionally provoked by a state seeking a pretext for a war. The strategy of provoking an enemy attack to gain justification for war seems to have a prominent place in American history. Both the first battle of the Mexican-American War and the Gulf of Tonkin incident appear to have been provoked by American actions and then misrepresented to Congress and the American public as unprovoked attacks. A compelling case has also been made that the Roosevelt administration tried to provoke German naval attacks in the Atlantic Ocean and the Japanese offensive in the Pacific for the purposes of getting Congressional and public approval for the United States to enter World War II (Schuessler 2010, Trachtenberg 2006). In a more recent example, there are reports that President George Bush and Prime Minister Tony Blair discussed baiting Saddam Hussein into shooting down a U.N. plane to justify beginning the invasion of Iraq (Van Natta Jr. 2006).

The repeated use and occasional success of this strategy presents something of a puzzle. First, why would an adversary respond to a provocation if he wishes to avoid providing an excuse for war, as is usually the case when one side is trying to provoke an incident? While staged attacks or fabricated stories require just one government to deceive the public, provoking an incident requires the opponent to "participate" in the deception. Second, why would the public find the incident to be convincing evidence of enemy aggression when their own government's actions provoked the incident in the first place?

In this paper, I seek to explain how and when the strategy of provoking an incident can be effective in both eliciting an adversary's response and misleading a domestic audience. I argue two main points: first, that a public or legislature can be deceived into supporting war

when it doesn't observe or is misinformed about the original provocative acts; and second, that the opponent can be lured into responding to the provocation, even when it doesn't want war, by the hope that the public or legislature *will* learn of the original provocative act. Both of these findings follow from the provocation itself being observed by the domestic audience with positive probability but not with certainty.

I develop this argument using a multi-method approach, combining formal theory, a survey experiment and a historical case study. The paper proceeds as follows. I begin by examining the previous literature on assigning blame for war in international relations and show that the desire to shift the blame for war has an important influence on state behavior. Then, I briefly review speeches by American presidents following the provoked incidents mentioned above to try to understand the purposes for which the government has attempted to exploit these incidents.

Next, I develop a formal model demonstrating that, in equilibrium, a government can successfully provoke an incident and mislead its legislature into war when the provocation is observed probabilistically. I also demonstrate that increasing the chance that the provocation is observed can have the opposite effects of increasing the probability that the provocation fails to shift the legislature's opinion but also increase the chance that the enemy is baited into responding to a provocative act.

Following the model, I conduct a survey experiment demonstrating that public opinion responds to the outbreak of war in a manner consistent with the model's predictions. In the survey experiment, respondents were randomly assigned different stories about the events leading to military action against Iran and asked whether they approve or disapprove of the military action. I find that public support for military action increases when preceded by an Iranian attack and decreases when preceded by American provocations. This latter effect is

particularly strong when the provocations occur before the Iranian attack.

Finally, I use the insights of the model to conduct a case study of government decision-making to explain why American policies before WWII failed to provoke a major incident with Germany but succeeded with Japan. I demonstrate that the failure against Germany followed from a combination of German restraint and public revelations about U.S. provocations, while the success against Japan followed from the Japanese government's hope that the American public would come to understand the attack as provoked and defensive.

These three lines of inquiry, formal, survey and historical, produce results supporting the two main points of my argument. The results also suggest that an increase in the visibility of a government's actions, and a restriction of its ability to conduct hidden provocations, could undermine this strategy and protect the public from deception. I return to this point in the conclusion.

2.1 Aggression and Self-Defense in International Politics

Although assigning blame and distinguishing between "aggressors" and "defenders" is important in international relations, it has not been central to mainstream theories of international relations. In fact, much of international relations theory has sought to blur or complicate the distinction. Neorealist theories explain acts of aggression as largely arising from the defensive motive of survival (Mearsheimer 2001, Waltz 1979). Important concepts in the field like the security dilemma, the spiral model, and the preventive motivation for war similarly view aggression as resulting from defensive motives, or even as being indistinguishable

from defensive actions altogether (Jervis 1978, Levy 1987). Most bargaining models show no distinction between the two whatsoever, with war resulting from the failure to reach a negotiated settlement that reflects relative power, irrespective of who stands to gain or who initiates the war (Fearon 1995, Powell 2006).

Still, the assignment of blame does seem to matter in many fields relevant to international relations. In international law, the distinction between aggressors and defenders is critical, as demonstrated in the language of the U.N. Charter. In political philosophy and just war theory, the distinction is equally critical when passing moral judgements on government decisions (Walzer 1977). Assigning blame also seems to matter in the public mind, and this has been reflected in the scholarship on the relationship between foreign policy and public opinion. For example, Jentleson (1992) shows that military actions that appear to be motivated by self-defense regularly enjoy the strongest support by the American public. Stein (2000) identifies a number of common justifications for war and similarly finds that self-defense is often the most compelling.

While blame may not matter in the direct practice of international politics, it is no surprise that it can influence state behavior given the role that it plays in international law and public debates over foreign policy. For example, states anticipating attack and considering preemption or prevention may nevertheless allow themselves to be attacked in order to place blame on the enemy. Reiter (1995) demonstrated that preemptive wars are rare occurrences, mainly because of the political costs of striking first, while Schweller (1992) and Levy (2008) both argued that democracies are less likely to engage in preventive war because of the public distaste for aggression. Prominent examples of states foregoing the first strike include France withholding fire along the Western Front in 1914 in the hope that Germany would fire first, and Israel deciding not to launch a preemptive strike on the eve of

war in 1973 (Albertini 1967, Druckman 2010). On multiple occasions, the United States has told foreign governments that American support would be more forthcoming if the foreign government allowed itself to be attacked first, including in 1939 to the Polish government and in 1967 to the Israeli government.[1]

In the above examples, the decision not to fire first is taken in anticipation of the enemy firing first. In cases where the enemy's fire is not forthcoming, some governments have shown that they are not above faking incidents completely. In Operation Himmler, the Nazi government shot prisoners who had been dressed as German soldiers in a series of faked Polish attacks along the border (Allen 2005). Even the United States government considered such plans against Cuba, with the Joint Chiefs of Staff drawing up a proposal to fake incidents to justify war against Cuba in 1962, and Robert Kennedy musing aloud during the Cuban Missile Crisis whether the U.S. could "sink the *Maine* again or something" in order to justify military action.[2]

The strategy examined in this paper is, in some ways, more puzzling than the above strategies. While allowing an aggressive enemy to fire first simply clarifies which side is the aggressor, and faking an enemy attack just requires one government to attempt to deceive its public, the strategy of provoking an attack requires the enemy to cooperate by responding to the provocation even as it seeks to avoid providing an excuse for war. The rest of the paper addresses this puzzle.

[1]See Dallek (1979, 197) and Oren (2002, 147). The literature on rally effects and diversionary war is also relevant here, though its focus on presidential approval and domestic politics limits its applicability to the question of generating public support for foreign policies. I return to this literature in the survey experiment section, which more directly measures a phenomenon akin to rally effects.

[2]See Bamford (2001, 82-91) and May & Zelikow (1997).

2.2 Provoking Attacks

In order to begin addressing this question, I first seek to understand specifically what beliefs governments want their publics and legislatures to adopt following provoked incidents. This will guide the analysis and inform the structure of the model I present in the next section. To accomplish this, I conduct a brief study into speeches made by American presidents following the provoked incidents discussed in the introduction. It is instructive to examine these speeches to understand what the government wanted the public to believe following these incidents, particularly given that we now know them to have been misleading.

I identified two common themes in these speeches. First, the president always portrays the attack as unprovoked, when the United States had in fact taken provocative actions. In his address to Congress asking for a declaration of war against Mexico, President Polk told Congress that American troops were sent to the Rio Grande River "to provide for the defense of that portion of our country" and were given "positive instructions to abstain from all aggressive acts toward Mexico or Mexican citizens" (Butler 1995, 67-71). Furthermore, in the ongoing political dispute with Mexico, he said the United States had carefully avoided "every expression that could tend to inflame the people of Mexico or defeat or delay a pacific result" and claimed that "we have tried every effort at reconciliation" (Butler 1995, 67-71). Polk kept hidden from his speech that the troops were in disputed territory, and that negotiations with Mexico had collapsed over the United States' demand for the cession of California (Schroeder 1973).

Roosevelt similarly portrayed the United States as a victim of unprovoked aggression, both in the U.S.S. Greer incident in the Atlantic and following Pearl Harbor. Speaking of the Greer incident, Roosevelt said, "I tell you the blunt fact that the German submarine fired first

upon this American destroyer without warning, and with deliberate design to sink her," even though he must have been aware that this was not true (Bailey & Ryan 1979). In his famous Day of Infamy speech, Roosevelt focused not so much on the unprovoked nature of the attack in which the United States was "suddenly and deliberately attacked by naval and air forces of the Empire of Japan," but on the political context (U.S. Senate 1941). He claimed that Japan "deliberately sought to deceive the United States by false statements and expressions of hope for continued peace" while the United States was "still in conversation with its government and its emperor looking toward the maintenance of peace in the Pacific" (U.S. Senate 1941). This was despite the fact that, in the weeks before Pearl Harbor, Roosevelt was anticipating a Japanese attack and believed that the issue was "how we should maneuver them [the Japanese] into the position of firing the first shot without allowing too much danger to ourselves (Dallek 1979, 307).

Johnson also accused North Vietnam of unprovoked aggression, calling the attacks in the Gulf of Tonkin "open aggression" and "deliberate attacks."[3] Johnson emphasized American innocence by claiming American ships to have been on the "high seas" and that they were "operating in international waters."[4] He didn't reveal the mission of the U.S. ships in the Gulf of Tonkin or the possibility that the attack was retaliation against U.S.-supported raids against North Vietnam (Moise 1996).

Second, the president always portrays the attack as the beginning of a major new policy by the enemy, rather than an isolated incident. Polk declared that Mexico had attacked with the purpose of initiating war, claiming that "Mexico has passed the boundary of the United States, has invaded our territory and shed American blood upon the American soil.

[3]See Siff (1999, 113) and U.S. Senate (1967, 120-122).
[4]See Siff (1999, 113) and U.S. Senate (1967, 120-122).

She has proclaimed that hostilities have commenced, and that the two nations are now at war" (Butler 1995, 67-71). Roosevelt explicitly made the case with regard to the attack on the U.S.S. Greer that "the incident is not isolated, but is part of a general plan" to "abolish the freedom of the seas, and to acquire absolute control and domination of these seas for themselves" as a step toward "domination of the United States [and] domination of the Western Hemisphere by force of arms" (Buhite & Levy 1992). Following Pearl Harbor, Roosevelt claimed that "our people, our territory, and our interests are in grave danger," as a result of the attack, calling the Japanese offensive a "premeditated invasion" (U.S. Senate 1941). This is despite the fact that Japan had almost no capability to actually threaten the United States homeland. Johnson called the Gulf of Tonkin attack "a new and grave turn" to the situation in Southeast Asia and repeatedly emphasized that the attack came in the context of broader Communist aggression in the region (Siff 1999).

What does this brief analysis tell us? In each case, the president was exaggerating a provoked and isolated incident, portraying it as an unprovoked and broad-based attack. The President clearly wanted to make use of some sort of event as a pretext for war, and the incidents were ambiguous enough that he was able to misrepresent them. The legislature, then, faced the dilemma of accepting or rejecting the President's interpretation of events. It is this dilemma that informs the model in the next section.

2.3 Model: Provocations, Incidents and War

In this section, I set up a simple model that captures the scenario described above. In the model, a legislature must decide whether or not to authorize war following an enemy attack, but may not be able to distinguish between an unprovoked, broad-based attack or

a provoked, isolated incident. An executive and an adversary must decide whether or not to provoke or attack, respectively, while anticipating how their decision will influence the legislature's decision. I use the model to understand the conditions under which an adversary will be baited by a provocation and a legislature will be mislead into authorizing war, focusing specifically on the key role played by the visibility of the provocation.

The model is a sequential game with three players, an executive (E), a legislature (L) and an adversary (A), with the executive and legislature both imagined as belonging to the same government. To start the game, nature assigns a type $w_A \sim f[0,1]$ to the adversary, with w_A being the adversary's payoff for war. The executive learns this type but the legislature does not, reflecting the legislature's lack of access to relevant information and intelligence.

The executive moves first by deciding to provoke or not provoke, $x_1 \in \{p, np\}$. The provocation can be thought of as any action that can impose a cost on the adversary and that the adversary can reverse with military action. This may include attacks against civilian or military targets, the violation of a country's airspace or territorial waters, or a blockade or embargo. The executive's move is observed by the adversary but not the legislature. The legislature is therefore uncertain both of the adversary's type and the executive's action, which may itself be conditioned on the adversary's type.

Following the executive's action, the adversary must choose to attack or not attack, $x_2 \in \{a, na\}$.[5] If the adversary does not attack, the game ends in the *status quo*, with the adversary suffering a cost if it was provoked. If the adversary does attack, the legislature then has the option of authorizing or not authorizing war in response, $x_3 \in \{w, nw\}$.[6] If

[5] I refer to an attack following a provocation as a *provoked attack*, and an attack following no provocation as an *unprovoked attack*.

[6] I omit the decision to authorize war or not authorize war following the adversary's decision not to attack since the legislature would always prefer peace following a decision by the adversary to not attack.

the legislature does not authorize war following a provoked attack, the game also ends in the *status quo* but without the adversary suffering a cost. This reflects the assumption that a provoked attack simply reverses the provocation without threatening further harm to the executive or legislature. However, if the legislature does not authorize war following an unprovoked attack, the game ends in *retreat*.

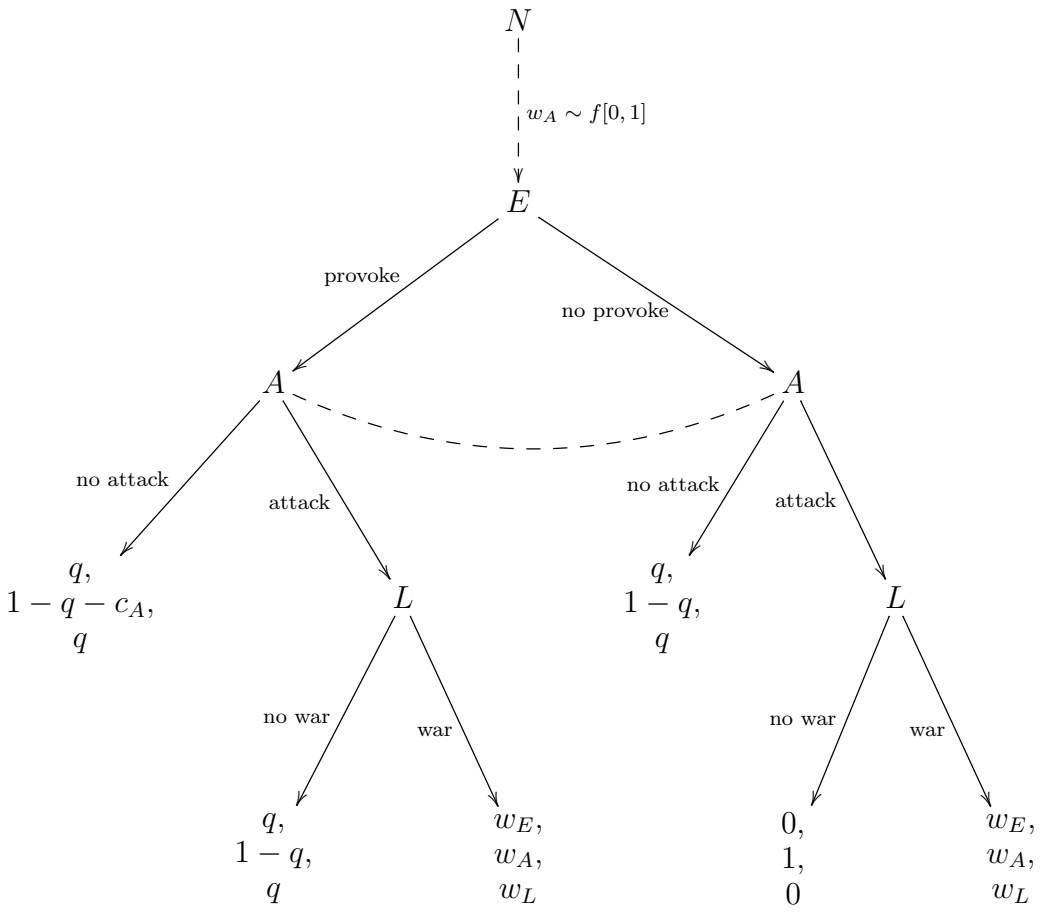
As mentioned above, the legislature does not observe the adversary's type or executive's action, so it may be uncertain about whether an attack was provoked or unprovoked. However, before the legislature moves, it receives a private signal $X \in \{p, np\}$ providing some information about the executive's action. The signal works as follows:

- If the executive chooses to not provoke ($x_1 = np$), the legislature receives a signal $X = np$.
- If the executive chooses to provoke ($x_1 = p$), the legislature receives a signal $X = p$ with probability π and a signal $X = np$ with probability $1 - \pi$.

Therefore, if a provocation occurs, it may be revealed to the legislature by the signal $X = p$, or it may remain indistinguishable from the scenario where a provocation doesn't occur.

The payoffs for the different outcomes are the following. The status quo payoffs are q for the executive and legislature and $1 - q$ for the adversary, where $q \in \{0, 1\}$. The adversary also will suffer a cost of $-c_A$ if he was provoked. If the legislature authorizes war, the payoffs for each player are w_i, where i indexes each player $i \in \{E, A, L\}$. The payoffs for retreat are 0 for the executive and legislature and 1 for the adversary. The executive is commonly known to prefer war to the status quo, $w_E > q$, while the legislature is commonly known to prefer the status quo to war, $w_L > q$. This creates the scenario of a war-seeking executive and a

Figure 2.1: Provocation Model Game Tree



defense-minded legislature that motivates the use of the strategy of provoking an incident. As explained above, the adversary's war payoff is between 0 and 1, so he may or may not prefer war to the status quo, but will always prefer retreat to both.

The game tree is in Figure 2.1.

2.3.1 Equilibrium Analysis

I begin the analysis by presenting the equilibrium in which the strategy of provoking an incident can be successful. The equilibrium concept is Perfect Bayesian Equilibrium and the

following equilibrium is in pure strategies.[7]

Proposition 1 *There exists an equilibrium in which the executive provokes some types of the adversary, the adversary responds by attacking, and the legislature authorizes war if it observes the signal $X = np$ when:*

$$w_L \geq (1 - \hat{\alpha}_{np})q \tag{2.1}$$

where $\hat{\alpha}_{np}$ is the legislature's estimate that an attack is unprovoked when it receives the signal $X = np$.

To understand the mechanism operating in this equilibrium, I begin by analyzing the legislature's decision to authorize or not authorize war following an attack. Define $\hat{\alpha}_p$ as the legislature's estimate than an attack is unprovoked when it receives the signal $X = p$, and $\hat{\alpha}_{np}$ as the legislature's estimate than an attack is provoked when it receives the signal $X = np$. $\hat{\alpha}_p = 0$ always holds because the legislature will only receive a signal $X = p$ if a provocation occurred. Since the legislature prefers the status quo to war, it will never authorize war in that scenario.

If the legislature receives the signal $X = np$, however, uncertainty may still remain as to whether a provocation occurred or not. The legislature then faces the dilemma of whether to authorize what may be an unnecessary war, or to not authorize war and be forced into retreat. The legislature will err on the side of authorizing war if equation (2.1) holds.

Assume for now that equation (2.1) holds. How will the adversary and executive behave? The executive knows that it can guarantee war by not provoking the types of adversaries

[7]The model has another pure strategy equilibrium in which the legislature never authorizes war and the adversary therefore always provokes to avoid the possibility of retreat. This is described in the appendix.

that will launch an unprovoked attack, $w_A \geq 1 - q$, because the legislature will receive a signal of $X = np$ with certainty and will authorize war. Since certain war is the executive's most preferred outcome, not provoking is a weakly dominant strategy against these types.

What about adversaries that will not attack unprovoked? Against these adversaries the executive may face the incentive to provoke an incident. In fact, the executive will prefer to provoke any adversary of type $w_A < 1 - q$ who will respond to a provocation. The executive knows that not provoking guarantees the status quo, while provoking will lead to war *if* the legislature receives the signal $X = np$, which occurs with probability $1 - \pi$.

Are there any adversaries that will respond to a provocation? Following a provocation, the adversary faces its own dilemma. If it does not attack, it receives $1 - q - c_A$, the status quo payoff minus the cost of the provocation. If it does attack and the legislature receives a signal $X = p$, the legislature will not authorize war and the game will end in the status quo, with a payoff of $1 - q$ for the adversary. The attack will have successfully reversed the provocation. However, if it launches a provoked attack and the legislature receives a signal $X = np$, the legislature will authorize war and the game will end in war with a payoff w_A. Since the probability that the legislature receives the correct signal is π, the adversary will launch a provoked attack if

$$w_A \geq 1 - q - \frac{c_A}{1 - \pi}. \tag{2.2}$$

The executive will therefore provoke all types of adversary whose type is $1 - q > w_A \geq 1 - q - \frac{c_A}{1-\pi}$, and those adversaries will respond. Finally, against types $w_A < 1 - q - \frac{c_A}{1-\pi}$, the game will end in the status quo no matter what, so the executive is indifferent between provoking and not provoking.

The final step is to confirm that equation (2.1) holds given these strategies so that the legislature won't deviate. It is enough to recognize that, with these strategies, the legislature may observe $X = np$ following both a provoked and unprovoked attack. Therefore, $1 > \hat{\alpha}_{np} > 0$, and there exist values of $w_L \in [0, 1]$ and $q \in [0, 1]$ such that equation (2.1) will hold. See the appendix for the full characterization of equilibrium and description of the value of $\hat{\alpha}_{np}$.

2.3.2 Misleading the Legislature and Baiting the Adversary

With this equilibrium, I can answer the questions asked in the introduction about how a legislature or public could be deceived into supporting a war and why an adversary would respond to a provocation even when it seeks to avoid war.

For the legislature, the fact that a provocation is not observed creates uncertainty whether an attack represents a true threat or is just a response to a provocation. When $\pi < 1$, provoked attacks may be observed as unprovoked attacks. Since the legislature is not always able to distinguish between a provoked and an unprovoked attack, and the unprovoked attack carries the possibility of retreat, the legislature may authorize war. The legislature understands that a provocation may have occurred, but the danger of being forced into retreat outweighs the risk of authorizing an unnecessary war.

For the adversary, the fact that the provocation might be observed can serve to lure it into responding to the provocation in the hope that the game won't end in war. This mechanism is not necessary to provoke adversary types of $w_A \geq 1 - q - c_A$. When $\pi = 0$, and the provocation will never be observed, these types will attack. The adversary knows in this case that any attack, even if provoked, will lead to war with certainty. However, it

attacks because it prefers war to tolerating a provocation.

When $\pi > 0$, adversaries of type $1 - q - c_A > w_A \geq 1 - q - \frac{c_A}{1-\pi}$ will also respond to the provocation, as the possibility of the provocation being observed lures them to respond. These are adversaries who prefer to avoid war even at the cost of tolerating the executive's provocation. However, they are lured into responding in the hope that the provocation will be observed and the legislature will not authorize war. This explains why even those state desperate to avoid war and aware that they are being provoked may still take the bait of responding to the provocation.

2.3.3 The Visibility of the Provocation and the Probability of Deception

In this section, I derive some comparative statics on the probability that the legislature will be mislead into authorizing an unnecessary war by focusing on the effect of varying the visibility and the cost of the provocation. This will help to identify some measures that may help legislatures and the public be less susceptible to the strategy of deception examined in this paper.

For this part of the analysis, I make two additional assumptions. First, I assume that the adversary's type is uniformly distributed, $w_A \sim U[0, 1]$. Second, I assume that $w_L \geq q(1-q)$, which ensures that the equilibrium will exist for all values of $\pi \in [0, 1]$.

Since the adversary is provoked if and only if it is *provocable*, of type $1 - q > w_A \geq 1 - q - \frac{c_A}{1-\pi}$, the probability that a provocation occurs is $\frac{c_A}{1-\pi}$ when $\pi < \frac{1-q-c_A}{1-q}$ and $1 - q$ when $\pi \geq \frac{1-q-c_A}{1-q}$. Multiplying this by the probability that the provocation is not seen, $1-\pi$, gives the following probabilities that the legislature will authorize a war it wishes to avoid,

defined here as ϕ:

$$\phi = c_A \tag{2.3}$$

when $\pi < \frac{1-q-c_A}{1-q}$, and

$$\phi = (1-q)(1-\pi) \tag{2.4}$$

when $\pi \geq \frac{1-q-c_A}{1-q}$.

One immediate implication of the above equations is that the higher the cost of suffering a provocation, the higher the probability that an adversary is provoked and a legislature mislead into authorizing a war it would prefer to avoid. When the cost of the provocation is higher, the adversary has a greater incentive to respond in the hope of reversing the provocation.[8] One implication is that restricting the executive's freedom of action to conduct operations, particularly those that can be kept out of the view of the public, will reduce its ability to successfully provoke an incident that could lead to war.

Now I turn to the effect of the provocation's visibility. Notice that the probability of deception is constant when π is low. This is because, under a uniform distribution, the increase in the probability that the legislature observes the provocation is offset by the increase in the proportion of adversaries that will respond to a provocation. However, once the visibility is high enough that all the adversary types are provocable, an increase in visibility only has the effect of reducing the probability that the legislature doesn't observe a provocation. This gives the following condition.

Proposition 2 *The probability that the legislature authorizes war after a provoked incident,*

[8]See appendix for proof.

ϕ, is weakly decreasing in the probability that the provocation is observed, π.

See appendix for proof.

What is the implication of this finding? It implies that decreasing the ability of a government to take hidden actions against the enemy will decrease the probability that a legislature or public will be mislead into authorizing war. This may indicate that more restrictions on executive power or institutional measures to increase transparency would decrease the probability that the government could provoke an incident as a means of misleading the public. However, as demonstrated, the increasing of transparency is not without its offsetting effects, specifically that it will lure more types of adversaries into responding to provocations rather than exercising caution and avoiding incidents that could be used as a pretext for war.

2.3.4 Extending the Model to Preemption

In some cases, provocative actions are not simply measures that impose a cost on the adversary, but are militarily threatening actions and precursors to attack. In this section I modify the model slightly to demonstrate that the same mechanism described above can provoke war when responding to a provocation is meant to preempt the beginning of a war.

Imagine that, following an adversary's decision to not respond to a provocative act, the game ends in war rather than peace. This could be the case if the provocation itself is a military preparation and the government is able to initiate war without the legislature's approval, or if the legislature will authorize war when military forces are already mobilized and ready to strike. I assign the payoffs for this outcome as the war payoffs w_i, plus the payoff k for the executive and legislature and minus the payoff k for the adversary. This extra payoff parameter k may represent the advantage in war that the executive and legislature

gain from striking first following a successful provocative action. It may also be negative, however, perhaps representing a political cost suffered by the executive and legislature from initiating the war. I examine both cases.

The following result follows from this modification.

Proposition 3 *A Perfect Bayesian equilibrium exists in which the executive will provoke some types of the adversary, and some types of the adversary will respond, when* $w_L \geq (1-\hat{\alpha})q$*, where*

$$\hat{\alpha} = \frac{P(1-q+\frac{k}{\pi} > w_A \geq 1-q)}{P(1-q+\frac{k}{\pi} > w_A \geq 1-q) + P(w_A < 1-q)(1-\pi)} \tag{2.5}$$

when $k > 0$*, and*

$$\hat{\alpha} = \frac{P(w_A \geq 1-q)}{P(w_A \geq 1-q) + P(w_A < 1-q-\frac{k}{\pi})} \tag{2.6}$$

when $k < 0$*.*

See appendix for full characterization. In these equilibria, the legislature will not authorize war after a provoked attack for the same reason as in the previous model, that it prefers to tolerate the provoked attack and return to the status quo. Therefore, the adversary that wants to avoid war may respond to a provocation in the hope that he preempts the executive's attempt to start a war and throws the decision into the legislature's hands.

In the scenario where $k < 0$, this may occur even though the adversary prefers that, if war is to occur, it not attack first. An adversary may launch a preemptive attack and risk being blamed for starting a war in the hope that the preemptive attack could avert war altogether by leading an informed legislature to decide against war.[9]

[9]Since the payoff for war after the preemptive attack is lower, it is possible, in theory, that a legislature

The implication here is that incidents can be provoked both by forcing an opponent into a situation where he has to fight back to reverse some cost or by forcing an opponent into a situation where he thinks a successful preemptive strike could avert war. Both of these can occur even if the adversary seeks to avoid war, or to avoid being blamed for starting a war, because a revelation that the adversary was provoked into its actions would spare the adversary from these outcomes.[10]

2.4 Survey Experiment: War with Iran

The above model demonstrated that an executive could successfully provoke an incident that leads to war if the legislature doesn't observe the provocation. The model also makes a number of predictions about the actions of the legislature following different histories of the game.

To test whether the behaviors predicted by this model hold true in the real world, I conducted an experiment embedded in a public opinion survey of a sample of the U.S. adult population. While it is not feasible to survey legislators, it is feasible and reasonable to test whether the model predicts the beliefs and actions of the general public. There are good reasons to believe that the public's beliefs are relevant to understanding the mechanism identified in the model. Often legislators are reflecting public beliefs rather than their own in their decision to act or not act as a veto player over a government decision to go to war. In fact, in attempting to provoke enemy attacks, the executive's goal is often to influence

would authorize a war that its own government started but would not authorize a war if the adversary had successfully attacked first.

[10]For example, the Japanese government may have believed that a successful attack on Pearl Harbor would have the latter effect. It would so increase the cost of a war for the United States that the American public would want to abandon the war, assuming the American public believed the attack to have been defensive rather than aggressive.

legislative opinion indirectly by influencing public opinion. In addition, the public can directly act as a veto player through voting, protesting, and refusing to serve in the military. In that case, the same mechanism identified in the model operates, with the public acting as the veto player rather than the legislature.

In the survey experiment, I presented respondents with a story about a series of events leading the President of the United States to take military action against nuclear and military targets inside Iran and asked whether they approved of the President's actions. To test the effect of different components of the story on their approval for military action, I independently and randomly varied 2 factors: 1) whether or not the crisis began with an Iranian attack on a U.S. ship in the Persian Gulf, and 2) whether major news outlets confirmed or did not confirm Iranian claims that the U.S. had been launching missile strikes inside Iran before the crisis began.[11]

I used a survey experiment in order to isolate the effects of different events on public opinion in a controlled manner. A number of studies based on observed public opinion have examined the effects of military incidents on presidential approval and support for war in the literature on the "rally around the flag" effect (Brody & Shapiro 1991, Parker 1995). Some have even compared rally effects across different rally events to isolate and identify the components that lead to greater or smaller rallies (Baker & ONeal 2001). However, these studies still suffer from a lack of experimental control and from a limited number of relevant rally events to compare. The survey experiment allows me to compare different events by controlling the factors that vary between the events. This method has been used in the international relations literature to test the existence of audience costs and mechanisms

[11]In terms of the model, the first factor represents whether the adversary committed an attack, while the second factor represents whether the public observed the executive committing a provocation.

generating the democratic peace (Tomz 2007, Trager & Vavreck 2011, Tomz & Weeks 2013). In a sense, this project could be considered the first experimental examination of the rally effect, even though the focus here is not on presidential approval but approval for military action. In addition to these considerations, the survey also allowed me to measure opinion on an issue important to contemporary political debates, with results that are of inherent interest to followers of U.S. foreign policy.

As I present in detail below, the results from the survey are consistent with the behaviors one would expect from the model. First, approval for military action increases significantly following an Iranian attack. Second, approval for military action decreases if news outlets confirm that the United States attacked first, but this effect is only statistically significant if Iran also attacked. This is consistent with the public interpreting the Iranian attack as a response to the U.S. attack and therefore not approving major military action in response, similar to the scenario in the model where the legislature learns than an attack is provoked.

2.4.1 Hypotheses

The model makes predictions about the legislature's behavior following different histories of the game and different signals the legislature can receive about its own government's actions. These predictions are not only dependent on the legislature's beliefs, but on the legislature's preferences for different outcomes. In the game, the legislature is assumed to prefer the status quo to war and war to retreat. In reality, a given individual may prefer war even if the alternative is maintaining the status quo, or an individual may prefer to avoid war even at the cost of retreat. In addition, the possible outcomes if war is not authorized may not be the discrete outcomes described in the model, but a continuum of possible outcomes in

which the cost of the outcome depends on the individual's estimate of the belligerence of the adversary. I attempt to ensure that the hypotheses I am testing are sensitive to these considerations.

First, the model predicts that the public should be more likely to support military action if the enemy government attacked. While the public's approval of military action in the model should only follow from an unprovoked attack, the public should have a higher estimate of the adversary's belligerence if the adversary attacks regardless of whether or not a provocation occurred. A provoked attack may not indicate the same level of belligerence as an unprovoked attack and may simply represent an isolated incident rather than a new policy of aggression, but the public should still increase its estimate of the enemy's belligerence following an attack under any circumstances. Therefore, the first prediction is the following:

> ***H1****: Public approval for military action should be higher in the treatment group in which Iran attacks the United States regardless of U.S. action.*

Second, as indicated, the model predicts that an attack following no provocation should indicate a higher level of belligerence and a higher level of danger to the public than an attack following a provocation, leading to a higher level of approval for military action. The public should also have a higher estimate of the enemy's belligerence if there is no attack, if no provocation occurs. This is because an adversary that peacefully tolerates a provocation should be viewed as particularly non-belligerent. Nevertheless, the effect of observing an American provocation should be stronger if Iran does launch an attack than if Iran does not launch an attack. In the case where Iran does attack, the provocation could be the reason, providing an explanation for the Iranian action other than aggressive intentions. This should more strongly undermine the case for war or the perception of enemy aggression. Therefore,

the second prediction is the following:

> ***H2****: Public approval for military action should be lower if the public receives a signal that the United States provoked Iran, but the effect of this should be stronger when comparing the cases where Iran attacked the United States than when comparing the cases where Iran didn't attack.*

2.4.2 Research Design

To test these predictions about public approval for military action, I conducted an online randomized survey experiment on a sample of adult U.S. citizens.[12] In the survey, respondents were asked to read a hypothetical situation describing a foreign policy crisis and then asked whether they approved or disapproved of the actions of the President. The story described a series of events in which:

1. negotiations between the United States and Iran over dismantling its nuclear program break down

2. Iran accuses the United States of having launched missile strikes within their country

3. the President of the United States announces that the only option left to deny the Iranian government nuclear weapons is military action

4. the President authorizes air strikes against nuclear and military targets in Iran and the deployment of troops to the Persian Gulf.

[12]This survey represented a convenience sample in order to test the propositions of the paper and was not representative.

Every survey contained these elements, but respondents were randomly assigned to one of six surveys in which other elements of the story were added or varied in order to test the hypotheses described above. Respondents were randomly assigned with equal probability to a survey in which they were told that negotiations broke down "[f]ollowing an incident in which Iran torpedoed an American ship in the Persian Gulf resulting in 11 American lives lost" or to a survey in which no incident was mentioned. Respondents were also randomly assigned to a survey in which "major newspapers and network news broadcasts" reported that the Iranian claims about U.S. missile strikes were either verified, dismissed as "totally unfounded," or neither verified nor disproven. Respondents were randomly and independently assigned to each group, creating a fully-crossed 2x3 survey design.

The scenario of U.S. military action against Iran was chosen for two reasons. First, it represents a possible real-world scenario that respondents would have some familiarity with and emotional investment in. Therefore, the answers that respondents provided are more likely to be valid representations of their true opinions that can be generalized to real-world behavior. In addition, the likelihood that such a survey would be deployed for the purposes of understanding the public's opinion on the issue of war with Iran serves to better mask the primary purpose of the experiment. Second, the responses to the survey do have some inherent interest to the reader beyond the role in testing the theory presented here. Since military action against Iran remains a possibility in the near future, I believe that some readers may be interested not only in the public's opinion on the subject, but on the sensitivity of public opinion to the different scenarios that could lead to war.

In terms of the model, the incident represented the decision by the adversary to either attack or not attack, and the reports about the Iranian claims represented the signal received by the public about its own government's actions. It is worth noting that the treatment in

which the Iranian claims were completely dismissed have no counterpart in the model, since in the model the public is never able to be certain that a provocation *didn't* occur. This was added to the survey despite its lack of correspondence with the model. I will show, however, that responses did not differ between the treatment in which the news dismisses the claim and the treatment in which the news could neither verify nor disprove the claim. This indicates that the respondents dismissed the idea that the news media could be confident that a provocation didn't occur and suggests that the model's specifications are correct. A complete sample of the survey text is provided in the appendix.

To complete the survey, I recruited 1,049 U.S. citizens using Amazon's Mechanical Turk website from August 7-8, 2014. Respondents were directed from the website to the online survey. As an internet sample, it suffered from some issues of being unrepresentative of the general population. In that sense, it shared problems with other convenience sampling methods such as using student samples. In this case the sample was more male and younger than the national average. Only 33% of respondents were female compared to 51% nationally and the median age was 27 compared to a national median of 37. The sample was also not politically representative, with almost half of the sample voting for Barack Obama in the 2012 presidential election, and less than a quarter voting for Mitt Romney. I address these issues in Section B of the appendix, but they don't effect the basic findings of the study.

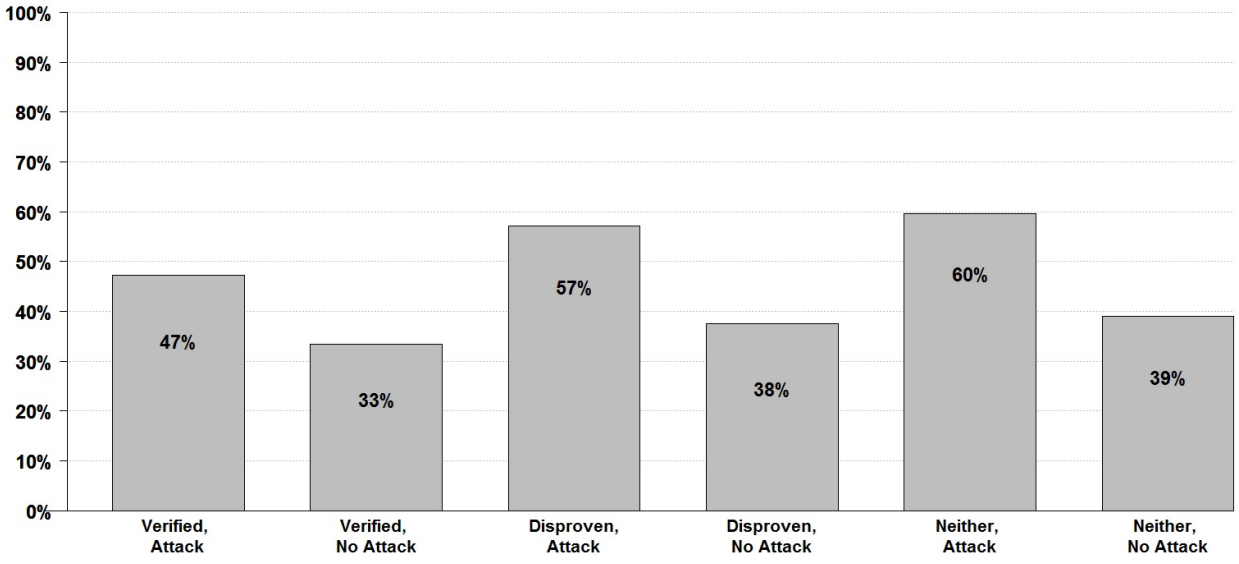
2.4.3 Results

Across all treatments, public approval for military action against Iran increased from 36.3% when Iran did not attack to 55.2% when Iran did attack. This is an increase of 18.9% caused by this treatment. To test whether the difference is significant, I used a Welch's

two-sample t-test comparing the treatment groups. The results show that the difference in means between the groups is statistically significant at the .05 level, with the 95% confidence interval ranging from 12.9% to 24.8%. An Iranian attack has a statistically significant effect of a sizeable magnitude.

To test how the signal the public receives about its government's actions affects approval for military action, I divided the sample into the six treatment groups and found the approval rate for each group. These results are displayed in the histogram below. The y-axis is the approval rate for each treatment groups. The six groups are divided by whether the news verified, disproved or neither verified nor disproved the Iranian claims about U.S. missile strikes and whether Iran attacked or didn't attack a U.S. ship.

Figure 2.2: Approval of Military Action by Treatment Group



The first thing to notice in this histogram is the effect of an Iranian attack on approval. As mentioned above, the effect is strong. In this histogram it is clear that it is strong across

all treatment groups. However, it does seem to be weaker in the treatment where the U.S. attack was verified than in the other groups. When the U.S. attack is verified, an Iranian attack only increases approval for military action by 14%, while when the U.S. attack is not verified, an Iranian attack increases approval for military action by about 20%. This indicates that the fact that the United States attacked Iran lead at least some people who would have otherwise approved military action to disapprove.

Another way to look at this is to compare the means across the different news media treatments while holding the Iranian attack treatment constant. When Iran attacks, approval for military action is 47.2% if the U.S. was verified to have struck first but 58.4% if the U.S. was not verified to have struck first. Therefore, the report that the U.S. attacked first seems to reduce the effect of the Iranian attack, perhaps because some people came to see the Iranian attack as provoked. However, when Iran doesn't attack, 38.3% approve military action if U.S. attacks are not verfied while 33.5% approve if U.S. attacks are verified. The report of U.S. attacks still reduces approval but less than in the case where it interacts with the Iranian attack treatment.

To confirm that this observation is significant, I conducted a Welch's two-sample t-test comparing the means of the "verified" and the combined "not verfied" treatments when Iran attacks, and then the means of the "verified" and the combined "not verfied" treatments when Iran doesn't attack. When Iran strikes the United States, the difference in means between the treatments where the U.S. is verified to have attacked and the treatment where the U.S. is not verified to have attacked is significant at the .05 level. The *p-value* is .024 and the 95% confidence interval is from 1.5 to 20.9%. However, the difference in means between the treatments were the U.S. is verified to have attacked and the treatment where the U.S. is not verified to have attacked is not significant at the .05 level when Iran doesn't strike

the United States. In that case the p-value is .249 and the 95% confidence interval is from -3% to 13%. While the imprecise results make it difficult to make strong statements about the magnitude or substantive significance of this effect, the results do show that reports of a U.S. provocation have a statistically significant effect only when Iran attacks. This supports the proposition that respondents are likely to change their interpretation of Iranian actions when informed that the U.S. acted first.

One final point to notice by observing the histogram is that there is very little, if any, difference between the treatments where the news media reports that it can neither verify nor disprove Iranian claims and where it reports that it has dismissed Iranian claims as unfounded. This implies that respondents made little or no distinction between these outcomes.[13]

2.4.4 Discussion

The results confirm hypotheses ***H1*** and ***H2***. Regarding ***H1***, the results clearly show that an Iranian attack increased support for military action against Iran across all treatment groups. Many respondents explicitly mentioned the Iranian attack in their explanations for their support for military action. Multiple respondents described the Iranian attack as "act of war," with one writing that the attack demonstrated that efforts "to resolve the matter peacefully" had clearly failed.

Regarding ***H2***, the U.S. provocation reduced support for military action in the case where Iran attacked the United States, indicating that many respondents reinterpreted the significance of the Iranian attack in light of the U.S. provocation. Over 20 respondents in

[13]For balance tests, parametric regression and estimation of first differences, see section B of the appendix. These tests showed largely the same results here, and also highlighted demographic variables of interest.

that treatment mentioned that the United States seemed to have provoked the Iranian attack through launching missile strikes. For example, one respondent wrote

> "In this story, it sounds like Iran had the same amount of 'reason' to attack as the US. It seems that the US started the fight by firing missile attacks and that Iran is responding in a similar way that the US would respond if they had the same thing happen to them."

These two findings are consistent with the prediction of the model and lend credence to the argument I make about the significance of military incidents and the visibility of the provocation.

One interesting result from the survey that I mentioned above is that respondents didn't make a distinction between the treatment where the news is unable to confirm or disconfirm the U.S. strikes and the treatment where the news dismisses the possibility of U.S. strikes completely. The latter treatment was included in the survey even though it didn't correspond with the components of the model. In the model, there is no signal that leaves the public certain that no provocation occurred, just a signal in which a provocation wasn't observed, leaving uncertainty as to whether it occurred or not. The fact that respondents ignored the possibility that the news could confirm that no provocation took place seems to validate the structure of the model. In fact, a number respondents wrote explicitly that they didn't believe that news report.

The survey also revealed some interesting findings that suggest that the model didn't capture all the relevant elements of this problem, suggesting routes for future research. First, even in the treatment where the United States provoked Iran, an Iranian attack increased support for war and many people still blamed Iran for starting the conflict in their expla-

nations. While the model predicts that the public's evaluation of Iran's belligerence will increase even if Iran is responding to a provocation, it does not predict that the public will necessarily increase their approval of military action, because the Iranian attack should be seen as a response to a provocation rather than an act of aggression. This indicates that the Iranian attack has some impact on public approval for military action independent of the mechanism identified in this paper.

Second, judging by the responses, a more severe Iranian attack may have increased support for military action even more. A number of respondents wrote that the 11 Americans killed did not justify large-scale military action. This leads me to believe that some people would support military action were the attack more severe.

Third, the definition of a "provocation" may be more flexible than what was allowed in the survey. For the survey, the provocation was U.S. missile strikes in Iran. However, many respondents who rejected military action wrote that they viewed the American presence in the Persian Gulf or the American attempts to coerce Iran to abandon its nuclear weapons program as provocative in and of themselves. They rejected military action based on the belief that we provoked the crisis by unduly interfering in another country's affairs. The particular provocation modeled here, of a naval attack, may only appeal to a certain segment of the population as a meaningful act, while the idea of some sort of provocation may have resonance across the population.

My final point in this discussion is that the survey results carry some inherent interest to the reader beyond their demonstration of theoretical principles. That was one of the reasons why I conducted the survey about a contemporary issue. While the sample isn't representative enough for it to accurately reflect public sentiment on this issue, the fact that an Iranian attack strongly influence public attitudes in favor of an attack is an important

thing to note. In this survey, approval for an attack increased by 20% and went above the 50% approval mark. Even in a sample that is, in all likelihood, less supportive of military action than the general population, a majority approve of military action after an Iranian attack.[14] This implies that the Iranian government, if it wants to avoid military action by the United States, should tread very cautiously in any potential military encounter. It also implies that the U.S. government, if it wants to take military action, will face a strong incentive to provoke an incident, and that any incident that does occur should be approached with a healthy dose of skepticism.

2.5 Case Study: U.S. Entry into WWII

One of the major examples of the United States government attempting to provoke incidents in order to have a pretext for war is American policy toward Germany and Japan prior to World War II. Trachtenberg (2006), Schuessler (2010) and others have made a compelling case that Roosevelt fought an essentially undeclared naval war against Germany and imposed an oil embargo on Japan in an attempt to provoke them into attacking the United States.[15] This case provides an opportunity to examine decision-making in the face of attempted provocations, helping us to understand why states allow themselves to be baited into incidents. In this section, I investigate why the United States failed to provoke war with

[14]Given that the sample for this survey was younger and more liberal than a representative survey, and that Pew Research Center surveys show 58% of the public willing to use military force to prevent Iran from developing nuclear weapons while only about 35% were willing in this survey with no Iranian attack, it is likely that the sample population from this survey was less hawkish than the American public (Pew Research Center 2012). This is confirmed in appendix B, where I show that ideology has a strong effect on approval for military action.

[15]In fact, this claim has provoked a recent debate in the international relations literature. See Reiter & Schuessler (2010), Reiter (2012) and Trachtenberg (2013).

Germany but succeeded with Japan.[16] I argue that, consistent with the theory in the rest of this paper, the visibility of the provocative policies played a crucial role. The revelation of Roosevelt's provocative policies undermined his ability to seize upon the incidents with Germany to justify war, while the hope that the provocative policies would be revealed and understood by the American public helped motivate the Japanese decision to attack the United States.

In months before American entry into the war, the United States became engaged with Germany in what has been described as an "undeclared" naval war in the Atlantic.[17] Beginning with the convoying of supplies to the U.K., Roosevelt extended the area in which U.S. ships could operate, tracked German ships that threatened U.S. convoys, and eventually gave orders that American ships were to fire on German and Italian ships on sight. While Roosevelt's purposes are still debated, a number of people that he discussed the policy with indicated that his objective in the Atlantic was to provoke an incident that would create public support for entry into the war.[18] Most revealing are statements by Winston Churchill that Roosevelt indicated "that he would wage war, but not declare it, and that he would become more and more provocative," and that "everything was to be done to force an incident . . . which would justify him in opening hostilities" (Reynolds 1981, 214,215).[19] Roosevelt

[16]In this sense my task is more modest than attempting to enter directly in the debate that currently exists over interpreting Roosevelt's pre-war policies. In some ways, one need not accept that Roosevelt was attempting to directly provoke war to accept my arguments, though it certainly strengthens them.

[17]For example, see the title of Langer & Gleason (1953).

[18]Long before the United States entered WWII, Roosevelt's advisors had concluded that only direct participation in the war could lead to the defeat of Germany, and Roosevelt himself had come to view Germany as a serious threat to the security of the United States (Trachtenberg 2006, 83-84). See also Trachtenberg (2013).

[19]Churchill said Roosevelt was doing this because he was "skating on pretty thin ice in his relations with Congress" and "if he were to put the issue of peace and war to Congress, they would debate it for three months" (Reynolds 1981, 214,215). Similar testimony comes from Roosevelt's advisors and confidants, who wanted more robust American action. Secretary of State Harold Ickes believed that Roosevelt was "waiting for the Germans to create an incident" rather than fire the first shot, while Stimson complained that the President was "waiting for the accidental shot of some irresponsible captain on either side to be the occasion

clearly misrepresented naval incidents provoked by American ships as unprovoked German attacks, particularly in the case of the *U.S.S. Greer*, using them to justify the escalation of American involvement in the war (Schuessler 2010).

So why did the policy against Germany fail to get the United States into war? While this may not have been Roosevelt's intention, it is unlikely that it would have succeeded either way. Germany adopted a policy of tolerating American provocations to avoid giving Roosevelt an excuse, calculating that this would be worthwhile to avoid a war with the United States while Germany was fighting in Russia. In the main case where an incident did occur and Roosevelt attempted to seize upon it, it was exposed as being provoked by the policies of the United States.

The German government tolerated the provocative acts in the belief that the provocations were being committed in order to create an excuse for war. Following the *U.S.S. Greer* incident, German Foreign Minister Ribbentrop messaged the Japanese Foreign Minister that he believed this to be the United States strategy (U.S. Department of State 1964, 504-505). In fact, Ribbentrop had been warning the Japanese about a U.S. policy of this sort for months (Ike 1967, 35). However, Germany was in the middle of their offensive against the Soviet Union, and to avoid an incident, the ships were given strict orders not to take any action that could give the United States an excuse to enter the war. Hitler repeatedly ordered his top naval advisors to avoid incidents with the United States to keep the United States out of the war while the offensive against the Soviet Union was underway, even in the face of flagging morale among U-Boat personnel and over the advice of Admiral Raeder

of his going to war" (Dallek 1979, 265). Similarly, Roosevelt told Ambassador to France William C. Bullitt that "we must await an incident" and that he was "confident that the Germans would give us an incident," and he told Secretary of Treasury Henry Morgenthau that he was "waiting to be pushed into this situation" (Hearden 1987, 196).

(Hearden 1987, 203).[20] Despite American harassment, Hitler clearly believed that tolerating the provocations was preferable to giving the United States an excuse to enter the war.

Despite these orders, incidents did occur, either by accident or by ships that believed themselves to be under attack.[21] The major incident that Roosevelt attempted to seize upon was the *U.S.S. Greer* incident. The *U.S.S. Greer* was a destroyer that was fired on by a German submarine after pursuing the submarine for nine hours and transmitting its coordinates to British bombers overhead (Bailey & Ryan 1979, 168-173). Roosevelt responded to the incident by delivering a scathing fireside chat against Hitler, blaming Germany for attempting to control the entirety of the Atlantic, and announcing his "shoot-on-sight" policy against German and Italian ships.

This incident failed to move the public or Congress, however, in large part due to suspicions – and then revelations – that it was provoked by aggressive American actions. While a majority of Americans approved of the shoot-on-sight policy, it failed to convince many in the opposition, who themselves believed that the administration had provoked the incident. The prominent isolationist Senator Gerald Nye accused Roosevelt of provoking the Greer incident and other naval incidents in order to have an excuse for his shoot-on-sight policy (Cole 1983, 446). Senator Tobey of New Hampshire objected that the American people were "being deceived in a gigantic conspiracy to drive them to war," while the isolationist *Chicago Tribune* accused the Roosevelt administration of seeking an incident as an excuse for war (Langer & Gleason 1953, 748). Even former President Herbert Hoover objected publicly to Roosevelt "edging our warships into danger zones" without the approval of Congress (Langer & Gleason 1953, 748).

[20] See also Thursfield (1948, 219-222,232-233).

[21] Given the U.S. actions, Hitler had conceded that commanders who attacked American ships by mistake would not be punished (Thursfield 1948, 220,222).

The German government, knowing the true story of the incident, seemed to understand that exposing the United States' provocation would undermine Roosevelt and played a role in pressing for a Congressional inquiry. Ribbentrop sent instructions to Hans Thomsen, the German Chargé d'Affairs in Washington, instructing him to contact isolationist members of Congress and ask them to hold a Congressional inquiry to expose the truth about the incident. He informed Thomsen that "there is no doubt that the statements from the German submarine are absolutely true" (U.S. Department of State 1964, 456). He believed that exposing "Roosevelt's war-mongering policy" would "deal it a decisive blow to the advantage of the isolationists" (U.S. Department of State 1964, 455). He also suggested that Congress recall the American destroyer and interrogate its crew. Thomsen responded three days later that he had contacted friendly isolationists and informed them of the German version of events, and that he received assurances that they would press for a Congressional investigation (U.S. Department of State 1964, 468).

Soon after, a Congressional hearing was held in which the true nature of the incident was revealed. During that inquiry, Admiral Harold Stark revealed that the Greer had been pursuing the German submarine before it fired on the American ship (Dallek 1979, 288).[22] It is unclear how much influence the German interference had, but the desired effect was achieved. These revelations undermined Roosevelt's ability to use it as justification for more aggressive policies, and he was not able to use the incidents in the Atlantic to carry the United States into war or implement more aggressive policies (Dallek 1979, 289). According to Bailey & Ryan (1979, 183), the episode became one of the most heavily criticized events of Roosevelt's time in office. Roosevelt's later difficulty in passing the revisions to the Neutrality Act demonstrated his failure to shift Congressional opinion with the Greer

[22]See also Cole (1983, 444).

incident (Langer & Gleason 1953, 757). The German case, then, demonstrates a failure of the strategy of provocation resulting from the German decision to tolerate the provocations rather than respond and the exposure of the American provocation in the case where an incident occurred.

Against Japan, Roosevelt imposed an oil embargo despite knowing that this would force Japan into a southern advance toward the Dutch East Indies, and then he adopted an intransigent position in negotiations to lift the embargo (Trachtenberg 2006). Some evidence suggests that the objective here was also to force Japan into attacking so that the United States could enter the war.[23] Like the German government, the Japanese government was willing to tolerate a lot to avoid war, including making major concessions.[24] It appears the major reason for this is that the Japanese clearly understood that they could not win a prolonged war against the United States (Sagan 1988).[25]

The puzzle in this case, then, is why the Japanese decided to start a war that they knew they couldn't win. The answer is that the Japanese had hoped that public opinion in the United States would force the war to end early, before the Japanese were defeated. This

[23]In the weeks before Pearl Harbor the administration anticipated that war would break out over the crisis, and rather than making efforts to resolve the crisis and avoid war, Roosevelt discussed the possibility of a surprise attack and said the issue was "how we should maneuver them [the Japanese] into the position of firing the first shot without allowing too much danger to ourselves (Dallek 1979, 307). Secretary of War Stimson had similarly written that "... we face the delicate question of the diplomatic fencing to be done so as to be sure that Japan was put in the wrong and made the first bad move - overt move" (Jun 1994, 340).

[24]For example, Ambassador to Japan Joseph Grew later wrote that "[Prince Konoye] had told me with unquestionable sincerity that he was prepared at that meeting to accept the American terms whatever they might be... It was clearly understood and admitted in Japan that the proposed agreement would inevitably entail the withdrawal of all Japanese troops from French Indochina and China as fast as they could practicably be withdrawn, with the mere face-saving expedient of leaving garrisons in Mongolia and North China" (Grew 1944, 4-5). Foreign Minister Toyoda also later wrote that "we were ready to settle everything on the spot, even the withdrawal of troops [from China]" (Jun 1994, 193). Trachtenberg (2006, 107-115) presents further evidence from both sides that Japan was willing to make major concessions to avoid war, and that even after Konoye was thrown from office, Prime Minister Tojo was willing to make concessions to avoid war.

[25]Also see Trachtenberg (2006, 107-109).

hope itself relied partially on the belief that the American public would recognize that the Japanese had been forced to attack by the provocative American actions.

In deciding to initiate a war that it knew it could not win, the Japanese government had hoped that American public opposition would ensure that the war was brief and would end in stalemate (Sagan 1988). For example, in a memorandum to the Emperor on September 6, 1941, the Japanese Cabinet and Supreme Command wrote that, while it "would be well-nigh impossible to expect the surrender of the United States... we cannot exclude the possibility that the war may end because of a great change in American public opinion" (Ike 1967, 153). Less than a month before the attack on Pearl Harbor, in a document titled "Proposal for Hastening the End of the War Against the United States, Great Britain and the Netherlands, and Chiang," the government wrote that the war's objective was to "destroy the will of the United States to fight," with a major part of this strategy being propaganda "persuading Americans to reconsider their Far Eastern policy and pointing out the uselessness of a Japanese-American war" (Ike 1967, 247). As part of this operation, Japanese agents in the United States had already been instructed to make contact with those who the Japanese government believed would undermine support for war (Sagan 1988).

The question that follows from this is why Japan attacked Pearl Harbor when their whole strategy hinged on the American public forcing the government to drop out of the war. Certainly some officials anticipated that the attack could galvanize the American public for the long war that the Japanese government hoped to avoid.[26] Evidence suggests that the

[26]President of the Privy Council Hara Yoshimichi told the Imperial Council a month before the war that "positive action against the United States" will ensure that "their indignation against the Japanese will be stronger than their hatred of Hitler" (Ike 1967, 237). Chief of Staff of the Eleventh Air Fleet Onishi Takijiro had recommended that "we should avoid anything like the Hawaiian operation that would put Americas back up too badly" (Agawa 1979, 229). Marquis Kido expressed similar worries the previous month during a talk with then-War Minister Tojo on the morning of October 14: "If hostilities should commence, Japan will have to take the initiative and attack the Philippines. As for America, since it will be the first time her domain will be attacked, the flaring up of her public sentiment is more than imaginable" (IMTFE reel 18,

Japanese hope was for Americans to recognize that Japan was forced into the war.

The Japanese government certainly viewed the war in this way, and the language of "self-defense" and "self-preservation" was ubiquitous in their discussions and statements.[27] In addition, Prime Minister Tojo explicitly stated that he was relying on this hope. He told the Imperial Council that "there is some merit in making it clear that Great Britain and the United States represent a strong threat to Japan's self-preservation... America may be enraged for a while, but later she will come to understand [why we did what we did]" (Ike 1967, 239).

One can also see this concern operating in the Japanese plan for Pearl Harbor. Japanese officials were concerned that the attack on Pearl Harbor not be seen as a sneak attack, but as the result of American pressure and intransigence. Ambassador Nomura, in recommending that Japan warn the United States before the Pearl Harbor attack, recognized that the attack could be used as "counterpropaganda" and that, if Japan did not break off negotiations and warn of the attack, the United States would blame Japan for rupturing the negotiations with the attack itself (Jun 1994, 346). Admiral Yamamoto shared similar views (Agawa 1979, 233,259). In the end, the warning was not given in time because of a miscommunication, and Roosevelt seized upon the attack exactly as Nomura feared he would.

Finally, given the Japanese hope that the American public would force the government out of the war, it is only logical that the public would have had to believe that exiting the war posed no direct threat to the American homeland. The Japanese strategy for the war was to set up a defensive perimeter in East Asia and simply hold out against American attacks. The Japanese government knew that they had no ability to threaten the U.S. homeland, and

Exhibit 1148-A, 5. Thanks to Steve Palley for this find).

[27]See Ike (1967).

must have hoped that the American public recognized that too. In fact, the Pearl Harbor attack led Americans to see their homeland as directly threatened and motivated the public to fight the war to the bitter end.

The Japanese strategy clearly failed. Despite their awareness of the embargo, the public was not aware of the position it put the Japanese government in or the intransigence of the United States in negotiating an end to the embargo. The public did not come to see the attack as provoked but as a sneak attack against a United States that was still willing to resolve any dispute peacefully. Sixty years later, in fact, we are still debating whether the attack was provoked, and at the time it was widely accepted as an unprovoked attack demonstrating Japanese militarism and even madness (Sagan 1988).

This case demonstrates a successful application of the provocation strategy. Japan still wanted to avoid war with the United States, but attacked in the hope that the American public would recognize that the Japanese attack was not an act of aggression, but a desperate act of self-defense. Instead, the attack galvanized the American public to fight a war until the bitter end. In neither the German nor the Japanese case did that government want war with the United States, and they sought to avoid it. Germany succeeded, largely through their own restraint and also because of the recognition by American isolationists that the U.S. provoked the incidents in the Atlantic. Japan, on the other hand, failed, attacking the U.S. in the hope that the public would recognize that American provocations, a hope that didn't materialize.

2.6 Conclusion

This paper demonstrates that the possibility that a provocative action will remain hidden explains why the strategy of provoking an incident can successfully be used to carry a country to war. The model demonstrates that a state can be provoked into an attack despite the fact that it could be blamed for aggression, and that a legislature can be mislead into blaming an adversary for aggression even though it was provoked into an attack. Even when the adversary anticipates this possibility and seeks to avoid war, it can be lured into responding by the hope that the provocation would be exposed and the war averted. The survey experiment shows that the public reacts to incidents by increasing support for military action and reacts to revelations of its own government's provocations by decreasing support for military action, particularly when combined with an incident, as predicted by the model. Finally, the case study demonstrates that states behave as predicted in the model, avoiding incidents when possible to avoid giving a pretext for war but also responding to provocations when they perceive enough of a chance for the provocation to be exposed and the strategy of misleading the legislature and public to be undermined.

Does this paper provide any insight for policy solutions to prevent governments from employing this kind of deceptive strategy? Much of international relations theory argues that it doesn't matter much who fired the first shot, with wars resulting from structural factors in the relations between states. Given that wars are usually fought over major stakes rather than small incidents, and given the sometimes misleading nature of these incidents, following the guidance of international relations theory and ignoring the circumstances surrounding the outbreak of war may seem the wisest course of action. However, it appears unlikely that the public will stop relying on the question of who fired first to guide its opinion.

As this paper demonstrates, it can be informative even though it can also be misleading. Therefore, I conclude that the best approach to preventing deception would be to increase the transparency of the government's military actions in order to remove this option from the hands of the executive.

The increased transparency suggested here would seem to indicate that the more democratic a country is, the less likely its public is to be deceived into war by a belligerent government. This speaks to a recent debate about deception and foreign policy decision-making in democracies. Schuessler (2010) challenged the claim by Reiter & Stam (2002) and others in the democratic peace literature that democracies will make better decisions about war and peace because of public debate and public constraint on the government. My paper strikes something of a middle ground in the debate about the public constraint on the government. I demonstrate that provocation and deception can be effective within limits, but that the ability to deceive the public can be moderated by transparency and institutional constraints on the government's freedom of action.

One factor not addressed in this model is whether a government may try to deceive its public into war not because it has more belligerent preferences, but because it has information about potential threats that the public does not have. In that case, deception can allow the government to avoid knowingly choosing bad policies in order to satisfy public opinion, as in models on pandering and posturing.[28] The possibility of deception for the purpose of passing good policy is an important one to consider. Many people would argue that the case of WWII fits this description, with Roosevelt attempting to bring the United States into the war because he had a sharper recognition of the danger from allowing a Germany victory in Europe than the public, rather than simply being more belligerent than the average

[28]For example, see Canes-Wrone, Herron & Shotts (2001), Stasavage (2004).

American. The model here already allows for the government to have better information on the opponent than the public or legislature, but in the case that the adversary is belligerent, the government's policy is to wait for the adversary to attack and reveal this belligerence for all to see. The current model does not allow for preventive or preemptive war against a belligerent adversary or the possibility that provocations or deceptions could be used to start such a war. Accounting for this possibility is worth addressing in future work.

2.7 Appendix A

2.7.1 Characterization of Equilibrium in Proposition 1

Begin by assuming that the legislature chooses $x_3 = w$ if it receives the signal $X = np$, and chooses $x_3 = nw$ if it receives the signal $X = p$.

Following $x_1 = np$, the adversary is certain that an attack will lead to war, and has the option of no attack which leads to the status quo. It is certain that an attack will lead to war because the legislature will receive a signal of $X = np$ with certainty and will follow its strategy assumed above. The adversary will therefore prefer to attack $x_2 = a$ if it prefers war to the status quo, $w_A \geq 1 - q$, and will prefer to not attack $x_2 = na$ if it prefers the status quo to war, $w_A < 1 - q$.

Following $x_1 = p$, the adversary knows that war will follow an attack if the legislature receives a signal $X = np$, which occurs with probability $1 - \pi$, and knows that the status quo will prevail if the legislature receives a signal $X = p$, which occurs with probability π. If the adversary chooses no attack, $x_2 = na$, the game will end in the status quo with a cost suffered by the adversary. Therefore, he will prefer war if

$$(1 - \pi)w_A + \pi(1 - q) \geq 1 - q - c_A \tag{2.7}$$

which is equivalent to equation (2).

The executive therefore faces three types of adversary. If $w_A \geq 1 - q$, the executive's choice is between $x_1 = np$ yielding war for certain and a payoff of w_E, or $x_1 = p$ yielding war with a probability $1 - \pi$ and the status quo payoff $1 - q$ with probability π. Since

$$w_E > w_E(1-\pi) + q\pi \tag{2.8}$$

due to the assumption that $w_E > q$, the executive will prefer $x_1 = np$.

If $1 - q \geq w_A > 1 - q - \frac{c_A}{1-\pi}$, the executive's choice is between $x_1 = np$ yielding the status quo for certain or $x_1 = p$ yielding war with a probability $1 - \pi$ and the status quo with probability π, as above. Since $w_E > q$, against these types the executive will prefer $x_1 = p$.

If $w_A < 1 - q - \frac{c_A}{1-\pi}$, the game ends in the status quo whatever the executive chooses, so he is indifferent, and any mix of $x_1 = p$ and $x_1 = np$ can sustain an equilibrium.

Given these strategies, it remains to be shown that the legislature will not deviate from the strategies assumed in the beginning of the characterization.

Define $\hat{\alpha}_{np}$ as the legislature's estimate that an attack is unprovoked when the legislature receives a signal of $X = np$ after an attack. This estimate will be the probability that an unprovoked attack occurs over the probability that an attack occurs and no provocation is observed. That is the following:

$$\hat{\alpha}_{np} \equiv \frac{P(x_1 = np, x_2 = a)}{P(x_1 = np, x_2 = a) + P(x_1 = p, x_2 = a)(1-\pi)} \tag{2.9}$$

The numerator, the probability that the executive doesn't provoke and the adversary attacks, is equal to the probability that the adversary is type $w_A \geq 1 - q$. The denominator, the probability that an attack occurs and no provocation is observed, is equal to the probability that the adversary is provocable and type $1 - q > w_A \geq 1 - q - \frac{c}{1-\pi}$ multiplied by the probability that the provocation is not observed by the legislature $(1 - \pi)$. Therefore, this equation can be rewritten:

$$\hat{\alpha}_{np} \equiv \frac{P(w_A \geq 1-q)}{P(w_A \geq 1-q) + (1-\pi)P(1-q > w_A \geq 1-q-\frac{c}{1-\pi})} \tag{2.10}$$

By assumption, the legislature prefers the status quo to war, $q > w_L$, but when the legislature observes $X = np$ following an attack it believes there to be an $\hat{\alpha}_{np}$ probability that not authorizing war would lead to retreat. Therefore, the legislature is willing to authorize war when $w_L \geq (1-\hat{\alpha}_{np})q$. Since the value of $\hat{\alpha}_{np}$ defined in equation 2.10 is between 0 and 1, there will exist values of $w_L \in [0,1]$ and $q \in [0,1]$ that don't violate $q > w_L$. Therefore, if I assume that the values w_L and q are such that $w_L \geq (1-\hat{\alpha}_{np})q$ holds, the legislature will not deviate and these strategies form an equilibrium.

2.7.2 Characterization of Second Pure Strategy Equilibrium

Begin by assuming that the legislature never wars, $x_3 = nw$.

In that case, $x_2 = a$ is a dominant strategy for all the executive's strategies. Following $x_1 = p$, the adversary will not deviate from attacking because $1-q > 1-q-c_A$. Following $x_1 = p$, the adversary will not deviate from attacking because $1 > 1-q$.

Given the executive's strategy of choosing $x_1 = p$ for all types w_A, he will not deviate because $q > 0$. Since the executive always chooses $x_1 = p$, the legislature will believe that $x_1 = p$ for any signal received. Therefore, $\hat{\alpha}_p = \hat{\alpha}_{np} = 1$ for all signals and game histories. The legislature will not deviate from x_3 for any signal or game history since $q > w_L$. Therefore, these strategies form an equilibrium.

2.7.3 Comparative Static that Higher c_A leads to weakly higher ϕ

From equation 2.3, $\phi = c_A$ when $\pi \in [0, \frac{1-q-c_A}{1-q}]$. Therefore, $\frac{dc_A}{d\pi} = 1 > 0$ and ϕ is increasing in c_A.

From equation 2.4, $\phi = (1-q)(1-\pi)$ when $\pi \in [\frac{1-q-c_A}{1-q}, 1]$. Therefore, $\frac{dc_A}{d\pi} = 0$, and ϕ is constant in c_A.

Finally, as c_A increases, $\frac{1-q-c_A}{1-q}$ decreases, and the threshold for ϕ to equal $(1-q)(1-\pi)$ decreases. Imagine that, through an increase in ϵ, the value of ϕ changes from c_A to $(1-q)(1-\pi)$. Since $(1-q)(1-\pi) \geq c_A$ for all $\pi \geq \frac{1-q-c_A}{1-q}$, ϕ will be weakly greater for any given value of c_A. Therefore, if increasing c_A changes the value of ϕ from c_A to $(1-q)(1-\pi)$, that value of ϕ cannot be lower.

Since ϕ is either constant or increasing in π for all values of $\pi \in [0,1]$, ϕ is weakly increasing in $\pi \in [0,1]$.

2.7.4 Comparative Static that Higher π leads to weakly lower ϕ

$\phi = c_A$ when $\pi \in [0, \frac{1-q-c_A}{1-q}]$. Therefore, $\frac{d\phi}{d\pi} = 0$ and ϕ is constant in π.

$\phi = (1-q)(1-\pi)$ when $\pi \in [\frac{1-q-c_A}{1-q}, 1]$. Therefore, $\frac{d\phi}{d\pi} = -1 < 0$, and ϕ is decreasing in π.

Since ϕ is either constant or decreasing in π for all values of $\pi \in [0,1]$, ϕ is weakly decreasing in $\pi \in [0,1]$.

2.7.5 Characterization of Equilibrium in Preemption Extension

Begin by assuming that the legislature adopts the strategy of $x_3 = w$ when $X = np$ and $x_3 = nw$ when $X = p$. Following $x_1 = np$, the adversary will not deviate from $x_2 = a$ if

$w_A \geq 1 - q$ and will not deviate from $x_2 = na$ if $w_A < 1 - q$, similar to the equilibrium in the previous model.

Following $x_1 = p$, the adversary anticipates that the legislature may or may not authorize war depending on the signal it receives, and will choose to attack if $\pi(1 - q) + (1 - \pi)w_A \geq w_A - k$. This can be written as

$$w_A \geq 1 - q + \frac{k}{\pi} \tag{2.11}$$

I start with the case where $k > 0$. If $w_A \geq 1-q$ and $w_A \geq 1-q-\frac{k}{\pi}$, an attack is guaranteed following either $x_1 = p$ or $x_1 = np$, and the executive will prefer not to provoke since this guarantees war. Since $1 - q + \frac{k}{\pi} > 1 - q$, there are some types that will attack following no provocation but not following a provocation. The executive prefers to provoke these types since the game will end in war following no response to a provocation and $w_E + k > w_E$. For the same reason, the executive will provoke types $1 - q > w_A$.

It remains to demonstrate that the legislature will not deviate given its beliefs. $\hat{\alpha}_{np}$ is given by equation 2.5, and $1 > \hat{\alpha}_{np} > 0$. As in the previous model, the legislature will authorize war if $w_L \geq (1 - \hat{\alpha}_{np})q$, and I assume that this condition holds for the value of $\hat{\alpha}_{np}$ given by equation 2.5.

Now I look at the case where $k < 0$. The difference is that, since $1-q+\frac{k}{\pi} \leq 1-q$, there are some types that will attack following a provocation but not following no provocation. The executive will prefer to provoke these types. $\hat{\alpha}_{np}$ is given by equation 2.6, and $1 > \hat{\alpha}_{np} > 0$. As before, the legislature will authorize war if $w_L \geq (1 - \hat{\alpha}_{np})q$, and I assume that this condition holds for the value of $\hat{\alpha}_{np}$ given by equation 2.6.

2.8 Appendix B

2.8.1 Survey Experiment Text

The following questions are about U.S. foreign policy. You will read about a hypothetical situation that the U.S. government may face in the future. We will describe one approach American leaders may take and ask whether you approve or disapprove of that approach.

Tensions have been high between the United States and Iran due to Irans pursuit of nuclear technology. The United States and its allies have been negotiating with Iran to reach an agreement to end to the Iranian nuclear program. Following

> an incident in which Iran torpedoed an American ship in the Persian Gulf resulting in 11 American lives lost, and
>
> **OR**
>
> [blank]

accusations of bad faith by both sides, negotiations were recently suspended. The United States government accused Iran of pursuing nuclear weapons and attempting to dominate the Persian Gulf region through force and intimidation. The Iranian government responded that the United States had been launching missile strikes against targets inside Iran from ships in the Persian Gulf that had resulted in loss of life,

> but major newspapers and network news broadcasts have dismissed these claims as totally unfounded.
>
> **OR**

with major newspapers and network news broadcasts reporting that they are unable to verify or disprove these claims.

OR

and major newspapers and network news broadcasts report that they have been able to verify the Iranian governments claim.

In a national address to the American public, the President announced a new policy. He said: The latest actions of the Iranian regime have given a new and grave turn to the already serious situation in the Persian Gulf. It has become clear that efforts to reach a peaceful solution have failed and that that the only option left for curbing Irans ambitions is military action.

Following the speech, the United States launches a major military operation of air strikes against military and nuclear targets in Iran. The United States also deploys extra troops to the Persian Gulf to respond to any retaliation against American forces stationed in the region or regional allies.

Summary

- Iran fires on an American ship in the Persian Gulf, resulting in 11 American deaths.

  OR

  [blank]

- Negotiations between the U.S. and Iran over their nuclear program break down.
- U.S. accuses Iran of seeking to dominate the region.

- Iran accuses U.S. of missile attacks,

  but major news sources dismiss this claim as totally unfounded.

  OR

  with major news sources unable to verify or disprove claim.

  OR

  and major news sources verify the Iranian governments claim.

- The President announces that efforts to reach a peaceful solution have failed.

- The United States authorizes military action against Iranian nuclear and military targets.

Do you approve or disapprove of the Presidents handling of this situation?

1. Approve

2. Disapprove

2.8.2 Balance and Parametric Regression

First, to test the balance of the survey experiment results, I ran two-sample t-tests across all treatment groups to test for difference of means in the demographic variables of sex, age and ideology. I found no significant difference of means for any of these variables, indicating a successful randomization in the experiment.

Second, to more formally verify the findings and examine the effect of demographics, I ran a parametric regression on the data. Using a logistic binomial regression, I estimated

Table 2.1: Logistic Regression on Survey Experiment Data

Coefficients	Estimate	Std. Error	Significance
Intercept	0.2112815	0.2690798	
Attack	0.7168304	0.1283942	***
Provocation	-0.3045903	0.1358739	*
Sex	-0.1078784	0.1405295	
Age	0.0005268	0.0060533	
Ideology	-0.1357689	0.0356495	***

the effects of an attack and a U.S. provocation on approval for military action, as well as the effects of sex, age and ideology. The results are below.

The regression confirms that the presence of an Iranian attack and a U.S. provocation has a statistically significant effect on approval, the attack at the 99.9 % level and the provocation at the 95 % level. In this case, the Iranian attack has the effect of increasing approval while the U.S. provocation has the effect of decreasing approval. This does not, however, fully confirm hypothesis ***H2*** that the provocation would have a stronger effect in the presence of an attack, which I discuss below. In addition, while sex and age had no effect on approval, ideology had a very strong effect. In particular, more conservative respondents were more likely to approval of military action than more liberal respondents. Since, as mentioned earlier, this sample was more liberal than the U.S. population, it is likely that the U.S. population would be more approving of military action than in the sample here. Nevertheless, even with a liberal population, the treatments had a strong effect.

To interpret these results, I found the distribution of first differences in expected values from 1000 simulations to figure out the effect of a one-point change in the value of the independent variables on approval for military action. The attack and provocation variables only take on a value of 0 or 1, as does sex, which is 0 for male and 1 for female. For age,

Table 2.2: First Differences of Expected Values for Change in Value of Independent Variable, in Percentages

	Mean	Std. Dev.	95% CI - Lower Bound	95% CI - Upper Bound
Attack	17.3	3.1	11.1	23.3
Provocation	-7.4	3.4	-13.9	-0.9
Sex	-2.5	3.6	-9.9	4.4
Age	0	0.1	-0.3	0.3
Ideology	-3.2	0.8	-4.6	-1.7

the variable takes on a value of the respondent's age, and for ideology it takes a value of a seven-point scale, with 7 being the most liberal. For all of the variables, the table shows the difference in approval from increasing the value of the variable by 1.

Consistent with the survey results reported above, the attack increases approval for military action by an estimated 17.3% and the provocation reduces approval for military action by an estimated 7.4%. Sex and age have no statistically significant effect, while ideology has a strong effect. A one-point change in ideology decreases approval by 3.2%, and since it is a seven-point scale, moving from "very liberal" to "very conservative" can be expected to reduce approval by 19.2%.

Finally, I note that I ran a logistic regression with an interaction variable of the provocation and attack to test whether the provocation had a stronger effect when an attack occurs, as I demonstrated earlier with the difference of means test. While it does decrease approval, the effect is not statistically significant. Given the small size of the effect and the limited number of respondents in the relevant treatment groups, it is surprising that this does not register a statistically significant effect. This points to the need for further research and a more wide-ranging survey to examine these effects.

Table 2.3: Logistic Regression with Interaction Variable

Coefficients	Estimate	Std. Error	Significance
Intercept	0.16759	0.20037	
Attack	0.78345	0.15781	***
Provocation	-0.21447	0.18438	
Attack*Provocation	-0.20550	0.27178	
Ideology	-0.13800	0.03537	***

Chapter 3

Fear, Appeasement and the Effectiveness of Deterrence[1]

3.1 Introduction

A central question in the study of deterrence has been how threats can be credible when they are meant to defend interests that do not immediately appear to be worth fighting over. While deterrence over some interests has "inherent credibility" because of those interests' value to the defender, much of deterrence theory was developed during the Cold War to understand how the United States could credibly threaten to use (possibly nuclear) force, "even when its stakes were low in a particular region" (Danilovic 2001). Many scholars have argued that defending states can bolster the credibility of their deterrent threats by making physical or verbal commitments to carry them out, such as placing "trip-wire" forces in a threatened area or giving public speeches before domestic audiences (Schelling 1966,

[1]Co-authored with Alexander V. Hirsch

Slantchev 2011, Fearon 1994). Others have focused on a defender's knowledge that failing to carry out a threat could damage their reputation for "resolve" in future interactions (Alt, Calvert & Humes 1988, Sechser 2010). Finally, some scholars have claimed that empirically, the question is irrelevant: by and large, the effectiveness of deterrence can be explained by the real value of the interests at stake.[2]

While existing theories can account for many instances of both successful and failed deterrence, it is still not clear that a concern for one's reputation or physical and verbal commitments are sufficient to make credible the threat of major war to defend marginal interests. Wars between states are major events, undertaken with enormous and sometimes catastrophic risk. Even if states feared the consequences of damaging their reputations, starting a war over a marginal interest would involve trading the risk of those consequences for the certainty of a catastrophic war. Furthermore, faced with such a high cost, it is difficult to imagine a commitment device sufficiently strong to make backing down even more costly. In fact, adversaries sometimes take advantage of this fact by engaging in low-level aggression below a defender's threshold for war, i.e. "salami tactics"(Schelling 1966, 66).

Still, minor transgressions *are* sometimes effectively deterred with the threat of a major war. Many of the most significant Cold War crises were over stakes that were, in and of themselves, relatively insignificant when compared to the costs and consequences of thermonuclear war; for example, in 1955 the Eisenhower administration prepared for war and even raised the possibility of using nuclear weapons in response to a Communist Chinese attack on the sparsely populated and strategically marginal islands of Quemoy and Matsu, ultimately deterring Chinese aggression (Chang 1988). More recently, the government of North Korea threatened war in response to both economic sanctions and an airstrike on

[2]See Danilovic (2001) for discussion.

their nuclear plant, and evidence suggests that the United States government took these threats seriously.[3] It is difficult to imagine that reputational considerations, audience costs, or the intrinsic value of a nuclear plant could induce the North Korean government to carry out a war that would almost certainly mean its own destruction.

In this paper, we develop a formal model demonstrating that the implicit threat of a major war can credibly deter an objectively minor transgression even without the existence of commitment devices or concerns about a defender's reputation. Credible deterrence over marginal interests is possible if a defending state fears his adversary's future intentions, and the defender prefers war sooner rather than later against any challenger that is unappeaseably belligerent.

The logic is simple. Suppose that a defender entertains even an infinitesimal fear that his adversary is unappeasable and intends to start a full-scale war. Could the defender learn that the adversary is unappeaseably belligerent if the adversary engages in a minor transgression, such as attacking a sparsely populated island? Intuition would suggest that he couldn't. However, if it is commonly understood that such a transgression will trigger a war, then the defender can make exactly this inference. The challenger's decision to transgress while expecting it to trigger a war reveals that the challenger does, in fact, have a preference for full-scale war over the status quo, a preference unlikely to be changed by successfully completing a minor transgression.

If the defender then believes war to be inevitable, he prefers to initiate it immediately

[3]The threat in response to economic sanctions was in 1994, and the threat in response to an airstrike was in 2003. See Wit, Poneman & Gallucci (2004) and KCNA News Agency (2003). Secretary of Defense William Perry and Assistant Secretary of State Robert Gallucci both wrote that they took the 1994 threat seriously, having believed that it would be "irresponsible" to treat it simply as bluff, and even in 1994 believed that an airstrike against the nuclear plant could trigger a war. See Carter & Perry (1999) and Wit, Poneman & Gallucci (2004).

rather than first be weakened by allowing the minor transgression. Understanding this, the adversary will believe the threat of war to be credible, and be deterred from transgressing unless she actually desires war. In this situation, deterrence works because even a minor transgression is an effective "test" of an adversary's belligerence when it is commonly believed that the defender's threat of war is credible.

An example helps to clarify the logic. During the early Cold War, the United States feared that the Soviet Union intended to launch a war to conquer Western Europe. A 1953 National Security Council report on possible U.S. responses to Soviet actions in Berlin begins by asserting that "control of Berlin, in and of itself, is not so important to the Soviet rulers as to justify involving the Soviet Union in general war"(NSC 173,1953). Thus, the report reasons that the Soviet Union will only attack West Berlin if they "decide for other reasons to provoke or initiate general war," and that the United States would "have to act on the assumption that general war is imminent." In other words, an invasion of Berlin must imply that the Soviet Union expects to trigger a wider war and affirmatively desires it, rather than imply that they think they can invade without triggering a war. Since an invasion would imply that a wider war is imminent, the United States was to respond with a full mobilization and "implementation of emergency war plans," thereby fulfilling the United States' commitment to fight in the event of an invasion.

Our model is a simple two-period game of aggression and deterrence between a potential challenger and a defender that captures this logic.[4] At the start of the game, the potential challenger can attempt a transgression that has *direct value* to her in the event that peace prevails, but also *military value* in the event that war breaks out. It is this transgression that the defender wishes to deter. The transgression could represent any number of prohibited

[4]Throughout the paper we refer to the defender as "he" and the challenger as "she."

actions that the challenger would desire even if her intentions were ultimately peaceful, such as occupying territory belonging to the defender or a protégé, enacting sanctions, or developing scientific technology that could be weaponized. However, it also generates an endogenous shift in military power (Fearon 1996, Powell 2006). If the challenger attempts to transgress, the defender may permit the transgression, or preemptively respond with war. If the transgression is permitted, then the challenger can either enjoy his direct gains and end the game peacefully, or initiate war under the more favorable military balance that results.

Crucially, we assume that the defender is uncertain of the challenger's willingness to go to war, and specifically fears that she is a type against whom war is inevitable. The payoffs in the model are very general, and can accommodate any number of reasons for why war may be inevitable after a transgression has been allowed; for example, war could be a commitment problem, or the leader of the challenging state may not fully internalize its cost.[5]

Unlike classical analyses of deterrence, we assume no uncertainty about the defender's preferences. Most earlier studies assumed that deterrent threats derive their credibility from the *possibility* that a defender may actually intrinsically prefer war to allowing a transgression and examined how a defender can credibly signal that preference when it exists.[6] Since the assumption that a defending state may intrinsically prefer war to appeasement is dubious in many of the most prominent motivating examples for deterrence theory, we assume that the defender is commonly known to be an *appeaser*, in that he definitively prefers to allow the transgression if it will avert war. The defender's core strategic dilemma is therefore closely related to Powell's analysis of "salami tactics" – he prefers to appease if possible, but prefers war sooner to war later if it is inevitable (Powell 1996).

[5]For examples of war caused by a commitment problem, see Powell (2004) and Powell (2006), and war caused by leaders not internalizing the costs, see Chiozza & Goemans (2004) and Jackson & Morelli (2007).

[6]See Slantchev (2011) for discussion.

Our analysis generates several insights about how and when fear can sustain credible deterrence. First, we derive a condition on the transgression itself such that the defender's implicit threat of war can be credible regardless of war's cost or the ex-ante probability that the challenger is belligerent. The condition is not that the transgression be of any objective magnitude, but rather that the transgression's military value *exceed* its direct value to a challenger who is initially indifferent between peace and war. When the transgression satisfies this condition, allowing it cannot effectively appease a challenger who is initially belligerent because it will increase her military capacity more than her payoff from peace.

This condition links our paper to the literature on bargaining with endogenous power shifts. These models have shown that, in the absence of uncertainty, similar conditions lead to war or the gradual elimination of one player (Fearon 1996, Schwarz & Sonin 2008). Our analysis demonstrates that with even a tiny amount of uncertainty about the challenger's intentions, this same condition can lead to peace with deterrence of even a very minor transgression with very high probability.

Second, our analysis shows that under additional mild assumptions, the probability that deterrence is successful is *increasing* in the *difference* between the transgression's military and direct value to the challenger. The logic is similar to the baseline result; when the transgression has a higher military value relative to its direct value, it is increasingly likely that an initially belligerent challenger will *remain* belligerent even after transgressing, making the defender more willing to respond with war and better able to credibly deter. Conversely, when the transgression has a lower military value relative to its direct value, it is increasingly likely that an initially belligerent challenger can actually be appeased by being allowed to transgress, making the defender him more willing to allow the transgression and easier to exploit.

This result has the surprising empirical implication that deterrence is actually *less* likely to succeed when the transgression is of greater direct value to the challenger. Even when the defender infers the challenger to be initially belligerent, he can maintain hope that allowing the transgression will effectively appease her. This may explain, for example, why the Allies were unable to effectively deter Hitler from annexing Austria or invading the Sudetenland. Since both territories had large populations of co-ethnics, the claim that the territories would satisfy Germany was plausible; Germany's actions therefore did not reveal that her ultimate intentions made war inevitable. In many applications it is also reasonable to assume that the challenger's direct value for the transgression is equal to the direct cost imposed on the defender by allowing it. In this case, increasing the direct value of the transgression can have a non-monotonic effect on the probability of deterrence; it first decreases as the defender becomes more willing to try appeasement, and then increases as the defender becomes more willing to fight a war over the transgression itself. In either case, our analysis demonstrates that empirical studies of deterrence that control for the "interests at stake" in a dispute are flawed because they fail to separately control for the military and nonmilitary value of a transgression.[7] Such studies are motivated by the premise that intrinsic values alone determine the willingness of states to carry out their threats. However, an equally important factor is states' expectations and inferences about their adversaries' future behavior – for this question, the *relationship* between the military and direct value of a transgression is essential.

Finally, an interesting feature of the model is that the defender may actually benefit from his uncertainty and fear about the challenger's intentions because it allows him to credibly deter a minor transgression with the major threat of war. We examine this property

[7]See Huth (1999) for summary of empirical literature.

by comparing the baseline model to an identical variant in which the challenger's type is known to the defender at the start of the game. We first show that the probability of deterrence indeed decreases when the challenger's type is revealed to the defender; absent uncertainty about the challenger, the defender can no longer prevent his known preference for appeasement from being exploited. A possible empirical implication of this result is that events introducing uncertainty about a potential challenger's preferences, such as a sudden regime change, may actually increase the likelihood that deterrence succeeds by creating plausible fear on the part of a defender. We then show that when the difference between the military value and the direct value of the transgression is sufficiently large, the defender actually benefits from her fear and uncertainty in expectation. It allows her to credibly deter, but rarely or never results in preventable wars because appeasement is unlikely to work when the challenger is actually belligerent.

The paper proceeds as follows. In the next section we present a simple mathematical example to illustrate the logic of our model. Next, we present the formal model and derive our main results. Following that, we present a case study of the crisis over the Turkish Straits in 1946 to demonstrate that this logic can help explain the United States' decision to defend Turkey from Soviet invasion. We then briefly discuss the robustness of our core results to a variety of extensions, including extending the game sequence, endogenizing demands, altering the bargaining sequence in each period, and two-sided uncertainty; these extensions are formally considered in a Supplemental Appendix. Finally, we conclude with a summary of our findings and questions for future research.

3.2 Example

To illustrate the intuition underlying our model, we first present a simple example. Suppose a challenger (C) and a defender (D) initially posses equal shares of a landmass of size and value equal to 1. The challenger may attempt to forcibly seize an additional sliver of size $\delta = \frac{1}{100}$ initially possessed by the defender; that is, to increase her holdings to $\frac{1}{2}+\delta$. The defender can only prevent the seizure with a war, in which the victor takes control of the entire landmass. If the defender allows the seizure, the challenger may then attempt to initiate war in an attempt to take control of the rest of the land mass. For the sake of intuition, imagine that each state's share is its respective homeland, and the sliver desired by the challenger is a mostly barren hilltop on the defender's border with some trivial strategic value in the event of war.

The direct value of possessing a given share of the landmass is common and equal to its size. The probability of victory in a war also depends on the prevailing division, but the challenger begins with a small edge. Specifically, the challenger's initial probability of victory is $p_q = \frac{1}{2} + \frac{1}{1{,}000}$, and should she successfully seize the hilltop becomes $p_r = \left(\frac{1}{2} + \delta\right) + \frac{1}{1{,}000}$. Since the probability of victory in a war increases by δ with seizure of the hilltop, this quantity is both the hilltop's *direct* and *military* value.

The defender's cost of war is known to be $c = \frac{1}{4}$. The challenger's cost of war is private information, and can either be "normal" ($\theta_C = c = \frac{1}{4}$) with probability $\frac{9{,}999}{10{,}000}$ or low ($\theta_C = 0$ with probability $\frac{1}{10{,}000}$. The low-cost type is willing to wage war to correct any imbalance between her share of the landmass and her military strength, and would therefore be willing to wage war after a concession in an attempt to seize the entire landmass. Ex-ante, however, the defender is almost certain that the challenger is not the low-cost belligerent type, i.e.

$P(\theta_C = 0) = \frac{1}{10{,}000}$.

Suppose that the challenger had the option to immediately seize the entire landmass. Clearly the defender would be willing to respond war rather than appeasement. Such a seizure would leave the defender with nothing while a war, although costly, would still have a positive payoff in expectation.

The central issue of our analysis, however, is whether it is reasonable for the defender to respond to an attempted seizure of the barren hilltop with full-scale war, which would deter any challenger who is a normal type and wishes to avoid war. Intuition suggests that he would not; were the defender to respond with war he would suffer an immediate cost of $\frac{1}{4}$, but only prevent a trivial change of $\delta = \frac{1}{100}$ in direct value and in the military balance. This tiny change in the military balance appears particularly irrelevant since the probability that the challenger is belligerent and will start a war is $\frac{1}{10{,}000}$. If this intuition were true, the challenger would always seize regardless of her type because she expects to make a gain without provoking war.

Nevertheless, this intuition is false; the mere presence of fear that the challenger is a low-cost type is sufficient to sustain credible deterrence. By attempting to seize the hilltop while anticipating that this action will result in war, the challenger reveals herself to be the low-cost belligerent type that will eventually go to war over the entire landmass. To see this, suppose that both sides believe that an attempted seizure of the hilltop will trigger war. Then the normal type of challenger will be deterred, since the tiny imbalance of $\frac{1}{1{,}000}$ between her initial holdings and her initial probability of victory is dwarfed by the cost of war $\frac{1}{4}$. Only the low-cost type will seize since she prefers war to the status quo.

Understanding this, the defender will infer that a challenger who seizes is a low-cost belligerent type. Although he'd still prefer to allow seizure if it would successfully appease

the challenger, he also knows that appeasement is impossible because gaining the hilltop will increase the challenger's payoff for war as much as it increases her payoff for peace. He will thus infer the inevitability of war from an attempted seizure and will respond preemptively. Anticipating war, it is strictly better for him to go to war before the hilltop is seized rather than after, since otherwise the challenger would gain a very slight military advantage δ. Since responding to a seizure with war is optimal, the normal-cost type of challenger will indeed be deterred, choosing to end the game peacefully rather than provoke a war.

Notice that the defender's ability to infer that the challenger is belligerent is conditional on his own strategy. If the challenger believed that the defender would permit seizure of the hilltop, then all types of challenger would opportunistically seize and the defender would learn nothing about the challenger's type. This points to the fact that this example has multiple equilibria. When the challenger believes that seizure will be permitted, all types seize, the defender infers nothing, and given the low ex-ante probability that the challenger is belligerent, permitting seizure is optimal. The challenger's decision to transgress only reveals her belligerence when she believes that the defender will respond with war, and war is an optimal response only when this belligerence is revealed.

This example presents the basic mechanism through which deterrence can be effective against a minor transgression by assuming the proper conditions. In the following section we present a simple general model that formalizes the logic of our example. The model allows us to extract comparative statics about when deterrence will hold by varying the military and direct values of the transgression and the defender's beliefs about the challenger's type.

3.3 The Model

The model is a simple two-period game of aggression and deterrence played between a potential challenger (C) and a defender (D). Throughout the paper we refer to the defender as "he" and the challenger as "she."

Sequence In the first period, the challenger chooses whether or not to attempt a transgression $x^1 \in \{a, \emptyset\}$ that has both *direct value* to her in the event that peace prevails, and *military value* in the event that war breaks out. The transgression could represent any number of prohibited actions that would shift the military balance toward the challenger, but also benefit her if her intentions vis-a-vis the defender were ultimately peaceful; it therefore presents the defender with an inference problem about the challenger's true intentions. Such actions could include occupying territory belonging to the defender or a protégé, enacting sanctions, or developing scientific technology that could be weaponized.[8] Worth noting is that developing military capabilities is usually intrinsically costly rather than beneficial to a challenger with peaceful intentions, and is therefore outside the scope of the model.[9] The challenger's attempt to transgress is perfectly observable to the defender, and therefore could also be interpreted as making a demand of the defender to allow it.

If the challenger does not attempt to transgress ($x^1 = \emptyset$), then the game ends with peace. However, if she does ($x^1 = a$), then the defender may either allow the transgression ($y^1 = n$) or resist it ($y^1 = w$). To make the case for our claim as difficult as possible, we assume that

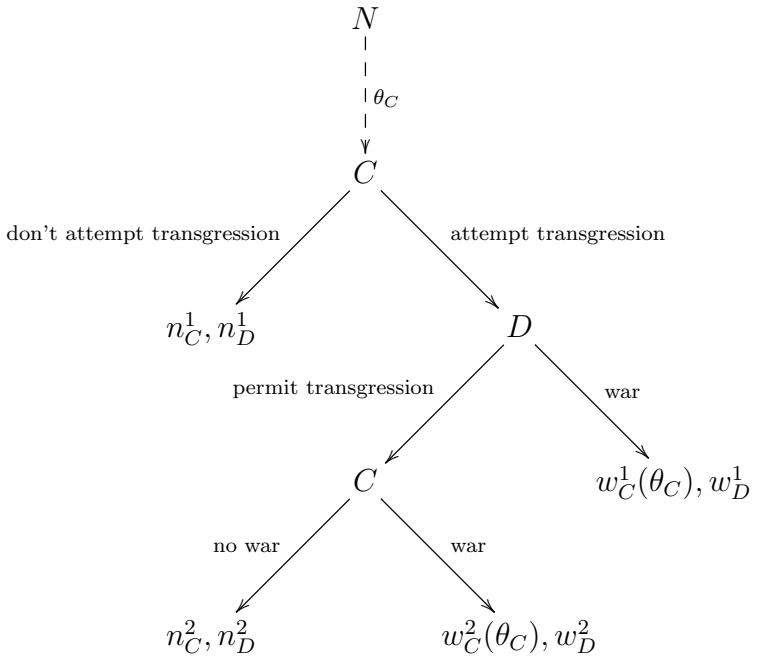
[8] In our analysis we do not distinguish between direct and extended deterrence or consider the incentives of a protégé in the latter case, as in Quackenbush (2006).

[9] According to Slantchev (2011) (an additional distinction with our model is that the defender has no opportunity to preempt a shift in military power). Worth noting, however, is that military investments could also satisfy the properties of the transgression in our model if they yielded benefits in interactions with internal political rivals or states other than the defender. In Baliga & Sjostrom (2008), acquiring nuclear weapons is modeled as having the properties of the transgression for these reasons.

the challenger's act presents the defender with a fait accompli; to resist the transgression he must respond with war.[10]

If the defender allows the challenger to transgress rather than respond preemptively with war, then the game proceeds to a second period. In the second period, the challenger's payoffs are assumed to be higher in the event of either peace or war as a result of having successfully transgressed, and the defender's are assumed to be lower. The challenger then decides whether to enjoy her direct gains and end the game peacefully ($x^2 = n$), or herself initiate war under the more favorable military balance ($x^2 = w$). The sequence of the game is depicted in Figure 3.1.

Figure 3.1: Deterrence Model Game Tree



[10]If alternatively the challenger had an additional opportunity to back down upon encountering resistance, credible deterrence would be easier; the defender could entertain the possibility that the challenger is bluffing. In the Supplemental Appendix we prove that whenever deterrence works in our baseline model it also works in such a variant.

Defender's Incentives Unlike classical analyses of deterrence credibility where the defender's payoffs are uncertain to the defender, we assume that the defender's payoffs are common knowledge and that he has a known preference for appeasement. We briefly note, however, that our main results are robust to the introduction of uncertainty about the defender, and in some cases it actually sharpens them; this extension is formally treated in the Supplemental Appendix.

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

1. allowing the transgression makes him worse off in both peace ($n_D^2 < n_D^1$) and war ($w_D^2 < w_D^1$),

2. allowing the transgression is strictly better than responding with war if the challenger will subsequently choose peace ($n_D^2 > w_D^1$).

Given these assumptions, the defender's optimal response to a transgression depends on his interim assessment β that a challenger who has attempted to transgress will initiate war even after she is allowed to do so. If war is to be inevitable then he prefers to avoid the cost $w_D^1 - w_D^2 > 0$ of allowing an unappeasably belligerent challenger to transgress; intuitively, this cost captures the (potentially small) endogenous shift in military power that results. However, if allowing the transgression would actually appease the challenger then he prefers to do so and avoid the cost $n_D^2 - w_D^1$ of a preventable war. The defender will therefore prefer to respond to the transgression with preemptive war when his interim belief β exceeds a threshold $\bar{\beta}$, where

$$\bar{\beta} = \frac{n_D^2 - w_D^1}{(n_D^2 - w_D^1) + (w_D^1 - w_D^2)} \in (0, 1)\,. \tag{3.1}$$

Crucially for our argument, $\bar{\beta} < 1$ always holds – if war is truly inevitable, then the defender prefers war sooner to war later regardless of its cost.[11]

The defender's core dilemma in our model is thus similar to Powell's analysis of "salami tactics"(Powell 1996). He is vulnerable to exploitation by the challenger because a small transgression is below his known threshold for war. However, his fear that the challenger's demands may in fact be far reaching, and his preference for war sooner rather than war later if it is to be inevitable, can potentially allow him to credibly deter in equilibrium. Indeed, although we model the transgression as an isolated act below the defender's threshold for war, it could equally represent a "slice" of a larger "salami" that the challenger has a potentially insatiable demand for in an unmodeled continuation game. In the Supplemental Appendix we consider a multi-period extension of the game with this property.

Finally, we note that omitting the defender's option to trigger a war if the defender chooses not to transgress is without loss of generality because the defender is assumed to have a known preference for peace (even at the cost of allowing the transgression).

Challenger's Incentives Because the defender is never intrinsically willing to fight a war to prevent the transgression in our model, the key factor sustaining his willingness to do so must be his *fear* of the challenger's future intentions. To model that fear, we assume the challenger's payoffs are unknown to the defender and depend on a *type* $\theta_C \in \Theta \subset \mathbb{R}$, where Θ is an interval. The defender's prior beliefs over the challenger's type is a continuous distribution $f(\theta_C)$ with full support over Θ.

To clarify our essential argument, we assume that the defender's payoffs are fixed and independent of the challenger's type, as would be the case if he was uncertain about the

[11]For example, if $n_D^1 = 0$, $w_D^1 = -10^{10}$, and $w_D^2 - w_D^1 = n_D^2 - n_D^1 = 1$, then $\bar{\beta} = \frac{10^{10}-1}{10^{10}}$

challenger's cost of war or her private valuation for certain war outcomes.[12] However, our main insights extend easily to variants where the challenger has private information about both parties' payoffs from war, as would be the case if she knew more about her probability of victory; this extension is formally treated in the Supplemental Appendix.[13] Without loss of generality we furthermore assume that the challenger's type affects only her war payoffs, and write her payoff as n_C^t if the game ends in period t with peace and $w_C^t(\theta_C)$ if the game ends in period t with war.

In the numerical example in the preceding section, the defender was uncertain about whether a challenger who attempted to seize the hilltop desired it for its own value, or to strengthen her position in a future war. To capture this uncertainty we assume that the challenger's payoffs satisfy the following properties.

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

2. In each period t the challenger's war payoff $w_C^t(\theta_C)$ is continuous and strictly increasing in θ_C. In addition, there exists a unique challenger type $\bar{\theta}_C^t$ strictly interior to Θ that is *indifferent* between peace and war (i.e. $w_C^t\left(\bar{\theta}_C^t\right) = n_C^t$) in each period t.

Together, these assumptions imply that all challenger types would value the transgression, that a challenger's type θ_C indexes her willingness to fight a war, and that in each period t there is positive probability that the challenger prefers peace to war $\left(\theta_C < \bar{\theta}_C^t\right)$ and war to peace $\left(\theta_C > \bar{\theta}_C^t\right)$.

[12] For examples of models with uncertainty about the cost of war, see Fearon (1995) or Sechser (2010). For uncertainty about private valuation, we note that it common feature of historical accounts of interstate relations is uncertainty about a challenger's value for possessing a defender's homeland. Glaser (1997) addresses this in his discussion of "greedy states."

[13] See Fey & Ramsay (2011) for an analysis of the consequences of the distinction between private and interdependent war values in a mechanism design context.

In our model, there is therefore always the possibility (potentially small) that the challenger cannot be deterred from transgressing because she prefers war to the status quo $\left(w_C^1(\theta_C) > n_C^1 \iff \theta_C > \bar{\theta}_C^1\right)$. In addition, once faced with a potential transgression, the defender always fears the possibility that the challenger is a type against whom war is inevitable; formally, these are types $\theta_C > \bar{\theta}_C^2$ who would unilaterally initiate war after being allowed to transgress.

Although challenger types $\theta_C > \bar{\theta}_C^2$ are modeled as unilaterally initiating war, this outcome could also represent an unmodeled continuation game where the challenger makes an additional demand against which the defender is willing to fight. Interpreted as such, a number of rationales for the defender's willingness to fight in the second period are possible; his threat over the subsequent demand could be "intrinsically" credible as in perfect deterrence theory, it could once again be driven by fear that war is inevitable, or some other commitment problem could lead to war.[14] In his seminal analysis of commitment problems and war, Powell (2006) discusses several prominent historical cases where it was the *resolution* of uncertainty about both sides' preferences that eventually led to war.[15]

Finally, we note that our results are robust to assuming that wars in the first period are always the result of a commitment problem, in the sense that peace in the second period is a Pareto-superior outcome for all types ($n_C^2 > w_C^1(\theta_C)\ \forall \theta_C$). This is the case in the example, where the low-cost belligerent type of challenger prefers seizing the hilltop peacefully to immediate war $\left(\frac{1}{2} + \delta > \frac{1}{2} + \frac{1}{1{,}000}\right)$, but cannot *not* commit not to exploit the subsequent

[14]Note that if the defender has the final decision of war vs. backing down in each period, then fear of inevitable war *alone* cannot sustain his willingness to fight at every stage – he must eventually anticipate a move by the challenger against which he is willing to fight for some other reason. Our multi-period extension in the Supplemental Appendix has this property. For overview of perfect deterrence theory, see Zagare (2004).

[15]Another possible rationale for war under complete information is leaders who do not fully internalize the cost of war, see Chiozza & Goemans (2004) and Jackson & Morelli (2007).

gain in military strength by initiating war $\left(\frac{1}{2}+\delta+\frac{1}{1{,}000}>\frac{1}{2}+\delta\right)$.

3.4 Results

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

Proposition 4 *A pure strategy equilibrium of the model always exists.*

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

$$\bar{\beta} \geq P\left(\theta_C \geq \bar{\theta}_C^2\right)$$

2. *There exists a* ***deterrence equilibrium****, in which the defender always responds to the transgression with war, and all types $\theta_C < \bar{\theta}_C^1$ who do not initially prefer war are deterred, i.f.f.*

$$\bar{\beta} \leq P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$$

When both pure strategy equilibria exist, there also exists a mixed strategy equilibrium, and the defender is best off in the deterrence equilibrium.

A pure strategy equilibrium of the model always exists, and when a mixed strategy equilibrium exists the defender is always better off in the pure strategy deterrence equilibrium. Because our focus is on the conditions under which the defender can achieve credible deterrence in equilibrium, we henceforth restrict attention to these equilibria.

Pure strategy equilibria of the model are of two types. The first type is a "no deterrence equilibrium." The challenger always attempts to transgress, and consequently the defender can infer nothing about the challenger's type simply from observing the transgression itself. He therefore decides how to respond on the basis of his prior $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ that the challenger is sufficiently belligerent to initiate war after transgressing. If that prior $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ is low and/or the defender's belief threshold $\bar{\beta}$ for responding with war is high, then this equilibrium will exist. Recall that $\bar{\beta}$ is determined by the cost $n_D^2 - w_D^1$ of an avoidable war relative to the cost $w_D^1 - w_D^2$ of allowing an unappeasably belligerent challenger to transgress. These conditions accord with the conventional wisdom for when deterrence should fail – when the benefit of avoiding war is high relative to the cost of appeasement, and the challenger is very likely ex-ante to be a peaceful type.

The second type of pure strategy equilibrium is a "deterrence equilibrium." In this equilibrium the defender always responds to the transgression with preemptive war. Consequently, the challenger is deterred from transgressing unless she is initially belligerent, in the sense of preferring war to the status quo (i.e. $\theta_C \geq \bar{\theta}_C^1$). Crucially, this deterrence allows the defender to draw an inference from observing the transgression itself even if it is objectively minor – precisely that the challenger is an initially belligerent type. As a result, he decides whether to respond with war *not* on the basis of his prior $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ that the challenger will initiate war after transgressing, but his *posterior* $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$ that allowing an *already belligerent* challenger to transgress will fail to appease her. This exceedingly simple observation is in fact our key insight. In the presence of fear that war may be inevitable, the primary factor determining the defender's ability to credibly deter in equilibrium is *not* the cost of war, the severity of the transgression, or the initial level of fear that the challenge is belligerent. The reason is that when deterrence is actually effective, the defender

can infer initial belligerence in equilibrium from observing the transgression itself. Instead, the primary factor is actually the *effectiveness of appeasement against an already belligerent challenger*.

The implications of this simple insight are surprisingly strong, as illustrated in the following corollary.

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

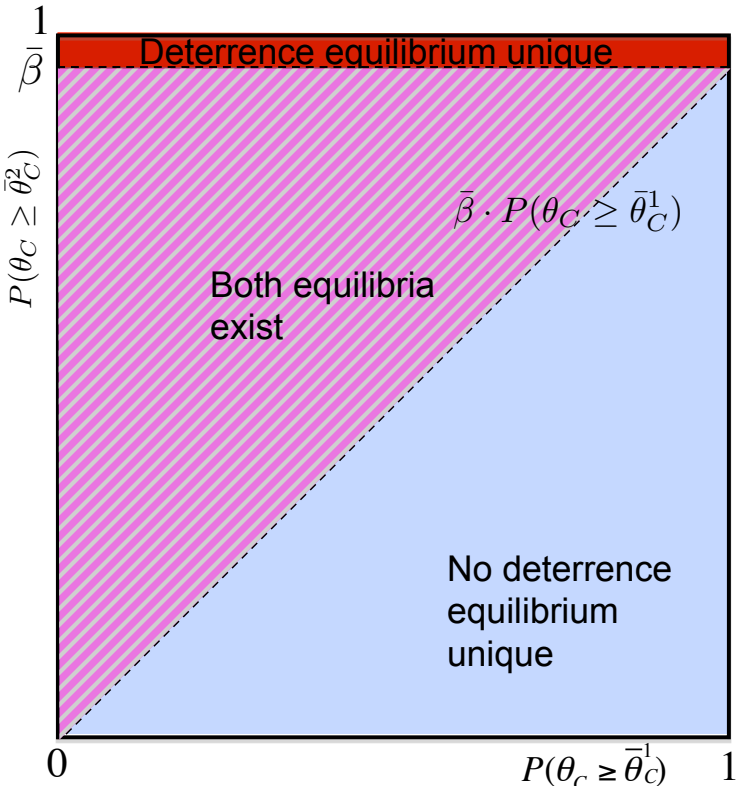
Thus, when appeasement is impossible against a belligerent challenger, the deterrence equilibrium always exists. This is true even when the "no deterrence equilibrium" also exists because of a high cost of war $n_D^1 - w_D^1$, a low cost of allowing the transgression in both direct $(n_D^1 - n_D^2)$ and military $(w_D^1 - w_D^2)$ terms, and/or a sufficiently low probability that the challenger is belligerent $P\left(\theta_C \geq \bar{\theta}_C^t\right)$ in both periods. The deterrence equilibrium remains the defender can use the transgression (however minor) as a *test* of the challenger's initial belligerence, knows that initial belligerence ensures future belligerence because appeasement is ineffective, and therefore prefers to respond with war upon observing the transgression. The challenger is thereby deterred unless she affirmatively prefers immediate war, fulfilling the defender's expectations.[16]

Figure 3.2 depicts the equilibrium correspondence for an example in which the defender's belief threshold $\bar{\beta}$ for responding with war is very high. The defender's prior $P\left(\theta_C \geq \bar{\theta}_C^1\right)$

[16]Baliga and Sjostrom's (2008) analysis of nuclear arms proliferation can also exhibit a qualitatively similar separating equilibrium when their assumption (3) fails and the "crazy type" values weapons sufficiently highly. However, there are also important differences. Because we micro-found the defender's willingness to attack a crazy type in the preference for war sooner vs. war later, she will do so in the equilibrium regardless of her war payoffs; our model also yields new comparative statics on when such an equilibrium will prevail.

that the challenger prefers immediate war to the status quo is on the x-axis, while the prior $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ that she would initiate war after transgressing is on the y-axis; both quantities are derived from the challenger's underlying payoffs and the distribution over her type. The figure demonstrates that the deterrence equilibrium can remain even when the probabilities that the challenger would be belligerent in either period are arbitrarily low, which can be seen by observing that the hatched triangle extends to the origin. Moreover, this property would persist even if the defender's threshold $\bar{\beta}$ were made arbitrarily high.

Figure 3.2: Deterrence Model Equilibria



The (In)effectiveness of Appeasement

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

In this section, we consider the question of what makes appeasement less effective, and consequently deterrence more effective; to answer this question we examine the payoff properties of the transgression itself. Recall that transgressing has both a military value $w_C^2(\theta_C) - w_C^1(\theta_C)$, which is the challenger's gain in the event of war, and a direct value $n_C^2 - n_C^1$, which is her gain in the event of peace. These quantities determine how the challenger's preference for war changes as a result of successfully transgressing, and we henceforth denote them using $\delta_C^m(\theta_C)$ and δ_C^d, respectively.

We begin with a simple condition that does not require any additional assumptions.

Lemma 1 *Appeasement is ineffective, and thus the deterrence equilibrium exists for all defender payoffs and probability distributions, if and only if* $\delta_C^m\left(\bar{\theta}_C^1\right) \geq \delta_C^d$.

Thus, a sufficient condition for the deterrence equilibrium to exist is that military value of the transgression $\delta_C^m(\cdot)$ exceed its direct value δ_C^d to a challenger of type $\bar{\theta}_C^1$ who is initially indifferent between peace and war. The logic is straightforward. For such a challenger type, $\delta_C^m\left(\bar{\theta}_C^1\right) \geq \delta_C^m$ means that the military gains from successfully transgressing increase her net benefit from war as much as the direct gains from transgressing reduce it. Since she initially weakly preferred war to peace, she and all types more belligerent than her must continue to

prefer war to peace after transgressing. Allowing the transgression therefore cannot appease any type of challenger who was initially belligerent (i.e. $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) = 1$), which by Corollary 1 implies that the deterrence equilibrium exists.

The condition in Lemma 1 turns out to be familiar from the literature examining complete information bargaining in the presence of endogenous shifts in military power. To our knowledge, however, it is absent from the literature (either empirical or theoretical) on deterrence.[17] The bargaining literature finds that this condition generally results in wars or the gradual elimination of one player. In contrast, we find that this same condition can lead to a fearful peace with deterrence of even a very minor transgression with very high probability. Both predictions are rooted in the same property; allowing the transgression cannot appease a belligerent challenger. However, the distinction arises from the difference in assumptions about whether the challenger is initially belligerent. In the complete information setting, the challenger's belligerence at the outset is assumed. In our model, the defender can believe that the challenger is very likely likely to be peaceful ex-ante; however, his *fear* that the challenger is unappeasably belligerent – however small – allows him to credibly deter.

Comparative Statics When credible deterrence is possible but difficult – either because the defender's threshold $\bar{\beta}$ for war is high and/or the probability the challenger will be belligerent $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ is low – the model generally features multiple equilibria. This complicates extracting testable comparative statics because the model cannot speak to how the players will form expectations about whether the transgression is being treated as a "red line" – it can only say when doing so would be an equilibrium.

Nevertheless, the consequences of our basic insight – that the effectiveness of appeasement

[17] See Fearon (1996) and Schwarz & Sonin (2008) for studies of complete information bargaining with endogenous power shifts.

influences the credibility of deterrence – can be examined by deriving comparative statics on the probability of deterrence under the assumption that the deterrence equilibrium prevails whenever it exists. With this assumption, the probability that deterrence is successful is 0 when the deterrence equilibrium does not exist, and is $P\left(\theta_C \leq \bar{\theta}_C^1\right)$ when it does.[18] The following Lemma considers this comparative static as a function of the challenger's payoffs, while holding the defender's payoffs fixed.

Lemma 2 *Suppose that,*

- *the deterrence equilibrium prevails whenever it exists,*
- *the transgression's military value is equal to δ_C^m for all challenger types,*
- *the challenger's first period payoffs are held fixed.*

Then appeasement is less effective (i.e. $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$ is increasing) the greater is the difference $\delta_C^m - \delta_C^d$ between the challenger's military and direct value for the transgression. Consequently, the probability that deterrence is successful is increasing in $\delta_C^m - \delta_C^d$.

With our additional assumptions, the probability of deterrence is increasing in the difference $\delta_C^m - \delta_C^d$ between the military and direct value of the transgression to the challenger. The intuition is similar to that of Lemma 1 – the greater is $\delta_C^m - \delta_C^d$, the more likely it is that appeasement will fail against a belligerent challenger, the more willing is the defender to respond with war conditional on deterrence, and the better able he is to deter. This effect is depicted in Figure 3.3; the left panel shows the probability of deterrence when the defender's

[18] This comparative static could also be interpreted as the probability that deterrence is feasible, rather than the actual empirical probability that it occurs.

payoffs are fixed, while the right panel depicts the probability when the defender's payoffs are initially drawn from a distribution.[19]

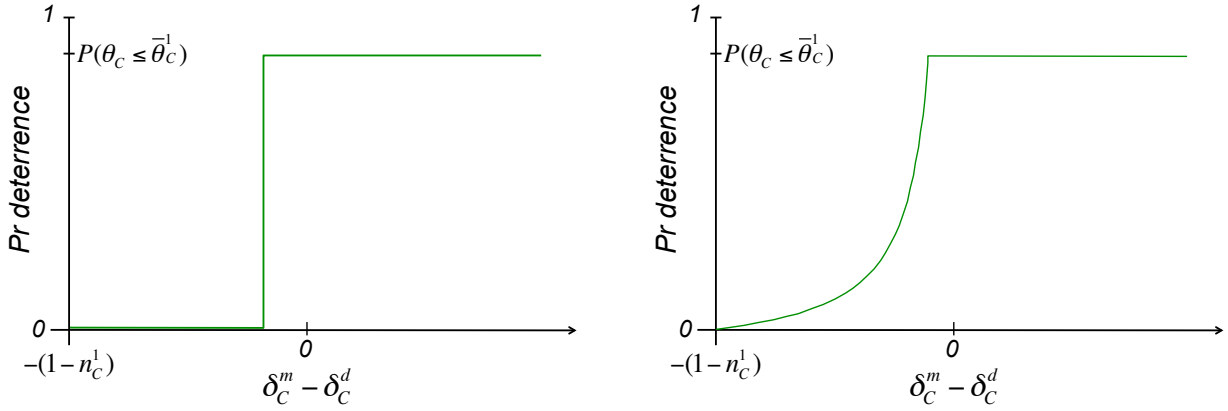


Figure 3.3: Probability of deterrence as function of $\delta_C^m - \delta_C^m$

Lemma 2 has surprising and counterintuitive implications for empirical studies of deterrence failure: it predicts that deterrence should be most effective against against transgressions that carry a high military value *relative to* their direct value for a challenger, regardless of their absolute value. This suggests, for example, that President Eisenhower's implicit threat of war over a Chinese communist invasion of the sparsely populated and strategically marginal islands like Quemoy and Matsu in 1955 may have actually been quite credible. Precisely because their thorough marginality could not appease a China intent on war, an invasion could have easily been perceived as an informative signal of both the present and future belligerence by the Chinese communists. Interpreted in the context of Lemma 2, the Allies' difficulty in deterring Hitler from annexing Austria and invading the Sudetenland prior to World War II is also easier to understand. While these territories were presumably

[19]Specifically, the right panel depicts $G\left(\frac{1-F(\bar{\theta}_C^2)}{1-F(\bar{\theta}_C^1)}\right) \cdot F\left(\bar{\theta}_C^1\right)$, where $G\left(\bar{\beta}\right)$ is the induced probability distribution over $\bar{\beta}$. To generate both figures we assume that $w_C^1\left(\theta_C\right) = \theta_C$, the transgression's military and direct cost to the defender's are equal to .1, and the challenger's type is uniformly distributed on $[0, 1]$. The left panel assumes that the defender's benefit from peace is $n_D^1 - w_D^1 = .55$, while the right panel assumes it is uniformly distributed on $[-.1, 1]$.

of significantly greater intrinsic value to the Allies than Quemoy and Matsu, they were also densely populated by German co-ethnics. Consequently, the notion that occupying them might satisfy a belligerent Germany was actually quite plausible, resulting in exploitation of the Allies' known preference for appeasement and deterrence failure.

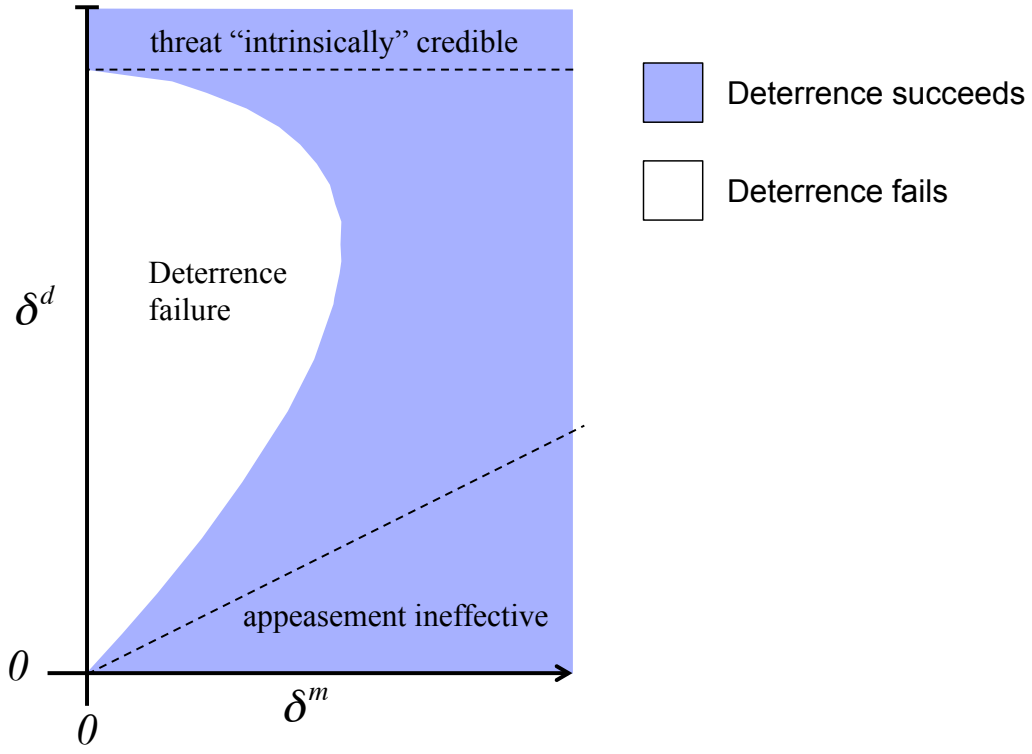
An additional complication worth noting is that the comparative static in Lemma 2 varies the challenger's values for transgressing while holding those of the defender fixed. However, in many applications it is reasonable to suppose that a transgression with greater direct or military value for the challenger is also one that imposes greater direct or military costs on the defender. This relationship is important for making empirical predictions about deterrence success because, while allowing a transgression with a higher direct value might more effectively appease, it is also more intrinsically worth fighting over. Figure 3.4 illustrates the probability of deterrence in a numerical example where the *values* to the challenger for transgressing are equal to the *costs* imposed on the defender. In the example, the probability of deterrence is always increasing in the transgression's military value (on the x-axis). However, increasing the transgression's direct value (on the y-axis) has a non-monotonic effect; the probability of deterrence first decreases due to the logic of Lemma 2, and then increases as the defender's intrinsic willingness to fight becomes the dominant factor.[20]

Lemma 2 and the preceding example jointly demonstrate the importance of distinguishing the military from the direct value of a transgression in empirical analyses of deterrence, a distinction that has been heretofore ignored.[21] Such studies are generally motivated by the premise that the absolute magnitude of intrinsic values alone determine states' willingness

[20]The figure is generated by assuming that $w_C^1(\theta_C) = \theta_C$ with $\theta_C \sim U[0, 1.1]$, and both the defender's benefit $n_D^1 - w_D^1$ and lowest type of challenger's benefit $n_C^1 - w_C^1(0)$ for avoiding war is .5.

[21]See Huth (1999) for a summary.

Figure 3.4: Probability of Deterrence when Challenger's Gains = Defender's Costs



to carry out their threats. However, our model demonstrates that an equally important factor is states' expectations and inferences about their adversaries' future behavior – for this question, the *relationship* between the military and direct value of a transgression is essential.

The Benefits of Fear

The preceding analysis demonstrates that counterintuitively, it can be the defender's *fear* – rather than a potential intrinsic willingness to fight – that allows him to credibly deter a minor transgression with the threat of a major war. This property suggests that the defender may actually benefit in expectation from his fear and uncertainty. In this final analytical section we examine this claim. To do so, we compare the baseline model to a variant that

is identical in every respect except that the challenger's type θ_C is revealed to the defender at the start of the game. This comparison allows us to isolate the effect of the defender's uncertainty without changing the probability distribution over the challenger's type. The following Lemma compares equilibrium outcomes and payoffs in the two games.

Lemma 3 *Suppose that the deterrence equilibrium prevails whenever it exists. Then,*

1. *the probability of deterrence would decrease if the defender knew the challenger's type.*
2. *when the probability* $P\left(\theta_C < \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right)$ *that appeasement is effective is below*

$$\left(\frac{P\left(\theta_C \leq \bar{\theta}_C^1\right)}{P\left(\theta_C \geq \bar{\theta}_C^1\right)}\right) \cdot \left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1}\right),$$

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

Lemma 3 first shows that the probability of deterrence decreases when the challenger's type is known to the defender; thus, it is specifically the defender's fear, and not just the possibility that the challenger may be unappeasably belligerent, that generates credible deterrence. To see why, suppose for simplicity that appeasement is completely ineffective, i.e. $\bar{\theta}_C^2 \leq \bar{\theta}_C^1$. If the defender knew the challenger's type, then she would be unable to deter her from transgressing whenever that type was outside of $\left(\bar{\theta}_C^2, \bar{\theta}_C^1\right)$. When $\theta_C > \bar{\theta}_C^1$ the challenger would be unappeasably belligerent and transgress, while when $\theta_C < \bar{\theta}_C^2$ it would be commonly known that the challenger would be peaceful after transgressing, and she would exploit the defender's known preference for appeasement. However, if the defender is uncertain of the challenger's type, then he can credibly maintain fear that a challenger who is in fact appeasable $\left(\theta_C < \bar{\theta}_C^2\right)$ may be unappeasably belligerent $\left(\theta_C > \bar{\theta}_C^1\right)$, infer that

unappeasable belligerence in equilibrium, be willing to respond with war, and deter types $\theta_C < \bar{\theta}_C^1$ from transgressing.

This insight suggests that events creating uncertainty about the intentions of a challenger can have important and surprising effects on the potential for credible deterrence. For example, a sudden regime change in a potential adversary may create fear on the part of a defender about that adversary's intentions that did not previously exist; understanding that fear and the potential inference the defender could draw from a transgression, the adversary could paradoxically be easier to deter.

Finally, it is indeed the case that the defender's uncertainty can sometimes benefit her in expectation; the second part of the Lemma states that this will be the case when the probability $P\left(\theta_C < \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right)$ that appeasement would work against a belligerent challenger be sufficiently low. The reason is that it is then unlikely that the defender's uncertainty will result in a preventable war. Moreover, when appeasement is ineffective $\left(\theta_C^2 < \bar{\theta}_C^1\right)$ the defender is unambiguously better off not knowing the challenger's type.

Counterintuitively then, the defender's fear can actually be source of strength. Thus, it is not surprising that states often claim to fear the unappeasable belligerence of their adversaries; for example, North Korea accompanied its warning in 2003 that an air strike on their nuclear plant would lead to "total war" by explicitly stating that such an attack would be viewed as a precursor to an invasion (KCNA News Agency 2003). This insight suggests interesting avenues for future work to which we return in the conclusion.

3.5 The Turkish Straits Crisis of 1946

To demonstrate that the logic we have identified sheds light on puzzling historical episodes, we examine the crisis over the Turkish Straits that occurred between the United States and the Soviet Union in the early days of the Cold War. In 1945 and 1946, the Soviet Union repeatedly demanded that Turkey allow it to place bases on the Turkish Straits, which was widely believed to imply a subjugation of Turkey by the Soviet Union as happened to the countries of Eastern Europe.[22] These demands, coupled with extensive Soviet military preparations in the Balkans, led American officials to prepare for armed aggression against Turkey. Truman eventually decided that the United States would fight a war to defend Turkey in the event of Soviet invasion. Although this commitment was never announced publicly, Stalin reversed course after learning about Truman's decision.

In this section, we argue that our model of deterrence helps to explain why the United States was willing to fight a major war to defend Turkey and why the Soviet Union found this commitment credible. We demonstrate that the incentives facing the United States were similar to those faced by the defender in the model, and that the willingness of the United States to defend Turkey can be explained by a logic similar to that in the model's deterrence equilibrium.

As in the model, American decision makers were unwilling to fight a major war solely for the defense of Turkey. From June to December 1945, before the United States believed itself to be in a worldwide confrontation with the Soviet Union, the United States had no plans for defending Turkey and no intention of intervening in the event of invasion despite

[22]The Soviets also demanded a revision of the Montreaux Convention governing use of the Straits, which all parties were willing to accept, and the cession of the provinces of Kars and Ardahan, a demand that was later dropped (Kuniholm 1980).

the belief that such an invasion was likely (Mark 1997). Given that WWII had just ended and that any war would have involved the Soviet army overrunning Western Europe and the United States bombing the Soviets Union with nuclear weapons, the American reluctance was understandable. However, as the Cold War developed the United States began to fear that the Soviet Union would undertake acts of aggression against other countries in the Middle East and even Western Europe, and it is clear that the United States would have preferred to fight a war immediately if Soviet ambitions were not confined to Turkey. In this scenario the United States would have viewed a wider war as inevitable, and Turkey occupied an important role in plans for such a war. Turkey was to be the first line of defense against a Soviet advance toward strategically vital areas of the Middle East, the loss of which would severely weaken the U.S. and its allies (Kuniholm 1980, Mark 1997).[23]

The Turkish Straits Crisis thus contained the fear and uncertainty about the challenger's intentions and the preference to fight sooner rather than later if the war was inevitable that are essential to our model. Historical accounts of the crisis also suggest that the Truman administration believed it could infer far-reaching Soviet ambitions from an invasion of Turkey *because* the Soviets understood that their actions would result in war, which is the key inference that sustains deterrence in equilibrium. Fearing unchecked Soviet aggression and believing that only the threat of U.S. intervention could deter the Soviets, the United States decided in August 1946 that it would defend Turkey if attacked. Most officials believed that an American warning would deter the Soviet Union from attacking Turkey, because it was generally thought that the Soviet Union did not desire a wider war (Mark 1997). Conversely, officials appear to have believed that the Soviet Union would only invade Turkey if

[23]These areas included the Suez Canal, which provided the British access to their Far Eastern colonies and provided the United States a base from which it was to conduct bombing runs against the Soviet homeland, and the Persian Gulf, which had large supplies of oil.

they did desire such a war. Undersecretary of State Dean Acheson believed that the Soviet Union would most likely be deterred, but he also argued that the United States would "learn whether the Soviet policy includes an *affirmative* provision to go to war *now*" if deterrence failed (Mark 1997, 400).[24] President Truman, when asked if he understood that the decision to defend Turkey may mean war, responded that "we might as well find out whether the Russians were bent on world conquest now as in five or ten years" (Mills 1951, 192).

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

While other deterrence mechanisms may have also been relevant, some of them fail to explain key features of the crisis. It is possible that reputational concerns were driving decision-making, and the United States was clearly interested in demonstrating its resolve to allies (U.S. Department of State 1969). Any explanation involving audience costs, however, would require threats to have been made publicly. While the United States did dispatch a naval force to the Mediterranean, it never publicly announced its decision to defend Turkey, instead communicating with the Soviet Union privately and downplaying the crisis in pub-

[24]Emphasis in original.

lic. In addition, there was no obvious commitment device that would have automatically engaged the United States in a conflict, such as military forces stationed in Turkey as a "trip-wire."Finally, there was nothing probabilistic about the Americans' threat; Truman clearly asserted that he would follow the recommendation to defend Turkey "to the end" (U.S. Department of State 1969, 840).

3.6 Robustness

In this section, we briefly discuss the robustness of our basic insight that fear can generate credible deterrence under seemingly extreme conditions. While our model is parsimonious to illustrate the main point, it is robust to a number of common complexities studied in the international relations literature. The robustness checks are formally conducted in a Supplemental Appendix and briefly described here.

The first robustness check considers a variant in which the defender's resistance in the first period does not result in immediate war, but instead the challenger has an opportunity to first back down. This extension actually makes the deterrence equilibrium of our model easier to sustain; the defender can entertain the possibility that the challenger is actually bluffing when he observes an attempted transgression. Indeed, in the motivating case where the challenger is very unlikely to be belligerent, justifying a defender's willingness to resist is easy; he simply expects the challenger to back down. In the Supplemental Appendix we prove that whenever deterrence works in the baseline model, it also works in this variant.

The second robustness check considers a variant of the game in which the challenger is also uncertain about the defender's intrinsic willingness to fight over the transgression, as in classical studies of deterrence. In this extension, Corollary 1 and Lemma 1 hold unaltered;

under the previously stated conditions the deterrence equilibrium always exists. Moreover, introducing even a small possibility that the defender is "intrinsically" willing to fight over the transgression can sometimes uniquely select this equilibrium. Intuitively, the reason is that "deterrence begets deterrence" – more deterrence increases the defender's interim assessment from a transgression that the challenger is unappeasable, makes him more willing to respond with war, generates a higher probability that the transgression will provoke him, and thereby results in more yet deterrence. Examining such "deterrence spirals" could be an interesting avenue for future work.

The third extension allows the defender to have private information about factors that affect the defender's payoff from war as well, such as the probability of victory. This possibility complicates the defender's inference problem upon observing a transgression; he can simultaneously infer that appeasement is less likely to work – making him more willing to fight – and that he would be weak in a war – making him less willing to fight. Nevertheless, Corollary 1 and Lemma 1 continue to hold unaltered with interdependent war values – the deterrence equilibrium always exists whenever appeasement is impossible. However, when this is not the case the equilibrium correspondence is more complex than in the baseline mode and can exhibit new patterns.

Finally, it is natural to wonder whether our insights depend on our simplifications to the bargaining protocol: the defender cannot scale down the size of her transgression, and in the second period the challenger does not make an explicit demand to which the defender can concede rather than fight. Would our logic continue to hold if either assumption were relaxed?

In the Supplemental Appendix we answer both questions in the affirmative. To do so, we consider a continuous and specialized version of the payoff environment. The challenger

and defender bargain and potentially fight over a landmass of size and value equal to 1. The challenger's ex-ante probability of victory in a war slightly exceeds her holdings; thus, there is a very small probability that her cost of war is sufficiently low that she is willing to fight. We assume a continuous mapping from holdings of the landmass to the probability of victory in a war that satisfies the property in Lemma 1 over a large range of values – advancement (slightly) increases the probability of victory more than the payoff from peace. Finally, we consider two variants of the game form. In the first the defender can choose the size of her transgression in the first period, but the sequence is otherwise unchanged. In the second, there is a sequence of exogenous transgressions, each below the defender's threshold for war, that would eventually allow the defender to occupy the entire landmass. In this extension the defender has the final decision in each period of whether to concede or fight.

In both variants, equilibria resembling the deterrence equilibrium exist due to the same logic as in the baseline model. When the demand is endogenous, being able to moderate it does not eliminate the deterrence equilibrium because it remains true that conceding to most demands would increase an already belligerent challenger's payoff from war more than her payoff from peace. Only very large demands can potentially sate a belligerent challenger's thirst for war, and against such demands the defender prefers to fight.

When there are many exogenous demands each below the defender's threshold for war, many equilibria resembling the deterrence equilibrium also exist. There is a final threshold at which the defender is intrinsically willing to fight because she anticipates that a challenger who advances beyond it will exploit salami tactics to eventually posses the entire landmass. At every threshold prior to this point, there exists an equilibrium where the defender is willing to fight due to the logic of our model. These equilibria somewhat resemble those of Powell's analysis of salami tactics in that the defender is willing to fight at the final threshold

for the same reason; however, the strategic rationale sustaining the willingness to fight at earlier thresholds is distinct (Powell 1996).

Finally, we note that this extension is only one possible microfoundation for the potential inevitability of war in the second period. However, as previously discussed many possible rationales for this inevitability can be found in the literature; a game of salami tactics may not be appropriate depending on the application. For example, further potential demands by the challenger could be "lumpy" and above the defender's threshold for war, or the challenger could fail to internalize war's cost or have an irrational preference for war. Alternatively, the challenger may actually be "satiable" and not require deterring to prevent further transgressions beyond a point. Our baseline model is therefore agnostic as to the rationale for the potential inevitability of war in the second period so as not to confuse the main points.

3.7 Conclusion

This paper examines a model of deterrence in which a defender is uncertain about the preferences of a challenger, and his fear that the challenger is unappeasably belligerent can sustain credible deterrence in equilibrium. We demonstrate that this fear can generate credible deterrence even when the probability that a challenger is belligerent is arbitrarily small and the value of the transgression being deterred is small relative to the cost of fighting a war. Unlike most previous analyses of deterrence, our theory does not assume that the defender is sometimes intrinsically willing to fight, or that she has access to commitment devices that help her to do so. Instead, our mechanism relies on the inference that the defender can make from a transgressive act taken in the face of an expectation of war.

After illustrating this simple insight, we derive several empirical implications about when

deterrence will be credible that are previously untested in the empirical deterrence literature. We show that transgressions that make effective "red lines" are not ones that are objectively large, but ones that carry a high military value *relative* to their direct value; the reason is that allowing such transgressions cannot appease an already belligerent challenger. We argue that this insight helps illuminate specific historical episodes of both successful and failed deterrence, such as the Turkish Straits Crisis, the First Offshore Islands Crisis over Quemoy and Matsu, and the Sudetenland Crisis. Finally, we show that events introducing uncertainty about a challenger's intentions can increase the likelihood of deterrence, and that the defender's fear can sometimes benefit her by allowing her to credibly deter at a negligible risk of avoidable wars.

The Cold War record furnishes additional examples of our logic driving states' willingness to respond to a relatively minor transgression with a major war. Senator Richard Russell appears to have used this logic when urging President Kennedy to attack rather than blockade Cuba during the Cuban Missile Crisis. He argued that the Soviet Union had revealed their intention to challenge the United States around the globe, and in all likelihood provoke war, by their willingness to place missiles in Cuba *after Kennedy had warned them* that such an act would carry major consequences (May & Zelikow 1997). Former Secretary of State Dean Acheson used a similar logic, writing that a conventional Soviet attack on Western Europe would provoke a U.S. nuclear response despite the destruction a nuclear war would cause the U.S. homeland, since such an attack would provide "compelling evidence that [the Soviets] had determined to run all risks and force matters to a final showdown, including a nuclear attack upon us" (Acheson 1958, 87).

Can the logic of our model help to explain North Korea's successful deterrence of limited actions by the United States as discussed in the introduction? Given the lack of documentary

evidence on contemporary American and North Korean decision-making, it is impossible to know for sure. Nevertheless, the available evidence suggests that both sides understand that North Korea is using certain actions as a test of the United States' intention to invade. As mentioned previously, Pyongyang's 2003 warning that an air strike on their nuclear plant would lead to "total war" explicitly stated that such an attack would be viewed a precursor to an invasion (KCNA News Agency 2003). Similarly, in recommending an airstrike against a North Korean missile testing site, William Perry and former Assistant Secretary of Defense Ashton Carter wrote that the United States must be careful to warn North Korea that the attack would only be against a specific target, not against their country or their military (Carter & Perry 2006). Special Envoy Jack Pritchard responded that, despite the warning, Pyongyang might very well interpret the air strike as the "start of an effort to bring down [their] regime" (Pritchard 2006). The incentive by North Korea to claim uncertainty about the United States' ultimate intentions, as well as the incentive by the U.S. to claim sharply limited goals, both follow directly from our logic.

These incentives, however, also point to some weaknesses in our analysis and interesting avenues for future work. The vast majority of the previous deterrence literature focuses on things that a defender can *do* – issue cheap talk claims, engage in costly signalling, employ commitment devices, etc. – to improve the credibility of her deterrent threats. Our analysis is different; we examine structural features of the environment outside defender's control that can sustain his credible deterrence in equilibrium – the defender's fear, and the payoff properties of the transgression itself.

The logic of our model and the North Korean case, however, clearly suggest actions that the defender would like to "do" to increase the credibility of her deterrent threat – to claim that he fears the challenger is unappeasably belligerent (even when he does not), and to

claim that he is using the transgression as a test of that belligerence to select the deterrence equilibrium when it exists. Our model, however, is insufficiently rich for such actions to affect equilibrium outcomes. Cheap talk cannot select equilibria, and there is nothing for the defender to signal – either he fears the challenger or he does not.

The history of the deterrence literature, however, suggests a way forward on this modeling conundrum. Classical deterrence theory conceives of the credibility of deterrence as rooted in an intrinsic willingness to fight. In order to understand strategic actions that a defender can take to increase his credibility, subsequent theorizing therefore assumed that a challenger was *uncertain* of that willingness. Our theory, in contrast, conceives of the credibility of deterrence as also rooted in fear; thus, the way forward to understand the previously-described incentives is to assume that the challenger is *uncertain of that fear*. In other words, understanding the incentives suggested by our model requires a modeling approach in which the challenger entertains "higher order uncertainty" about whether a defender actually fears her potential belligerence, or is merely claiming to do so. To our knowledge, no prior models in international relations have considered higher order uncertainty of this form.[25]

Such a modeling approach could potentially eliminate the issue of multiple equilibria presented by the model. Moreover, classical mechanisms for improving the credibility of deterrence could be understood in a new light. Under what conditions could cheap talk about *fear* increase the effectiveness of deterrence? What restrains a defender's willingness to claim that he fears a defender's future intentions? What sorts of costly signals most credibly communicate fear, and can such signals backfire on the defender and result in avoidable wars? Relatedly, how can a challenger credibly communicate limited aims and thus exploit a defender with a known preference for appeasement? Could such claims backfire by making

[25] **?** consider higher order uncertainty about a buyer's valuation for a good in an economic transaction.

her appear sufficiently weak to be exploited herself? To the extent that fear is an important component of credible deterrence, understanding such incentives is an important avenue for future work.

3.8 Appendix

Proof of Proposition 4 The defender's strategy consists of a probability of responding to the transgression with war, which we denote α. The challenger's utility from not transgressing is n_C^1, and from transgressing is $\alpha \cdot w_C^1(\theta_C) + (1-\alpha) \cdot \max\{w_C^2(\theta_C), n_C^2\}$. The latter is strictly increasing in θ_C and greater than n_C^1 for all α when $\theta_C = \bar{\theta}_C^1$. Thus, the challenger's strategy must be to always transgress, or to transgress i.f.f her type is above a cutpoint $\bar{\theta}_C \leq \bar{\theta}_C^1$ at which she is indifferent between transgressing and not.

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

$$P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C\right) = \bar{\beta} \tag{3.2}$$

i.e. the defender's posterior belief that the challenger will initiate war if allowed to transgress is equal to his threshold belief $\bar{\beta}$. The left hand side approaches $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ as $\bar{\theta}_C$ approaches the lower bound of the type space, is equal to 1 at $\bar{\theta}_C = \bar{\theta}_C^2$, and is strictly increasing in between. Thus, a cutpoint satisfying (3.2) exists i.f.f. the no deterrence equilibrium exists ($P\left(\theta_C \geq \bar{\theta}_C^2\right) < \bar{\beta}$). We denote this cutpoint $\bar{\theta}_C^*$, which must be $< \bar{\theta}_C^2$.

We now check conditions such that there exists some $\alpha^* \in (0,1)$ that induces the challenger to play the cutpoint strategy $\bar{\theta}_C^* < \bar{\theta}_C^2$. A necessary condition and sufficient condition

is that this type be indifferent between transgressing and not, i.e. there exists an α^* s.t.

$$\alpha^* \cdot w_C^1\left(\bar{\theta}_C^*\right) + (1-\alpha^*) \cdot n_C^2 = n_C^1. \tag{3.3}$$

If $\bar{\theta}_C^* > \bar{\theta}_C^1 \iff P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right) < \bar{\beta}$ (i.e. the deterrence equilibrium does not exist) then the condition cannot be satisfied since this would imply that both $w_C^1\left(\bar{\theta}_C^*\right)$ and n_C^2 are greater than n_C^1. Conversely, if $\bar{\theta}_C^* < \bar{\theta}_C^1$ then an α^* satisfying (3.3) exists and is unique.

Thus, a unique mixed strategy equilibrium exists i.f.f. both the no deterrence and deterrence equilibria exist, and the equilibrium strategies $\left(\alpha^*, \bar{\theta}_C^*\right)$ are uniquely characterized by (3.2) and (3.3). We now show that when there are multiple equilibria, i.e. $\bar{\beta} \in \left[P\left(\theta_C \geq \bar{\theta}_C^2\right), P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right)\right]$, the defender is strictly better off in the deterrence equilibrium than in either the no deterrence or mixed strategy equilibrium. The defender's utility in the deterrence equilibrium is $U^{de} = P\left(\theta_C \leq \bar{\theta}_C^1\right) \cdot n_D^1 + P\left(\theta_C > \bar{\theta}_C^1\right) \cdot w_D^1$. His utility in the mixed strategy equilibrium is

$$\begin{aligned} U^{ms} &= P\left(\theta_C \leq \bar{\theta}_C^*\right) \cdot n_D^1 + P\left(\theta_C > \bar{\theta}_C^*\right) \cdot \left(P\left(\theta_C \leq \bar{\theta}_C^2 \,|\, \theta_C > \bar{\theta}_C^*\right) \cdot n_D^2 + P\left(\theta_C > \bar{\theta}_C^2 \,|\, \theta_C > \bar{\theta}_C^*\right) \cdot w_D^2\right) \\ &= P\left(\theta_C \leq \bar{\theta}_C^*\right) \cdot n_D^1 + P\left(\theta_C > \bar{\theta}_C^*\right) \cdot w_D^1 \qquad \text{by def'n of } \bar{\theta}_C^*. \end{aligned}$$

This is less than U^{de} since $\bar{\theta}_C^* < \bar{\theta}_C^1$ by construction $\rightarrow P\left(\theta_C \leq \bar{\theta}_C^*\right) < P\left(\theta_C \leq \bar{\theta}_C^1\right)$, and

$n_D^1 > w_D^1$. Finally, his utility in the no deterrence equilibrium is

$$\begin{aligned}
U^{nd} &= P\left(\theta_C \leq \bar{\theta}_C^2\right) \cdot n_D^2 + P\left(\theta_C > \bar{\theta}_C^2\right) \cdot w_D^2. \\
&= P\left(\theta_C \leq \bar{\theta}_C^1\right) \cdot \underbrace{\left(P\left(\theta_C \leq \bar{\theta}_C^2 \mid \theta_C \leq \bar{\theta}_C^1\right) \cdot n_D^2 + P\left(\theta_C > \bar{\theta}_C^2 \mid \theta_C \leq \bar{\theta}_C^1\right) \cdot w_D^2\right)}_{< n_D^1 \text{ since } n_D^1 > n_D^2 > w_D^1 > w_D^2} \\
&\quad + P\left(\theta_C > \bar{\theta}_C^1\right) \cdot \underbrace{\left(P\left(\theta_C \leq \bar{\theta}_C^2 \mid \theta_C > \bar{\theta}_C^1\right) \cdot n_D^2 + P\left(\theta_C > \bar{\theta}_C^2 \mid \theta_C > \bar{\theta}_C^1\right) \cdot w_D^2\right)}_{< w_D^1 \text{ since the deterrence equilibrium exists}} \\
&< P\left(\theta_C \leq \bar{\theta}_C^1\right) \cdot n_D^1 \; + P\left(\theta_C > \bar{\theta}_C^1\right) \cdot w_D^1 = U^{de}
\end{aligned}$$

Proof of Lemma 1

$$\begin{aligned}
\delta_C^m\left(\bar{\theta}_C^1\right) \quad &\geq \quad \delta_C^d \iff w_C^1\left(\bar{\theta}_C^1\right) \geq n_C^1 + \left(\delta_C^d - \delta_C^m\left(\bar{\theta}_C^1\right)\right) \iff w_C^2\left(\bar{\theta}_C^1\right) \geq n_C^2 \\
&\iff \quad \bar{\theta}_C^2 \leq \bar{\theta}_C^1 \iff P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) = 1 \quad \blacksquare
\end{aligned}$$

Proof of Lemma 2 Holding $\bar{\theta}_C^1$ (i.e. the challenger's first period payoffs) fixed, the ineffectiveness of appeasement $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$ is decreasing in $\bar{\theta}_C^2$. Thus for the first claim, it suffices to show $\bar{\theta}_C^2$ is decreasing in $\delta_C^m - \delta_C^d$. By assumption, $n_C^2 = n_C^1 + \delta_C^d$ and $w_C^2\left(\theta_C\right) = w_C^1\left(\theta_C\right) + \delta_C^m$ and $n_C^2 = w_C^2\left(\bar{\theta}_C^2\right)$, which together imply that $n_C^1 = w_C^1\left(\bar{\theta}_C^2\right) + \left(\delta_C^m - \delta_C^d\right)$. This implies the desired property since $w_C^1\left(\theta_C\right)$ is increasing in θ_C.

To show that the probability of deterrence is increasing in $\delta_C^m - \delta_C^d$, note that with the assumed equilibrium selection and by Proposition 4, the probability of deterrence is 0 if $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) < \bar{\beta}$ and $P\left(\theta_C \leq \bar{\theta}_C^1\right)$ otherwise. Holding $\bar{\beta}$ (the defender's payoffs) fixed, the probability of deterrence is therefore (step-wise) increasing in $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$. Since this is increasing in $\delta_C^m - \delta_C^d$ the result is shown. $\blacksquare$

Proof of Lemma 3 If the defender knew the challengers type, then he would respond with war i.f.f. $\theta_C > \bar{\theta}_C^2$, and thus the challenger would be deterred i.f.f. $\theta_C \in \left(\bar{\theta}_C^1, \bar{\theta}_C^2\right)$. The probability of deterrence would therefore be $\max\left\{F\left(\bar{\theta}_C^1\right) - F\left(\bar{\theta}_C^2\right), 0\right\}$. If the deterrence equilibrium exists i.e. $\iff P\left(\theta_C > \bar{\theta}_C^2 \,|\, \theta_C > \bar{\theta}_C^1\right) > \bar{\beta}$, then the probability of deterrence is $F\left(\bar{\theta}_C^1\right) > \max\left\{F\left(\bar{\theta}_C^1\right) - F\left(\bar{\theta}_C^2\right), 0\right\}$. If the deterrence equilibrium does not exist then the probability of deterrence is 0. Since a necessary condition for this is $\bar{\theta}_C^2 > \bar{\theta}_C^1$, the probability of deterrence would also be 0 if the defender knew the challenger's type.

Now, the defender is always better off knowing the challenger's type if appeasement is ineffective $\left(\bar{\theta}_C^2 \leq \bar{\theta}_C^1\right)$; types $< \bar{\theta}_C^2$ are deterred when they otherwise would not be, and for all other types the outcome is identical. If appeasement could be effective $\left(\bar{\theta}_C^1 < \bar{\theta}_C^2\right)$ then she gains $n_D^2 - n_D^1$ against peaceful types $< \bar{\theta}_C^1$ who would have otherwise transgressed, but loses $n_D^2 - w_D^1$ by fighting preventable wars against appeasable types $\theta_C \in \left(\bar{\theta}_C^1, \bar{\theta}_C^2\right)$. Against all other types the outcomes are identical. Thus, in expectation the defender is better off not knowing the challenger's type if and only if

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

i.e. the benefit of deterring types $< \bar{\theta}_C^1$ exceeds the cost of preventable wars $n_C^2 - w_C^1$ against appeasable types. It is straightforward to that show this condition reduces to the condition stated in the Lemma. ■

3.9 Supplemental Appendix

3.9.1 Robustness to challenger backing down

Lemma 4 *Consider an alternative game* Γ' *form in which the challenger can back down in the first stage if the defender resists. Whenever the deterrence equilibrium exists in the original game* Γ *it also exists in* Γ'*.*

Proof: In Γ' the deterrence equilibrium takes the following form; the defender always resists, the challenger is deterred unless she prefers immediate war, and when she transgresses she also fights upon resistance. Now if the defender always resists, challenger types $\theta_C \geq \bar{\theta}_C^1$ still prefer to transgress because the defender will resist and they will then proceed with war. Challenger types $\theta_C < \bar{\theta}_C^1$ cannot get away with the transgression because the defender always resists, can back down upon encountering resistance, and are therefore indifferent between transgressing and not; they are thus willing to play the required strategy of not transgressing. Upon observing a transgression the defender therefore continues to infer that the challenger is of type $\theta_C \geq \bar{\theta}_C^1$, and in this case resisting is equivalent to unilaterally initiating war himself; his incentives and inferences are unchanged and he is therefore willing to carry out his equilibrium strategy. ■

3.9.2 Game with interdependent war values

Both players' payoffs in the event of war depend on the challenger's type $\theta_C \in \Theta \subset \mathbb{R}$ that is unknown to the defender but known to the challenger, where Θ is an interval and θ_C has a prior distribution $f(\theta_C)$ with full support over Θ. The challenger's type is therefore to be interpreted as a state of the world that affects both players' payoffs over which the

challenger has private information. Our notation and assumptions for the challenger's payoffs are unchanged. For the defender, we now express the dependence of his war payoff on the challenger's type using $w_D^t(\theta_C)$, and make the following slightly-modified assumptions.

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

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

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

Challenger Incentives In the second period, the challenger transgresses i.f.f. $\theta_C \geq \bar{\theta}_C^2$. In the first period, challengers of type $\theta_C \geq \bar{\theta}_C^1$ always transgress. Challengers of type $\theta_C < \bar{\theta}_C^1$ transgress i.f.f.,

$$\alpha \cdot w_C^1(\theta_C) + (1-\alpha) \cdot \max\left\{n_C^2, w_C^2(\theta_C)\right\} \geq n_C^1.$$

For each such type, there exists a unique interior probability $\hat{\alpha}(\theta_C)$ that would make them indifferent between transgressing and not, and given that probability the challenger would

play a cutpoint strategy at θ_C. It is simple to verify that for $\theta_C \leq \bar{\theta}_C^1$, $\hat{\alpha}(\theta_C)$ is always well defined, strictly increasing in θ_C, strictly interior to $(0,1)$, and $\hat{\alpha}\left(\bar{\theta}_C^1\right) = 1$.

Defender's Incentives Suppose that the challenger uses a threshold for transgressing equal to $\hat{\theta}_C$. Then upon observing a transgression, the defender's payoff from war is

$$\int_{\hat{\theta}_C}^{\infty} w_D^1(\theta_C) \frac{f(\theta_C)}{1 - F\left(\hat{\theta}_C\right)} d\theta_C$$

and from appeasement is,

$$\int_{\hat{\theta}_C}^{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}} n_D^2 \frac{f(\theta_C)}{1 - F\left(\hat{\theta}_C\right)} d\theta_C + \int_{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}}^{\infty} w_D^2(\theta_C) \cdot \frac{f(\theta_C)}{1 - F\left(\hat{\theta}_C\right)} d\theta_C.$$

Hence she will prefer to respond to the transgression with war i.f.f.

$$\int_{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}}^{\infty} \left(w_D^1(\theta_C) - w_D^2(\theta_C)\right) \cdot \frac{f(\theta_C)}{1 - F\left(\hat{\theta}_C\right)} \geq \int_{\hat{\theta}_C}^{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}} \left(n_D^2 - w_D^1(\theta_C)\right) \frac{f(\theta_C)}{1 - F\left(\hat{\theta}_C\right)} d\theta_C$$

Now it is straightforward to show that the condition above is satisfied i.f.f.

$$\bar{\beta}\left(\hat{\theta}_C\right) \leq P\left(\theta \geq \bar{\theta}_C^2 \,|\, \theta \geq \hat{\theta}_C\right), \tag{3.4}$$

where

$$\bar{\beta}\left(\hat{\theta}_C\right) = \frac{n_D^2 - E\left[w_D^1(\theta_C) \,|\, \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]}{\left(n_D^2 - E\left[w_D^1(\theta_C) \,|\, \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]\right) + E\left[w_D^1(\theta_C) - w_D^2(\theta_C) \,|\, \theta_C \geq \bar{\theta}_C^2\right]} \tag{3.5}$$

Intuitively, $n_D^2 - E\left[w_D^1(\theta_C) \,|\, \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]$ is the benefit from appeasement conditional on

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

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

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

Lemma 5 *Equilibria of the model with interdependent values are as follows.*

- *The deterrence equilibrium exists i.f.f.*

$$\bar{\beta}\left(\bar{\theta}_C^1\right) \leq P\left(\theta \geq \bar{\theta}_C^2 \mid \theta \geq \bar{\theta}_C^1\right)$$

- *The no deterrence equilibrium exists i.f.f.*

$$P\left(\theta_C \geq \bar{\theta}_C^2\right) \leq \bar{\beta}(-\infty)$$

- *A mixed strategy equilibrium in which the challenger uses threshold* $\hat{\theta}_C^* < \min\left\{\bar{\theta}_C^1, \bar{\theta}_C^2\right\}$

exists i.f.f

$$\bar{\beta}\left(\hat{\theta}_C^*\right) = P\left(\theta \geq \bar{\theta}_C^2 \,|\, \theta \geq \hat{\theta}_C^*\right)$$

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

The most important observation from the above characterization is the following: because $\bar{\beta}\left(\hat{\theta}_C\right)$ is interior for all $\hat{\theta}_C$ (meaning that war sooner is better than war later), our basic insight holds unaltered. When appeasement is ineffective $\left(\bar{\theta}_C^2 \leq \bar{\theta}_C^1\right)$, the deterrence equilibrium exists for all distributions over the challenger's type θ_C and functions $w_D^t\left(\theta_C\right)$ mapping the challenger's type into the defender's payoff from war that satisfy the initial assumptions. Thus, Corollary 1 and Lemma 1 continue to hold unaltered with interdependent values.

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

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

ple equilibria: with higher thresholds, the defender can find appeasement less likely to be effective, but simultaneously more desirable if it would be effective.

3.9.3 Game with two-sided uncertainty

The defender is now assumed to have a type θ_D upon which his war payoffs in each period depend, so we write $w_D^t(\theta_D)$ to express this dependence. We maintain the assumption that payoffs in peace for both players are fixed and common knowledge, and make new assumptions on the defender's type that mirror those of the challenger. Specifically, θ_D also belongs to an interval, has some prior distribution $g(\theta_D)$ with full support, and is distributed independently of θ_C. Thus, war values are private and each side's uncertainty may be interpreted as about the opponent's cost of war. We modify the assumptions the defender's payoffs as follows:

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

2. In each period t the defender's war payoff $w_D^t(\theta_D)$ is continuous and strictly increasing in θ_D. In addition, there exists a unique defender type $\bar{\theta}_D^t$ that is indifferent between peace and war in period t.

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

The first assumption extends the properties of the transgression to a setting where the defender's payoffs can vary, and the second mirrors the assumptions made on the challenger's type. Importantly, it implies that with strictly positive probability the defender's threat is

"inherently" credible in that he is willing to go to war solely to prevent the transgression. Formally, for both players let $\bar{\theta}_i^{s,t}$ denote a player indifferent between peace in period s and war in period t – since $n_D^2 < n_D^1$ we have $\bar{\theta}_D^{2,1} < \bar{\theta}_D^{1,1}$ and types in between are willing to fight a war over the transgression.

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

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

Defender's Incentives The defender's war payoffs now depend on her type θ_D; moreover, because types are independent the threshold $\hat{\theta}_C$ that the challenger uses for transgressing only affects her payoffs through the interim assessment β that the challenger would initiate war after being allowed to transgress. He therefore prefers to respond to the transgression

with war when $\beta \geq \bar{\beta}(\theta_D)$, where

$$\bar{\beta}(\theta_D) = \frac{n_D^2 - w_D^1(\theta_D)}{(n_D^2 - w_D^1(\theta_D)) + (w_D^1(\theta_D) - w_D^2(\theta_D))}. \tag{3.6}$$

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

Equilibrium Characterization

Lemma 6 *Equilibria of the model with two-sided uncertainty are as follows.*

- *The deterrence equilibrium exists i.f.f.*

$$\bar{\beta}(-\infty) \leq P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta \geq \bar{\theta}_C^1\right)$$

- *The no deterrence equilibrium exists i.f.f.*

$$P\left(\theta_D \in \left[\bar{\theta}_D\left(P\left(\theta_C \geq \bar{\theta}_C^2\right)\right), \bar{\theta}_D^1\right] \,|\, \theta_D \leq \bar{\theta}_D^1\right) \leq \hat{\alpha}(-\infty)$$

- *An interior equilibrium with challenger threshold $\hat{\theta}_C^* < \min\left\{\bar{\theta}_C^1, \bar{\theta}_C^2\right\}$ exists i.f.f.*

$$P\left(\theta_D \in \left[\bar{\theta}_D\left(P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \hat{\theta}_C^*\right)\right), \bar{\theta}_D^1\right] \,|\, \theta_D \leq \bar{\theta}_D^1\right) = \hat{\alpha}\left(\hat{\theta}_C^*\right)$$

or equivalently

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

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

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

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

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

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

3.9.4 Robustness to Alternative Protocols

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

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

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

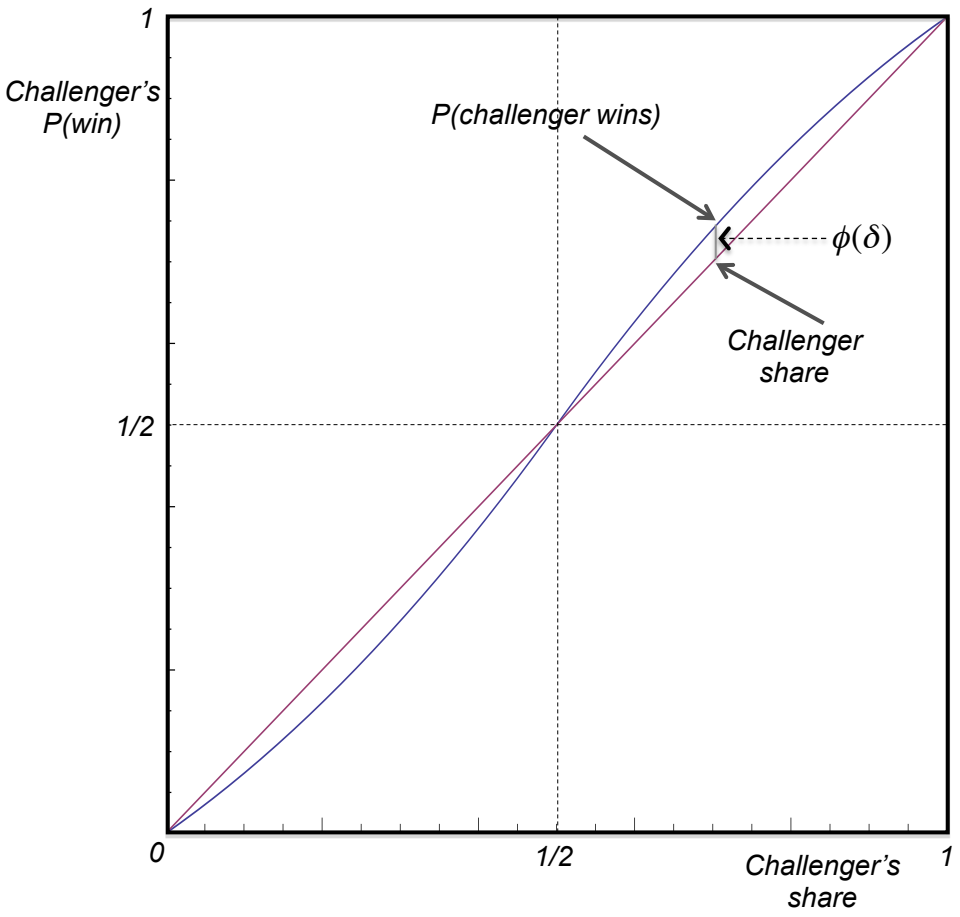
$$p(\delta_t) = \left(\frac{1}{2} + \delta_t\right) + \phi(\delta_t)$$

where $\phi(\delta_t) = \frac{2\delta_t(1-2\delta_t)}{Z}$ *and* Z *is very large. So the advantaged party has a military strength that* ***exceeds*** *her share of the landmass by a very small amount* $\phi(\delta_t)$.[26] *Finally, suppose that the defender's cost of war is commonly known to be* $c_D = \frac{1}{4}$. *The challenger's type* θ_C *is unknown and uniformly distributed over* $\theta_C \sim U\left[-\frac{1}{4}, 0\right]$, *and her cost of war is* $c_C = -\theta_C$.

[26] We require at least $Z > 6$ for $p(\delta_t)$ to be strictly increasing in δ_t.

The challenger's probability of victory in a war as a function of her position is depicted in Figure 3.5. The generalized example captures the three essential features of the original example. First, the advantaged party has a military advantage $\phi(\delta_t) = \frac{2\delta_t(1-2\delta_t)}{Z}$ that exceeds (by a tiny amount) her share of the landmass. Second, the excess military advantage is *nondecreasing* in the challenger's excess holdings over the interval $\left[0, \frac{1}{4}\right]$, so there is a region where allowing further advancement cannot appease an already belligerent challenger. Third, in any period it is very unlikely ex-ante that the challenger is willing to go to war, since this requires $c_C < \frac{2\delta_t(1-2\delta_T)}{Z}$ which occurs with very small probability when Z is large.

Figure 3.5: Probability that Challenger Wins in Generalized Example



We now present the first extension.

Extension 1 (Finite Exogenous Transgressions) *Consider a $T \geq 3$ period game, and a $T+1$-length series of increasing values $0 = \delta_0 < \delta_1 ... < \delta_T = \frac{1}{2}$. Each δ_t represents the challenger's* ***excess*** *holdings above $\frac{1}{2}$ if the game advances to period t. Assume that the challenger is initially advantaged ($\delta_1 > 0$), and that the increments of advancement $\delta_t - \delta_{t-1}$ are less than the defender's cost of war c_D for all $t < T$. In each period t, the challenger decides whether or not to attempt to advance from δ_t to δ_{t+1}. If she doesn't attempt to advance the game ends. If she does attempt to advance, the defender chooses whether or not to respond with war, and the challenger's probability of victory is $p(\delta_t)$.*

In this extension there are a finite number of exogenously fixed positions to which the challenger can advance, each advancement represents a transgression of fixed size, and each increment $\delta_t - \delta_{t-1}$ of advancement short of possessing the entire landmass is less than the defender's cost $c_D = \frac{1}{4}$ of war. Thus, the defender is always vulnerable to "salami tactics." When $\delta_1 < \delta_2 < c_D = \frac{1}{4} < \delta_3$, the game essentially reduces to the baseline model; the reason is that the defender knows she will be unable to credibly resist any advancement beyond δ_2, anticipates that conceding at δ_2 will result in a concession of size $1 - \delta_2 > \frac{1}{4}$, and is therefore there willing to fight.[27]

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

Proposition 5 *For any t^* such that $\delta_{t^*} < c_D = \frac{1}{4}$, there exists an equilibrium in which all types of challengers advance to δ_{t^*}, challengers with cost $c_C < \phi(\delta_{t^*})$ attempt to advance to δ_{t^*+1}, and the defender always responds with war.*

[27] When $\delta_1 < \delta_2 < c_D = \frac{1}{4} < \delta_3$, the game maps to the baseline model by letting the defender's payoffs be $n_D^1 = \frac{1}{2} - \delta_1$, $n_D^2 = \frac{1}{2} - \delta_2$, $w_D^1 = \left(\frac{1}{2} - \delta_1\right) - \phi(\delta_1) - c_D$, $w_D^2 = \left(\frac{1}{2} - \delta_2\right) - \phi(\delta_2) - c_D$, the challenger's payoffs be $n_C^1 = \frac{1}{2} + \delta_1$, $n_C^2 = \frac{1}{2} + \delta_2$, $w_C^1 = \left(\frac{1}{2} + \delta_1\right) + \phi(\delta_1) + \theta_C$, $w_C^2 = \left(\frac{1}{2} + \delta_2\right) + \phi(\delta_2) + \theta_C$, and $\theta_C \sim U\left[-\frac{1}{4}, 0\right]$.

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

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

$$\left(\frac{1}{2} - \delta_t^*\right) - P\left(c_C < \phi\left(\delta_t^*\right)\right)\left(\phi\left(\delta_t^*\right) + c_D\right). \tag{3.7}$$

In words, the defender's equilibrium expected holdings are $\frac{1}{2} - \delta_t^*$, with probability $P\left(c_C < \phi\left(\delta_t^*\right)\right)$ war occurs in period t^*, and when this occurs the defender suffers the challenger's excess military advantage $\phi\left(\delta_t^*\right)$ and the cost of war c_D. If instead the defender responds with war in period t, his payoff is $\left(1 - p\left(\delta_t\right)\right) - c_D = \left(\frac{1}{2} - \delta_t - \phi\left(\delta_t\right)\right) - c_D$, which is $<$ eqn. (3.7) i.f.f.

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

Since $\phi\left(\delta_t\right) \to 0 \; \forall \delta_t$ as $Z \to \infty$, the r.h.s. approaches $\delta_t^* - \delta_t < \frac{1}{4}$ (by assumption) as $Z \to \infty$. So since $c_D = \frac{1}{4}$ there exists a Z sufficiently large such that the inequality is satisfied for all $t < t^*$. Intuitively, we can scale down the excess military advantage function $\phi\left(\delta_t\right)$ by

increasing Z sufficiently so that the calculation essentially reduces to whether the cost of war exceeds the foregone share of the landmass from allowing the challenger to advance from t all the way to t^*. This will always be true since (by assumption) the cost of war exceeds the challenger's excess holdings in the period where war occurs ($\delta_{t^*} < c_D$).

Now consider a period $t \geq \bar{t}$ in which the challenger attempts to advance. This is off path, but we do not need beliefs about the challenger's type since if she is allowed to advance the strategies are for her to continue to advance and the defender to permit it. So if the defender allows advancement in t the challenger will eventually possess the entire landmass and the defender's payoff will be 0. If instead he responds with war his payoff is $\left(\frac{1}{2} - \delta_t - \phi\left(\delta_t\right)\right) - c_D$. Since $\delta_{\bar{t}} < \frac{1}{4}$ and $\delta_t > \frac{1}{4}$ $\forall t > \bar{t}$, for Z sufficiently large it will be optimal for the defender to respond with war in $\bar{t}$ but not in $t > \bar{t}$. In words, at $\bar{t}$ the remaining landmass just exceeds the defender's cost of war, so he will respond with war knowing that should he allow advancement he will also allow it in all future periods. For $t > \bar{t}$, the challenger is already sufficiently advanced that letting her take the remaining landmass is optimal.

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

with war in $\hat{t}$ is optimal, since permitting advancement will only weaken the defender in the inevitable war. ■

Proposition 5 demonstrates that our main result about the ability to deter minor transgressions with a costly war is not an artifact of assuming a single fixed transgression. Specifically, in the generalized example *any* "red line" against further advancement δ_{t^*} that is less than the defender's cost of war c_D can be sustained. This holds even if the fixed increments of advancement $\delta_t - \delta_{t-1}$ are arbitrarily small relative to the defender's cost of war $c_D = \frac{1}{4}$.

The reason is precisely that illustrated in the two-period model. At the equilibrium red line δ_{t^*}, the challenger expects the defender to respond to further advancement (however small) with war, so the defender can infer in equilibrium that a challenger who attempts to advance beyond δ_{t^*} in fact desires war in period t^*. Because the challenger's probability of victory $p(\delta_t)$ is such that advancing makes war *relatively* more attractive when $\delta_t < \frac{1}{4}$, the probability of appeasing an already-belligerent challenger by allowing further advancement beyond the red line is 0. Hence, responding with war at the equilibrium red line δ_{t^*} is optimal.

We now consider the second extension, in which the challenger makes an endogenous "demand" δ_2 of how far to advance. As in the baseline model, in this example the defender can allow a positive demand or respond with war, and if she advances the challenger can exploit her gains afterward by unilaterally initiating war.

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

whether to unilaterally initiate war or enjoy her gains. After either choice the game ends.

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

Proposition 6 *When the challenger's excess share δ_1 under the status quo is less than $\frac{c_D}{2} = \frac{1}{8}$, there exists an equilibrium in which the defender responds to a strictly positive demand $\delta_2 \in \left(\delta_1, \frac{1}{2}\right]$, however small, with war.*

Proof: We construct an equilibrium where all demands are on-path. Challengers with cost $c_C > \phi(\delta_1)$ demand the status quo ($\delta_2^*(c_C) = \delta_1$), it is accepted, they do not initiate war in period 2, and the game ends. All challengers with cost $c_C \leq \phi(\delta_1)$ mix identically over all positive demands $\delta_2 \in \left(\delta_1, \frac{1}{2}\right]$ and the defender always responds with war. Should any such demand be accepted (off path), challengers with cost $c_C < \phi(\delta_2)$ unilaterally initiate war in the second period.

To see this is an equilibrium, consider first the defender's strategy. If he sees no demand ($\delta_2 = \delta_1$), he infers that the challenger will initiate war in the second period with probability 0 and so maintaining the status quo is optimal. Should he see a positive demand ($\delta_2 > \delta_1$), he can infer that the challenger's cost is below $\phi(\delta_1)$ but no more, since all such challengers mix identically over all positive demands. If the demand he receives satisfies $\delta_2 \in \left(\delta_1, \frac{1}{2} - \delta_1\right)$, then $\phi(\delta_1) < \phi(\delta_2)$, and since challengers with cost $c_C < \phi(\delta_2)$ will unilaterally initiate war in period 2, the probability of appeasing an already belligerent challenger by accepting such a demand is 0. Thus responding with war is optimal. If instead $\delta_2 \in \left[\frac{1}{2} - \delta_1, \frac{1}{2}\right]$, then even if allowing the demand would appease the challenger for sure the defender prefers to respond with war, since accepting such a demand will leave the defender with no more than $\delta_1 < \frac{1}{8}$, while responding with war leaves him with $\left(\frac{1}{2} - \delta_1 - \phi(\delta_1)\right) - c_D > \frac{1}{8} - \phi(\delta_1)$ which is $> \frac{1}{8}$ for sufficiently large Z.

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

Proposition 6 further demonstrates that our result is not an artifact of having a fixed size of the transgression. When the status quo division is sufficiently close to an even division, there exists equilibria in which the defender responds to *any* positive demand, however small, with war. The logic is again identical to the two-period model. At the status quo, the challenger expects the defender to respond to any positive demand with war. Hence, the defender can infer in equilibrium that a challenger who makes such a demand desires war under the status quo. Because the challenger's probability of victory $p(\delta_t)$ is such that advancements $\delta_2 \in \left(\delta_1, \frac{1}{2} - \delta_1\right)$ make war relatively more attractive, the probability of appeasing an already-belligerent challenger by permitting such an advancement is 0. Alternatively, while advancements $\delta_2 \in \left[\frac{1}{2} - \delta_1, \frac{1}{2}\right]$ have some hope of successful appeasement, they are so large that the defender prefers to suffer the cost of war.

Chapter 4

War Initiation Before a Domestic Audience

Scholars and practitioners of international relations have both praised and disparaged the effects of democracy on the conduct of foreign affairs. Enthusiasts have argued that public influence over foreign policy, particularly over the government's ability to declare war, restrains excessively belligerent leaders and ensures that policy better serves the interests of the people.[1] Skeptics, on the other hand, have argued that effective decision-making in foreign policy requires information and expertise that the public does not possess, and that requiring public consensus undermines the ability of leaders to make wise and timely decisions.[2] While echoes of the skeptical perspective persist, the enthusiasts perspective dominates the

[1] See, for example, Lake (1992), Russett (1993), and Reiter & Stam (2002).

[2] This perspective forms part of what is known as the "Almond-Lippmann Consensus," named after two prominent skeptics from the immediate post-war period (Holsti 1992). Major statements from these authors can be found in Lippmann (1955) and Almond (1950), and other prominent statements of this view from the same era are Morgenthau & Thompson (1993) and Kennan (1951). Although there are many papers in the contemporary literature that give reasons to be skeptical of the effects of democratic institutions on policy outcomes, there does not seem to be a revival of skepticism specifically about foreign policy-making, even though many of the insights from the contemporary literature should apply (see Ashworth (2012) for review).

contemporary literature on the topic.[3]

Recent work has shown that electoral constraints on democratic officials can distort decision-making, as leaders who are more concerned with influencing public beliefs than choosing good policies engage in pandering and posturing (Ashworth 2012). Following this work, I revisit the debate about democracy and foreign policy. This paper asks whether and when it is possible that democracy in foreign policy decision-making can reduce the public's welfare. I address the question formally by examining the specific scenario where a government may have to decide whether or not to initiate preemptive or preventive war. In that scenario, a government anticipates that its enemy will initiate war in the near future, and it has an incentive to initiate war itself and gain a first-strike advantage. A democratic government may face the disincentive that preempting will result in it being blamed for starting a war by a public interested in restraining excessive belligerence.[4] This creates the possibility that democratic constraints will undermine the government's ability to make prudent decisions in the face of an imminent threat.

To understand government decision-making and public welfare in this scenario, I develop a formal model of war initiation before a domestic audience. The model has two strategic players, states A and B, and a non-strategic audience.[5] A and B have private values for their war payoffs, with higher types being more belligerent. A receives a private signal informing him of B's type, and must then decide whether or not to initiate a war in which he gets a first-strike advantage. If he chooses not to, B must then decide whether to initiate war or end the game peacefully. State A's payoff in the event of war depends not only on his value

[3]Some have also written about the possibility of an emotional public pushing a government into an unwise war, for example Mansfield & Snyder (1995). Still, the bulk of the literature seems to believe that democracy reduces the government's propensity to initiate war more than it promotes it.

[4]See Schweller (1992) and Levy (2008) for further discussion of democracy and preventive war.

[5]I refer to state A as "he" and state B as "she".

for war, but also on the audience's belief about the enemy's belligerence and the necessity of preemption.

I find that public opposition to war initiation is possible in equilibrium, and that it can prevent both wars of aggression by one's own government and necessary preemptive strikes against a belligerent adversary. In the case where an enemy attack is imminent, if democracy is strong and the public has a high threshold for approval of war, the government may knowingly allow itself to be attacked so that the public will become convinced of the enemy's belligerence and support war.

To determine whether democracy leaves the public better or worse off, I analyze the public's welfare for different levels of public influence over its own government. I find that strengthening democracy can improve public welfare by reducing the likelihood of wars of aggression. I also show that, since a democratic public may approve of war initiation when conditions warrant it, a strong democracy does not necessarily imply a hamstrung government.

However, I also identify two scenarios where reducing the influence of the public can improve the public's welfare. First, I identify conditions under which freeing the government to ignore public opposition to war can make the public better off. It is possible that reducing the cost of public opposition to the government can lead to a sudden increase in the willingness of all types of governments to launch necessary preemptive wars at the price of only a small increase in the likelihood of unnecessary aggressive wars.

Second, I find that weakening democracy can also lead the public to change from opposing to approving war initiation in situations where this is in the public's interest. The public may oppose war initiation when it would be better off approving it because of its expectations about its own government's behavior in a disapproval equilibrium. Specifically, the public

may expect that its government will only attack first against peaceful adversaries and will wait to be attacked by belligerent ones. Reducing the influence of the public can free the government to also attack against belligerent adversaries, resulting in an increase in the public's estimate of the adversary's belligerence following an attack, leading to approval of war.

Complementing these findings, I show that both of these scenarios are more likely to exist as the public's initial estimate of the adversary's belligerence increases. These findings show that democracy can reduce public welfare by preventing the government from promptly responding to threats, even though democracy will also have some effect in deterring one's own government from aggressive wars. Furthermore, as the security environment becomes more hostile, insulating the government and the implementation of its policies from public influence becomes more likely to benefit the public. These findings challenge some of the existing enthusiasm about democracy and foreign policy and suggest that analyses of democracy need to recognize that the effect of democracy on foreign policy is contingent on the security environment.

Finally, my analysis produces an interesting and unexpected theoretical finding that I label the *security dilemma reversal*. The traditional "security dilemma" expectation is that the combination of a first-strike advantage and uncertainty about an adversary should make preemptive attacks more likely. However, under the security dilemma reversal, the government will be more likely to initiate war when certain that an adversary is peaceful than when uncertain of the adversary's intentions. When the public opposes preemptive war and the government places a high value on public support, there exists a "second-strike" advantage that can lead countries to prefer waiting for an adversary attack when they believe such a possibility exists.

The paper proceeds as follows. First, I review the relevant literature on democracy, foreign policy and war initiation. Second, I present the model of war initiation before a domestic audience and explore its implications. Third, I analyze the welfare effects of increasing the executive's responsiveness to the audience's opposition to the policy. Fourth, I discuss the security dilemma reversal. Finally, I conclude by reviewing the implications of my findings and providing suggestions for future research.

4.1 Democracy and War

There is a long history of debate over whether democracies are more or less successful than other governments in their conduct of foreign policy. Both perspectives are supported a sizeable body of literature and opinion. One of the major themes running through this literature is that empowering the public will restrain the war-like impulses of government elites. The classic statement comes from Immanuel Kant, who argues that the public is more averse to war than elites because they have to bear the costs (Desch 2008). Similar arguments about the public aversion to war can be found in contemporary works on the democratic peace such as Lake (1992) and Russett (1993). Reiter & Stam (2002) have argued that this popular distaste for war also leads to better war outcomes for democracies, because those governments are only willing to engage in wars with low costs and a high probability of success. The related literature on audience costs has similar findings about democracies' success in foreign policy crises (Partell & Palmer 1999, Gelpi & Griesdorf 2001).

On the other side, critics have written that the public's lack of knowledge and expertise in foreign affairs can slow a government's ability to react to external threats, as the government must first build consensus for action. Classic statements from this perspective come from

as far back as Alexis de Toqueville, John Locke, and even Plato, and were common among American officials and foreign policy writers in the immediate post-war era such as George Kennan, Walter Lippmann and Hans Morgenthau (Desch 2008). This point of view is less commonly represented in contemporary international relations scholarship, though echoes of it can be found in the popular realist literature.[6]

This debate is mirrored in a large literature, mostly applied to American politics, on principal-agent problems between the public and elected officials (see Ashworth (2012) for review). Many of these papers contain the essential elements described here, with leaders who may have divergent preferences from those of their constituents but who may also have better information than their constituents with which to make policy decisions. Some IR papers, mostly examining international institutions, such as Stasavage (2004) and Fang (2008), have taken a similar approach.[7] I follow this approach, modeling a scenario in which leaders may have both divergent preferences and better information.[8]

My model's findings are largely consistent with that literature. When the constraints on the government are weak, the equilibria are "informative" and the government will act according to its interest (Stasavage 2004). In the model, regardless of whether the public approves or opposes war, weak constraints imply more wars, whether aggressive or preemptive. This will be to the benefit of the public when the government's interests are aligned with theirs, since they will not be forced to pander to public opinion and will instead use

[6]Not everyone agrees that these are the dominant effects of democracy. While both of these effects tend to pacify democratic governments, consistent with work by Schweller (1992) on democracy and preventive war, Mansfield & Snyder (1995) argue that young democracies may be less pacific because they are particularly vulnerable to nationalistic appeals. Levy (2008) and Desch (2008) both seem to believe that democracy doesn't have a particular bias toward pacifism or belligerence.

[7]The "two-level games" concept from Putnam (1988) is an informal forerunner of this literature, and was applied to international trade negotiations.

[8]In fact, this class of model shows that the second problem cannot exist without the first, since the need to build consensus would only exist if there was a possibility that the government is making decisions against the public's interest.

their private information to select the best policy. It may not benefit the public when the government is biased, however, because the government may choose the biased policy.

When the public constraints on the government are strong, however, the government's decisions will not reflect its policy preferences as closely, because the government must also take actions to influence public beliefs. The public's influence has both an accountability effect, restraining biased governments, and a distorting effect that can prevent the selection of the best policy (Ashworth 2012). In particular, both unbiased and biased governments may "pander" to the public by knowingly ignoring the information they receive. In the model, this means ignoring the signal and waiting to be attacked. If the government initiates war, the public's belief about the adversary's belligerence may remain low because of the possibility that the war was due to its own government's belligerence. Therefore, the government may wait to be attacked to ensure a significant change in public beliefs about the adversary, at the cost of sacrificing the first-strike advantage. In some cases, this pandering may have the effect of reducing the public's overall welfare (Maskin & Tirole 2004, Fox 2007).

My model differs from previous work in significant ways. There is no negotiating between the states nor any opportunity to hide actions such as bargaining offers (e.g. Stasavage (2004)). Instead, states make a simple war or peace choice that is fully observable by all players. While the issue of the sincerity of pre-war bargaining is important, and has been addressed in informal studies such as Montgomery (2013), I do not address it here.

One important difference is in my focus on public support for the sake of the policy rather than for the career incentives of the elected officials. Most papers in this literature focus on the career concerns of the leaders and the preference of the public to have competent and unbiased representatives (Ashworth 2012). In this paper, the public's belief is about the adversary's belligerence rather than the government's bias, and this belief only matters to

the government's payoff in the case of war. Although I model the public as a non-strategic actor, the best way to think about this is as a model of public effort or mobilization for war. The policy itself requires that the public support it in order for it to be effective, and the public will only support it if it believes the war to be necessary, which is the case when the opponent is believed to be belligerent.

The results in this new scenario are consistent with existing findings. Because a lack of public support reduces the government's payoff for war, the government may forego initiating war when it knows the public will oppose it. This is the accountability mechanism in action, though it operates through the payoff for the policy rather than the value of staying in office. When the government does pander, it does so not to convince the public about its own type, but instead because it wants to give the adversary an opportunity to reveal its type. Still, the effect is the same, since the price the government pays for allowing that revelation is selecting a bad policy.

The model's finding that a government can be prevented from initiating war by public opinion is consistent with work by Schweller (1992) and others that democracies are less likely to initiate preventive war. The alarming result is that this can backfire on the public. It would not be surprising to some to learn that democracies may not always be able to take prompt and timely action because of public opposition. Schuessler (2010), for example, argues that Roosevelt wanted to enter WWII because of the threat he perceived from Germany and Japan, but due to Congressional opposition had to wait until Pearl Harbor, and even attempt to provoke a Japanese or German attack. Israel decided against launching a preemptive strike on the eve of the 1973 October War so that the international community would clearly recognize Egypt as the aggressor, the opposite of the decision Israel had made six years earlier to begin the Six-Day War (Oren 2002, Druckman 2010). This kind of

intentional maneuvering to not fire first is not uncommon at the beginning of wars.

As these examples demonstrate, the issue is not just academic but is relevant for policy-making. In the United States, much of the debate surrounding Congress' war powers has been about the tension between the desire for the legislature to check the executive's power and the legislature's inability to effectively direct foreign affairs (Fisher 2004). This issue also ends up informing policy debates, such as the debate over Congressional approval of the Iraq War.

4.2 Model

4.2.1 Sequence and Payoffs

The model is a sequential game of war initiation with three players: state A, state B, and a non-strategic audience, to be thought of as state A's public.

The game begins with nature assigning types to A and B. Nature assigns to state A the type w_A, randomly drawn from a continuous distribution with finite support on the interval $[0, 1]$, where w_A corresponds to state A's war payoff. Nature also assigns to state B the type w_B, also drawn from a continuous distribution on the interval $[0, 1]$, where w_B corresponds to state B's war payoff. These types are private values for the players. Before the game begins, A receives a private signal $\omega = w_B$ informing him of B's type. This reflects the government's informational advantage over the audience, and the audience remains uncertain about either state's type. In a later section, I modify the signal so that it may be uninformative and also leave A uncertain.

The sequence of the game is straightforward and only involves two moves. A moves first

and chooses either war or peace ($x_1 \in \{W, P\}$). If he chooses war, the game ends in war with A receiving a first-strike advantage. A receives a payoff $V_A = w_A + k$ and B receives $V_B = w_B - k$, where $1 \geq k > 0$ represents the first-strike advantage. If A chooses peace, B moves second and also chooses between either war or peace ($x_2 \in \{W, P\}$). If B chooses war, the game ends in war with the payoffs $V_A = w_A$ and $V_B = w_B$. If B chooses peace, the game ends in peace with the payoffs $V_A = p_A$ and $V_B = p_B$, where $p_i \in [0, 1]$.

If the game ends in war, A also pays an *audience opposition cost* defined by the function $c(\hat{\beta}_t)$, which is based on the audience's beliefs about state B's type. State A only pays this cost in the event of war, and it can be thought of as the public's influence on the implementation of the policy. For example, if an anti-war public is less willing to make sacrifices on behalf of the war effort, or more likely to demand that the war end early, the war payoff will be lower for A. The function is based on B's type because it is B's belligerence that determines, for the audience, whether opposition to preemptive war is warranted.

Because the audience does not observe state A's or B's type or state A's signal, it must generate its belief about state B by observing the play of the game. Define β as the prior probability that state B's war payoff w_B is greater than or equal to state B's peace payoff, p_B, and define $\hat{\beta}_t$ as the audience's estimate that $w_B \geq p_B$ in the event of war, with $t = \{1, 2\}$ indexing which round the war occurs in. If $w_B \geq p_B$, I label B's type as *belligerent*; otherwise, B is *peaceful*. Further, take $\bar{\beta}$ as an exogenously determined parameter between 0 and 1, exclusive. The audience opposition cost takes the following form:

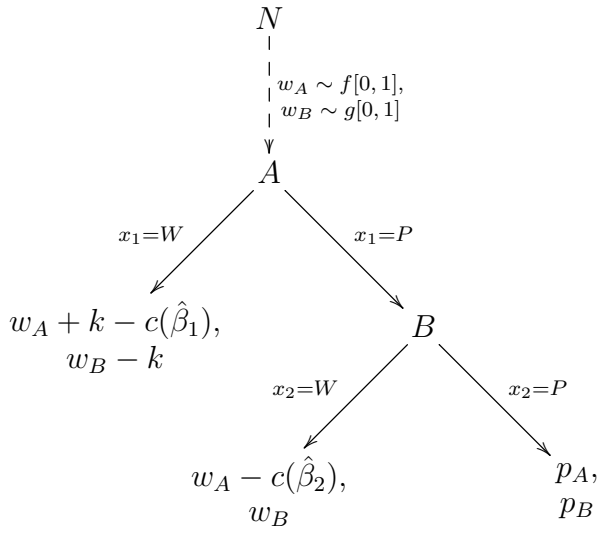
$$c(\hat{\beta}) = \left\{ \begin{array}{ll} 0 & \text{if} \quad \hat{\beta}_t \geq \bar{\beta} \\ c & \text{if} \quad \hat{\beta}_t < \bar{\beta} \end{array} \right\} \tag{4.1}$$

where $1 \geq c > 0$. The assumption here is that the audience opposes war against an

adversary who it believes posed no threat, and only supports a war if it has a sufficiently strong belief that the adversary intended war. In the welfare results section, I make more explicit assumptions about the audience's preferences. State A receives a payoff of $-c$ if the audience's belief about state B's type is below the exogenously given threshold $\bar{\beta}$. Given the restriction $\bar{\beta} \in (0,1)$, $c(1) = 0$ and $c(0) = c$. State A's total utility function is then $U_A = V_A - c(\hat{\beta})$.

The game tree is below.

Figure 4.1: War Initiation Model Game Tree



4.2.2 Equilibria

The equilibrium concept used is Perfect Bayesian Equilibrium. In any *PBE* of this game, state B will follow the same strategy for any possible on- or off-path belief about state A's type. State B chooses war if $w_B \geq p_B$ and peace if $w_B < p_B$ for every possible belief about A's type.[9] These strategies are optimal for all of B's possible beliefs.

[9]In equilibrium, I assume that state B chooses war if indifferent.

Given these strategies, in any equilibrium in which B's decision is on-path, the other actors will believe that B is belligerent with certainty if they observe B choosing war. Therefore, state A can expect audience support ($c(\hat{\beta}_2) = 0$) and a payoff of $U_A = w_A$ in the event that he chooses peace and is subsequently attacked. When B's decision is off-path, however, the other players' beliefs at these outcomes remain unspecified, even though B's off-path strategy will be to choose war if belligerent and peace otherwise.

Here I specify the audience's off-path belief for any Perfect Bayesian Equilibrium of the game. I assume that, in any game where state B's decision is off-path, the audience will believe that state B is belligerent with certainty if the outcome is war ($P(w_B \geq p_B) = 1$) and state B is peaceful with certainty if the outcome is peace ($P(w_B \geq p_B) = 0$). This specification assigns the audience beliefs in the off-path outcomes that are identical to the audience's beliefs for the corresponding on-path outcomes in any Perfect Bayesian Equilibrium.[10]

Given these strategies and beliefs, two possibilities exist for any pure strategy equilibria. The audience could either support or oppose a decision by state A to initiate war in the first round, corresponding to $c(\hat{\beta}_1) = 0$ and $c(\hat{\beta}_1) = c$, respectively. I identify one pure strategy equilibrium in which the audience supports war and two in which the audience opposes war. I first examine the equilibrium in which the audience supports war. (See appendix for all proofs.)

Proposition 7 *There exists a pure-strategy equilibrium in which the audience will not op-*

[10]This does not correspond to any particular equilibrium refinement found in the literature but is reasonable given B's off-path strategy in any *PBE*.

pose state A's decision to initiate war in the first round if

$$\bar{\beta} < \gamma_1 \equiv \frac{\beta}{\beta + (1-\beta)\alpha_1} \tag{4.2}$$

where $\alpha_1 = P(w_A \geq p_A - k)$.

I label this equilibrium *E1*. In this equilibrium, state A will initiate war when he receives a signal that state B prefers war, $\omega \geq p_B$, and, if he is a type $w_A \geq p_A - k$, when state B prefers peace, $\omega < p_B$. Otherwise, state A will choose peace.

Since the audience supports war initiation by its own government, state A has no incentive to wait against a belligerent enemy and will always preempt. A will only choose peace when he prefers peace to initiating war and is certain that B does also. This equilibrium will exist when the audience's threshold for approval $\bar{\beta}$ is below its belief about B's type when A attacks, $\hat{\beta}_1 = \gamma_1$.

There exist two equilibria in which the audience opposes war initiation. In one, state A preempts state B's attacks despite the audience opposition, and in the other, state A never preempts. I begin with the equilibrium in which state A does preempt.

Proposition 8 *There exists a pure-strategy equilibrium in which the audience will oppose A's decision to initiate war in the first round when*

$$\bar{\beta} \geq \gamma_2 \equiv \frac{\beta}{\beta + (1-\beta)\alpha_2} \tag{4.3}$$

and

$$k \geq c \tag{4.4}$$

where $\alpha_2 = P(w_A \geq p_A - k + c)$.

I label this equilibrium *E2*. In this equilibrium, state A will initiate war when he receives a signal that state B prefers war, $\omega \geq p_B$, and, if he is a type $w_A \geq p_A - k + c$, when he receives a signal that state B prefers peace, $\omega < p_B$. Otherwise, state A will choose peace.

In this equilibrium, even though the audience opposes preemptive war, state A preempts because the value of the first strike k is greater than the cost of the audience's opposition c. As in the previous equilibrium, A will only choose peace when he prefers peace to war and is certain that B does, though in this case, aggressive war is less appealing because it carries with it the cost of the audience's opposition. This equilibrium will exist when the audience's threshold for approval $\bar{\beta}$ is above the audience's belief about B's type when A attacks, $\hat{\beta}_1 = \gamma_2$. $\gamma_2 > \gamma_1$, so there is no overlap with equilibrium *E1*.

Finally, I describe the equilibrium in which state A does not preempt due to the opposition of the public.

Proposition 9 *There exists a pure-strategy equilibrium in which the audience will oppose A's decision to initiate war in the first round when*

$$k < c. \tag{4.5}$$

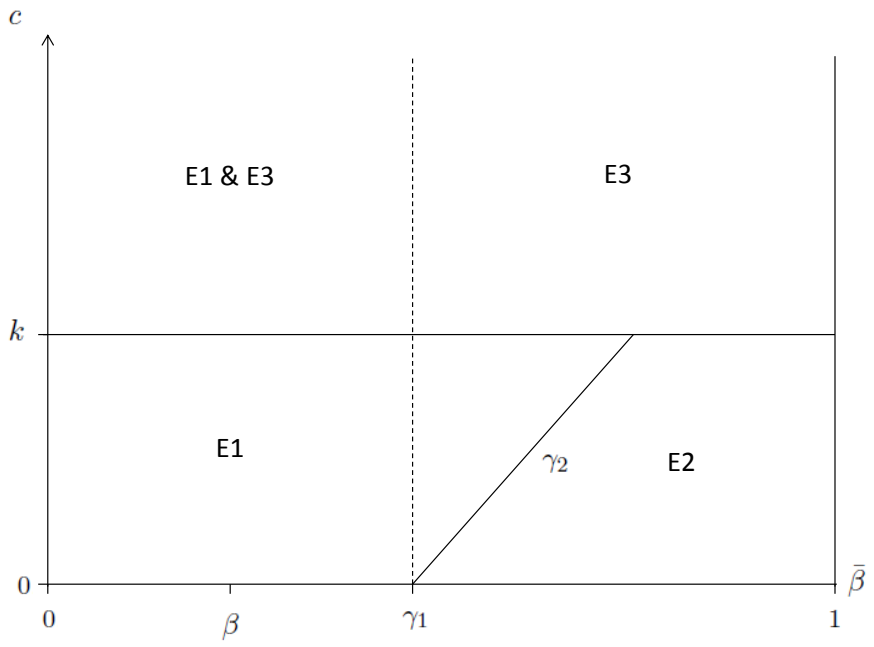
I label this equilibrium *E3*. In this equilibrium, state A will only initiate war when he receives a signal that state B prefers peace, $\omega < p_B$, and he is a type $w_A \geq p_A - k + c$.

In this equilibrium, state A never preempts state B, even when he is certain state B will attack. Instead, he knowingly waits to be attacked in order to win the public's support for war. This does not occur in the other equilibria. For this equilibrium to exist, state A must be willing to sacrifice the first-strike advantage to gain audience support, so $k < c$.

Since state A only initiates war when he is certain that B will not, preferring to wait for B to attack otherwise, the audience's belief when A initiates war is $\hat{\beta}_1 = 0$. Any attack by one's own government will indicate that the enemy is peaceful. This means that the equilibrium can exist for any value of $\bar{\beta}$.

Below, I show the areas within which the equilibria exist, with the cost of audience opposition c on the y-axis and the threshold for audience approval $\bar{\beta}$ on the x-axis. This graph will be a useful reference for the remainder of the paper.

Figure 4.2: War Initiation Equilibria



4.3 Discussion

4.3.1 Equilibrium Comparison

Both approval of and opposition to war initiation are possible in this game depending on which equilibria exist. In *E1*, the audience does not oppose state A initiating war, while in *E2* and *E3*, the audience does oppose state A initiating war. By comparing these equilibria, one can identify the effects of public opposition and democracy on war initiation.

First, public opposition can prevent aggressive war against an adversary that is not belligerent. When state B is peaceful, state A is more likely to attack her when the public approves than when it does not. State A will initiate war against a peaceful state B with probability α_1 in E1 and α_2 in E2 and E3, where $\alpha_1 > \alpha_2$. This is because the cost of the audience's opposition c deters any state A who is the type $p_A - k + c > w_A > p_A - k$. In addition, within *E2* and *E3*, an increase in c leads to a decrease in α_2, so A is less likely to initiate war. This reduces the overall likelihood of war because state B will also choose peace.

Second, public opposition can prevent necessary preemptive attacks against a belligerent adversary when democracy is strong. In *E1*, state A will launch preemptive war when he is certain that state B is belligerent, knowing that the audience will approve. In *E2*, state A will launch preemptive war when he is certain that state B is belligerent, despite the fact that the audience will oppose it. In *E3*, however, state A will intentionally wait to be attacked, knowing that the audience will oppose preemptive war but will not oppose war after being attacked. E3 will only exist when $c > k$, so that state A values public support more than a first-strike advantage.

These findings show that, when a government has better information than its public but

may also be biased, democracy can restraint both belligerent and prudent policies. Given these assumptions about the government's information and preferences, the arguments of both the enthusiasts and skeptics of democracy have merit. These findings are consistent with the literature on government accountability, which says that democracy can create accountability, but also pandering that is against the government's best information. However, it is not immediately clear from this analysis whether increased public influence over the government will benefit or hurt the public. I examine this question in the next section.

4.3.2 Welfare Analysis

To determine whether democracy is beneficial to the public or not, I analyze the welfare of the audience when the constraint on state behavior is stronger and weaker. In order to conduct this analysis, extra assumptions are needed and new variables must be defined to provide structure to the model.

First, I define payoff variables for the audience. Up to this point in the analysis, the audience had no need for payoff functions since it is a non-strategic actor whose beliefs enter directly into state A's payoff function. In order to analyze the audience's welfare, however, the audience must have payoffs for the different possible outcomes. Labelling the audience as player C, I define the audience's peace payoff as p_C and the war payoff as w_C. The audience shares state A's value for a first strike, k. I assume all three are greater than zero and that $p_C > w_C + k$ to ensure the audience prefers peace to initiating war. Otherwise, it would be unnecessary to convince the audience of state B's belligerence to gain support for war initiation.

Second, I define a method for deriving the threshold $\bar{\beta}$ from the audience's payoffs.

The threshold represents the minimum belief about state B's belligerence that the audience requires to approve war. It is reasonable to derive this threshold from the belief at which the audience is indifferent between initiating war and waiting in the hope that the game ends in peace. This means $\bar{\beta}$ can be derived from the equality $w_C + k = \bar{\beta}w_C + (1 - \bar{\beta})p_C$. Rearranging terms, $\bar{\beta}$ is defined as the following

$$\bar{\beta} = \frac{p_C - w_C - k}{p_C - w_C} \tag{4.6}$$

So, for example, if $\bar{\beta} = \beta$, the audience is indifferent between initiating and foregoing war given its prior belief about state B's belligerence. Any increase in its estimate of state B's belligerence above its prior would lead it to support war, and any decrease below the prior would lead it to oppose war.

For each equilibrium, I define the audience's expected payoff as θ_E, with $E \in \{1, 2, 3\}$ representing the different equilibria. The full equations can be found in the appendix. To understand the welfare effects of increasing or decreasing the influence of democracy in foreign affairs, I examine the welfare effects of raising and lowering the variable c, the cost that the audience is able to impose on the government through public opposition. Lowering c can be thought of as insulating the government from public opinion in its ability to implement policy.

Positive Features of Democracy

Strengthening democracy can be welfare-enhancing when it deters aggressive wars without affecting decision-making in the scenarios demanding preemptive war. I find that increasing c can improve welfare within *E2* and *E3*, where the public opposes war.

Proposition 10 *Increasing c increases θ_2 and θ_3, holding all other parameter values constant,*

Within these equilibria, increasing c increases the cost state A pays for initiating war.[11] This reduces α_2, the probability that state A prefers war over peace, and fewer types initiate aggressive war. It has no effect on state A's strategy against a belligerent adversary, which only changes when changing equilibria. This implies that, for some parameter values, strengthening democracy can improve the public's welfare by deterring unnecessary war.

I also find that the audience is better off approving of war when the equilibrium in which the audience approves war, *E1*, overlaps the equilibrium in which the audience opposes war, *E3*. This implies that, if E1 prevails when the two overlap, the audience will approve war initiation when it is in their interest, even in a strong democracy.

Proposition 11 *For every value of $\bar{\beta}$ where E1 and E3 overlap, $\theta_1 > \theta_3$ for any value of c, holding the other variables constant. Since θ_1 is invariable to c, this result also holds for every value of $\bar{\beta}$ where E1 exists even if it does not overlap E3.*

When $\bar{\beta} \leq \gamma_1$, so that *E1* exists, it is always better to be in *E1* than in *E3*, which also exists within that range for $c \geq k$. The benefit from deterring aggressive war will always be outweighed by the benefit of allowing necessary preemptive strikes. $\bar{\beta} \leq \gamma_1$ includes $\bar{\beta} = \beta$, so the audience is better off approving war when it is indifferent between war and peace given its prior beliefs about state B's belligerence. This is because an attack by state A will increase the public's estimate from its prior β to $\bar{\beta} = \gamma_1$.

An immediate implication of this finding is that, even when democracy is strong, it may be in the public's best interest to allow its government the freedom to initiate war. If the

[11] The value θ_1 is invariant to c.

equilibrium *E1* prevails when it overlap *E3*, the audience will do just that. (I explore below what may happen when *E3* prevails, referring back to last part of this proposition). This finding highlights the fact that democracy itself does not necessarily have to be limited for the government to be freed to make prudent policy decisions based on its best information.

Welfare-Enhancing Effects of Restricting Democracy

While democracy can be beneficial to the public, I also identify two scenarios where reducing the influence of the public upon the government's decision-making can actually improve their welfare.

The first scenario is in moving between the two equilibria in which the public opposes war initiation. While increasing c improves welfare within the equilibria *E2* and *E3*, it can also move the audience between equilibria, and the welfare effect of moving between these two equilibria is more complicated. While increasing c can be improved for certain parameter values, it may be worsened for others.

Proposition 12 *There will always exist some* $c' < k$ *where* $\alpha_2 = P(w_A \geq p_A - k + c')$*, and some* $c'' > k$ *where* $\alpha_3 = P(w_A \geq p_A - k + c'')$*, such that* $\theta_2 > \theta_3$ *if*

$$\bar{\beta} < \frac{\beta}{\beta + (1-\beta)(\alpha_2 - \alpha_3)}, \tag{4.7}$$

unless c'' *is large enough that* $\alpha_3 = 0$.

In both *E2* and *E3*, the public will oppose war initiation because it can never be sufficiently convinced of the adversary's belligerence. However, by moving from *E3* to *E2*, governments of all types may become willing to ignore the public's opposition and launch

preemptive war when it is warranted. This can be achieved at the price of only making aggressive war slightly more likely. For example, if $c' = k + \epsilon$ and $c'' = k - \epsilon$, moving from c' to c'' can cause an instant welfare increase of βk, the benefit from preemptive war, at only a very slight increase in the probability of aggressive war. Therefore, for certain values of c' and c'' and $\bar{\beta}$, it may be in the public's interest to reduce their ability to punish the government for initiating war. There will always be some values for which this condition holds unless democracy can be strengthened so much that the probability of aggressive war in *E3* is 0.

The intuition behind this result is the following. When the government anticipates an imminent attack, the relevant tradeoff is between the political benefit of public support and the military benefit of a first strike, since war itself is inevitable. Reducing the public's influence so that the government is willing to preempt will effect the government whether it is belligerent, peaceful, or anything in between. However, when the government doesn't anticipate an attack, the relevant tradeoff is between war and peace. The same reduction in public influence may therefore only effect governments of a certain level of belligerence, who are now willing to suffer public opposition in order to start a war. Governments who were already so belligerent that they wanted war, or who were so peaceful that the public opposition was irrelevant, will still make the same choice between war and peace.

The second scenario is moving from the equilibrium *E3*, in which the audience opposes war, to *E1*, in which the audience approves of war. As mentioned in proposition 11, *E1* and *E3* may overlap, and *E1* will always be preferred when this occurs. Still, E3 will exist, even though the public is better off not restraining its government. Because the audience expects that state A will allow state B to initiate war in *E3*, the audience will infer that state B must be peaceful if it observes state A start the war.

If E3 prevails when the *E1* and *E3* overlap, then it will be in the audience's interest to move into a parameter space where only *E1* exists. As proposition 11 showed, the audience will be better off in *E1* for any given value of $\bar{\beta}$ no matter what the value of the other parameters. Reducing c so that *E3* no longer exists for the relevant parameter values leave only *E1* remaining. In that case, reducing or eliminating democracy becomes a means through which to eliminate the equilibrium in which the public disapproves any war initiation and deters preemptive war.

Intuitively, *E3* can be thought of as an equilibrium where a strong norm of non-aggression exists, and preemption never occurs even when warranted. The government does not preempt because it expects it would be blamed for the war, and the public blames the government because it knows that any war initiation could not possibly be preemptive. Reducing democracy can undermine this norm by eliminating the expectation that the government will wait to be attacked for the sake of public support. By eliminating the norm, the public would then expect that its government would launch preemptive war when warranted, and since the public prefers preemption given its payoffs, it will support its government.

For example, Kimball (2004) writes that the American public had become convinced that American entry into WWII was likely and even advisable, but were still unwilling to have the United States fire the first shot. Perhaps the American people had the understanding that the United States government was determined to allow the enemy to fire the first shot, and was only would only infer that the enemy posed a threat in that circumstances. As it turned out, the enemy did attack and draw the United States into the war. However, measures to insulate the government from the effect of public opinion may have also been a means to escape the equilibrium in which the government was unwilling to launch a preemptive strike and the public was left worse off.

These findings demonstrate that reducing the government's responsiveness to the public can improve the public's welfare in the right circumstances. Next, I show that these circumstances become more relevant when the prior likelihood that state B is belligerent, β, increases.

Proposition 13 *As β increases, γ_1 increases, and E1 exists for a wider range of $\bar{\beta}$. In addition, as β increases, the left-hand side of equation 4.7 increases, and $\theta_2 > \theta_3$ for higher values of the threshold $\bar{\beta}$.*

This proposition can be interpreted as follows. First, as the prior probability that state B is belligerent increases, the range of $\bar{\beta}$ within which *E3* is preferred to *E2* grows smaller. Second, as the prior probability that state B is belligerent increases, the equilibrium in which the audience approves of war will exist for a wider range of $\bar{\beta}$. Remember that *E1* is preferred to *E3* for any $\bar{\beta}$ where *E1* exists, and that therefore reducing the influence of the public can benefit the public's welfare. This implies that increasing the likelihood that the adversary is belligerent increases the range within which democracy can become a hindrance to effective policy.

The other parameter that can be thought of as influencing the "security environment," the value of the first strike k, has more ambiguous effects. Increasing k can make one's own government more belligerent, and make the government less responsive to the public's influence, but also make the public more likely to prefer war when uncertain. Which effect dominates would be reliant on the functional forms assigned to the probability distributions and cannot be estimated given the current specifications.

The broad implication of these findings is that reducing the public's influence and the strength of democracy in foreign affairs can be beneficial to the public in a hostile security

environment, specifically when the adversary is likely to be belligerent. For example, in the case of a country like Israel, where the public is widely convinced that its enemies have belligerent intentions, the public would not only be worse off by not allowing its own government the option of a preemptive strike, but could potentially be made better off by insulating the government from the public's influence altogether.

4.3.3 Security Dilemma Reversal

A final interesting and unexpected condition can be derived from these equilibria that I label the *security dilemma reversal*. To demonstrate this principle, I modify the signaling component of the model. In the modified model, the signal that state A receives is $\omega = w_B$ with probability π or $\omega = \emptyset$ with probability $1 - \pi$. That means that there exists the possibility that state A remains uncertain of the opponent's type and has to make the decision between war or peace given that uncertainty.

Given this modification, there exists an equilibrium that I label *F1* that is equivalent to the equilibrium *E3* in the previous model, the equilibrium in which disapproval deters preemptive strikes. In it, state A does not initiate preemptive war when he receives a signal that state B is belligerent and initiates aggressive war if he receives a signal that state B is peaceful and $w_A \geq p_A - k + c$. These strategies are the same as in *E3*. In addition, state A initiates war if he remains uncertain when $w_A \geq p_A + \frac{c-k}{1-\beta}$. I provide a proof of the existence of this equilibrium in the appendix.

The security dilemma reversal is stated in following proposition:

Proposition 14 *In equilibrium F1, state A is more likely to initiate war when he receives a signal that B is peaceful than when he receives no signal at all, and is more likely to initiate*

war when he receives no signal than when he receives a signal that B is belligerent.

In the other equilibria in this model, state A is more likely to initiate war when he is uncertain of state B's intentions than when he is certain state B is peaceful. He will also preempt with certainty if he receives the signal that state B is belligerent with certainty. This is consistent with traditional international relations concepts like the security dilemma and the spiral theory. A state may be motivated to start a war due to the fear that the adversary has aggressive intentions for the future, but reassurance that the adversary is peaceful would lead the state to choose peace.

However, in *F1*, where the audience opposes war and the value of audience approval is greater than the value of a first strike, the situation is reversed. In this case, it becomes so important for state A to demonstrate state B's belligerence to the audience that he becomes less likely to start the war the more likely he thinks state B is to start the war. State A will be most likely to preempt when he is certain that B has peaceful intentions, and least likely when he is certain that B has belligerent intentions. Rather than a first-strike advantage creating the incentive to preempt, a second-strike advantage creates an incentive to wait. Since some types of state A who would have otherwise preferred war will forego a preemptive attack when uncertain, war can be averted if B turns out to be peaceful.

This is contrary to the standard logic in international relations laid out in arguments about the security dilemma and the spiral model that uncertainty makes preemption more likely (Jervis 1978). This has the positive implication that the need to demonstrate peaceful intentions to one's own public or an international audience can not only make preemptive war less likely, as Reiter (1995) argued, but can eliminate some of the mutual fear of preemptive attack that creates instability in international politics. Just as a strong first-strike advantage

could lead to unnecessary war out of fear of being attacked, a strong second-strike advantage can lead to peace as either belligerent or fearful states decide to wait to be attacked.

4.4 Conclusion

In this paper I examined the effect of a democratic constraint over a government's ability to initiate war. I found that this constraint can both prevent unnecessary wars initiated by an overly belligerent government, but can also prevent necessary wars initiated by a government better informed than about an external threat its public. In particular, it can lead to a government knowingly allowing itself to be attacked in order to win public support. I also found another effect of this constraint called the *security dilemma reversal*, in which a state is less likely to start war when uncertain rather than certain about its adversary's peaceful intentions.

In addition, I examined the welfare effect of increasing the government's responsiveness to public opinion. I found that increasing the government's responsiveness can improve public welfare, but can also leave the public worse off in certain scenarios. Specifically, reducing democracy can improve welfare when the public disapproves of war initiation but would be better off approving, and it can free the government to ignore public opposition and launch preemptive war under certain conditions. Furthermore, these scenarios are more likely to exist the more likely the adversary is to be belligerent.

These results show that democracy's value in foreign policy is contingent on the security environment, and that insulating the leadership from public pressures can improve decision-making. They also show that analyses of the value of democracy in foreign policy-making cannot be conducted without reference to the national security problems facing individual

states. This suggests that conclusions such as those reached by Reiter & Stam (2002) and others about the value of democracy in foreign affairs needs to at least be qualified to take account of the security challenges facing the state.

Furthermore, these carry an implication for studies of democracy and foreign affairs. Since democracy can hinder foreign policy in cases where the security environment is hostile, one may even expect that democracy would be less likely to be adopted in those scenarios. One may expect that democracy is reduced during periods of threat, especially as one of the findings here is that reducing democracy is a means of escaping a suboptimal equilibrium. Gibler (2010), Thompson (1996), and others have made the argument that decentralized decision-making and even democracy itself cannot thrive except in environments of relatively low national security risk. This means that there is likely to be endogeneity in studies that compare the foreign policy behavior of democracies to non-democracies because the external environment may have had an influence on the extent to which foreign policy decision-making in those states is democratized.

Going forward, research on the relationship between democracy and foreign policy may benefit from more detailed analyses understanding the interaction between the national security environment and constraints on the government's ability to direct national security policy, rather than trying to examine the effect of democracy too broadly. This paper points to examining factors like the adversary's likely belligerence and the size of the first-strike advantage.

For those concerned with the concentration of power over decisions of war and peace, it is necessary to recognize that different environments will require striking a different balance between protecting the public from external threat and checking the executive's decision-making power. Understanding the tradeoffs should allow for a better answer to questions

such as whether reform of the United States' current national security apparatus, much of which was developed during the Cold War, is necessary in the current security environment.

4.5 Appendix

4.5.1 Proof of Proposition 1: Equilibrium *E1*

State A's strategies depend on its belief $\hat{\beta}_A$ that $w_B \geq p_B$. A's belief on its turn will depend on the signal it receives.

1. If $\omega = w_B \geq p_B$, then $\hat{\beta}_A = 1$.

2. If $\omega = w_B < p_B$, then $\hat{\beta}_A = 0$.

Assume the following: the threshold $\bar{\beta} \leq \gamma_1$; the audience opposition functions $c(\hat{\beta}_1) = c(\hat{\beta}_2) = 0$; A adopts the strategy $x_1 = W$ when (1) holds, $x_1 = W$ when (2) holds if $w_A \geq p_A - k$ and $x_1 = P$ otherwise; B adopts the strategy $x_2 = W$ when $w_B \geq p_B$ and $x_2 = P$ otherwise.

This forms an equilibrium. B will not deviate because $U_B(w_B) \geq U_B(p_B)$ when $w_B \geq p_B$ and $U_B(w_B) < U_B(p_B)$ when $w_B < p_B$. This is invariant to B's belief about A's type and will hold for any belief about A's type $\hat{\alpha} \in [0, 1]$, so it is *PBE*.

A will not deviate when (1) holds because deviation is followed by $x_2 = W$ and $U_A = w_A$, which is less than the payoff $U_A = w_A + k$ since $k > 0$ by assumption; when (2) holds because deviation is followed by $x_2 = P$ and $U_A = p_A$, which is less than the payoff $U_A = w_A + k$ when $w_A \geq p_A - k$ and greater when $w_A < p_A - k$.

The final step is to show that audience's beliefs sustain the audience opposition functions $c(\hat{\beta}_1) = 0$ and $c(\hat{\beta}_2) = 0$. In the second round, the belief $\hat{\beta}_2 = 1$ always holds because $x_2 = W$ iff $w_B \geq p_B$. The belief $\hat{\beta}_2 = 1$ will also hold if the outcomes are off-path because of the specification of off-path beliefs made above.

Given the strategies above, Bayes Rule tells us that

$$\hat{\beta}_1 = \gamma_1 \equiv \frac{\beta}{\beta + (1-\beta)\alpha_1} \tag{4.8}$$

if state A chooses war.

$c(\hat{\beta}_1) = 0$ if $\bar{\beta} \leq \gamma_1$.

4.5.2 Proof of Proposition 2: Equilibrium *E2*

Assume the following: $k \geq c$; the threshold $\bar{\beta} > \gamma_2$; the audience opposition function $c(\hat{\beta}_1) = c$, and the function $c(\hat{\beta}_2) = 0$; A adopts the strategy $x_1 = W$ when (1) holds, $x_1 = W$ when (2) holds if $w_A \geq p_A - k + c$ and $x_1 = P$ otherwise; B adopts the strategy $x_2 = W$ when $w_B \geq p_B$ and $x_2 = P$ otherwise.

This forms an equilibrium. B will not deviate by the same logic from proposition 1.

A will not deviate when (1) holds because deviation is followed by $x_2 = W$ and $U_A = w_A$, which is less than the payoff $U_A = w_A + k - c$ since $k > c$ by assumption; when (2) holds because deviation is followed by $x_2 = P$ and $U_A = p_A$, which is less than the payoff $U_A = w_A + k - c$ when $w_A \geq p_A - k + c$ and greater when $w_A < p_A - k + c$.

The audience's beliefs sustain the audience opposition functions $c(\hat{\beta}_1) = c$ and $c(\hat{\beta}_2) = 0$. In the second round, the belief $\hat{\beta}_2 = 1$ always holds by the logic in proposition 1.

Given the strategies above, Bayes Rule tells us that

$$\hat{\beta}_1 = \gamma_2 \equiv \frac{\beta}{\beta + (1-\beta)\alpha_2} \tag{4.9}$$

if state A chooses war.

$c(\hat{\beta}_1) = c$ if $\bar{\beta} > \gamma_2$.

4.5.3 Proof of Proposition 3: Equilibrium *E3*

Assume the following: $k < c$; the audience opposition function $c(\hat{\beta}_1) = c$, and the function $c(\hat{\beta}_2) = 0$; A adopts the strategy $x_1 = P$ when (1) holds, $x_1 = W$ when (2) holds if $w_A \geq p_A - k + c$ and $x_1 = P$ otherwise; B adopts the strategy $x_2 = W$ when $w_B \geq p_B$ and $x_2 = P$ otherwise.

This forms an equilibrium. B will not deviate by the same logic from proposition 1.

A will not deviate when (1) holds because deviation gives $U_A = w_A + k - c$, which is less than the payoff $U_A = w_A$ when $k < c$; when (2) holds because deviation is followed by $x_2 = P$ and $U_A = p_A$, which is less than the payoff $U_A = w_A + k - c$ when $w_A \geq p_A - k + c$ and greater when $w_A < p_A - k + c$.

The audience's beliefs sustain the audience opposition functions $c(\hat{\beta}_1) = c$ and $c(\hat{\beta}_2) = 0$. In the second round, the belief $\hat{\beta}_2 = 1$ always holds by the logic in proposition 1.

Given the strategies above, Bayes Rule tells us that

$$\hat{\beta}_1 = 0 \tag{4.10}$$

if state A chooses war.

$c(\hat{\beta}_1) = c$ for all possible values of $\bar{\beta}$.

4.5.4 Proof of Proposition 4: First Welfare Result

The following are the audience welfare values for each equilibrium.

$$\theta_1 = \beta(w_C + k) + (1 - \beta)\alpha_1(w_C + k) + (1 - \beta)(1 - \alpha_1)p_C \tag{4.11}$$

$$\theta_2 = \beta(w_C + k) + (1-\beta)\alpha_2(w_C + k) + (1-\beta)(1-\alpha_2)p_C \tag{4.12}$$

$$\theta_3 = \beta w_C + (1-\beta)\alpha_2(w_C + k) + (1-\beta)(1-\alpha_2)p_C \tag{4.13}$$

$\alpha_2 = P(w_A \geq p - k + c)$, so $\frac{d\alpha_2}{dc} < 0$. Therefore, $\frac{d\theta_1}{dc} = (1-\beta)(w_C + k)\frac{d\alpha_2}{dc} - (1-\beta)p_C\frac{d\alpha_2}{dc}$. Since $p_C > w_C + k$, $\frac{d\theta_2}{dc} > 0$. Therefore, increasing c also increases the public's welfare in *E2*.

The same analysis holds for θ_3.

4.5.5 Proof of Proposition 5: Second Welfare Result

$\theta_1 > \theta_3$ when $w_C + k(\frac{\beta+(1-\beta)(\alpha_1-\alpha_2)}{(1-\beta)(\alpha_1-\alpha_2)}) > p_C$. This is equivalent to

$$\frac{k}{p_C - w_C} > \frac{\beta + (1-\beta)(\alpha_1 - \alpha_2)}{(1-\beta)(\alpha_1 - \alpha_2)}. \tag{4.14}$$

Since $\bar{\beta} < \gamma_1$ when E1 exists, by definition $\frac{p_C - w_C - k}{p_C - w_C} < \frac{\beta}{\beta+(1-\beta)\alpha_1}$. This is equivalent to

$$\frac{k}{p_C - w_C} > \frac{\beta + (1-\beta)(\alpha_1)}{(1-\beta)(\alpha_1)}. \tag{4.15}$$

The right hand side of equation 4.15 is always greater than the right hand side of equation 4.14. This equation can be reduced to $\alpha_2\beta \geq 0$, which always holds because $\alpha_2 \geq 0$ and $\beta \geq 0$. Therefore, if equation 4.15 holds, then equation 4.14 must also hold.

In addition, since the value c only enters into θ_3, even if the value of c is different for the different equilibria, θ_1 will be invariant to changing this value, holding all other variables constant. Therefore, even when equilibria *E1* and *E3* don't overlap, equilibrium *E1* will be preferred to *E3* for a given $\bar{\beta}$ and a given set of payoffs for states A and B.

4.5.6 Proof of Proposition 6: Third Welfare Result

$\theta_2 \geq \theta_3$ if $w_C + k(\frac{\beta+(1-\beta)(\alpha_2-\alpha_3)}{(1-\beta)(\alpha_2-\alpha_3)}) > p_C$, which is equivalent to equation 4.7 given equation 4.6. The right hand side in equation 4.7 is always equal to or greater than γ_2, so some values of $\bar{\beta}$ will exist that equation 4.7 holds for unless $\alpha_3 = 0$.

4.5.7 Proof of Proposition 7: Fourth Welfare Result

$\frac{d\gamma_2}{d\beta} = \frac{\alpha_2}{(\beta+(1-\beta)\alpha_2)^2} > 0$ for all $\beta \in (0,1)$.

$\frac{d}{d\beta} = \frac{\alpha_2-\alpha_3}{(\beta+(1-\beta)(\alpha_2-\alpha_3))^2} > 0$ for all $\beta \in (0,1)$.

4.5.8 Proof of Proposition 8: Security Dilemma Reversal

The equilibrium F1 is identical to the equilibrium E3 in proposition 3, except for A's behavior when it receives the signal $\omega = \emptyset$. When A receives that signal, he chooses war if $w_A \geq p_A + \frac{c-k}{1-\beta}$.

Preemption is least likely when A is certain that B is belligerent, more likely when A is uncertain of B's type, and most likely when A is certain that B is peaceful. When $\omega \geq p_A$, the probability of preemption is 0. When $\omega = \emptyset$, the probability of preemption is $P(w_A \geq p_A + \frac{c-k}{1-\beta})$. When $\omega < p_A$, the probability of preemption is $P(w_A \geq p_A + c - k)$. Because $c \geq k$ in this equilibrium, $p_A + \frac{c-k}{1-\beta} > w_A \geq p_A + c - k$ and $P(w_A \geq p_A + \frac{c-k}{1-\beta}) < P(w_A \geq p_A + c - k)$.

Chapter 5

Conclusion

The primary focus of the crisis bargaining literature has been on how states demostrate resolve or otherwise manage the opponent's expectations.[1] Even theories that incorporate third parties have done so largely in order to examine the effect of these parties on signaling to an opponent (Fearon 1994, Schultz 1998). This dissertation begins with the premise that certain crisis behaviors cannot be understood in this traditional crisis bargaining framework in which states are only concerned with signaling to an adversary.

A number of behaviors have been observed which are better understood as being for the purpose of influencing a domestic audience. There are a wide range of these behaviors. They include provocations, exaggerations of incidents, waiting for enemy attacks, false flag operations, explaining policies to the public, "counterfeit" diplomacy, and seeking the approval of international institutions. While some of these behaviors have been studied, the crisis bargaining literature has not been systematically extended to incorporate behaviors aimed at third party audiences.

[1] See Kurizaki (2007) and Slantchev (2010) for papers about managing expectations without necessarily demonstrating resolve.

I began to fill this gap by examining a family of strategies oriented toward influencing the public's beliefs about the adversary's belligerence: provocations, exaggerations and the forgoing of preemptive attacks. All three of these behaviors have been observed before war initiation, and all three appear to be aimed at winning public support for war.

In chapter two, I examined the strategy of provocations. I modeled a crisis bargaining and war initiation scenario in which a legislature had to authorize war. I showed that this introduces an incentive for the adversary to practice restraint, even in the face of provocations by the government. I found that it is possible for an adversary to let go of its restraint and be baited into attacking, even if it desires to avoid war, if it believes that the provocative action might be revealed. One implication of this is that increasing transparency makes the strategy of provocation less likely to succeed.

To test the findings of this model, first I conducted an online survey experiment. I surveyed 1,000 respondents about their approval for military action with Iran. I found that respondents were much more likely to approve military action following an Iranian attack, but that this effect could be mitigated if the attack was revealed to have been a response to an American provocation. These findings are consistent with the model's. I also conducted a historical case study of the German and Japanese reaction to American provocations before World War II. I found that both states attempted to practice restraint. I showed that, when Germany failed to practice restraint, the American provocation was revealed and the United States government was unable to successfully seize upon the incident. I also argue that Japan was motivated to attack the United States partly in the hope that it would be recognized that the United States forced them into it.

In chapter three, we examined exaggerations in the form of disproportionate responses to minor incidents. We modeled a crisis scenario in which a defender was uncertain of

the aggressor's intentions. We showed that, in equilibrium, even a minor transgression could reveal far-reaching ambitions by the aggressor, justifying a response of major war by the defender. We also showed that the difference between the transgression's military and inherent value were determining factors for the effectiveness of deterrence against the transgression. The implication of these findings is that minor incidents can legitimately trigger major war, and are may not only be used as excuses to implement policies already decided upon.

We conducted a historical case study of the Turkish Straits Crisis. We found that the mechanism identified in our model can help explain how the United States was able to successfully deter the Soviet Union from invading Turkey. Most importantly, we show that the United States was prepared to infer far-reaching Soviet ambitions from a willingness of the Soviet Union to start a major war for the sake of subjugating Turkey.

In chapter four, I examined the strategy of preemption when there is a domestic audience. I modeled a scenario in which a government may anticipate war and want to preempt, but also has to convince the public to support the war. I found that, in equilibrium, the government may allow the enemy to attack so that the public becomes convinced of the enemy's belligerence. I also show that this can be welfare-reducing for the public in certain scenarios, and that these scenarios are more likely to exist when the adversary is more likely to be belligerent. The implication is that weakening democracy can be in the public's interest in certain security environments.

5.1 Implications

These studies carry major implications for the study of international crises and crisis bargaining. The first implication is that, in international crises, demonstrating resolve to the adversary may not be the only or even the primary objective. Dealing with domestic political constraints may be just as important as managing enemy behaviors and expectations, and strategies adopted in international crises may be more significant for their influence on public beliefs.

Observers should therefore not always understand statements or actions in crises as sincere. Often, states take actions or make statements that are intended for domestic consumption. Sometimes, crises themselves are provoked or drawn out for the sole purpose of winning public support, and they serve no purpose of signaling resolve to an adversary. This may modify our understanding of historical crises and our understanding of crises as they are occurring.

This should also influence analysis and advice given to leaders about how to manage crises. Any analysis of bargaining that only accounts for the signals sent to the opponent cannot realistically provide guidance to political leaders. Many of the papers about signaling resolve would seem to suggest that the key to prevailing in a crisis is in winning the competition of escalation and demonstrating resolve. This may require large demands, the prolongation of a crisis, the mobilization of troops and deployment of forces, the escalation of risk of an accident, and commiting oneself through uncompromising public pronouncements.

But actions meant to increase one's international bargaining power may not be advisable if the state must also mobilize public support by shifting blame for the impasse to the opponent. Schelling (1966) recognized this implicitly when writing that acting crazy can be an

asset in the international arena, but that the United States government cannot be expected to behave in such a seemingly irresponsible manner. The chapters of this dissertation suggest that states may do better by escalating crises more slowly, demonstrating a willingness to reach a negotiated compromise, and letting belligerent enemies expose themselves. States must also be ready to engage in diplomacy, whether through negotiations with enemy governments, international conferences, or the United Nations, even if they don't believe that these are in their interests.

A better understanding of enemy behaviors should also lead to better advice for leaders. Often, distinguishing between statements and behaviors meant to send signals to other governments and statements and behaviors meant for domestic consumption is difficult for leaders. An understanding of crisis behavior that incorporates a domestic audience should allow leaders to better understand the true intentions of enemy governments and devise strategies catered to those intentions. It may even help leaders devise strategies that assist enemy governments in dealing with their populations or internal political problems when this is in their interest.

The second implication is that democracy may not always influence crises positively. The contemporary consensus is that democracy improves state performance in international crises. The audience cost and opposition signaling literature both show that the participation of a domestic audience in crisis bargaining can increase the credibility of threats (Fearon 1994, Schultz 1998). These insights have been used in studies that show that democracies tend to perform better in international crises (Partell & Palmer 1999, Gelpi & Griesdorf 2001).

In this dissertation, I found that democracy may not always improve state performance or public welfare in international crises. The need to win public support may lead states to take deceptive actions, such as attempting to provoke crises as a pretext for war. Increasing

transparency and taking steps to ensure that the government's preferences are more congruent with the public's do reduce these incentives, so strengthening democracy can prevent these kinds of practices. However, I also found that states may forego prudent national security decisions, particularly initiating preemptive or preventive war, in order to win public support. Again, strengthening democracy can improve the public's welfare by preventing excessive belligerence, but there are scenarios in which weakening democracy can improve the public's welfare. Weakening democracy can increase the willingness of all types of governments to ignore public opposition and launch preemptive wars when they are warranted. Under the right conditions, this can occur at a very small cost of only slightly increasing the probability of aggressive war. Weakening democracy can also help the public escape a scenario in which their government doesn't initiate war because it anticipates being blamed even though the public is better off approving war initiation.

Both of these scenarios are more likely to exist the more belligerent the adversary is. When the security environment is such that the government may need to make prompt policy decisions based on information unavailable to the public, insulating the government from the need to build a public consensus can produce better policy outcomes. Therefore, it can be said that the value of democracy is contingent on the national security environment. The contingent nature of the value of democracy is an important thing to understand, especially for contemporary studies which often compare democracies to autocracies without controlling for important factors which may make one regime type more or less advantageous than the other.

A further implication of this is that the existence of democracy may also be contingent on the security environment. Considering the fact that democracy can undermine national security policy in hostile environments, and that this can actually harm the public's welfare,

it would not be surprising if countries under threat were less likely to be democratic than countries that didn't face major threats. Gibler (2010), Thompson (1996), and others have made similar arguments in different contexts. This endogeneity further complicates studies which compare the foreign policy performance of democracies and autocracies.

What this points to is that the relationship between regime type and security environment needs to be examined more closely. Regime types will perform differently when facing different security challenges, and the security environment will have an effect on whether certain regime types can survive or not. Understanding these relationships better would better inform debates in the United States about the concentration of power in the executive and the war powers of the Congress. Simply arguing that we need more or less democracy without reference to the particular security challenges facing the United States is probably not helpful. However, being to evaluate how the security environment effects the need for the government to be either held accountable or made more insulated from political pressure would be useful. This is particularly true in the current post-Cold War world, where many of the government's national security powers have been inhereted from the Cold War.

The final implication of this dissertation is about how the public should understand crises events, how they should evaluate certain acts by their own government and foreign governments, and specifically whether they should place importance on the question of which side fired first. I have shown that the state that fires or appears to fire first is not always the state that can legimitately be labeled the "aggressor." States can be provoked into military incidents, incidents which should be treated as isolated events but are exaggerated into signifying major acts of aggression and precursors to war. States may also launch preemptive attacks. While this is not news for international relations scholars, I have explored the consequences for this tactic when the public may infer that a preemptive attack was an

act of aggression. I have shown that, in that situation specifically, labeling the state that fires first as the "aggressor" can backfire on the public. While the public may be genuinely interested in restraining excessive belligerence by its own government, it may instead prevent its own government from making prompt and prudent national security decisions.

On the other hand, the main implication of this dissertation is that judging crisis events is complicated and inherently uncertain. When faced with incomplete information, observers cannot simply ignore attacks in making their judgements. Instead, they face genuine dilemmas created by a lack of information. This dissertation has shown that a relatively minor attack can, in fact, signal dangerously aggressive intentions by an enemy. The public cannot simply ignore those attacks or dismiss them as isolated incidents. If uncertain about the nature of the attack, the public or the legislature may also face the genuine dilemma of deciding whether or not to respond promptly to an attack and risk authorizing an unnecessary war, or deciding whether to restrict the government's war powers and risk weakening the government's war effort. This basic dilemma is the same whether the public is deciding how to respond to an incident, as shown in chapter two, or whether the public is simply deciding whether or not to allow its government the freedom to initiate war, as shown in chapter four.

Faced with these dilemmas, the solution to the challenge of judging crisis events must lie in institutional changes. In this case, they are the same institutional changes discussed earlier as serving public welfare under democracy. Increased transparency and government accountability can prevent deception, while insulating the government from the public can, under certain circumstances, either reduce the impact of the public's uninformed judgments on policy or even change the public's judgment from disapproval to approval of war initiation when this is in the public's interest.

5.2 Further Research

A great deal more research that needs to be done about international negotiations and crisis bargaining that incorporates the role of the domestic audience. I mentioned a number of phenomena that are clearly aimed at influencing domestic audiences that have not been studied systematically or formally. These include "false flag" operations, "counterfeit diplomacy" and even the simple act of verbally justifying policies to the public.

Other crisis behaviors need to also be understood in terms of the role of the domestic audience. States will often negotiate behind closed doors, or even keep concessions secret, despite the fact that the negotiated settlement the two sides agreed to would be preferred to war by the majority of the public. Kennedy famously removed U.S. missiles from Turkey as a concession in the Cuban Missile Crisis and told the Soviet government that keeping this fact a secret was a condition of the deal, even though the public would have surely preferred to make that concession rather than going to war.

Much has also been written about states following codes of honor in international negotiations and ensuring that enemies can make concessions while "saving face" (O'Neill 1999). This leads governments to use diplomatic language, to be careful not to escalate crises, to make offers and concessions secretly, and can also trap them in scenarios where they feel war is necessary because of honor considerations when they would otherwise wish to avoid fighting. It is likely that many of these strategic choices can be understood in terms of influencing public beliefs.

Going forward, incorporating third parties into formal analyses of international behavior may prove to be a productive endeavor. Bargaining models with third parties that examine issues like secrecy, diplomatic language and face-saving concessions may shed light on actual

crisis behavior that cannot be understood simply by considering the interactions of two unitary state actors.

Bibliography

Acheson, Dean. 1958. *Power and Diplomacy.* Cambridge, Mass.: Harvard University Press.

Agawa, Hiroyuki. 1979. *The Reluctant Admiral: Yamamoto and the Imperial Navy.* Tokyo: Kodanasha International Ltd.

Albertini, Luigi. 1967. *The Origins of the War of 1914, Vol. 3.* London: Oxford University Press.

Allen, Martin. 2005. *Himmler's Secret War.* London: Robson.

Allison, Graham. 1969. "Conceptual Models and the Cuban Missile Crisis." *American Political Science Review* 63(3):689–718.

Almond, Gabriel A. 1950. *The American People and Foreign Policy.* New York: Harcourt Brace.

Alt, James, Randall Calvert & Brian Humes. 1988. "Reputation and Hegemonic Stability: A Game-Theoretic Analysis." *American Political Science Review* 82(2):445–466.

Ashworth, Scott. 2012. "Electoral Accountability: Recent Theoretical and Empirical Work." *Annual Review of Political Science* 15:183–201.

Bailey, Thomas A. & Paul B. Ryan. 1979. *Hitler vs. Roosevelt: The Undeclared Naval War*. New York: The Free Press.

Baker, William & John R. ONeal. 2001. "Patriotism or Opinion Leadership?: The Nature and Origins of the 'Rally Round the Flag' Effect." *Journal of Conflict Resolution* 45(4):661–687.

Baliga, Sandeep & Tomas Sjostrom. 2008. "Strategic Ambiguity and Arms Proliferation." *Journal of Political Economy* 116(6):1023–1057.

Bamford, James. 2001. *Body of Secrets: Anatomy of the Ultra-Secret National Security Agency from the Cold War Through the Dawn of a New Century*. New York: Doubleday.

Brody, Richard A. & Catherine R. Shapiro. 1991. The Rally Phenomenon in Public Opinion. In *Assessing the President: The Media, Elite Opinion and Public Support*, ed. Richard A. Brody. Stanford, CA: Stanford University Press.

Bueno de Mesquita, Bruce, James D. Morrow & Ethan R. Zorick. 1997. "Capabilities, Perception and Escalation." *American Political Science Review* 91(1):15–27.

Buhite, Russell D. & David W. Levy. 1992. *FDR's Fireside Chats*. Norman, OK: University of Oklahoma Press.

Butler, Steven R. 1995. *A Documentary History of the Mexican War*. Richardson, TX: Descendents of Mexican War Veterans.

Callaghan, Karen J. & Simo Virtanen. 1993. "Revised Models of the Rally Phenomenon: The Case of the Carter Presidency." *The Journal of Politics* 55(3):756–764.

Canes-Wrone, Brandice, Michael C. Herron & Kenneth W. Shotts. 2001. "Leadership and Pandering: A Theory of Executive Policymaking." *American Journal of Political Science* 45(3):532–550.

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

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

Chang, Gordon H. 1988. "To the Nuclear Brink: Eisenhower, Dulles and the Quemoy-Matsu Crisis." *International Security* 12(4):96–122.

Chapman, Terrence L. 2007. "International Security Institutions, Domestic Politics, and Institutional Legitimacy." *Journal of Conflict Resolution* 51(1):134–166.

Chiozza, Giacomo & H.E. Goemans. 2004. "International Conflict and the Tenure of Leaders: Is War Still 'Ex Post' Inefficient?" *American Journal of Political Science* 48(3):604–619.

Cole, Wayne S. 1983. *Roosevelt and the Isolationists, 1932-1945*. Lincoln, NE: University of Nebraska Press.

Dallek, Robert. 1979. *Franklin D. Roosevelt and American Foreign Policy, 1932-1945*. New York: Oxford University Press.

Danilovic, Vesna. 2001. "The Sources of Threat Credibility in Extended Deterrence." *The Journal of Conflict Resolution* 45(3):341–369.

Desch, Michael C. 2008. *Power and Military Effectiveness: The Fallacy of Democratic Triumphalism*. Baltimore, MD: The Johns Hopkins University Press.

Downs, George W. & David M. Rocke. 1994. "Conflict, Agency and Gambling for Resurrection: The Principal-Agent Problem Goes to War." *American Journal of Political Science* 38(2):362–380.

Druckman, Yaron. 2010. "Morning of Yom Kippur War: Cabinet rejects call for preemptive strike." *Yediot Aharonot*, 10 June 2010. Accessed from http://www.ynetnews.com/articles/0,7340,L-3965041,00.html.

Edwards, Robert. 2006. *White Death: Russia's War on Finland, 1939-1940*. London: Weidenfeld and Nicolson.

Fang, Songying. 2008. "The Informational Role of International Institutions and Domestic Politics." *American Journal of Political Science* 52(2):304–321.

Fearon, James. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(3):577–592.

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

Fearon, James. 1996. "Bargaining over Objects that Influence Future Bargaining Power." Unpublished Manuscript, Chicago, IL.

Fearon, James. 1998. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

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

Fisher, Louis. 2004. *Presidential War Power, Second Edition Revised.* Lawrence, KS: University Press of Kansas.

Fox, Justin. 2007. "Government Transparency and Policymaking." *Public Choice* 131:23–44.

Gelpi, Christopher & Michael Griesdorf. 2001. "Winners or Losers?: Democracies in International Crises, 1918-1994." *American Political Science Review* 95(3):633–647.

Gibler, Douglas M. 2010. "Outside-In: The Effects of External Threats on State Centralization." *Journal of Conflict Resolution* 54(4):519–542.

Glaser, Charles. 1997. "The Security Dilemma Revisited." *World Politics* 50(1):171–201.

Grew, Joseph. 1944. *Ten Years in Japan: A Contemporary Record Drawn from the Diaries and Private and Official Papers of Joseph C. Grew, United States Ambassador to Japan, 1932-1942.* New York: Simon and Schuster.

Hearden, Patrick J. 1987. *Roosevelt Confronts Hitler: America's Entry into World War II.* DeKalb, IL: Northern Illinois University Press.

Hetherington, Marc J. & Michael Nelson. 2003. "Anatomy of a Rally Effect: George W. Bush and the War on Terrorism." *PS: Political Science and Politics* 36(1):37–42.

Holsti, Ole R. 1992. "Public Opinion and Foreign Policy: Challenges to the Almond-Lippmann Consensus." *International Studies Quarterly* 36(4):439–466.

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

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

Ike, Nobutaka, ed. 1967. *Japan's Decision for War: Records of the 1941 Policy Conferences*. Stanford, CA: Stanford University Press.

International Military Tribunal for the Far East. 1948. "Court Papers, Journal, Exhibits and Judgements of the International Military Tribunal for the Far East.". United States Government, Washington D.C. Microfilm reel 18.

Jackson, Matthew & Massimo Morelli. 2007. "Political Bias and War." *American Economic Review* 97(4):1353–1373.

Jentleson, Bruce. 1992. "The Pretty Prudent Public: Post Post-Vietnam American Opinion on the Use of Military Force." *International Studies Quarterly* 36(1):49–73.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton, NJ: Princeton University Press.

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30(2):167–214.

Jun, Tsunoda. 1994. Leaning Toward War. In *The Final Confrontation: Japan's Negotiations with the United States, 1941*, ed. David Morley. New York: Columbia University Press.

KCNA News Agency. 2003. "North Korea warns 'total war' if USA attacks 'peaceful' plant." BBC Summary of World Broadcasts, February 6, 2003. Lexis-Nexis Academic: News. Available from ⟨ http://www.lexisnexis.com/us/lnacademic ⟩. Accessed April 10, 2010.

Kennan, George F. 1951. *American Diplomacy, 1900-1950.* Chicago: University of Chicago Press.

Kimball, Warren F. 2004. "Franklin D. Roosevelt and World War II." *Presidential Studies Quarterly* 34(1):83–99.

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

Kurizaki, Shuhei. 2007. "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." *American Political Science Review* 101(3):543–558.

Kydd, Andrew. 1997. "Sheep in Sheep's Clothing: Why Security Seekers Do Not Fight Each Other." *Security Studies* 7(1):114–155.

Lake, David A. 1992. "Powerful Pacifists: Democratic States and War." *American Political Science Review* 86(1):24–37.

Langer, William L. & S. Everett Gleason. 1953. *The Undeclared War, 1940-1941.* New York: Harpers and Brothers Publishers.

Levy, Jack S. 1987. "Declining Power and the Preventive Motivation for War." *World Politics* 40(1):82–107.

Levy, Jack S. 1998. "The Causes of War and the Conditions of Peace." *Annual Review of Political Science* 1:139–165.

Levy, Jack S. 2008. "Preventive War and Democratic Politics." *International Studies Quarterly* 52:1–24.

Lippmann, Walter. 1955. *Essays in the Public Philosophy.* Boston, MA: Little Brown.

Luard, Evan. 1986. *War in International Society.* London: I.B. Tauris and Co. Ltd.

Mansfield, Edward D. & Jack Snyder. 1995. "Democratization and the Danger of War." *International Security* 20(1):5–38.

Mark, Eduard. 1997. "The War Scare of 1946 and its Consequences." *Diplomatic History* 21(3):383–415.

Maskin, Eric & Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94(4):1034–1054.

May, Ernest R. & Philip D. Zelikow, eds. 1997. *The Kennedy Tapes: Inside the White House During the Cuban Missile Crisis.* Cambridge, MA: Belknap Press.

Mearsheimer, John J. 2001. *The Tragedy of Great Power Politics.* New York: W.W. Norton.

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

Moise, Edwin E. 1996. *The Tonkin Gulf and the Escalation of the Vietnam War.* Chapel Hill, NC: University of North Carolina Press.

Montgomery, Evan Braden. 2013. "Counterfeit Diplomacy and Mobilization in Democracies." *Security Studies* 22:33–67.

Morgenthau, Hans & Kenneth Thompson. 1993. *Politics Among Nations, 6th ed.* New York: McGraw-Hill.

Mueller, John. 1973. *War, Presidents and Public Opinion.* New York: Wiley.

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

NSC 173. 1953. "US Policy to Counter Possible Soviet or Satellite Action against Berlin." Issued Dec. 1, 1953, declassified May 19, 1981. Reproduced in Declassified Documents Reference System. Farmington Hills, MI: Gale, 2010.

O'Neill, Barry. 1999. *Honor, Symbols and War.* Ann Arbor, MI: University of Michigan Press.

Oren, Michael B. 2002. *Six Days of War: June 1967 and the Making of the Modern Middle East.* New York: Oxford University Press.

Parker, Suzanne L. 1995. "Toward an Understanding of the 'Rally' Effect." *The Public Opinion Quarterly* 59(4):526–546.

Partell, Peter J. & Glenn Palmer. 1999. "Audience Costs and Interstate Crises: An Empirical Assessment of Fearon's Model of Dispute Outcomes." *International Studies Quarterly* 43(2):389–406.

Pew Research Center. 2012. "Public Takes Strong Stance Against Iran's Nuclear Program." *Pew Research Center for the People and the Press*, 15 February 2012. Accessed from http://www.people-press.org/2012/02/15/public-takes-strong-stance-against-irans-nuclear-program/.

Powell, Robert. 1987. "Crisis Bargaining, Escalation, and MAD." *American Political Science Review* 81(3):717–735.

Powell, Robert. 1996. "Uncertainty, Shifting Power and Appeasement." *American Political Science Review* 90(4):749–764.

Powell, Robert. 2004. "The Inefficient Use of Power: Costly Conflict with Complete Information." *American Political Science Review* 98(2):231–241.

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

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

Putnam, Robert D. 1988. "Diplomacy and Domestic Politics: The Logic of Two-Level Games." *International Organization* 42(3):427–460.

Quackenbush, Stephen L. 2006. "Not Only Whether but Whom: Three-Party Extended Deterrence." *Journal of Conflict Resolution* 50(4):562–583.

Ramsay, Kristopher W. 2004. "Politics at the Water's Edge: Crisis Bargaining and Electoral Competition." *The Journal of Conflict Resolution* 48(4):459–486.

Reiter, Dan. 1995. "Exploding the Powder Keg Myth: Preemptive Wars Almost Never Happen." *International Security* 20(2):5–34.

Reiter, Dan. 2012. "Democracy, Deception, and Entry into War." *Security Studies* 21(4):594–623.

Reiter, Dan & Allan C. Stam. 2002. *Democracies at War*. Princeton, NJ: Princeton University Press.

Reiter, Dan & John M. Schuessler. 2010. "Correspondence: FDR, U.S. Entry into World War II, and Selection Effects Theory." *International Security* 35(2):176–185.

Reynolds, David. 1981. *The Creation of the Anglo-American Alliance, 1937-1941: A Study in Competitive Cooperation*. Chapel Hill, NC: University of North Carolina Press.

Russett, Bruce M. 1993. *Grasping the Democratic Peace: Principles for a Post-Cold War World*. Princeton, NJ: Princeton University Press.

Sagan, Scott D. 1988. "The Origins of the Pacific War." *The Journal of Interdisciplinary History* 18(4):893–922.

Sartori, Anne E. 2002. "The Might of the Pen: A Reputational Theory of Communication in International Disputes." *International Organization* 56(1):121–149.

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

Schroeder, John H. 1973. *Mr. Polk's War*. Madison, WI: University of Wisconsin Press.

Schuessler, John M. 2010. "The Deception Dividend: FDR's Undeclared War." *International Security* 34(4):133–165.

Schultz, Kenneth A. 1998. "Domestic Opposition and Signaling in International Crises." *American Political Science Review* 92(4):829–844.

Schwarz, Michael & Konstantin Sonin. 2008. "A Theory of Brinkmanship, Conflicts and Commitments." *Journal of Law, Economics and Organization* 24(1):163–183.

Schweller, Randall L. 1992. "Domestic Structure and Preventive War: Are Democracies More Pacific?" *World Politics* 44:235–269.

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64(4):627–660.

Siff, Ezra Y. 1999. *Why the Senate Slept: The Gulf of Tonkin Resolution and the Beginning of America's Vietnam War*. Westport, CT: Praeger Publishers.

Slantchev, Branislav L. 2010. "Feigning Weakness." *International Organization* 64(3):357–388.

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

Smith, Alastair. 1995. "Diversionary Foreign Policy in Democratic Systems." *International Studies Quarterly* 40:133–153.

Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92(3):623–638.

Stasavage, David. 2004. "Open-Door or Closed-Door? Transparency in Domestic and International Bargaining." *International Organization* 58(4):667–703.

Stein, Arthur A. 2000. The Justifying State: Why Anarchy Doesn't Mean No Excuses. In *Peace, Prosperity and Politics*, ed. John Mueller. Boulder, CO: Westview Press.

Thompson, Alexander. 2006. "Coercion Through IOs: The Security Council and the Logic of Information Transmission." *International Organization* 60:1–34.

Thompson, William R. 1996. "Democracy and Peace: Putting the Cart Before the Horse?" *International Organization* 50:141–174.

Thursfield, H.G., ed. 1948. *Brassey's Naval Annual 1948*. New York: The MacMillan Company.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4):821–840.

Tomz, Michael & Jessica L. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(3).

Trachtenberg, Marc. 2000. "The 'Accidental War' Question." Unpublished Draft.

Trachtenberg, Marc. 2006. *The Craft of International History: A Guide to Method*. Princeton, NJ: Princeton University Press.

Trachtenberg, Marc. 2013. "Dan Reiter and America's Road to War in 1941." Unpublished Draft.

Trager, Robert F. 2010. "Diplomatic Calculus in Anarchy: How Communication Matters." *American Political Science Review* 104(2):347–368.

Trager, Robert F. & Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55(3):526–545.

Treisman, Daniel. 2004. "Rational Appeasement." *International Organization* 58(2):345–373.

U.S. Department of State. 1964. "Documents on German Foreign Policy, 1918-1945." United States Government Printing Office, Washington, DC.

U.S. Department of State. 1969. "Foreign Relations of the United States, 1946. Vol. VII: The Near East and Africa." United States Government Printing Office, Washington, DC.

U.S. Senate. 1941. "Joint Address to Congress Leading to a Declaration of War Against Japan.". SEN 77A-H1, Records of the United States Senate; Record Group 46; National Archives. Accessed from http://www.archives.gov/historical-docs/.

U.S. Senate. 1967. "Senate Committee on Foreign Relations, 90th Congress, 1st Session." Background Information Relating to Southeast Asia and Vietnam (3d Revised Edition). U.S. Government Printing Office, Washington, D.C.

Van Natta Jr., Don. 2006. "Bush Was Set on Path to War, British Memo Says." *New York Time*, 27 March 2006. Accessed from http://www.nytimes.com/2006/03/27/international/europe/27memo.html.

Vehvilainen, Olli. 2002. *Finland in the Second World War: Between Germany and Russia*. New York: Palgrave.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Reading, Mass.: Addison-Wesley Publishing Company.

Walzer, Michael. 1977. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York: Basic Books.

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

Zagare, Frank C. 2004. "Reconciling Rationality with Deterrence: A Re-examination of the Logical Foundations of Deterrence Theory." *Journal of Theoretical Politics* 16(2):107–141.