

SPLIT SIGNALS OF CREDIBILITY:
IMPACT OF REGIME TYPE AND POLARIZATION
ON AUDIENCE COSTS

A Master's Thesis

by
ÖMER DENİZ ÖZDEMİR

Department of
International Relations
İhsan Doğramacı Bilkent University
Ankara
September 2025

ÖMER DENİZ ÖZDEMİR

SPLIT SIGNALS OF CREDIBILITY:

Bilkent University 2025

IMPACT OF REGIME TYPE AND POLARIZATION ON AUDIENCE COSTS

To my family



SPLIT SIGNALS OF CREDIBILITY:
IMPACT OF REGIME TYPE AND POLARIZATION
ON AUDIENCE CCOSTS

The Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

by

ÖMER DENİZ ÖZDEMİR

In Partial Fulfillment of the Requirements for the Degree of
MASTER OF ARTS IN INTERNATIONAL RELATIONS

THE DEPARTMENT OF
INTERNATIONAL RELATIONS
İHSAN DOĞRAMACI BİLKENT UNIVERSITY
ANKARA

September 2025

SPLIT SIGNALS OF CREDIBILITY: THE IMPACT OF REGIME TYPE
AND POLARIZATION ON AUDIENCE COSTS

By Ömer Deniz Özdemir

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of International Relations.

İhsan Efe Tokdemir
Advisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of International Relations.

Çağlayan Başer
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of International Relations.

Onur Erpul
Examining Committee Member

Approval of the Graduate School of Economics and Social Sciences

Refet S. Gürkaynak
Director

ABSTRACT

SPLIT SIGNALS OF CREDIBILITY: THE IMPACT OF REGIME TYPE AND POLARIZATION ON AUDIENCE COSTS

Özdemir, Ömer Deniz

M.A., Department of International Relations

Supervisor: Assoc. Prof. Dr. İhsan Efe Tokdemir

September 2025

International threats vary: some deter, others provoke, with consequences for escalation, bargaining leverage, and leaders' survival. Audience-cost research links credibility to domestic punishment yet typically collapses regime diversity into a democracy–autocracy binary and sidelines polarization. This thesis addresses both gaps by comparing democracies, anocracies, and autocracies and testing how affective polarization reshapes credibility within each. Theory: regime type sets a baseline for audience costs—high in democracies, low in autocracies, and intermediate in anocracies—while polarization diffuses punishability in democracies (weakening credibility), concentrates it in anocracies (strengthening credibility), and leaves autocracies largely unchanged due to insulation. Design merges MID with COW covariates and V-Dem regime and polarization measures to build a dyadic panel of 1,491 dyad–years (1816–2014) with standard controls (capabilities, contiguity,

alliances, trade, GDP, target regime). Logistic regressions evaluate whether targets reciprocate (low perceived credibility) or comply (high perceived credibility) as a proxy for audience-cost likelihood; hypotheses address baseline regime differences and polarization's moderating effect within regime families. Results in the main models are statistically significant: democracies display higher inferred audience-cost potential than autocracies, anocracies fall between; polarization erodes credibility in democracies, amplifies it in anocracies, and shows no systematic effect in autocracies. Scope conditions include proxy measurement of audience costs, regime coding choices, and polarization indices; these constraints qualify but do not undermine the contribution linking domestic punishability to international threat credibility. Definitions, coding thresholds, and data limitations are stated to maintain transparency and interpretive caution across cases and time.

Keywords: domestic audience costs, regime type, polarization, crisis bargaining, threat credibility

ÖZET

GÜVENİRLİĞİN BÖLÜNMÜŞ SİNYALLERİ: REJİM TÜRÜ VE KUTUPLAŞMANIN İZLEYİCİ MALİYETİ ÜZERİNE ETKİSİ

Özdemir, Ömer Deniz

Yüksek Lisans, Uluslararası İlişkiler Bölümü

Tez Danışmanı: Doç. Dr. İhsan Efe Tokdemir

Eylül 2025

Uluslararası tehditler farklılaşır: bazıları caydırır, bazıları direnişi tetikler; bu da tırmanmayı, pazarlık gücünü ve liderlerin siyasal bekasını etkiler. Seçmen maliyeti araştırmaları inandırıcılığı iç cezalandırmaya bağlasa da rejim çeşitliliğini çoğu kez demokrasi–otokrazi ikiliğine indirger ve kutuplaşmayı arka plana iter. Bu tez, demokrasiler, anokrasiler ve otokrasileri karşılaştırarak ve duygusal kutuplaşmanın her biri içinde inandırıcılığı nasıl yeniden şekillendirdiğini test ederek her iki boşluğu adresler. Kuramsal olarak, rejim tipi seçmen maliyetleri için bir temel belirler—demokrasilerde yüksek, otokrasilerde düşük, anokrasilerde ayırt edici biçimde orta—; kutuplaşma demokrasilerde cezalandırılabilirliği dağıtarak (inandırıcılığı zayıflatır), anokrasilerde yoğunlaştırarak (inandırıcılığı güçlendirir) ve otokrasilerde yalıtım nedeniyle büyük ölçüde değiştirmeyerek işler. Araştırma tasarımı, MID’i COW

eşdeğişkenleri ve V-Dem rejim/kutuplaşma ölçümleriyle birleştirerek standart kontrollerle (kapasite, bitişiklik, ittifaklar, ticaret, GSYH, hedef rejim) 1816–2014 arasında 1.491 dyad-yıl gözleminden oluşan bir dyadik panel kurar. Lojistik regresyonlar, hedeflerin karşılık verip vermediğini (düşük algılanan inandırıcılık) ya da uyum gösterip göstermediğini (yüksek algılanan inandırıcılık) seçmen maliyeti olasılığının dolaylı göstergesi olarak değerlendirir; hipotezler hem temel rejim farklılıklarını hem de rejim aileleri içinde kutuplaşmanın düzenleyici etkisini kapsar. Bulgular, ana modellerde istatistiksel olarak anlamlıdır: demokrasilerde çıkarılan seçmen maliyeti potansiyeli otokrasilerden yüksektir, anokrasiler aradadır; kutuplaşma demokrasilerde inandırıcılığı aşındırır, anokrasilerde güçlendirir, otokrasilerde sistematik bir etki göstermez. Kapsam koşulları, seçmen maliyetlerinin dolaylı ölçümü, rejim kodlama tercihleri ve kutuplaşma endekslerine ilişkindir; bu kısıtlar, iç cezalandırılabilirliği uluslararası tehdit inandırıcılığına bağlayan kuramsal katkıyı zayıflatmaz.

Anahtar Kelimeler: izleyici maliyetleri, rejim tipi, kutuplaşma, kriz pazarlığı, tehdit inandırıcılığı

ACKNOWLEDGMENTS

Coming into this new, great field and completing my master's meant learning a new way of thinking from the ground up, while making mistakes and returning the next day to do a little better. If I have managed surviving up to this point, I did so by stepping on some amazing giants' shoulders. Thus, what I now offer in these pages is a mere depiction of my boundless gratitude to them.

Among those who helped me the most, my advisor, Dr. İhsan Efe Tokdemir stands first. Long before there were these polished pages, he treated the clumsy intern with too many questions as someone worth investing and believing in. He modeled what good work looks like and made space to try, fail, and try again with stronger tools. When the thesis wavered, he helped me find its center; when PhD applications felt out of reach, he turned them into realistic plans. For time given without hurry, for the unspoken firm message—“you can do this, but do it right”—and never stop trusting me, thank you, Efe Hocam. I hope one day to become an academic worthy of your example.

The firmest claims in this thesis owe much to Dr. Çağlayan Başer. As my thesis defense committee member, her feedback arrived as careful attention rather than blunt correction—finding the core of an argument, protecting it from clutter, and

showing where a cleaner line would carry more weight. For precision delivered with calm, thank you, Çağlayan Hocam.

From the first worries about getting into the master's program to the long stretch of PhD applications, Dr. Onur Erpul has been there—solid, available, and kind. As a committee member, you brought a rare mix of precision and patience: returning questions with better questions, making methods feel like choices that carry responsibility, and offering suggestions specific enough to act on yet broad enough to reshape the project. For the calm on difficult days and the rigor that improved every chapter, thank you, Onur Hocam.

I am also grateful to the International Relations Department for never letting me feel like a guest. From day one to graduation, hallway chats turned into guidance; course debates turned into senses of belonging. To my cohort—my brothers and sisters in arms—thank you for shared notes, jokes, and late-night pep talks when deadlines stacked up. And in the middle of it all, Ece abla kept the engine running with a calm that changed the day's weather—solving all trivial crises before they grew. For making a demanding place feel like home, thank you to each of you.

To all the great professors whose courses I was fortunate to take: thank you all for opening intellectual doors. Among many, I owe special thanks to two very dear people, Dr. Seçkin Köstem and Dr. Ekrem Başer. Seçkin Hocam's classroom raised

the bar without making it feel out of reach; he asked for careful reading, fair engagement, and the courage to refine an argument until it held. As Ekrem Hocam's assistant, I learned the craft up and close: he guided my ideas with frank, generous advices—sharpening plans, and turning stressful moments into firm steps forward. For expecting better and helping me reach it, thank you, Seçkin Hocam; and for all I learned from you, thank you, Ekrem Hocam.

For the ones that opened the doors I never knew that were out there to begin with, I thank the CFPPR community which became the academic family I did not know I needed. First, I offer my heartfelt thanks to Dr. Ersel Aydınli and Dr. Julie Aydınli for building up a place where novice people like me could find room to grow because they are willing to do so. You welcomed an overtly enthusiastic intern and made space for a developing voice, showing that warmth and excellence are not opposites.

Along with them, I extend my gratitude to the mentors— or more like the big brothers and sisters I never had—whom I found at CFPPR. Faruk Aksoy offered a compass when details threatened to swallow the map, generous with practical fixes and the deeper habits that keep research honest. Firuze Sezgin brought steadiness and a listening ear, turning worry into a plan. İlker Kalin shared workflows and quiet advice that turn intention into finished pages. Metin Koca asked questions that lingered and improved both prose and thinking. And Onur Hocam, again, opened

doors and cheered small wins as steps toward bigger ones. I arrived as an intern; I leave a proud affiliate because of you.

At the center of it all stands my family, both close and extended. Fatma Özdemir—my mother—kept, sending strength in ordinary ways: a call, a meal, a reminder to rest. Ömer Özdemir—my father—offered a quiet steadiness that made long weeks bearable and far goals believable. Can Özdemir—my brother—brought humor and the nudge to keep going when the page would not move. You all celebrated small milestones like final victories and treated setbacks as part of the road. For love that asks only, “How can we help?”—thank you Özdemir family.

Finally, there is Özge. When confidence thinned, she lent hers; when days grew heavy, she shouldered the weight without keeping scores. She reminded me to be a person while I was trying to be a researcher. She met all my small crises with her practical ingenuity. She believed when I could not, and that belief turned plans into achievements. This thesis and the road opening ahead—none of it would be realized without her beside me. For the endless care that made the all possible, thank you, Özge.

TABLE OF CONTENTS

ABSTRACT	I
ÖZET	III
ACKNOWLEDGMENTS	V
TABLE OF CONTENTS	IX
LIST OF TABLES	XI
LIST OF FIGURES	XII
LIST OF ABBREVIATIONS	XIII
CHAPTER 1	1
INTRODUCTION	1
CHAPTER 2	12
LITERATURE REVIEW	12
2.1. What is the Theory of Domestic Audience Costs	13
2.1.1. Who is the Audience?	15
2.1.2. What is the Cost?	17
2.2. Three Waves of ACT Studies	19
2.2.1. First Wave: Operationalizing ACT	21
2.2.2. Second Wave: Dichotomy of Regime Types	23
2.2.2.1. Disaggregating the Binary: Anocracies	26
2.2.3. Third Wave: Three Types of ACT Micro-Foundations	27
2.2.3.1. Systemic Factors	28
2.2.3.2. Situational Factors	30
2.2.3.3. Societal Factors	32
2.2.3.4. The Overlooked Variable: Affective Polarization	34
2.3. Recapping the Literature and Blind Spots	35
CHAPTER 3	37
THEORETICAL MECHANISM	37
3.1. Political Regimes and Stability	37
3.1.1. Audience Cost and Democratic Stability	41
3.1.2. Audience Cost and Autocratic Stability	45

3.1.3. Audience Cost and Anocracies' Instability.....	48
3.1.4. Recapping Stabilities across Regimes.....	52
3.2. Audience Cost and Affective Polarization.....	53
3.2.1. Audience Cost in Democracies with Polarization.....	57
3.2.2. Audience Cost in Autocracies with Polarization	61
3.2.3. Audience Cost in Anocracies with Polarization.....	66
3.2.4. Recapping Polarization across Regimes	70
CHAPTER 4	75
RESEARCH DESIGN	75
4.1. Data	76
4.2. Variables	78
4.2.1. Dependent Variable.....	79
4.2.2. Independent Variable	83
4.2.2.1. Interaction Term.....	87
4.2.3. Control Variables	89
4.3. Operationalization.....	93
CHAPTER 5	96
RESULTS	96
5.1. Empirical Approach: Logistic Regression	97
5.2. Findings.....	99
5.2.1. Logistic Regression for Regime types	99
5.2.2. Logistic Regression of Polarization Moderation on Polity Scores	104
CHAPTER 6	109
CONCLUSION.....	109
6.1. Summary of Findings.....	111
6.2. Implications.....	115
6.3. Limitations	117
6.4. Future Studies	119
REFERENCES.....	123
APPENDICES	144

LIST OF TABLES

Table 1. AC Incurrence Likelihoods across Regimes and Polarization Levels	73
Table 2 Summary Statistics of all Variables	93
Table 3 Expected Significant Results	98
Table 4 Logistic Regression Results for Polity Score.....	100
Table 5 Logistic Regression Results for Polity Score moderated by Polarization ..	104

LIST OF FIGURES

Figure 1 Threat Response (IV proxy indicator) Distribution.....	83
Figure 2 Distribution of Regime Types	87
Figure 3 Impact of Threatener's Polity Score on Target's Reciprocation	103
Figure 4 Impact of Threatener's Polarization on Target Reciprocation	107
Figure 5 Naïve Model for Polity Score.....	144
Figure 6 Naïve Model for Polarization's Conditional Impact.....	144

LIST OF ABBREVIATIONS

ACT – Domestic Audience Costs Theory

AC – Audience Costs

CINC – Composite Index of National Capability

COW – Correlates of War Datasets Project

MCT – Militarized Compellent Threats

MID – Militarized Interstate Disputes

V-Dem – Varieties of Democracy

CHAPTER 1

INTRODUCTION

*I always needed to maintain my resolve in foreign policy,
so, the people would not throw accusations in my face...*

Nikita Khrushchev, July 1955

(Khrushchev, 2007, pp. 52)

Soviet Premier, Khrushchev's above remark aptly captures a persistent certainty about international politics: incarnating through whether elections in democracies or coups in non-democracies, public sentiment towards the people in charge can decisively influence—and at times dictate—foreign policy choices (Popkin, 1991; Zaller, 1992). At the heart of this dynamic lies a plain reality: incumbents are political survivalists, acutely aware that their ability to achieve any kind of external goal depends on sustaining a sufficient level of domestic trust, especially among their followers, to keep holding onto their office. Given this tacit prerequisite of incumbent survival over foreign affairs, few leaders will risk the loss of face or the perception of incompetence due to a misstep abroad that can erode trust in them back at home. (Geddes & Zeller, 1989; De Mesquita & Siverson, 1995; Smith, 1998).

This necessary condition of internal survival is also valuable for one of the most common tools in international relations: issuing threats. From formal ultimatums to military mobilizations, threats are commonplace signals in international politics to

convey certain messages of demands, yet a persistent puzzle remains—why some threats secure compliance while others are dismissed or even provoke escalation. Conventional explanations attribute this variation to factors like material capabilities, prior records of following through, or the salience of the issues at stake (Schelling, 1956; 1966; 1982; Press, 2005; Johnson, 2015; Sechser, 2018). While these do matter, they all assume an *ex-ante* condition: that leaders can statically hold offices to act on their threats without any problem. However, linking back to the incumbent survival pre-requisite, leaders can be ousted and if a certain leader is taken out the picture, even overwhelming military strength or an impeccable history of resolve offers little deterrent value.

Domestic Audience Costs Theory (ACT) enters the picture here by notably theorizing this domestic–international link of survival in answering the threat credibility dilemma as a function of the domestic political price a leader expects to be paid if they issue a threat and then back down (Fearon, 1994a; 1994b). These “costs” differ across systems—in democracies, they may take the form of electoral defeats; in autocracies, they may entail elite defections leading up to coups. The theory’s core insight is unsophisticated yet far-reaching: the greater the expected domestic penalty for a retreat from an international threat is, the more credible a leader’s threat appears to an adversary internationally, making domestic political survival the hidden linchpin of effective coercive diplomacy. While initially considered as a mechanism most relevant to democracies, where institutional checks and public scrutiny are well-pronounced (Fearon, 1994a; Eyerman & Hart, 1996; Schultz, 2001; Tomz, 2007), later work has shown repeatedly that autocratic leaders, too, incur audience costs (AC) (Geddes, 1999; Slantchev, 2006; Weiss, 2013; Weeks,

2008, 2012, 2014). Yet scholarship still tends to confine the phenomenon of ACT within a democracy–autocracy binary, leaving underexplored how AC operate in-between these regimes, namely anocracies, and how regimes impact can be moderated through critical micro-factors such as polarization of society. To illustrate why the considering anocracies and polarization matters for ACT, two contrasting cases from 1970s Latin America offer a useful comparison.

In August 1975, Peruvian leader General Juan Velasco Alvarado—who had taken power through a military coup in 1968—entered a high-stakes confrontation with Chile. Motivated by long-standing nationalist grievances over territories lost in the War of the Pacific nearly a century earlier, Velasco issued a strikingly blunt threat: “Chileans should stop with the bullshit or tomorrow, my troops and I shall eat breakfast in Santiago” (Masterson, 1991). This rhetoric was shortly backed by rapid military mobilization in a single day, with tanks parading in cities like Ica near the Chilean border, creating the impression of imminent war. Yet, within the same day, Velasco recognized the severe risks of actual conflict and pulled back. By the time Velasco retreated, however, Chile reciprocated by moving to secure ground along the border. This swift reversal and concurrent reciprocation to the threat were interpreted as a humiliating loss of face back at home, given Velasco’s public belligerence beforehand. Less than a month later, he was overthrown in a coup orchestrated by his own right-hand man—a drastic cost levied primarily for retreating after issuing high-profile threats and desecrating Peru’s international prestige by getting reciprocated (Drinot, 2017; Hurtado-Torres & Fernandois, 2020).

In December 1978, in another country not so far away, Argentina's military ruler General Jorge Rafael Videla—who had similarly come to power through a coup only two years ago—faced a comparable confrontation with Chile, once again regarding territorial disagreements, but this time over the Beagle Channel Islands. Like Velasco, Videla couched his claims in uncompromising nationalist terms, warning that Argentina would take “all measures necessary, deadly or not” to defend its sovereignty (Ferrara & Hunt, 1979). His forces were mobilized along the Chilean border, and the situation escalated to the brink of armed conflict in a single day, much like Velasco. However, just as in the Peruvian case, Videla's forces stood down within the same day and event ceased as a mere militarized threat. Unlike Velasco, though, Videla extracted partial compliance from Chile over the islands thanks to his threats by the time he pulled troops back. Despite the retreat from a public threat, Videla was insulated from the kind of domestic backlash that unseated Velasco. At the time, he faced only minor protests, which he crushed, and remained in power until voluntarily stepping down in 1981, handing the incumbency to another non-democratic leader (Villar-Gertner, 2014; Lisinska, 2019; Magnani & Barreto, 2022).

On the surface, the both leaders—Velasco and Videla—operated under remarkably similar circumstances. Yet, despite these similarities, the domestic consequences diverged sharply: Velasco ousted, whereas Videla survived. At a first glance, the ACT literature offers clear starting points for explaining these kinds of divergences in cost outcomes. Work since Fearon (1994a; 1994b) has shown that even non-democratic leaders can incur AC, particularly when their “audience” consists of military or party elites with the capacity to depose them (Geddes, 1999; Weeks,

2008; Ezrow & Frantz, 2011). Scholars have further refined these arguments by distinguishing subtypes of autocracies, arguing that personalist–military hybrids, or regimes with cohesive elite coalitions, are more vulnerable to elite-led punishment after failed coercive threats (Frantz, 2018; 2024). Others have looked at micro-level institutional features—such as varying levels of media openness, multiplicity of channels for elite defection, or informal coup-proofing arrangements—that can amplify or dampen AC effects (Kinne & Marinov, 2012; Weiss, 2013).

However, when applied to Velasco and Videla, these approaches fall short in fully explaining the difference in outcomes. Retrospectively, both leaders were military rulers who had seized power via coups; both issued uncompromising threats backed by visible military mobilization; both escalated nationalist disputes with same country, Chile, right up to the brink of war; and both ultimately backed down before open hostilities commenced with full force. In each instance, the retreat was public and unequivocal. Furthermore, from the standpoint of ACT literature, both leaders fall under the same non-democracies with highly-cohesive military audience category in the Geddes, Wright, and Frantz (2014) dataset, suggesting similar vulnerability profiles. Besides, both operated in contexts where elite coordination was feasible and where military factions were decisive political actors, while media openness was highly restricted. By the logic of prevailing ACT literature, both should have faced roughly comparable risks of removal after backing down. Yet, as stated, while one being ousted, the other prevailed.

Given the inaptness of the existing ACT works to elaborate these cases, one wonders what else accounts for, then, the difference in outcomes in terms of AC incurrence.

In addressing this inquiry, two useful details stand out. First, Velasco's Peru in 1975 was already an entrenched autocracy, with a Democracy score of -8 . Videla's Argentina in 1978, by contrast, was in the midst of sliding from anocracy toward full autocracy—moving from -4 at his takeover in late 1976 to -6.5 by the time of the crisis in 1978, just above the anocracy threshold. Second, the public polarization profiles of the two countries at the time of threats could not have been more different: Peru's polarization was relatively low, while Argentina's was at an all-time high (Coppedge et al, 2020). The interaction between regime types, especially in-between types like anocracies, and the degree of polarization is rarely employed as an explanatory lens in the AC literature. Yet this perspective helps clarify why above two otherwise comparable cases diverge: in one, polarization enables grievances to be weaponized within a gray-zone regime, whereas in the other, the absence of polarization in a more suppressive regime obstructed the emergence of such costs

These disclose the two vital research gaps in ACT literature. First gap is about the fact that while ACT scholarship has expanded beyond its democratic origins to consider authoritarian subtypes (Weeks, 2008; Slantchev, 2006), it has remained largely anchored in a democracy–autocracy binary. Anocracies—the fluid, hybrid regimes that occupy the middle with features of both types—have received comparatively little systematic attention, despite being common and more prone to regime-based instabilities like regime changes by coups or elections (Marshall & Cole, 2014). This omission matters because anocracies, by virtue of their transitional character, may present distinctive AC dynamics not captured by either democratic or authoritarian models. A second gap lies in the treatment of polarization. Polarization has been studied in relation to ACT before, however, only in democratic contexts via

experiments embedded in public opinion surveys (Levendusky & Horowitz, 2012; Evers, Fisher, & Schaaf, 2019). These studies tend to find limited or no effect, but they are spatially and temporally bounded, with little consideration of how polarization might function in autocracies or anocracies. Yet, polarization is deeply consequential in such settings—not merely a byproduct of authoritarian or hybrid rule but a tool for regime survival through a mentality of divide and rule (McCoy, Rahman & Somer, 2018). Its effects on audience behavior in the wake of international threats and retreat may therefore diverge sharply from patterns observed in democracies.

By placing the Velasco and Videla episodes under this light, the relevance of both regime position (with a focus on anocracies) and polarization becomes clearer. The cases suggest that combinations of regime hybridity and polarization levels can shape how threats are perceived and thus how much political risk leaders face when backing down. This thesis builds on that insight by systematically comparing democracies, anocracies, and autocracies, and by examining how polarization interacts with each regime type to influence the likelihood of AC incurrence. In doing so, it addresses two underexplored questions: how AC incurrence likelihood comparatively alters across regimes, including anocracies, and how polarization of society moderate AC incurrence propensity across all three of these regime types. In answering these inquires, this study conducts a large-N quantitative analyses utilizing existing observational datasets to form a novel dataset. The primary datasets for this analysis will be the Militarized Interstate Disputes (MID) dataset (Palmer et al, 2022) for AC incurrence likelihood (DV), Varieties of Democracy (V-Dem) dataset for

both regime types (IV) and societal polarization level (interaction term) (Coppedge et al, 2020).

In terms of DV, it should be noted that AC is like the “dark matter” of international relations—hard to observe but central to theoretical models of public opinion’s role on foreign policy (Schultz, 2012). This approach implies that it is rather hard to directly measure AC out in the open as it also involves making counterfactual observations like the non-incurrence of backlash. Thus, several studies on ACT use the MID dataset’s target state threat response variable to proxy AC incurrence likelihood as the perceived credibility of threats to gauge whether the leader face higher or lower likelihood of AC for backing down (Weeks, 2008; Potter & Baum, 2014; Quek & Johnston, 2017; Crisman-Cox, & Gibilisco, 2018; Yu, Whang, & Lee, 2022). These studies show that target states’ responses—coded binary as 1 for reciprocation and 0 for compliance or no response—serve as signals of the challenger’s perceived threat credibility and so, its likelihood of AC incurrence. Notably, Weeks (2008, pp. 50) explains the idea behind this proxy measure for AC incurrence likelihood as follows, “one should expect that in line with a survivalist mentality, initiators with a high ability to generate audience costs should be less likely to face resistance and more likely to receive compliance than states with a low ability to generate audience costs.” Building on these, the dependent variable, AC, will be proxied by the target state’s response to threats.

Coming to the other main variables, regime type (IV) is measured on a continuous – 10 to +10 polity scores index from V-Dem. Scores from –6 to –10 is coded as autocracies, –5 to +5 as anocracies, and +6 to +10 as democracies. Affective or

public polarization will be used as an interaction term, measured using the V-Dem dataset, providing a continuous scale of ideological divergence and social polarization from 1 to 10. The interaction term will examine the combined effect of each regime type and polarization on AC. In total, the V-Dem dataset will classify regimes, polarization levels will be extracted from the same dataset, and responses to threats will be measured using the MID dataset as proxies for AC. These datasets will be merged using common identifiers like country codes and incident year while unit of analysis will be crises-year (Coppedge et al, 2020).

Several control variables will be included to account for other factors influencing state behavior in interstate disputes, drawing exclusively from the V-Dem dataset and the MID dataset. Economic indicators will be measured using V-Dem's GDP per capita and trade data. Military capabilities will be captured through MID measures of military personnel and expenditure. Political stability and governance will be controlled using V-Dem's regime stability scores and additional institutional indices. Historical conflict experience—including previous disputes and alliance membership—will be taken from the MID dataset. Geopolitical considerations such as geographical contiguity and border dispute history will also rely on MID data. Demographic size will be measured through V-Dem's population estimates, while ethnic and social fractionalization will be drawn from V-Dem's social division indicators.

A series of logistic regression models will be used to examine how regime type and societal polarization shape the likelihood of AC incurrence, proxied by target state responses. The analysis proceeds in two stages: first, it estimates differences in AC

likelihood across regime types alone; second, it tests how polarization interacts with each regime type to alter that likelihood. Logistic regression is well-suited here because the dependent variable—target state response—is binary, coded as either reciprocation or compliance. As used in the AC literature, compliance will be interpreted as a signal that the challenger’s threat is credible, indicating a higher likelihood of AC incurrence if they were to back down, coded 0; reciprocation will indicate the opposite, coded 1. Without accounting for polarization, the expectation is that democracies will display the highest likelihood of AC incurrence (and thus threat credibility), followed by anocracies, then autocracies. Once polarization is introduced, the expected pattern shifts: democracies’ likelihood of AC incurrence decreases, anocracies’ increases, and autocracies remain largely unchanged. This design allows for a direct test of whether polarization systematically reshapes threat credibility across all three regime types.

After this introduction, the thesis is structured as follows: In Chapter 2, I present the survey of the scholarly land for this theory by first explaining what Domestic AC Theory is and how it functions through signaling mechanisms in international disputes. I then explore the three waves of study over ACT across. This chapter also addresses gaps in the literature. Chapter 3 constructs a theoretical mechanism based on political instability across regimes and potential detrimental consequences of polarization, incorporating insights from social psychology on in-group favoritism and out-group hostility. Chapter 4 details the research design, data sources, and variable operationalizations. In Chapter 5, I present the empirical analysis and findings. Finally, Chapter 6 concludes with key insights, limitations of the study, and future research directions.



CHAPTER 2

LITERATURE REVIEW

This chapter is organized into two interrelated halves. The first introduces phenomenon of ACT, explaining how the credibility of international threats depends not only on conventional means like material capabilities but also on the domestic political price leaders risk paying if they back down. It then examines two foundational questions at the heart of this theory—who constitutes the “audience,” and what forms these “costs” take. Following this conceptual foundation, the second half traces the evolution of ACT scholarship through three interrelated waves: first, early empirical efforts to test whether AC can be observed beyond theory; second, regime-type comparisons that have largely treated the world as a democracy–autocracy binary; and third, more recent work on micro-level factors, such as elite cohesion, leader characteristics, and media freedom, under three categories. Along with these, this review lays the foundation for the thesis’s contribution to advancing the ACT literature by addressing both the lack of studies on anocracies and the underexplored role of polarization.

2.1. What is the Theory of Domestic Audience Costs

ACT, developed by James Fearon (1994a; 1994b), seminally redirects the study of international threat credibility from purely international factors to the domestic political arena, linking a leader's external commitments to their vulnerabilities at home. The central claim is straightforward: when leaders escalate a crisis by issuing threats, they risk incurring political penalties, called AC, if they subsequently back down or do not follow through their threats. Importantly, these costs are imposed not by foreign adversaries but by domestic actors interpret retreat as incompetence and punish the leader through various means, ranging from reduced political support on polls up to outright removal from office by a coup.

The ingenuity of Fearon's theory lies in that fact that rather than treating threat credibility solely as a function of material capabilities, reputational history, or the stakes involved, ACT emphasizes that credibility also depends on whether domestic political consequences make backing down prohibitively costly. Put differently, if a retreat from a threat is likely to trigger significant domestic political backlash, then even a reluctant leader may be compelled to follow through—making the threat appear more credible to the target state. The logic underlying ACT unfolds as follows: once a leader issues a threat, it becomes tied to their political standing at home, especially if threat is publicly made, because they have staked both personal and national prestige on this commitment's fulfillment. Backing down, then, risks damaging this standing by signaling poor judgment or incompetence, inviting criticism from political rivals, eroding public support, and ultimately, undermining domestic legitimacy. Anticipating such consequences, leaders are incentivized to

avoid making empty threats in the first place and to carry out those they have made. From the target's perspective, what makes a threat intimidating is not simply the leader's personal resolve or capabilities, but the perception that backing down would trap them in severe domestic political fallout. The greater the risk of humiliation, loss of authority, or even removal from power at home, the more credible—and thus more compelling—the threat appears abroad. (Fearon, 1997; Schultz, 1998, 2001; Smith, 1998; Ramsey, 2004; Acharya & Grillo, 2019).

Historical cases bring the AC mechanism into a better view. During the 1790 Nootka Island crisis, British Prime Minister William Pitt took a confrontational public stance against Spain by mobilizing the Royal Navy, a critical form of public threat. Facing a vocal political opposition and looming elections, backing down would have severely damaged Pitt's domestic standing. As Fearon notes, Pitt "would have faced serious domestic political costs for backing down after his threat, much larger than if he had chosen initially to pursue a less public and aggressive line" (1994a, 580). Despite their similar relative material capabilities at the time, Spain ultimately yielded and accepted Pitt's terms of agreements, implicitly validating the domestic binds Pitt had created for himself. The crisis was resolved without conflict, but only because Pitt's threat was taken credible—made so by his political vulnerability at home.

It should also be noted that ACT is neither deterministic nor all-encompassing. Theory never claims that all international threats produce significant AC, nor that credibility always hinges on the prospect of domestic punishment to begin with. Rather, ACT isolates a key mechanism through which internal political constraints

shape external behavior. The perceived likelihood of such costs depends on contextual variables such as regime type and audience composition. Given these, ACT clearly fills a critical gap. It does not deny the importance of power, reputation, or other related factors, but it emphasizes that these tools matter only if the leader manage first to remain in office, meaning domestic survival is the foundation upon which all other strategies are built. In simpler terms, as de Mesquita et al. notes “the will to politically survive is so overwhelming that it fundamentally motivates the selection of all political policies, both internally and externally, as the lack thereof impedes all political acts.” (2003, 79). Thus, ACT shows how ontological imperatives can make international commitments stick. To understand how this theory works in practice, it is essential to detail two core components of the theory: who forms the “audience” and what exactly are the “costs” that binds leaders to their threats.

2.1.1. Who is the Audience?

The concept of “the audience” is central to the theory, but it resists a narrow definition. While the audience is often equated it with the public—especially voters in democratic systems—this reading underestimates the flexibility of Fearon’s argument. In ACT, the audience is not firmly defined by institutional identities, but is equated mostly with its functional capacity to hold leaders accountable. In essence, the audience is any actor—or coalition of actors—that can punish a leader politically for backing down after a public commitment at the international level. This broader view of the audience allows ACT to operate across diverse political contexts without relying on fixed regime types. The audience might consist of voters, yes—but also

party cadre, rival factions, influential media, or military elite can have a say in the game. What matters is not who the audience is formally, but whether they can translate perceived weakness into politically detrimental consequences for the leader. In this sense, Fearon's theory is less about faces and more about ideas: it is the presence of a domestic gaze capable of punishing retreat that drives credibility (Fearon, 1994a; 1994b).

Furthermore, the audience need not even act in visible ways. Meaning that, its influence often operates through anticipation. Leaders weigh not only how an adversary will respond to their threats, but also how their domestic observers might interpret—and react to—concessions or reversals. Once a threat is made, it invites scrutiny and creates expectations by the audience since a threat is not just a political instrument to compel adversaries but also a mechanism through which leaders signal resolve or capabilities to their audience. As Fearon writes, “threatening, by nature, begets the possible costs for backing down but also establishes resolve for the domestic audiences” (1994a, 577). That possibility alone can alter behavior. Importantly, the audience is not a fixed entity across crises. Its composition, attentiveness, and potential to act can vary with the stakes or the content of the issue and the visibility of the threat. But its role remains consistent: to monitor the leader's behavior and judge whether they deliver on commitments. This transforms foreign policy into a kind of domestic performance. Even when aimed at external actors, threats are watched and interpreted at home as emblematic gestures of incumbent's capabilities or its lack thereof.

As a crucial historical example, it is not hard to trace ACT's logic in the Cuban Missile Crisis. When President John F. Kennedy faced a crisis with the USSR, he went out to the American public and addressed the nation on live television, announcing the U.S. naval quarantine of Cuba and possibility of nuclear missiles being utilized, an open threat. By doing so, he was not only signaling resolve to the Soviet Union—he was locking in a domestic audience. Paradoxically, any sign of retreat would have risked political backlash from Congress, the press, and the public. In other words, once Kennedy's position was tied to a visible and dramatic commitment, the political cost of reversal became deeply personal and immediate, causing credibility of his threats against USSR to be increased. Eventually, the commitment made via threats led Kennedy to maneuver in such a way that Khrushchev got compelled to remove nuclear missiles both from Cuba and Turkey (Allison, 1969; Lebow, 1983). Given this example, the audience in ACT is more than a singular face but the domestic mechanism through which inconsistency becomes politically dangerous and commitments become tokens of persuasion against enemies.

2.1.2. What is the Cost?

While the audience is like the force that observes and judges, the cost is the effect of that judgment when it turns negative. The “costs,” just like “audiences,” are not fully defined but taken as the domestic political penalties a leader receives or anticipates to receive when they escalate crises by threats and then retreat. These are not external sanctions or battlefield losses—they are internal consequences that can shake the leader's ability to remain in power. Fearon emphasizes that these costs come in varied forms (1994a; 1994b). They may include electoral defeat, media

condemnation, elite withdrawal, or legislative revolt. It can also take time for these costs to occur or they need piling up to have a hitting impact, meaning that costs can also gradually erode leaders position until the final blow of ousting. In more severe cases, they can precipitate resignations, cabinet collapses, or coups. What unites these scenarios is not their institutional form but their effect: they undermine the leader's authority by making them appear weak, indecisive, or untrustworthy, even causing them to lose their seats.

In this sense, costs operate as political taxes levied on inconsistency: each escalation raises the stakes for retreat. Eventually, the domestic consequences of backing down outweigh even the international risks of escalation. This is what makes AC binding. Leaders escalate with the hope of signaling resolve, but once they commit threats, they are no longer free to de-escalate without consequences of some dent on their authority. The threat that once may have served as a bargaining chip becomes a political obligation. Even if new information emerges, or the adversary offers a face-saving compromise, the domestic cost of reversal may now be too high to pay in a given case (Fearon 1994a).

A striking historical illustration of the variety and potency of costs is the 1982 Falklands War. When Argentina occupied the islands, British Prime Minister Margaret Thatcher responded with bold rhetoric and the deployment of a naval task force. What followed was not merely a military confrontation, but a political one. For Thatcher, retreat after such public escalation would have been seen as a sign of weakness, damaging her authority and potentially costing her re-election as an election was looming. The domestic price she faced was electoral: diminished public

support and the erosion of political capital. On the other hand, for Argentine leader Leopoldo Galtieri, the calculus was different but no less severe. Leading a military junta, his authority was tied directly to perceptions of strength among his small cadre of fellow officers. When it became clear that Britain would not back down, and further escalation risked devastating military defeat, Galtieri chose to retreat even after making decisive threats to keep the invasion of Falklands going. But this decision—interpreted as weakness—provoked backlash. He was removed from power by the same military elites, who had backed his regime, and later exiled. His cost was not electoral, but institutional and existential: removal by coup (Weeks, 2014; Mercau, 2019).

This contrast reveals that costs can differ in forms—votes in democracies, coups in autocracies—but their function is constant. They turn public threats into commitments by making retreat damaging for the incumbent’s survival. If the audience is the gaze that watches, the cost is the blow that follows when leaders fail to deliver. Together, they bind leaders not necessarily out of desire, but out of political necessity. And therein lies the paradox of Fearon’s theory: threats are credible not because leaders are strong, but because they are vulnerable. A state can make its commitment more believable by making it more politically dangerous to back away. Put simply, it is not resolve but entrapment—crafted by domestic exposure—that gives a threat its edge.

2.2. Three Waves of ACT Studies

Having now traced the conceptual foundations of ACT, it is useful to situate its scholarly evolution over time. Since Fearon's seminal work in 1994, ACT has developed in three interconnected waves. The first focused on translating the theory from formal models to empirical analyses. In this phase, scholars attempted empirically testing whether domestic punishments for backing down—AC—could be observed, measured, or falsified. This primary wave managed to anchor ACT within the empirical toolkit of international relations, relying on survey experiments, case studies, and utilization of Large-N observational datasets.

The second wave revisited Fearon's core assumption that AC primarily emerge in democracies. While many confirmed this logic, others challenged it, showing that autocratic leaders can also incur costs—albeit via different ways than their democratic counterparts, e.g. elite defection, party rebellion, or coups. However, this led to a persistent democracy–autocracy dichotomy, sidelining the middle gray era of anocracies. These regimes, now strikingly raising in numbers, combine institutional ambiguity with selective accountability—features that warrant independent analysis. This thesis intervenes by observing anocracies through a comparative lens as its first contribution (Mufti, 2018; Trinkunas, 2024).

The third wave has turned to micro-foundations of ACT. Yet one crucial factor remains underexamined: affective polarization—hostility between political groups that possibly distorts public accountability. Given its importance and commonality, it is pertinent to elaborate on the moderative impact of affective polarization across regimes. To do so, this thesis addresses that gap, offering a dual contribution: exploring ACT in anocracies and integrating the interacting impact of affective

polarization into its theoretical framework. Before doing these, remainder of this literature review will present a survey of the three waves of ACT studies.

2.2.1. First Wave: Operationalizing ACT

Following Fearon's coining of the ACT (1994a; 1994b), an initial wave of scholarship sought to move the theory from the realm of formal modeling into empirical analysis. Fearon's core argument—that leaders who issue public threats and then back down can face domestic political punishment—provided a novel explanation for credibility in crisis diplomacy. Yet to assess its validity, scholars first needed to operationalize AC and determine whether they could observe them in actual political contexts.

One main way for early researchers was survey experiments, primarily conducted in democratic settings. These studies tried to simulate foreign policy crises and measure public reactions in the face of retreat. These studies generally found that leaders who retreated after issuing threats faced notable declines in domestic support—especially when their threats had been publicly made (Eyerman & Hart, 1996; Smith, 1998; Partell & Palmer 1999; Gelpi & Griesdorf, 2001; Schultz 2001; Tomz, 2007; Tomz & Weeks, 2013). Building on this foundation, scholars later explored how media coverage, framing, or justification of threat mediated public reactions to retreat within the same survey context (Potter & Baum, 2010; Levy, 2012; Levy et al., 2015; Kertzer & Brutger 2015; Crisman-Cox & Gibilisco, 2018).

Despite these, empirical disputes persisted. Some challenged these experimental findings with the historical case studies, arguing that AC do not consistently appear over the long history of crises across decades (Trachtenberg, 2012). Yet, these scholars are later discredited for cherry-picking historical events to unreasonably refute ACT (Moon & Souva, 2016). Some others raised concerns about the generalizability of experimental results beyond democratic states, or questioned whether publics follow foreign policy closely enough to impose costs for inconsistency in every democratic setting (Mercer, 2012; Gartzke & Lupu, 2012). Scholars also began identifying contextual conditions that may shape when and how costs materialize which became more vocalized with the third wave (Morrow, 2000; Wagner, 2000; Prins, 2003; Uzonyi et al., 2012).

As a response to these critics, researchers increasingly incorporated observational methods along with experimental ones, especially by using MID dataset at first and then Militarized Compellent Threats (MCT) dataset to strengthen the empirical grounding of the theory (Sechser, 2011). By examining crises outcomes and signaling patterns beyond the spatial and temporal boundaries of survey experiments, these studies—such as those by Prins (2003), Weeks (2008), and Uzonyi et al. (2012)—inferred the perceived behavioral effects that AC should produce by observing the responses given by the target states, even in the absence of direct public opinion data. Although such observational-data-driven studies are contested for not being able to fully capture AC or overestimate it, they are widely used and still offer a crucial window into how audience incentives might shape leader behavior across real-world cases via Large-N data (Secher, 2011; Fuhrmann & Sechser, 2014).

Together, these experimental and observational strategies helped transition ACT from a game-theoretical world to an empirically-grounded one. They did this by establishing methodological toolkits and demonstrated that the theory could generate testable predictions, even if its mechanisms were difficult to isolate directly. This foundational work made it possible for subsequent research to shift the question from whether AC exist to where, how, and for whom they are most likely to occur—setting the stage for a second wave focused on regime type and systemic contexts.

2.2.2. Second Wave: Dichotomy of Regime Types

With the empirical foundations of ACT gradually being solidified in the first wave, a second wave of research emerged that turned its focus to a core premise in Fearon's original formulation: AC is more likely to be generated in *democratic* systems. This assumption flowed logically from the mechanisms outlined in the original model. Democracies are systems of the institutional transparency, electoral accountability, and press freedom, facilitating domestic audiences to monitor and then, punish, if need be, a leader's foreign policy missteps. As a result, leaders of democracies are expected to incur costs for retreating more likely, thus making their threats more credible (Fearon 1994a; 1994b).

A significant body of research set out to test and, in many cases, confirm this presumption. Scholars during this phase focused overwhelmingly on institutional openness as the key facilitator of AC incurrence. They argued that the existence of competitive elections, active opposition parties, and independent media enables

democratic audiences to monitor their leaders and hold them accountable for failures in foreign policy much easier than their autocratic counterparts. Studies by Schultz (2001), Gelpi & Griesdorf (2001), Tomz (2007), Kurizaki & Whang (2015), and Chiozza (2017) operationalized these mechanisms using surveys, concluding that democratic institutions do, in fact, create the preconditions for audience-based punishment. In this way, ACT became tightly linked to the institutional machinery of democracies.

However, this “consensus” did not go unchallenged. An alternative strand of scholarship started questioning whether AC were exclusive to democracies, or whether other regime types can also face similar domestic constraints. On this issue, Jessica Weeks (2008; 2014) offered one of the most influential interventions. Her work disaggregated authoritarian regimes into subtypes—such as military, single-party, and personalist—and showed that AC were more likely in regimes with institutionalized leadership selection processes and elite coordination mechanisms. Even in autocratic regimes where leaders depend on a loyal, ruling coalition, albeit how small it is, or military endorsement, retreat in foreign policy can signal weakness, erode confidence, and even trigger removal from office.

Building on this framework, scholars like Frantz (2003; 2007a; 2007b), Conrad, et al., (2014), and Geddes, et al. (2014; 2018) examined leadership turnover and regime durability in authoritarian systems. Their findings underscored, authoritarian leaders, while not directly accountable to voters, are often vulnerable to elite defection, party fragmentation, or military coups—especially when foreign policy failures reveal diminished competence. The audience in these regimes may be narrower, but the

costs of failure can be more immediate and existential. This strand of the literature emphasized that AC need not be democratic to be real; they can arise wherever leaders are embedded in relationships of political dependency. Other scholars extended the authoritarian turn in more nuanced directions. Weiss (2013) explored how Chinese leaders, situated within an autocratic state, use state-sanctioned public protest to send signals of resolve to foreign adversaries, but also noted that such mobilization creates domestic pressure not to back down. Weiss and Dafoe (2019) formalized this with the concept of “AC capacity,” arguing that what matters most is not regime type per se, but whether any actors—public or elite—have the capacity and incentive to punish a leader for inconsistency. Li and Chen (2021), drawing on Chinese survey data, demonstrated that even under conditions of media control, nationalist publics could impose reputational costs on leaders perceived as weak or inconsistent. Similarly, Toepfl (2020) and Piotrowska (2020) highlighted how even in highly repressive settings, autocratic leaders must carefully manage public narratives, often through attempts of reframing their threat retreats, and symbols of strength, lest they suffer internal challenges or legitimacy crises.

These contributions collectively broadened the theoretical scope of ACT. They challenged the view that AC require free elections or a robust public sphere. Instead, they revealed that different regime types generate different kinds of audiences and costs, from coups and purges to elite ostracism or institutional marginalization. This addition of autocracies into the mixture has widened the empirical domain of the theory, making it more inclusive—but also reinforcing a new systemic binary. By reframing the literature around a contrast between democracies and autocracies, the field began to reproduce a different kind of conceptual rigidity: the idea that ACT

can be cleanly mapped between two regimes, each with distinct internal logics.

Within this binary, anocracies—also referred to as hybrids—remained conspicuously absent.

2.2.2.1. Disaggregating the Binary: Anocracies

Anocracies are regimes that straddle the line between democracy and autocracy, combining elements of both with the possibility of one day evolving into either one of them via democratization or democratic backsliding. They may hold elections but undermine them through repression or fraud. They may tolerate some media but suppress dissenting voices through the same tools. They often lack full electoral accountability, but neither do they enjoy the cohesive elite structures of consolidated authoritarianism (Diamond 2002; Bogaards 2009; Gilbert & Mohseni, 2011; Mufti, 2018). Yet despite their empirical prevalence, anocracies have been swept under the authoritarian umbrella or excluded from analysis altogether due to conceptual ambiguity or coding difficulties, which can now be overcome thanks to V-Dem's democracy score measures.

Yet, this omission is significant. Anocracies are no longer marginal or transitional; they now make up a substantial portion of the global political landscape as they steadily increased in number since Cold War's end (Levitsky & Way, 2002; 2010).

As Diamond (2002) observes, while democratic backsliding has become more common and efforts toward democratization gradually faltering globally, the distinction between regime types has grown increasingly blurred, moving away from a simple democracy-versus-autocracy dichotomy, commonly made for a long time.

Despite this, ACT research has yet to develop a comprehensive or comparative approach on anocracies. Meaning that, these burgeoning regimes remain conceptually sidelined and empirically underexamined, caught in the shadow of a debate that remains fixated on the democratic-autocratic divide. This thesis addresses that gap. By foregrounding anocracies as a meaningful regime type in their own right—and by comparing their dynamics with both democratic and authoritarian counterparts—it seeks to expand the empirical scope and conceptual precision of ACT.

2.2.3. Third Wave: Three Types of ACT Micro-Foundations

As ACT evolved, a third wave of scholarship emerged that shifted focus from macro-level regime typologies to the micro-foundations. Rather than asking which regime types can produce AC, ACT literature probed deeper into the conditions under which domestic audiences perceive the issued threats, evaluate backdowns, and respond with political consequences. This marked a shift away from structure and toward agency—placing greater emphasis on the internal determinants of audience behavior and their variability across different conditions. Given the plethora of studies on micro-foundational aspects, these certain factors can better be observed once they are compartmentalized under three categories, namely systemic, situational, societal factors. Following subsections will first detail three types of micro-foundations studied in the literature and then underline one crucially missing micro-foundation, affective polarization.

2.2.3.1. Systemic Factors

Systemic explanations provide a nuanced understanding of how differences between political systems, including but not limited to regime type, and leaders in charge of them influence the likelihood and severity of AC. One critical aspect is the nature of a leader's threats—whether they are rhetorical or involve military action. Leaders issuing public or military threats face higher AC because such actions have higher costs in terms of signals, demonstrating a serious commitment (Kertzer & Brutger, 2016). Unlike rhetorical threats, which may be dismissed as mere posturing, military threats are viewed by both the public and international actors as more binding. Once made, these threats create a significant reputational cost if the leader backs down, as retreating would signal a failure to fulfill a public promise.

Another key factor influencing AC is the leader's ability to justify their retreat. As Wong (2024) observes, leaders who are adept at framing their backdown as a calculated or strategic decision—either by emphasizing international pressure or portraying it as a victory in disguise—can reduce the AC they incur. Moreover, by managing through public perceptions, these leaders can transform a perceived weakness into a show of pragmatism or long-term planning, effectively shielding themselves from immediate political fallout. Regime stability is also pivotal in determining AC. In more stable regimes, leaders can absorb the costs of retreating more easily because they often have stronger institutions and a more secure grip on power (Yu, Whang, & Lee, 2022). Stability provides leaders with the leeway to navigate foreign policy setbacks without immediate risk to their political survival. In contrast, leaders in unstable regimes are highly vulnerable to AC. These leaders often

face intense scrutiny, making it easier to exploit any signs of weakness, making the cost of backing down greater (Hauenstein, 2020; Kohama et al., 2023).

How a leader ascended to power also plays a crucial role in shaping AC. Leaders who rise onto personalist or military regimes are particularly vulnerable because their authority is often based on displays of strength and decisiveness (Geddes, & et al., 2014; 2018). In these contexts, backing down from a public commitment not only risks losing international credibility but also undermines the leader's domestic legitimacy, as they are expected to maintain an image of control and dominance. Leaders' track record in previous crises also plays a significant role in shaping future AC dynamics. Leaders who have repeatedly backed down from past crises are likely to face greater scrutiny in subsequent disputes, as their domestic and international credibility may have already eroded (Nomikos & Sambanis, 2019). On the other hand, leaders who have consistently followed through on their commitments in earlier crises may benefit from greater leeway, as their threats or promises are perceived as more credible. Interestingly, threat sending leader's gender also comes into play. Given the stereotypical expectation of "female leaders acting/being weak," their threats' credibility gets lowered yet still audience is more willing to punish a female leader when she fails to follow through (Scwartz & Blair, 2020).

Lastly, the leader's ability to monitor and suppress dissent significantly affects the potential for AC. In regimes where the government exercises tight control over the media and political opposition, leaders can limit public criticism and, therefore, reduce AC (Weeks, 2008; 2012). When dissent is easily suppressed, leaders can avoid the full domestic consequences of retreating from an international crisis.

Similarly, regimes with robust monitoring capabilities can preempt or neutralize political backlash by targeting potential critics or controlling the narrative (Brown & Marcum, 2011). Another fundamental systemic factor is the economic condition at the time of threat signaling. If there would be high inflation rate, high unemployment rate, or economic stagnation, leaders end up necessarily sticking with their treats since public would already be dissatisfied with the incumbent and willing to punish the leader given the opportunity (Suziki, 2018). Thus, systemic explanations highlight the importance of leadership style, regime stability, and control mechanisms in determining AC's severity.

2.2.3.2. Situational Factors

Situational explanations of AC focus on the characteristics of the crises or threats themselves. A key factor under this category is the direction of critique, which determines whether dissatisfaction is aimed directly at the leader or at the broader situation (Levy, 2012). Leaders are more likely to face AC if the public believes their decisions directly contributed to the crisis. In contrast, if the complexity of the issue or external circumstances can be blamed, the leader may avoid significant domestic costs. Issue salience plays an equally important role. Highly salient crises, such as those involving national security, territorial integrity, or deeply held economic interests, raise public expectations for decisive leadership (Clare, 2007; Hutchison & Gibler, 2013; Janusch, 2021). The more important the issue is to the public, the higher the potential costs of retreat for a leader. In such cases, backing down can be perceived as a failure to defend critical national interests, leading to political

backlash at home. A related factor is the cost of commitment. When leaders publicly commit to a certain, hefty course of action (e.g. mobilizing troops instead of a sanction), the stakes increase, as retreating from that commitment incurs reputational damage (Trager & Vavreck, 2011; Kertzer & Brutger, 2016; Zellman, 2020). The more a leader's credibility is tied to following through on a threat or promise, the more damaging a retreat becomes. Leaders who consistently fail to deliver on their commitments risk undermining their own authority, both domestically and internationally. The presence of outside observers further complicates AC dynamics (Li & Chen, 2021).

On top of these, if the number international actors, such as rival states or global organizations, that can closely monitor or involve in the given crisis increases, the domestic audience is more likely to scrutinize the leader's actions due to increasing significance of the issue (Fearon, 1994a; 1994b; Baum, 2004; Chaudoin, 2014; Brutger & Kertzer, 2018). Put simply, higher numbers of external observers add pressure to maintain a strong stance, as backing down can lead to perceptions of weakness both at home and abroad. International visibility intensifies the potential for AC, especially when the leader's actions are subject to global judgment. In certain cases, leaders may use policy substitution to manage AC. Rather than backing down entirely, they shift focus to alternative strategies or even creating alternative less costly and smaller crisis, such as economic sanctions or diplomatic disputes, to demonstrate that they are still actively addressing the crisis in different forms (Lin-Greenberg, 2019). This approach allows leaders to defuse tense situations while maintaining an image of strength and engagement, reducing the potential for AC without fully retreating from the issue at hand. Another key factor is about how the

threat is delivered. Yarhi-Milo (2013), McMannus & Yarhi-Milo (2017), and Takei (2024) find that threats privately sent (e.g. through envoy) but later publicized are less likely to generate AC, while public threats have a higher likelihood of generating costs. In conclusion, situational factors highlight the critical role that the specific characteristics of a crisis play in determining the likelihood and severity of AC. Whether through the direction of critique, the salience of the issue, or the leader's track record, these factors underscore how context shapes domestic reactions to a leader's foreign policy actions during international disputes. These situational factors, when combined with others, provide a more comprehensive understanding of how AC operates in shaping foreign policy decisions.

2.2.3.3. Societal Factors

Societal explanations focus on the characteristics and behavior of the audiences, emphasizing, for instance, how their political and ideological orientations influence the likelihood or magnitude of AC. One key factor is the existing ideological disposition of the audience. When a leader's base aligns ideologically with a strong or hardline foreign policy, backing down becomes particularly risky, as these groups expect firm stances in international disputes (Toepfl, 2020). Leaders in such contexts are more likely to face AC if they deviate from these expectations. Another significant factor is audience political awareness of the audience. In societies where the public is well-informed and engaged with politics, particularly foreign policy, leaders are more likely to incur AC for backing down (Slantchev, 2006; 2012; Potter & Baum, 2014; Quek & Johnston, 2017). Politically aware citizens tend to hold their leaders accountable for perceived failures, especially when these involve

international crises that the public follows closely. In contrast, lower political engagement can offer leaders more flexibility, as less-informed audiences may not fully grasp the stakes of the situation. In this sense, media exposure also plays a role in shaping AC. In line with this, audience party affiliation also matters. If members of the audience belong to the opposition, they are more likely to impose costs, whereas those affiliated with the incumbent party are less likely to do so (Smetana, 2025). In regimes where audiences can access to independent or foreign media more, leaders face greater scrutiny, as such media can expose inconsistencies or failures in foreign policy without the possibility of interference by the incumbent (Choi & James, 2007; Kurizaki, 2007; Croco & Hanmer, 2020; Piotrowska, 2020). This contrasts with tightly controlled media environments, where leaders can manage the narrative and reduce the risk of backlash by controlling information flow. The political preferences of the audience further influence the likelihood of AC. Leaders governing in polarized environments may face AC from opposing factions, especially if those factions are organized and capable of mobilizing support against the government (Smetana, 2025). In such cases, opposition forces can use foreign policy failures as leverage to challenge the leader's legitimacy.

The presence of opposition or rival groups within society is another critical element. Organized opposition groups, whether political parties or elite factions, can amplify the domestic costs of backing down, particularly if they frame the leader's retreat as a sign of incompetence or weakness (Uzonyi et al., 2012). Strong opposition forces are often more vocal in criticizing foreign policy failures, increasing the pressure on leaders to maintain a hardline stance. Lastly, the size of the audience supporting the regime is crucial. Leaders with a large and loyal base of supporters are better able to

manage AC because their core audience is more likely to excuse or downplay a retreat in foreign policy (Brown & Marcum, 2011). This loyal support base can serve as a buffer, protecting leaders from political fallout and allowing them more room to navigate international crises. Overall, societal explanations emphasize that the composition and behavior of the domestic audience significantly affect the likelihood and severity of AC. Leaders must consider the political awareness, ideological leanings, and organizational capacity of their domestic audience when making foreign policy decisions, as these factors play a decisive role in determining whether a leader will incur costs for backing down in an international crisis.

2.2.3.4. The Overlooked Variable: Affective Polarization

Despite the breadth of micro-foundational work in the third wave of ACT scholarship, one critical societal dynamic remains largely overlooked: affective or public polarization. Unlike ideological disagreement, affective polarization reflects identity-based divisions rooted in emotional attachments and partisan loyalty among citizens. Citizens under such conditions are more likely to interpret international crises through a zero-sum lens of group identity rather than policy consistency, turning basically the audience more hostile and grounds of costs more detrimental. This dynamic alters how publics respond to both threats and retreats—potentially heightening instability, amplifying demands for resolve, or reshaping accountability pressures in unpredictable ways (Tajfel, 1978; Brewer, 1991; Levy, 1998; Iyengar et al., 2019). This omission is striking because affective polarization has direct implications for the very mechanisms ACT seeks to explain. It affects the societal micro-foundations of AC by distorting how publics assign blame and credit,

potentially overriding traditional expectations of consistency. It reconfigures the situational foundations of AC by inflating issue salience, particularly when international disputes activate group identity or symbolic stakes. It intersects with systemic foundations, shaping how leaders gauge domestic reactions and calibrate signals under the shadow of emotionally charged, loyalist publics.

Given these implications, it is striking that affective polarization has received limited attention in ACT. Where it has appeared—such as in Levendusky and Horowitz (2012), Evers et al. (2019), Croco et al. (2021), and Park (2022)—research remains narrowly focused on experimental studies within the United States. This poses three key limitations. First is spatial: findings based solely on the U.S. context limit external validity across other democracies, let alone anocracies or authoritarian regimes. Second is generalizability: experimental designs, while internally robust, cannot fully capture the institutional or contextual complexity of real-time crises. Third is a temporal constraint: experiments assess present attitudes but cannot trace how polarization has shaped AC across past crises or over time. As a result, affective polarization remains under-theorized as a cross-regime variable and under-examined in empirical studies that go beyond controlled settings. Bridging this gap requires a comparative, real-world approach that connects affective polarization with regime type and historical crisis behavior.

2.3. Recapping the Literature and Blind Spots

To sum the parts reviewed so far, while ACT has clearly deepened the understanding of how domestic politics constrain international signaling in the cases of threats,

crucial dimensions remain overlooked. The theory has evolved through three waves—empirical validation, regime-type debates, and micro-foundational explorations—but still relies on a democracy–autocracy binary, leaving anocracies conceptually and empirically overlooked. At the same time, affective polarization—a powerful societal force that can shape how publics respond to threats and retreats—remains insufficiently theorized by mostly being confined to experimental research in democratic settings.

The following chapter addresses the two central gaps by developing a theoretical mechanism in which affective polarization operates as an interaction term across regime types. That is, the impact of democratic, anocracy, or autocratic institutions on AC incurrence will be approached dynamically contingent on the levels of affective polarization within a society. This interactional perspective offers a more layered account of how credibility is built or undermined, and more precisely, under which political configurations leaders are more or less likely to incur domestic punishment for backing down in international crises.

CHAPTER 3

THEORETICAL MECHANISM

This chapter constructs the theoretical mechanism first by defining what political regimes are and how regime stability can affect coherence of governance. It then turns to explore how the dynamics of stability operate differently across democracies, anocracies, and autocracies, shaping each regime's propensity to incur AC. With this institutional backdrop in place, the third part will introduce what affective polarization is and depict it as a moderating variable that conditions the relationship between each three regimes and their AC incurrence likelihoods. By examining how emotionally charged public divisions interact with regime-specific vulnerabilities, this section of theoretical mechanism will showcase a set of hypotheses explaining how the impact of polarization on AC varies systematically across regime types comparatively.

3.1. Political Regimes and Stability

Given this thesis's focus on affective polarization's moderating impact over different regimes in shaping their AC incurrence, it is first necessary to clarify what constitutes a political regime. One way to conceptualize political regimes is that "several interconnected social arrangements by which national communities govern their political affairs" (Gurr, 1974, 1483). These arrangements consist of both formal institutions—such as electoral rules, executive constraints, and judicial

independence—and informal norms that guide leadership behavior and citizen expectations. Together, they form “the humanly-devised rules of the game” that structure political life, primarily including, but not limited to, leadership selection, policymaking, and modes of citizen participation (Hurwitz, 1973; North, 1990). In this sense, political regimes serve as the broader institutional architectures through which authority is exercised, diffused, constrained, or consolidated.

Over decades, one foundational concern within the study of political regimes has been the extent to which these arrangements endure over time—or more precisely, their degree of stability. Gurr (1974, 1484) captures the essence of this concern when he writes, “from a functional standpoint, a certain persistent degree of civil order is a prerequisite for the effective operation of any political environment.” In other words, the ability of political systems to form and implement decisions—whether in routine governance or in the face of international crises—relies on a baseline level of internal coherence. When this coherence of governance is fragile or disrupted by factors like institutional decay, elite defection, or internal/external shocks, the regime’s capacity to act credibly and decisively deteriorates, both internally and externally.

Given this cruciality of regime stability, it is pertinent to unpack this term along with its Janus-faced companion, instability. In his seminal work, Huntington (1968) approaches political stability not as stasis or the absence of dissent, but as a regime’s capacity to absorb pressure, process demands, and respond to ever-emerging grievances without triggering systemic breakdown due to their overaccumulation. This foundational insight—stability as resilience in the presence of stress—continues to shape both empirical assessments and theoretical models of regime performance.

Goldstone et al., (2004; 2006; 2010) further this understanding by identifying core mechanisms of the ways through which grievances can be steamed out—such as institutionalized leadership turnovers via free and fair elections. These mechanisms do not suppress or resolve ever-emerging pressures of grievances but channel them out, allowing contestation to be relieved without episodes of destabilization. Thus, in stable regimes, grievances are persisted but not feared as they are expected to be politically absorbed.

On the flip side, instability refers to a regime's departure from the regularized order sustained by institutionalized political patterns, arising from its inability to absorb and resolve accumulating grievances. In this context, Huntington (1968) defines instability as a condition in which grievances build up more rapidly than institutions can manage, ultimately overwhelming the regime's buffering capacity. As this disequilibrium deepens, individuals whose grievances remain unaddressed increasingly seek and eventually prefer fundamental changes in the existing political order through a variety of means. The manifestations of such instability can occur both gradually and abruptly. As Fearon and Laitin (2003) emphasize, instability often begins with latent dysfunction—leadership crises, elite fragmentation, constitutional paralysis, or unresolved internal tensions—that slowly erode regime coherence beneath a seemingly intact institutional surface. If left unattended long enough, regardless of the regime type, these slow-burning pressures can ignite through catalytic events such as coups, mass mobilizations, contested elections, or targeted political violence (Bueno de Mesquita et al., 2005; Hegre et al., 2001; Gates et al., 2006). Though the immediate triggers may differ, the outcome remains consistent:

the disruption of authority structures and a heightened uncertainty over who governs and how.

In parallel, Dahl (1971) and Eckstein (1992) argue that regimes are most vulnerable—and most prone to take whatever steps are necessary to keep the system intact—when mounting social demands outrun the system’s procedural capacity to process them. As strain intensifies, institutions lose their buffering function: leaders become exposed to backlash and to disruptive shocks such as elite defection, mass mobilization, or even collapse. In that environment, even modest policy adjustments or de-escalatory moves take on outsized political risk—not because the moves are inherently radical, but because on shaky institutional ground they can be read as weakness and trigger cascading challenges to authority. Leaders under these conditions face a sharper international dilemma over threats and audience costs: backing down may reduce external tensions, yet domestically it can embolden rivals, alienate supporters, and unravel fragile coalitions. As institutional resilience erodes, the perceived domestic price of compromise rises, nudging incumbents toward escalation, rigidity, or brinkmanship—even when prudence would otherwise counsel restraint.

The same stress logic maps differently onto regime types. In well-institutionalized democracies, vulnerability is predictable: accountability via elections, a free media, and routine oversight makes the penalty for bluffing and retreat salient and legible. Because that domestic sanction is expected, targets treat democratic threats as more binding. In fragile hybrid systems, vulnerability is erratic—punishment may be severe but is uncertain and uneven—so external audiences cannot reliably infer a

domestic cost to backing down. In insulated autocracies, elite compacts and coercive capacity dampen mass pressures; public disapproval seldom converts into leadership turnover, keeping the expected domestic cost of retreat low. Taken together, this yields a clear directional expectation: as regimes become more democratic, the expected domestic penalty for retreat increases in a systematic, publicly observable way, and external audiences anticipate it—making leaders’ threats appear more credible and reducing the likelihood that targets reciprocate.

H1: *The more democratized a regime gets, meaning having higher polity scores, higher the likelihood of AC incurrence becomes, making their threats to be perceived more credible by targets.*

Given the importance of regime stability and instability, the next subsegment of this thesis’s theoretical mechanism turns to the stability dynamics within three major regime types: democracy, anocracy, and autocracy. While these categories share some overlapping features, each interacts with stability in their own distinct ways. These regime-specific dynamics help illuminate why the domestic political consequences of backing down from international threats—what ACT seeks to explain—do not operate uniformly across all political systems. As the analysis proceeds, these foundational differences in institutional durability will provide essential context for understanding how affective polarization may amplify or constrain the political repercussions of foreign policy decisions.

3.1.1. Audience Cost and Democratic Stability

In terms of political stability, democracies are widely regarded as the most normatively constrained and steady regime type (Eckstein, 1961; Feng, 1997; Brownlee & Miao, 2022). But what accounts for this stability? The answer lies in their intricate institutional architecture. To elaborate, democratic regimes are equipped with a broad array of mechanisms that cater public grievances without over-accumulating and escalating them into systemic upheavals over time. In this sense, their internal stability stems from rule-based arrangements that distribute political power across multiple actors and hold leaders accountable through routine processes of oversight and competition. Structural features such as free and fair elections, judicial independence, parliamentary scrutiny, and a vibrant civil society serve as stabilizing forces by regulating executive authority and reinforcing legal constraints (Dahl, 1971; Huntington, 1996). In these systems, incumbents are required to operate within codified rules, while opposition parties are institutionalized as legitimate competitors rather than existential threats. Importantly, these arrangements allow political grievances to be channeled through legal and electoral pathways, thereby preventing their eruption into regime-threatening disruptions, meaning instabilities. In this sense, democracies offer robust opportunity structures—institutionalized mechanisms through which dissent and demands are processed without undermining the regime's survival (McAdam, 1999).

The political constraints placed on democratic leaders—such as mass suffrage, separation of powers, and judicial review—prevent the concentration of authority and reduce the likelihood of executive abuse. As Maoz and Russett (1993) emphasize, these domestic constraints not only stabilize internal governance but also shape how democratic leaders behave internationally. The expectation that power is

exercised through restraint and procedural legitimacy at home extends outward, informing how democracies conduct foreign policy. Democratic leaders are therefore seen as operating within systems that discourage erratic escalation or arbitrary decision-making. As a result, democracies are widely perceived as highly stable regimes—characterized by institutional continuity, predictability, and rule-bound authority.

This foundation of institutionalized opportunity directly shapes how democratic regimes signal during international crises and to what extent they are susceptible to AC incurrence. In these contexts, political transparency, media scrutiny, and institutional oversight make foreign policy actions highly visible to domestic audiences. According to Fearon (1994a), threats issued by democratic regimes' leaders function as public commitments: they are more observable, politically salient, and difficult to retract without consequences. If a leader subsequently retreats, domestic actors—voters, opposition parties, and the press—may interpret the reversal as a failure of resolve, prompting reputational costs, electoral backlash, or formal sanction. What enables this process, as Weingast (1997) underscores, is the shared societal expectation that institutional breaches or signs of weakness will provoke unhindered collectivized resistance. This widespread anticipation of and capacity for political consequences creates the domestic environment in which AC is incurred. Democratic leaders are deterred from bluffing not just by abstract norms, but by institutionalized accountability and public exposure.

When democratic institutions function coherently—featuring transparent governance, well-defined political roles, and a politically engaged citizenry—the mechanisms of

accountability remain intact. In these democracies—those with transparent decision-making, active civil society, and robust partisan competition—retreat from an international commitment provokes severe political costs, as voters, the media, and opposition parties can trace responsibility and mobilize disapproval. As a result, democratic threats retain their weight as highly credible. This credibility stems not only from the institutional capacity to impose costs but also from the predictability of how those costs will be applied: leaders know their standing will suffer if they renege, and audiences know they have effective channels to punish them. This shared understanding generates a self-reinforcing loop, where leaders are deterred from issuing empty threats and opponents abroad are deterred from dismissing them as bluffs. In effect, the coherence of democratic institutions transforms domestic accountability into an external resource, projecting reliability onto the international stage. Consequently, the very features that sustain democratic legitimacy at home—transparency, competition, and civic engagement—also elevate the credibility of signals abroad, making democratic states more effective in coercive bargaining than regimes where accountability is diluted or distorted.

H1a: *Regimes with high polity scores, namely Democracies, have a higher likelihood of AC incurrence compared to anocracies and autocracies, making their threats to be perceived highly credible by targets.*

Building on these, H1a positions the thesis to double-check a cornerstone claim that runs from Fearon's original formulation through the second wave of ACT research: namely, that democracies are systematically more likely to incur audience costs, and thus their public threats are more credible than those issued by autocracies or

hybrids. Fearon's logic ties credibility to predictable domestic penalties for retreat—penalties that are more visible and enforceable where elections, a free press, and opposition oversight make foreign-policy reversals legible to voters and rivals. The second-wave literature translated that logic into testable propositions and found supportive evidence using diverse strategies—elite cues and domestic openness (Schultz, 2001), experimental and survey designs tracing voters' willingness to punish bluffing (Gelpi & Griesdorf, 2001; Tomz, 2007), and structural estimators that recover latent costs even when they are not directly observed, with related work reinforcing the institutional channels (Chiozza, 2017). This thesis revisits and strengthens that claim on three fronts.

3.1.2. Audience Cost and Autocratic Stability

Autocratic regimes are built on an entirely different foundation of political stability than democracies. These regimes can display considerable longevity, but the foundations of their political stability differ markedly from those of democracies. Stability in autocratic systems is not built on institutionalized channels for resolving conflict or airing dissent. Rather, it rests on the capacity to suppress opposition and to prevent demands from surfacing at all in the first place. These regimes are structurally designed to concentrate power in the hands of the executive, while subordinating other institutions such as legislatures, judiciaries, and the media to regime control (Svolik, 2012; Geddes, et al. 2014; 2018). Similarly, civil society is either co-opted to fabricate legitimacy or repressed, and political opposition is either delegitimized or eliminated.

Despite these, autocracies can remain stable for long periods. However, this stability is not a product of resilient institutional pluralism but of coercive insulation. Leaders maintain power through a combination of repression, clientelist networks, loyalty monitoring, and elite co-optation (Frantz & Kendall-Taylor, 2014). The size of the autocrat's support coalition is often narrow, and regime survival depends on maintaining loyalty within that inner circle (Bueno de Mesquita et al., 2005). Political opportunity structures—defined as the institutional mechanisms for expressing and absorbing dissent—are nearly non-existent outside this elite stratum. For the general public, grievances cannot be voiced without repression; for insiders, dissent must be channeled through informal, high-stakes bargaining. Because grievance expression is so costly due to excessive repression, most autocratic regimes avoid political compromise and operate under a logic of preemption. The suppression of dissent can create a deceptive appearance of stability. As Escribà-Folch (2013) and Geddes, et al. (2018) argue, the inability to process discontent through regularized mechanisms makes instabilities to emerge not incrementally but through abrupt shocks—coups, elite defection, or security crises—especially when the ruling coalition begins to fracture or when the leader appears weak.

These domestic dynamics carry direct implications for international signaling and threat credibility. ACT, as articulated by Fearon (1994a), posits that leaders are more credible in foreign crises when domestic audiences can punish them for backing down. Since autocracies lack clear public accountability mechanisms such as free and fair elections, free media, or protest rights, they have often been viewed as unable to generate AC. However, more recent scholarship contests this assumption. Weeks (2008, 2012) shows that authoritarian regimes can, under certain conditions,

generate AC through elite-based mechanisms. In regimes where political survival depends on maintaining the confidence of powerful insiders—military officers, party officials, or security chiefs—retreating from a public threat can signal weakness, embolden challengers, or provoke internal defection. This logic is particularly acute in personalist regimes, where the leader’s domestic authority and international posture are closely bound together. Foreign threats and militarized disputes are used not only to advance strategic goals but also to reinforce internal legitimacy. In such cases, backing down can create reputational costs that reverberate across the elite network. Guriev and Treisman (2019) describe these systems as “informational autocracies,” where survival rests on projecting competence, nationalism, and strength through carefully curated narratives. Any public reversal that undermines this image risks exposing the regime to destabilizing vulnerabilities.

When it comes to issuing threats, the contrast with democracies is clear. Democratic leaders face the prospect of public backlash—channeled through elections, media, and institutional checks—whereas autocratic leaders operate in a fundamentally different signaling environment. The public sphere is tightly controlled, and avenues for dissent are actively suppressed, meaning the broader public rarely imposes meaningful political costs for foreign policy inconsistency. Yet, as the literature shows, this absence of mass accountability does not eliminate AC altogether. Instead, the relevant audience shifts upward: regime insiders, elite coalitions, and coercive apparatuses. These actors may not openly challenge the leadership, but they are attentive to signs of faltering resolve. A failed threat or visible retreat can be read as weakness, triggering elite defections, coup plotting, or recalibrations of loyalty.

For external observers, this structure of risk can still lend credibility to autocratic threats—not because leaders are accountable to voters, but because their survival is contingent on sustaining elite loyalty under precarious conditions. However, the likelihood of AC incurrence in autocracies is contingent on the type of autocracy. In personalist systems, where authority and image are fused, AC risks may be higher; in party-based or military regimes, elite dynamics may diffuse or constrain them differently. Autocratic stability, in this sense, is suppressive rather than participatory—it masks latent instability beneath the surface of coercion. While traditional AC mechanisms tied to public contestation are absent, the logic of credibility persists, albeit in muted and conditional form.

H1b: *Regimes with low polity scores, namely Autocracies, have a lower likelihood of AC incurrence compared to democracies and anocracies, making their threats to be perceived less credible by targets.*

3.1.3. Audience Cost and Anocracies' Instability

Anocracies present the most ambivalent and distinct picture compared to their counterparts. They operate in the transitional space between democratic and autocratic rule, blending features of both but mastering neither. Elections are often held, yet they are routinely manipulated through fraud. Civil society and media pluralism exist but only within narrow limits, shrinking whenever dissent grows threatening. These regimes are not simply “on the way” to becoming democracies or

dictatorships; rather, they constitute a distinct type marked by shallow institutionalization and deep contradictions (Diamond, 1999; Wigell, 2008; Bogaards, 2009; Fuller, 2016). Far from evolving in a linear fashion, many remain trapped in an unstable equilibrium, swaying between democratization or democratic backsliding.

This instability stems from structural incoherence. Pate (2020) describes anocracies as exhibiting “dual tendencies,” simultaneously concentrating and diffusing power. Gates et al. (2006) similarly label them “institutionally inconsistent”—incapable of enforcing cohesion through repression, yet also incapable of resolving disputes through inclusive democratic mechanisms. The consequences are visible in empirical patterns: Hegre et al. (2001) and Regan & Bell (2010) demonstrate that anocracies experience political instability at an annual rate of 3.7%, substantially higher than both democracies (0.8%) and autocracies (1.3%). Huntington’s (1968) insight offers an explanation: instability emerges when citizens’ demands accumulate faster than political institutions can process them, in return seeking regime-wide changes via acts like protests.

In practice, leaders of anocracies confront a mounting backlog of unresolved grievances. When faced with such pressure, concessions do not necessarily stabilize the regime; instead, they often embolden opposition forces and heighten regime fragility. This is because, as McAdam (1982), Muller & Weede (1990), and Gandhi & Vreeland (2004) highlight, the coexistence of visible mobilization opportunities with weak institutions creates fertile ground for unrest. Under these conditions, leaders often lean toward riskier strategies: as Düveroğlu and Tokdemir (2024) note,

leaders of anocracies, especially the ones with fragile tenure and short time horizons, frequently calculate that escalation abroad is safer than retreat at home, since concession risks triggering collapse.

The personalization of responsibility further reshapes the logic of audience costs in anocracies. In democracies, responsibility for foreign policy failure can be dispersed through institutional mechanisms such as legislatures, party competition, or independent media scrutiny. In autocracies, dissent is largely contained within elite circles, and the broader public remains excluded from meaningful influence. By contrast, leaders of anocracies lack both the insulating coercive power of autocracies and the institutional buffers of democracies. As Zarpli (2022) observes, the partial presence of democratic channels—however defective—means publics retain the ability to mobilize when opportunities arise, especially under the stimulus of external shocks. Sanctions, military failures, or diplomatic defeats can therefore provide opposition groups with a rare window to overcome collective action problems. This dynamic creates stronger incentives for leaders of anocracies to act on their threats than autocrats, since any perception of weakness risks rapid mobilization against them. Yet at the same time, regime fragility means that foreign policy failure or retreat cannot be offloaded onto institutions; it reflects squarely on the leader's personal legitimacy. Frantz (2012), Pate (2020), and Marshall (2023) all show how leaders of anocracies frequently dismantle institutional constraints, concentrating authority in their own hands. This personalization leaves them directly exposed to the political consequences of external reversals.

The logic of risky external behavior is further reinforced by the search for legitimacy. In such contexts, foreign policy adventurism often becomes a compensatory tool. Onega (2024) argues that insecure regimes pursue status-seeking abroad to offset domestic fragility, using external displays of resolve to bolster internal legitimacy. Yet this strategy paradoxically raises the stakes of credibility: once foreign policy resolve becomes intertwined with regime survival, backing down in an international crisis threatens not only foreign policy standing but also domestic coalitions and the leader's very grip on power. Retreat in such circumstances is not merely interpreted as political weakness but as betrayal, carrying the potential to trigger elite defection, mass protest, or even regime collapse.

Thus, while Fearon (1994a) theorized audience costs primarily through the democratic mechanisms of transparency and accountability, anocracies generate credibility through vulnerability. Their threats appear credible not because institutions support them, but because institutions cannot cushion the consequences of failure. Audience costs in these regimes are neither as low as in autocracies, where repression insulates leaders from mass backlash, nor as systematically high as in democracies, where institutions and public scrutiny ensure accountability. Instead, they are contingent and mid-level, produced by the chronic mismatch between accumulated grievances and insufficient institutional capacity to process them. This leaves leaders of anocracies with few safe exits: escalation often appears the less dangerous option compared to concession. In this sense, audience costs in anocracies arise less from democratic-style accountability than from the inherent peril of concession in a fragile political order—an order where retreat risks unraveling the regime itself.

H1c: *Regimes with mid-range polity scores, namely Anocracies, relatively have higher AC incurrence likelihood than Autocracies but lower than Democracies, making their threats to be moderately credible.*

3.1.4. Recapping Stabilities across Regimes

Taken together, the stability profiles of democracies, autocracies, and anocracies provide a baseline for understanding how audience costs take shape. Democracies, grounded in institutionalized competition and transparency, tend to generate the highest likelihood of AC, though their force ultimately hinges on the degree of political cohesion. Autocracies, by contrast, rely on repression to shield leaders from mass contestation, producing the lowest propensity for AC, with costs surfacing primarily in episodes of elite fracture or crises of legitimacy. Even within autocracies, the logic of AC is contingent on subtype—almost absent under personalist rule, channeled through elite cohesion in party or military regimes, and tied to dynastic legitimacy in monarchies. Anocracies occupy the unstable middle ground: institutional incoherence and volatility expose leaders to a broad and unpredictable spectrum of domestic punishments, leaving their threats only moderately credible and highly context-dependent. These regime-specific logics prepare the ground for the next section, where affective polarization is introduced as a pivotal intervening force—one capable of dulling, magnifying, or fundamentally recasting the credibility of international threats.

3.2. Audience Cost and Affective Polarization

While nations or their incumbents operate through specific institutional and normative configurations designed to manage the accumulation of grievances, the actors embedded within these systems—namely the public—are far from passive. Put differently, public opinion, which forms the bedrock of ACT, is not simply a byproduct of state's institutional machinery. Rather, it works through that machinery via intricate emotional and cognitive attachments, often rooted in partisan and social identities. In line with this, because these alignments are concurrently working intertwined with regime-specific institutions, the same social stimulus of backing down can generate different audience reactions depending on a regime's baseline accountability and its capacity to absorb grievances. In other words, audiences judge foreign-policy behavior not only by institutional rules or procedural norms but also through the cognitive prisms that shape their political alignments. Given this, the domestic political consequences of backing down shift accordingly, prompting a recalibration of how international threats are conveyed and interpreted by target states for each sender.

In this sense, polarization—especially the affective or public kind—stands out as a powerful yet understudied force in shaping audiences' reactions to leaders' actions or inactions on the international stage. As emphasized in the literature review, although a growing body of ACT scholarship has examined several micro-factors, the role of affective polarization in shaping AC remains under-theorized—particularly when considered alongside regime-specific conditions. This thesis therefore asks not only how regime type affects the likelihood of AC, but also how affective polarization

interacts with each regime's institutional features in moderating that likelihood. In short, polarization here is treated as a conditioning force whose impact depends on the regime into which it enters rather than as a uniform amplifier or dampener across all systems.

To situate affective polarization conceptually, it is useful to recall its roots in social psychology. As early as the 1970s, Tajfel et al. (1971; 1978; 1979) showed through social identity theory that individuals derive self-worth from group membership, distinguishing between benign in-groups ("us") and malign out-groups ("them") even in arbitrarily assigned settings. These foundational studies demonstrated that alignments of loyalty prompted by group cohesion serve as powerful interpretive lenses for, shaping how individuals perceive both their own actions and those of others. Building on this, Lipset and Rokkan's (1967) seminal work shows how durable social cleavages (based on matters like class, religion, region) can channel political engagement away from rational topic-based deliberation and toward often hostile identity-based contestations, such that group interests regularly trump issues themselves. Contemporary work similarly finds that partisan identities increasingly produce animus that exceeds policy disagreements, with opponents treated in more antagonistic, even moralized terms (Iyengar, Sood, & Lelkes, 2012). When these identity lenses overlay regime institutions, they reweight how domestic audiences evaluate foreign-policy resolve, retreat, and the legitimacy of punishment.

This elevated animosity and restructuring of within regime dynamics are directly tied to the "risky shift" phenomenon, a detrimental upshot of affective polarization for group decision-making processes in the face of high-stake issues. As demonstrated

first by Stoner (1968), when individuals deliberate in groups—especially under high stakes involving prestige, honor, or survival—they tend to favor riskier courses of action than they would independently. This tendency is especially pronounced when alternative decision-making channels are blocked, distrusted, or absent. Several mechanisms jointly produce this shift: responsibility diffuses when many share the burden of a risky decision, reducing perceived personal costs (Myers, 1978); repeated interaction encourages members to match or exceed the prevailing stance to signal loyalty and commitment (Cartwright, 1971); and echo-chamber dynamics can reinforce moral certainty and a sense of consensus (Moscovici & Zavalloni, 1969). Extending this logic to politics, Smith et al. (2024) show that citizens in highly polarized contexts—precisely because they are tightly cohesive, morally convinced, and shielded from countervailing information—become more willing to endorse confrontational or even unlawful collective action that they would not risk individually doing so, such as mass protests, once faced especially with high-stakes issues. In such environments, risky shift is not merely a psychological curiosity but a driver of political behavior, pushing group decisions toward uncompromising, high-stakes actions that individuals might otherwise avoid and increasingly framing politics as a zero-sum contest where retreat is equated with defeat.

These dynamics travel directly to AC incurrences. Regimes differ in their baseline ability to absorb grievances and maintain institutional stability; they also differ in how they translate public pressure into political punishment. Under polarization, these differences sharpen. AC is not generated by an abstract, unified “public” but by groups whose willingness to punish retreat is shaped by emotional loyalties and identity-based narratives. In polarized societies, groups are not just divided; they are

locked into us-versus-them antagonism—often moralized—such that in-groups may defend even failed threats as necessary while out-groups condemn successes if they accrue to the “wrong” side. When this antagonism is coupled with risky-shift dynamics—heightened in echo chambers and reinforced by the diffusion of responsibility—polarized groups, especially where alternative channels are weak or distrusted, can behave like cornered actors. United in purpose yet constrained in choice, they lean toward more extreme and zero-sum actions. To external targets, this configuration signals that domestic retreat could be politically perilous for the leader, raising the perceived likelihood of follow-through and, in turn, the credibility of the threat.

However, it should also be reminded that one core point for this thesis is that polarization does not simply “add” or “subtract” AC in the same way everywhere. It reworks how each regime’s institutions process public pressures. In one system, heightened polarization may disperse audience reactions into various preexisting channels making the AC incurrence perception a noisy signal; in another, it may detrimentally channel pressure into elite defection or street mobilization; in still another, it may be blunted or redirected by coercive controls. Accordingly, once polarization enters the picture, meaningful comparison is best made within each regime: start from that regime’s ordinary, low-level polarization baseline and then ask how the likelihood of AC changes when affective polarization runs high. Cross-regime contrasts can still provide context, but the theory and tests that follow center on how polarization alters each regime’s internal logic relative to its own starting point, rather than on a single ranking that would blur these conditioning effects.

With that framing in place, the next three subsections proceed regime by regime. For each, we first describe the baseline: how audiences and institutions ordinarily generate or mute AC under ambient polarization. We then show how heightened affective polarization modifies that internal process—where it amplifies pressures, where it redirects them, and where it short-circuits them. The hypotheses are stated accordingly as within-regime expectations about movement from the baseline. This structure retains the comparative sensibility of ACT while ensuring that, once polarization is introduced, regimes are assessed first and foremost against themselves.

3.2.1. Audience Cost in Democracies with Polarization

At normal times democracies signal highly credible threats because institutional channels diverse grievances into a coherent sanction. Even if some low level of polarization exists, hostility is bounded by shared procedures and norms; the ordinary frictions of politics are contained by rules that facilitate aggregation rather than splintering. Forums of voice—parliamentary scrutiny, investigative journalism, civil society mobilization, and, ultimately, elections—tend to add up rather than talk past one another. Foreign targets observing this pattern see a sender backed by an audience whose sanction, if the leader retreats, is both likely and legible. In short, free media, legitimized opposition parties in the congress, independent oversight bodies, and competitive elections ordinarily work in tandem: thus, retreats from external commitments predictably invite a political price. For this reason, democracies are often assumed to have a high propensity for incurring AC (Fearon, 1994a; 1994b). This is the ambient-polarization picture: the system's multiplicity of

channels functions as a bundle, converting diffuse discontent into a single, intelligible political cost.

Heightened affective polarization reworks that internal process in ways that chokeholds the system. This change is especially consequential because it scrambles the ordinary hierarchy of democratic values. Ordinarily, citizens in democracies accept that leaders may come and go, as elected servants of public's interest, subject to replacement, while the grievance-addressing channels of the system must endure. Stability of the rules takes precedence over the fate of any single leader or party (Rawls, 1971; Bowles, 2016; Carugati, 2020). Under high polarization, that hierarchy blurs: defending the camp substitutes for defending the system (McCoy & Somer, 2019). Loyalists reinterpret retreats as prudence or strategic patience, while opponents recast the same events as proof of weakness (Arbatli & Rosenberg, 2020). Public-opinion evidence mirrors this asymmetric filtering: incumbent camps discount bad news; opposition camps amplify it (Cilizoglu, Tokdemir, & Zarpli, 2025). Echo chambers magnify certainty, and diffusion of responsibility lowers the perceived personal cost of endorsing riskier collective stances. The upshot is deeper entrenchment of factional loyalty at the expense of the credibility ordinarily supplied by institutional stability.

As hostilities mount right up to the risky-shift incentives, the very multiplicity of democratic channels that once bundled discontent into a single sanction begins to splinter across partisan battlegrounds. Parliamentary inquiries are reframed as ritualized theater and answered with minority walkouts; investigative coverage is met with counternarratives and media siloing; oversight hearings become venues for

identity affirmation rather than fact-finding; protests meet counter-protests; and even elections are cast as existential plebiscites on group survival rather than as routine instruments of accountability. Pressure that once added up becomes offsetting and diffused across parallel partisan tracks. Leaders contemplating retreat now face a cacophony of cross-pressures they can navigate or survive with loyalist cover—helped along by the diffusion of responsibility that allows blame to be spread and by echo-amplified narratives that reinterpret retreat as tactical realism for one side and unforgivable weakness for the other. Targets, watching from the outside, no longer infer a single, authoritative penalty waiting on the other side of a democratic leader’s reversal.

The result is a two-picture contrast within the same regime. In the ambient-polarization picture, leaders who back down confront a concentrated, unified sanction that foreign targets can anticipate and trust thanks to diverse arenas aggregating into a decisive cost. In the high-polarization picture, leaders face contested, offsetting punishments distributed across multiple institutions; hostility is higher, but it is channeled within camps rather than across the system, risky-shift dynamics push both sides further out, echo chambers harden interpretive frames, and diffusion of responsibility makes escalation feel safer. A retreat may embolden one camp and enrage the other, yet neither necessarily crystallizes into a regime-wide penalty. The very channels that normally guarantee coherent accountability now hobble it. From the perspective of foreign targets, the signal of credibility shifts accordingly: instead of a sender backed by a unified audience poised to impose robust costs, they see a fragmented chorus whose “costs” are contested, delayed, reversible, and easier for an incumbent to survive by leaning on in-group loyalty

Taken together, these dynamics preserve the insight that democracies are—at baseline—robust in generating AC through coordinated institutions, while also showing how high affective polarization disperses that cost as people take sides more aggressively. AC does not vanish in polarized democracies; it disperses and becomes noisier as accountability turns from system-preserving and cumulative to camp-centered and offsetting. Compared to the ambient baseline, where democratic accountability aggregates across forums and produces a single, intelligible sanction, high polarization makes accountability centrifugal and divisive, lowering the perceived likelihood that retreat triggers a decisive political price. Foreign targets update in line with this within-regime shift: threats that would have been read as credible under ambient polarization look safer to test when polarization is high.

H2a: *The more polarized a democracy is, the lower the propensity to incur AC is perceived, compared to a low polarized democracy, making democratic threateners more likely to be reciprocated by the targets.*

A vivid example of polarization's such malice impact over democracies is Obama's 2013 "red line" threat over Syria's potential use of chemical weapons. After declaring that Assad's use of such weapons would trigger U.S. military action, Obama ultimately did not follow through this commitment when these weapons are used. If U.S.'s democracy at the time had been working as it should have, multiple channels of accountability—Congress, media, and public opinion—would have converged into imposing a clear penalty for such inaction, thereby strengthening the credibility of threats, as ACT envisions. Yet polarization, which was rather high at

the time, turned these channels into venues of partisan competition. Democrats largely defended Obama and framed restraint as prudence, while Republicans split between calling for military action and opposing intervention altogether. Media outlets mirrored these divides, turning the episode into a partisan dispute rather than a test of national honor, as some praised Obama, others called him incompetent. Even the broader public, still weary of Iraq and Afghanistan, failed to mobilize a unified cost for inaction like mass protest (Taddonio, 2015). The multiplicity of democratic channels thus produced dispersion rather than concentration under polarization. This happened as audience, that was expected to levy costs, moved into camps, and found ways to battle over their camp's interests rather than punishing for the retreat. On the side of threat's target, Assad could see that the United States' audience was inconsistent and divided, reducing the credibility of Obama's threat due to the faltering condition of punishment. His eventual repeated crossings of the "red line" with even more chemical attacks illustrated how polarization transforms institutional abundance into a dissonance, leaving threats under-reinforced by AC (Da Vinha, 2021).

3.2.2. Audience Cost in Autocracies with Polarization

In autocracies, the dynamics of AC diverge sharply because the foundations of stability and accountability are fundamentally different. Whereas democracies disperse accountability through elections, legislatures, courts, and the press, authoritarian regimes consolidate control by suppressing dissent and narrowing the channels through which grievances could otherwise be expressed. As a result, the ordinary public has little capacity to punish leaders who back down after making

international threats. AC are therefore structurally limited, since the mechanisms that would normally transform discontent into sanction are deliberately neutralized (Weeks, 2008; 2014). From an external vantage point, credibility hinges less on mass pressures than on the opaque equilibrium of elite control behind the leader. Adversaries update accordingly: autocratic threats are evaluated through signals about elite cohesion, coup-proofing, surveillance and coercive capacity, and the reliability of patronage networks rather than through expectations of a mass-level backlash.

Under ambient polarization, this baseline does not change in a way that would create a mass-driven AC mechanism. Everyday animus or in-group bias may exist in society, but hostility is contained by censorship, surveillance, and the regime's control of assembly and information. Echo chambers can be manufactured or encouraged by state media, yet they mainly reproduce regime narratives rather than aggregate discontent upward. Risky-shift dynamics—where like-minded groups push toward more extreme stances—are pre-empted by repression or siphoned into regime-sanctioned outlets that pose little danger to incumbents. Diffusion of responsibility within the mass public further lowers the likelihood that any one group bears costly mobilization in the absence of protective institutions. The ambient-polarization picture in autocracies is thus one of blocked aggregation: even if low-level hostility or in-group loyalty is present, the architecture of control prevents these sentiments from becoming a coherent, system-level sanction. AC remain low and primarily elite-mediated.

When affective polarization intensifies, the internal process still does not yield a mass-driven AC channel. High polarization can deepen us-versus-them hostility, energize risky-shift incentives inside like-minded clusters, harden echo-chamber certainties, and spread blame thinly through diffusion of responsibility. Yet without independent media, autonomous civil society, or permissive protest rights, these psychological and group-dynamic forces remain politically inert at the level that matters for AC: they do not translate into institutionalized penalties for leaders who retreat. Often the regime absorbs or instrumentalizes them. Loyalist echo chambers can be mobilized to reframe retreat as strategic prudence or necessary sequencing; diffusion of responsibility relieves individuals of the perceived cost of toeing the regime line; opposition echo chambers, lacking safe channels for aggregation, struggle to convert indignation into enforceable punishment. As with the ambient case, the credible threat of sanction—if it exists—originates among actors capable of removal or sidelining: party bosses, security chiefs, military cliques, and key economic patrons (Frantz & Kendall-Taylor, 2014). The high-polarization picture, then, looks like the ambient one in its bottom line: mass affect does not move AC unless it penetrates the ruling coalition or forces open channels the regime otherwise keeps shut.

It is important here to reclarify scope before drawing the contrast. This thesis centers on public (affective) polarization; if the focus were instead political/elite polarization, the within-autocracy picture could look different across regime types. Elite polarization—fragmentation among top military, party, or security actors—can sharpen removal risks in personalist or military juntas (making retreats more dangerous and thus potentially raising AC), while in single-party autocracies strong

organizational cohesion may buffer leaders against elite splits, keeping AC low even amid policy failures. Likewise, elite-level informational polarization (selective exposure inside the ruling coalition, strategic misinformation among principals) could distort assessments of foreign-policy risks and alter coup-coordination thresholds, shifting AC through elite—rather than mass—channels. In short, were the object of analysis elite polarization rather than public polarization, we would expect meaningful variation across autocracy types in how “backing down” translates into removal risk. With that scope clarified, the resulting two-picture contrast for autocracies under public polarization is distinctive: in the ambient-polarization picture, mass sentiments—though present—are prevented from aggregating; AC are low and hinge on elite cohesion. In the high-polarization picture, mass hostility, risky-shift tendencies, echo chambers, and diffusion of responsibility may intensify, but they remain contained or harnessed by coercive and censorial architecture. The locus of potential sanction thus stays where it began: among elites with the autonomy and means to impose consequences. From the perspective of foreign targets evaluating credibility, the signal is therefore stable across levels of public polarization: what matters is not whether the public is angrier or more loyal, but whether the elite coalition is unified, disciplined, and coup-proof—or fracturing and primed for defection.

Circling back to the resulting two-pictures, the contrast within autocracies is therefore distinctive. In the ambient-polarization picture, mass sentiments—though present—are prevented from aggregating; AC are low and hinge on elite cohesion. In the high-polarization picture, mass hostility, risky-shift tendencies, echo chambers, and diffusion of responsibility may all intensify, but they remain contained or

harnessed by the coercive and censorial architecture. The locus of potential sanction stays where it began: among elites with the autonomy and means to impose consequences. From the perspective of foreign targets evaluating credibility, the signal is accordingly stable across public-polarization levels: what matters is not whether the public is angry or loyal, but whether the elite coalition is unified, disciplined, and coup-proof—or fracturing and primed for defection. Taken together, autocracies thus exhibit within-regime invariance with respect to mass polarization: baseline AC are limited by design, and heightened public polarization does not, on its own, create channels that convert discontent into enforceable political costs. Threat credibility in these regimes’ rests less on public affect and more on the fragile equilibrium of elite loyalty, surveillance, and the constant risk of factional recalculation.

H2b: *The changes in polarization level in an autocracy would not significantly impact the likelihood of AC incurrence, as targets threat responses will be unaffected by polarization.*

The downfall of Leopoldo Galtieri after the Falklands War illustrates this mechanism clearly. Leading a personalist military regime, Galtieri expected the conflict to consolidate his legitimacy. When the looming possibility of defeat exposed his threats as hollow, it was not masses that forced him out—citizen polarization was politically irrelevant under repression. Instead, the military elite themselves sanctioned his failure, removing him within weeks (Slantchev, 2006). This case demonstrates why some autocracies like personalist regimes are particularly susceptible to retreats: once elites perceive retreat as intolerable, leaders may face

immediate removal despite an absence of visible societal backlash. The Galtieri example therefore underscores the point that external actors judging credibility in autocracies must track elite dynamics rather than assume public polarization operates as it does in democracies.

3.2.3. Audience Cost in Anocracies with Polarization

Anocracies occupy a precarious middle ground between democracies and autocracies, with institutions that exist on paper but rarely perform as intended. Elections may be held yet manipulated, courts sit without impartiality, and legislatures debate without enforcement power. In these settings, grievances outpace the system's capacity to address them; channels of participation are distrusted; and episodes of instability recur. AC, accordingly, are moderate and volatile: leaders' threats carry some credibility because accumulated discontent can spill into backlash, but any sanction is uneven and hard to sustain because the institutions that would normally process and signal punishment are thin and selectively applied. It is a slow-burn fire—the heat is real, but containment sometimes holds.

Affective polarization adds fuel to this fire. By tightening in-group bonds and reinterpreting politics through an “us vs. them” lens, it lowers hesitation to act collectively and converts private frustration into shared conviction. Within like-minded camps, risky-shift dynamics push preferences toward bolder stances; echo chambers provide moral certainty and filter out moderating cues; and diffusion of responsibility dulls the personal risks of joining collective action. In democracies, resilient institutions can absorb and deflect these surges; in autocracies, coercion and

elite cohesion often smother them. In anocracies, where channels already struggle, polarization strips away what little credibility they retain. Opposition actors conclude that formal avenues are pointless; loyalists treat the same institutions as tools to bend or bypass. Grievances stop moving through committees, courts, or regulated forums and instead accumulate in extra-institutional networks that are primed for confrontation—sometimes contentious, sometimes violent—because that is where participants believe action has consequences.

For leaders, this is more than a political problem; it is an ontological one born of hybridity. Democratic incumbents can retreat and pay mainly at the ballot box; autocrats can often lean on cohesive security forces. Anocracy leaders have neither reliable legitimacy nor reliable insulation. In a polarized climate, where hostility is high and risky-shift incentives reward toughness, retreat reads not as ordinary weakness but as betrayal—proof that the center cannot hold. With echo-amplified narratives and distrusted institutions, there is no convincing “off-ramp” to translate anger back into orderly contestation. A step back can trigger cascading backlash: crowds that swell rather than disperse, mid-level security actors who hesitate or defect, and rival elites who re-price the risk of standing still. The leader cannot count on the system to absorb the blow; the system is precisely what people no longer trust.

The within-regime contrast is therefore stark. Under ambient polarization, anocracies generate moderate, inconsistent AC: sometimes a retreat draws sanction through ad hoc mobilization or elite wobble, sometimes it does not, and targets abroad see ambiguity. Under high polarization, the same retreat looks like a spark in dry tinder. Hostility hardens, echo chambers tighten consensus inside camps, risky-shift

pressures push organizations toward uncompromising tactics, and institutional distrust removes peaceful outlets and credible narratives for reversal. The domestic price of backing down rises sharply, not because institutions are strong (as in democracies) or because repression can guarantee order (as in autocracies), but because the leader has no dependable insulation and the public has little faith left in diffusion mechanisms. To observers abroad, the signal flips: a threat from a polarized anocracy appears more credible than the same threat under ambient conditions, because retreat now looks domestically untenable.

Two clarifications keep this picture careful and clear. First, the effect is thresholded and path-dependent: modest polarization may leave the baseline intact; credibility climbs once polarization so erodes remaining trust in channels that identity and regime fate feel fused. Second, efforts to reassert control—rapid crackdowns, rushed purges, theatrical legal moves—can backfire, validating grievances and widening elite fissures, thereby amplifying AC rather than dampening them. In short, where anocracies already smolder because grievances outpace institutional answers, polarization both aggravates the heat and emboldens actors to act—sometimes violently—at the very moment when the leader’s protective insulation is thinnest. That is why, within this regime type, polarization does not merely increase AC; it reconfigures them, transforming sporadic, containable costs into consequences that leaders cannot easily escape—and that targets abroad are increasingly unwilling to test.

H2c: *The more polarized an anocracy is, the higher the propensity to incur AC is perceived, compared to a low polarized anocracy, making threats more likely to be complied by the targets.*

The 2009 Manas air base ultimatum of Kyrgyzstan illustrates how AC incurrence in anocracies can be transformed from moderate levels into existential by polarization. At the time, Kyrgyzstan under Bakiyev was an anocracy: parliament still held votes, opposition deputies occupied seats, and an election was scheduled, yet these institutions were weak, manipulated, and distrusted. The political environment was sharply polarized, with protests both against and in support of Bakiyev surfacing nationwide (Nanovsky & Knox, 2024). In this volatile setting, Bakiyev threatened that the United States would be forced to leave Manas Air Base within six months and reinforced his stance by securing a parliamentary vote of 78 to 1, which rendered the ultimatum a public commitment that could not be easily reversed (Marat, 2009). With the looming July 2009 election, polarization further magnified the risks of retreat, as both supporters and opponents were primed to treat compromise as betrayal. Concurrently, pre-election protests framed Manas as a matter of national honor, showing how grievances spilled outside institutional channels (Mankoff, 2013). Polarization thus bounded individuals into collective camps more willing to mobilize than they might alone, a clear manifestation of the risky shifts dynamic. Recognizing that Bakiyev faced unusually high domestic costs, the United States treated his threat as credible and complied—offering an annual lease tripling payment to \$60 million and conceding Kyrgyz control over base security (Leonid, 2007). After the disputed July election, opposition leaders alleged fraud, protests reignited, and within months Bakiyev fled into exile in Minsk, confirming that U.S.

leaders had read the situation correctly. This case illustrates how threats are to be taken as credible by targets of polarized anocracies because, for these regimes, retreat risks unleashing the very domestic pressures that can consume incumbents far beyond the loss of an election.

3.2.4. Recapping Polarization across Regimes

Taken together, the above discussion shows that polarization does not operate uniformly across regimes; it reconfigures how AC are produced by altering three vitally interrelated features—deflectibility, types, and magnitude of costs—within each regime as they move from low to high polarization. The point of comparison is, thus, within each regime’s own baseline, not a single cross-regime ranking: polarization conditions the familiar channels of accountability and control, sometimes fragmenting them, and sometimes hardening them into sharper, less escapable sanctions. As these features shift, so do target perceptions of credibility—the operational proxy for AC in this study—because outsiders read threats through the likelihood that a retreat would trigger a coherent, enforceable penalty. What follows distills this logic for democracies, autocracies, and anocracies, showing how changes in deflectibility, forms, and the severity of those together trace a clear path from low to high polarization in each regime.

In democracies, the baseline features low deflectibility because multiple venues—media scrutiny, opposition oversight, courts, and elections—tend to aggregate blame into a single, legible sanction. The types of costs are predominantly institutional such as electoral losses but not existential outcomes that are seen mostly in autocracies or

anocracies. The magnitude of costs, in low polarization, is substantial and predictable enough to deter retreat or acquire compliance thanks to the potential converge of backlash coming from a multiple channels. Under high polarization, risky-shift dynamics push each side toward tougher stances, and diffusion of responsibility makes loyalists more comfortable absorbing leader-friendly narratives. The net effect is higher deflectibility: blame can be redirected into offsetting partisan frames across venues. The types of costs remain mostly institutional but become contested. And the magnitude declines relative to the baseline because accountability disperses rather than converges. Targets infer that democratic threats are easier to test when polarization runs high.

In autocracies, mass-level deflectibility is largely irrelevant at baseline: censorship, surveillance, and coercion prevent public blame from aggregating into enforceable sanctions, so leaders seldom need to diffuse fault to the public. The types of costs that matter are elite-centric (demotion, sidelining, purge, or removal by party or military actors), with staged or absent elections offering little leverage to mass audiences. The magnitude of costs is driven mainly by autocracy subtype (e.g., higher peril in brittle personalist/military juntas; lower in disciplined single-party systems), rather than by public sentiment. With high polarization, mass hostility, risky-shift tendencies, and echo-chamber effects in society remain politically inert without channels to translate them into sanction; regimes often contain or even instrumentalize them. Thus, deflectibility at the mass level stays essentially unchanged, too, types remain elite-bound, and magnitude continues to hinge on elite cohesion—not on the level of public polarization. Yet, as noted before, if one were

analyzing elite polarization, patterns could differ across autocracy types; here the focus is public polarization.

In anocracies, under-performing institutions and low trust make baseline deflectibility already inconsistent: sometimes leaders can spread blame across weak courts, partisan legislatures, or fragmented security organizations, but these exits are unreliable. The types of costs are mixed (episodic protests, uneven electoral penalties where contests still matter, intermittent elite wobble), and the magnitude is moderate yet volatile, swinging with shocks rather than following stable rules. With high polarization, already fragile channels are disrupted further. People who are weary of institutions that fail to process complaints gain more grievance and courage; risky-shift dynamics push organizations toward bolder tactics. In such cases, deflectibility drastically gets lowered—leaders have few credible ways to offload or reframe retreat in arenas no one trusts. The types of costs expand and harden (sustained mobilization, violent clashes, defections in the security apparatus, even coups, alongside any electoral losses), and the magnitude jumps from political penalties to existential stakes—office, status, wealth, and potentially personal safety. To outside targets, this combination reads as a sharp rise in AC: a polarized anocracy looks more likely to punish retreat, making its threats appear markedly more credible than in its own baseline.

Given the alteration of these three cost-related factors, Table 1 reports the expected likelihood of AC incurrence by regime and polarization level. Because target perceptions of credibility are our operational proxy for AC incurrence likelihood, the pattern also indicates how outsiders read threats in each cell: more check marks or a

positive polarization effect imply stronger credibility; fewer checks or a negative effect imply safer opportunities to reciprocate. While the right-hand column captures our primary within-regime inference, holding polarization constant also allows cautious across-regime comparisons that follow the check counts.

Table 1. AC Incurrence Likelihoods across Regimes and Polarization Levels

Without polarization, democracies show the high likelihood, anocracies are moderate, and autocracies are low of AC incurrence. When polarization rises, the

AC Incurrence Likelihoods	without Polarization	with Polarization
Democracies	✓✓✓	Negative
Anocracies	✓✓	Positive
Autocracies	✓	No Impact

within-regime effects diverge: in democracies the effect is negative—greater cost diffusivity disperses sanction across partisan venues, reducing the chance that a retreat yields a single, authoritative penalty; in anocracies the effect is positive—polarization collapses diffusivity, expands the types and raises the magnitude of costs as mobilized camps and weak institutions convert retreat into instability; in

autocracies there is no impact—mass polarization is politically inert and AC remain elite-bound.



CHAPTER 4

RESEARCH DESIGN

This chapter sets out the empirical strategy for answering two interrelated inquiries of this thesis: first, how the likelihood of AC incurrence varies across political regimes—including democracies, autocracies, and the overlooked anocracies—and second, how public polarization moderates the propensity for AC across these regime types. To address these questions, the chapter is organized into three sections. The Data section introduces the empirical sources, highlighting the scope of the observations, timespan, and the structure of the datasets used. The Variables section defines the dependent, independent, interaction, and control variables, linking each to the broader theoretical framework developed earlier. The final section, Operationalization, outlines the process of assembling and merging the data into a form suitable for statistical analysis, culminating in the logistic regression models that test the six hypotheses developed in the theory chapter. Together, these steps provide the foundation for tracing how regime type and polarization interact to shape the credibility of international threats.

4.1. Data

The empirical analysis in this thesis relies on a large-N quantitative research design that draws on culmination of multiple datasets, each contributing distinct yet complementary dimensions to the study. The core unit of analysis is the dyad–year, and the unit of observation is the directed dyad–year, meaning that each observation reflects state A’s behavior toward state B in a given year. The temporal scope extends from 1816 to 2014, encompassing 1,491 directed dyadic interactions across 193 unique countries, which provides both historical depth and contemporary relevance. The temporal scope extends from 1816 to 2014, encompassing 1,491 directed dyadic interactions across 193 unique countries, which provides both historical depth and contemporary relevance. That said, coverage is uneven across time. Earlier decades frequently have missing values for variables such as societal polarization and, in some cases, regime type; accordingly, models that include polarization draw on a smaller estimation sample (1,414 interactions) due to listwise deletion. Moreover, the study’s contemporary definition of democracy—free and fair elections, institutionalized party competition, and a free media—rarely applies to 19th-century polities, many of which were monarchies or empires. As a result, the early period clusters near the autocratic end of the Polity scale, while regime variation and the measurement quality of polarization increase markedly after the mid-20th century. Interpretations should therefore treat the long historical window as offering baseline contrast, with the substantive dynamics most informative for the modern era where concept measurement and data density are strongest.

The dependent variable, AC incurrence, is proxied via the binary variable of target state threat response, acquired from the MID, a standard source in international ACT research. To capture variation across regime types and societal polarization, the independent variable and the interaction term are drawn from the V-Dem, which provides measures of regime characteristics and polarization at the monadic level, as well as logged yearly population and logged GDP per capita for control purposes. These datasets are then systematically merged with a series of dyadic data sources from the Correlates of War project (COW), which supply the necessary contextual and structural controls.

The COW datasets provide several key dyadic measures that form the backbone of the control set, as they provided IV. The Diplomatic Exchange dataset (v2006) identifies whether dyads maintain formal diplomatic relations below or above the ministerial level, while the Direct Contiguity dataset (v3.2) captures the presence of land or river borders between states. The Formal Alliances dataset (v4.1) records whether dyads are bound by any kind of formal alliances, thereby reflecting institutionalized cooperation or mutual commitments that might affect conflict dynamics. Similarly, the International Trade dataset (v4.0) provides annual bilateral trade figures between sovereign states from 1870 to 2014, which are included as logged totals to control for economic interdependence. At the national level, the National Material Capabilities dataset (v6.0) furnishes information on relative capabilities, including composite measures such as the Composite Index of National Capability (CINC), which serve as critical controls for state size and development. In addition, war involvement is controlled for by distinguishing whether state A is

engaged in either internal or external conflict in a given year. Finally, the regime type of the target state is incorporated as a control variable.

Together, these sources create a comprehensive dataset that links dyadic conditions, regime characteristics, and structural factors into a format suitable for statistical testing of the thesis's core questions. The breadth of the temporal and spatial coverage ensures that findings are not limited to a narrow historical moment or narrow regional contexts, while the integration of monadic and dyadic information enables a nuanced analysis of how regime type and polarization jointly shape the likelihood of audience cost incurrence. By combining MID, V-Dem, and the COW family of datasets into a single design, the analysis situates AC at the intersection of domestic structures and dyadic interactions, providing a robust empirical foundation for testing the theoretical expectations advanced earlier in the thesis.

4.2. Variables

This section introduces the key components of the empirical model by outlining the dependent, independent, interaction, and control variables. Each of these is situated within the theoretical framework developed in the earlier chapter, ensuring conceptual consistency between theory and measurement. The discussion emphasizes how these variables capture the dynamics of AC across regime types and under varying levels of polarization. In doing so, the section bridges the theoretical expectations with the empirical strategy that follows.

4.2.1. Dependent Variable

For its dependent variable, this study follows the often-used convention in the ACT literature of treating the likelihood of AC incurrence as a perceived or indirectly observed phenomenon. As Schultz (2012) memorably puts it, audience costs are the “dark matter” of IR—hard to see directly yet central to the dynamics of signaling, credibility, and public opinion in foreign-policy crises. The ambiguity is inherent: AC are embedded in counterfactuals. At bottom, the concept asks what would have happened had a leader behaved differently—would domestic audiences have punished a retreat? In other words, AC hinge on whether the *possibility* of domestic backlash for backing down convinces adversaries to take the threat seriously. That counterfactual nature makes empirical capture especially challenging.

This difficulty is not incidental; it follows from the game-theoretic structure of audience-cost models (Fearon, 1994a; 1994b). In crisis-bargaining games, leaders first decide whether to issue public threats and, if they do, whether to carry them out. Because backing down after a public commitment invites domestic punishment, forward-looking leaders *select out* of situations likely to trigger that punishment. Two implications follow. First, the key counterfactual—*would the public have punished a retreat?* —is rarely realized because leaders avoid visible failure. Second, what we observe is a selected set of crises in which leaders either do not bluff or engineer outcomes that obscure a clear backdown. Ironically, the same strategic logic that generates audience costs also suppresses direct observation of them.

Given this selection problem, the literature leans on the game's structure to infer AC from observable implications. Schultz (2001) shows that tests based only on realized retreats systematically understate both the average level of audience costs and cross-regime differences; he motivates designs that use domestic political conditions—how public the commitment is, the strength of opposition, the degree of electoral accountability—to detect AC even when they are not literally paid. Kurizaki and Whang (2015) address the problem head-on by embedding a structural statistical model in the crisis game to recover the *size* of audience costs, including those never actually incurred. In a similar vein, Crisman-Cox and Gibilisco (2018) estimate an infinitely repeated crisis game to recover country-specific audience costs without relying on crude democracy proxies, showing that higher audience costs both commit initiators to stand firm and deter potential opponents—making the long-run system more peaceful on net.

Consistent with this tradition, the present study relies on observable implications of credibility—specifically, target state behavior. Building on the MID dataset, which records 9,901 incidents of international threats backed by military force for 196 states between 1816 and 2014, I treat the target's reciprocation (rather than compliance) as evidence of *lower* perceived credibility and, by implication, a *lower* likelihood that audience costs would be incurred by the initiator. This operationalization translates the latent mechanism into an opponent's observable choice, aligning the measurement strategy with both the theoretical logic and leading solutions to the non-measurement problem.

Specifically, the variable *recip* is employed as a proxy for AC incurrence likelihood. This binary indicator captures whether the target state responded to an initiating state's threat with its own militarized action (*recip*=1) or refrained from doing so and fully or partially complied with demands, if there are any, (*recip*=0). The logic underlying this operationalization is straightforward but powerful. If a challenger threatens and the target responds with reciprocation—mobilizing troops, deploying forces, or otherwise escalating—the target is signaling disbelief in the challenger's commitment to follow through. Such reciprocation implies that the challenger's threats are *uncredible*, suggesting that the leader of the challenger is not constrained by high AC.

In other words, the likelihood of AC incurrence in these cases is perceived low for the threat sender because the target believes the leader can back down without incurring severe political repercussions at home, based on the calculus of incumbent survival. By contrast, when the target chooses not to reciprocate and instead complies with the challenger's demands, it is revealing its belief that the challenger's threats are credible. This credibility, in turn, is anchored in the perception that the challenger's leadership is constrained by domestic audiences that would punish them if they failed to act consistently with their stated threats. Thus, compliance reflects a situation where the threatening state is judged to have a high likelihood of AC incurrence. Leaders are seen as unable to back down without suffering political damage, whether through loss of public support, elite backlash, or broader institutional consequences. Put differently, the higher the probability of target compliance, the stronger the expectation that the threatening state would incur audience costs if it were to retreat.

This logic of inference has been widely adopted in empirical studies of ACT. Weeks (2008), Potter and Baum (2014), Quek and Johnston (2017), Crisman-Cox and Gibilisco (2018), and Yu, Whang, and Lee (2022), among others, employ the MID's *recip* variable as a proxy for AC because it captures observable manifestations of credibility from the perspective of the target state. What these studies demonstrate is that the credibility of threats—and therefore the underlying likelihood of AC—can be reasonably inferred from whether targets resist or comply. While imperfect, this approach provides an indirect yet tractable measure that empirically explains the theoretical foundations of ACT. Although AC cannot be measured directly, the use of *recip* as a proxy allows this thesis to engage empirically with one of the most important yet elusive concepts in international relations. By focusing on how target states respond to threats, the analysis taps into the observable consequences of credibility that lie at the heart of audience cost theory. This strategy not only aligns with a well-established methodological precedent but also enables a rigorous test of the dissertation's central questions: how the likelihood of AC incurrence varies across regime types and how polarization moderates this relationship.

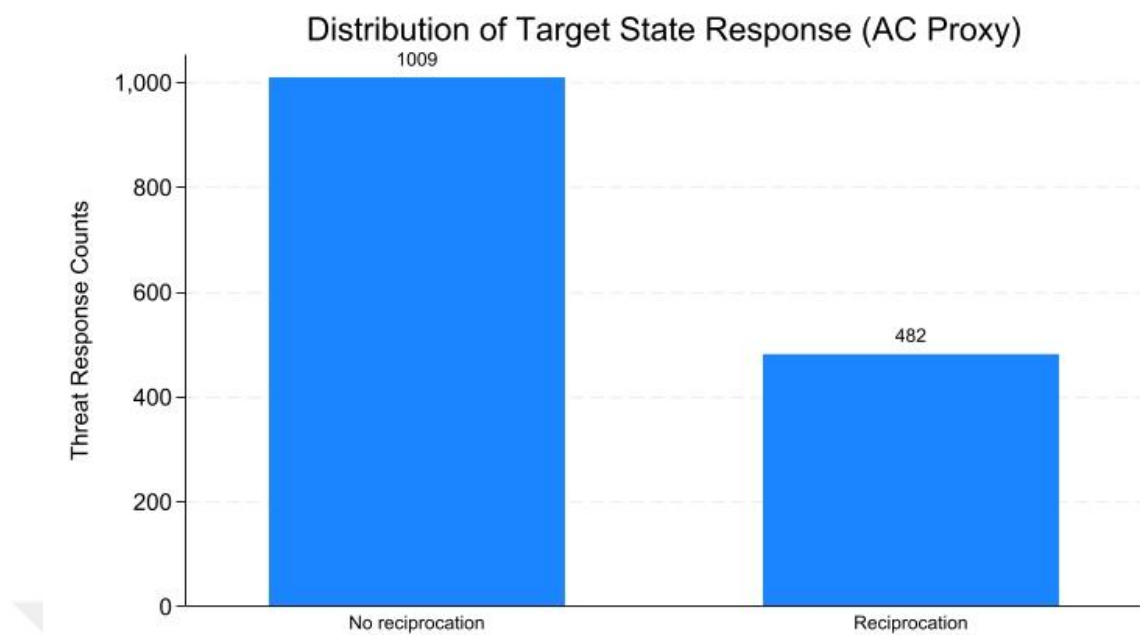


Figure 1 Threat Response (IV proxy indicator) Distribution

4.2.2. Independent Variable

The central independent variable in this thesis is regime type, measured through the widely used Polity score. The Polity score is computed by subtracting an autocracy score from a democracy score, producing a single unified index that ranges from -10 (hereditary monarchy, representing strongly autocratic rule) to $+10$ (consolidated democracy, representing strongly democratic rule). Unlike typologies that treat autocracy and democracy as strictly dichotomous and mutually exclusive categories, the Polity scheme is distinctive in that it simultaneously evaluates the coexistence of democratic and autocratic institutional characteristics within a regime. In doing so, it conceptualizes regime type as a spectrum of authority, one that runs from fully institutionalized autocracies at one end, through regimes that are incoherent and internally contradictory—what are called “anocracies”—and finally to fully

institutionalized democracies at the other. This continuum captures not only the institutional makeup of political authority but also the dynamic tension between competing elements of rule, making it particularly useful for comparative studies of how regime characteristics shape international behavior.

The theoretical justification for relying on the Polity score as IV is that the Polity score is not limited to formal democratic procedures, such as elections, but instead captures broader institutional qualities of governance that directly affect leaders' vulnerability to domestic punishment for foreign policy missteps. These include the openness of executive recruitment, the degree of constraints on executive authority, and the competitiveness of political participation. Each of these components is directly tied to the likelihood of audience costs: the greater the institutionalized capacity for opposition forces to challenge leadership decisions, the more credible threats are likely to appear to external observers

In operational terms, the Polity score can be translated into a three-part categorization that is commonly used in large-N international relations research. Regimes with scores ranging from -10 to -6 are classified as autocracies, those from -5 to $+5$ as anocracies, and those from $+6$ to $+10$ as democracies. This trichotomous classification retains the full richness of the underlying continuous measure while permitting cross-regime comparisons in a more structured form. Importantly, this categorization allows for the inclusion of anocracies as a distinct type rather than merely a transitional category or an intermediate point on the autocracy–democracy continuum. While democracies and autocracies have been studied extensively in the audience cost literature, anocracies have received considerably less attention. Yet,

their hybrid institutional character—marked by weak constraints on executive power, inconsistent competition, and often volatile legitimacy structures—makes them especially relevant for understanding variations in credibility and signaling in crisis settings. By treating anocracies as a substantive category in their own right, this thesis directly addresses a long-standing gap in the literature on audience costs and regime type.

In the empirical analysis, the Polity score functions as the primary independent variable against which the dependent variable—audience cost incurrence, proxied through target state responses in the MID dataset—is evaluated. By incorporating this measure, the thesis is able to directly test whether regimes positioned at different points along the autocracy–anocracy–democracy spectrum exhibit systematically different patterns of credibility in international crises. The expectation, building on the theoretical framework developed earlier, is that democracies should generally exhibit the highest likelihood of audience cost incurrence because of institutionalized accountability and transparency; autocracies should exhibit the lowest likelihood because of the repression or co-optation of opposition; and anocracies should occupy a distinctive middle position where regime instability, incoherence, and contested legitimacy produce unique credibility dynamics.

The temporal and spatial scope of the data allows this analysis to evaluate regime effects over a long historical period, with observations covering the years 1816 to 2014 and more than 190 unique states. This scope ensures both variation in regime type and sufficient numbers of autocratic, Anocracy, and democratic cases to permit systematic comparison. Importantly, the dyadic structure of the dataset ensures that

each observation reflects the directed behavior of state A toward state B in a given year, making it possible to evaluate not only the type of the initiator's regime but also the reaction of the target state as a proxy for audience cost dynamics. In this way, the Polity score is not merely an abstract classification but becomes an empirically grounded independent variable capable of revealing how institutional structures shape the credibility of threats and the likelihood of domestic punishment for backing down.

Taken together, the Polity score provides a robust and theoretically consistent measure of regime type as the independent variable in this thesis. It captures the full continuum of governance authority, distinguishes among autocracies, anocracies, and democracies, and aligns with the theoretical expectations regarding audience cost dynamics across regime types. In the subsequent section, this measure will be paired with the moderating factor of polarization to investigate how domestic societal divides interact with regime structures to shape the credibility of leaders' threats in international crises.

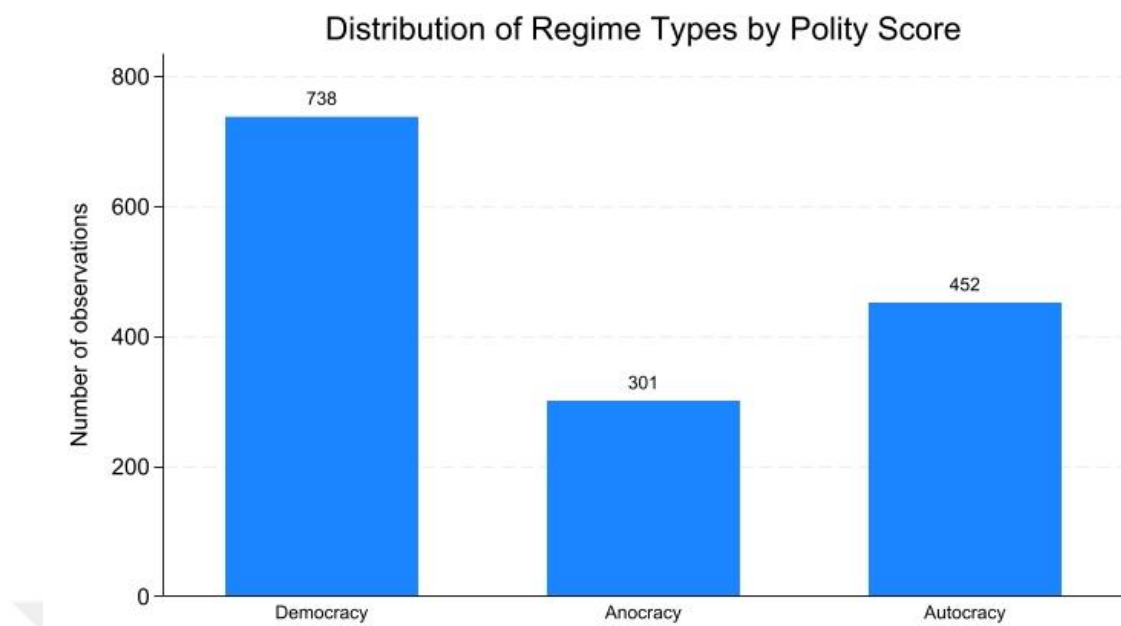


Figure 2 Distribution of Regime Types

4.2.2.1. Interaction Term

The key moderating factor in this study is affective polarization, which captures the extent to which societal divisions over politics translate into durable and emotionally charged cleavages. Unlike simple ideological disagreement, polarization here refers to the intensity with which opposing views are held and the degree to which these divisions shape collective behavior to the extent of deep-seated animosity. Drawing on the V-Dem dataset, polarization is measured on a continuous 1 to 10 scale, where higher values represent more severe societal polarization. The time coverage spans 1950 to 2024, enabling an investigation of both long-term historical variation and contemporary patterns. The measure asks how differences of opinion on major political issues are perceived within a society, ranging from general agreement on the direction of development to deeply entrenched clashes of views across nearly all

issues. This continuous interval measure provides an analytically tractable scale that reflects gradations in the polarization of public opinion.

The interaction of this polarization measure with regime type (polity IV) constitutes a central contribution of the research design. While the Polity score allows regimes to be located on a spectrum between autocracy, anocracy, and democracy, polarization reflects the degree to which public divisions constrain or embolden leaders. Bringing these together makes it possible to assess whether the likelihood of AC incurrence is heightened or dampened not only by the formal institutional structures of regimes but also by the social context in which leaders operate. For instance, the same institutional environment may produce different outcomes under high versus low polarization: a democracy characterized by broad consensus may exhibit more stable signaling patterns than one fractured by deep partisan divides, while anocracies—already institutionally incoherent—may see polarization amplify their instability. Autocracies, in contrast, might display muted or nonlinear effects, given their weaker channels of public accountability.

Theoretically, this interaction term allows the analysis to test whether polarization acts as a magnifier or dampener of regime-specific audience cost dynamics. It directly engages with the broader argument of this thesis, which posits that institutional configurations alone cannot fully explain variation in signaling credibility; rather, the interplay between regime type and polarization determines whether threats are perceived as credible, whether leaders are punished for backing down, and ultimately whether AC is incurred. By structuring the analysis around this interaction, the empirical strategy aligns with recent scholarship emphasizing the

micro-foundational role of public opinion in international conflict while extending the debate to include anocracies, a regime type often overlooked in audience cost theory.

4.2.3. Control Variables

Beyond the theoretical mechanism outlined in the preceding section, it is necessary to acknowledge that the likelihood of AC incurrence and the dynamics of interstate disputes are influenced by a broader set of contextual factors. States do not operate in a vacuum, and their decisions to issue, escalate, or retreat from threats are often conditioned by structural, economic, and relational variables that shape both external perceptions and domestic expectations. To account for these potentially confounding influences, this study incorporates a set of control variables that previous scholarship has consistently identified as important determinants of militarized conflict behavior. These factors—ranging from diplomatic relations and alliance commitments to relative capabilities, trade flows, demographic size, economic prosperity, and ongoing wars—help ensure that the relationship between regime type, polarization, and AC is not spuriously driven by other well-established drivers of conflict. In doing so, the analysis isolates the specific contribution of regime variation and societal polarization to the credibility of threats while remaining attentive to the wider strategic environment that influences state behavior.

Diplomatic relations (Bayer, 2006) serve as the first control variable. Although originally coded ordinally, reflecting varying levels of representation, it has been recoded into a binary measure in which ties at or above the ministerial level are

coded as 1 and lower levels as 0. The rationale for including this control lies in the fact that the existence of formal diplomatic channels already establishes a baseline relationship between states that may shape the credibility of threats. When two countries maintain high-level diplomatic ties, issuing militarized threats risks damaging a pre-existing institutional and political connection, thereby raising the potential reputational cost for leaders. Conversely, in the absence of such relations, threats are less constrained by pre-existing commitments. In the logic of AC, the destruction of valued diplomatic relations could provoke domestic backlash, as the public may perceive leaders as recklessly squandering established avenues of engagement.

Contiguity (Stinnett et al, 2002) is a second important control variable, measured as a binary indicator of whether two states share a border, either by land or significant waterways like rivers. Contiguity is theorized to matter because proximity inherently lowers the costs of military escalation. If a state threatens a neighbor, the logistical feasibility of following through with militarized action increases, which can shape the target's assessment of credibility. For AC, this means that leaders threatening contiguous neighbors face different incentives: their publics may expect a higher probability that threats can be carried out and thus punish backing down more harshly. Conversely, threats to geographically distant states might be perceived as less credible because the material and strategic costs of escalation are higher, reducing the domestic expectation of enforcement and muting potential backlash.

Formal alliance (Gibler, 2009) variable capture whether two states are officially bound by agreements such as defense pacts, neutrality treaties, non-aggression

agreements, or ententes. This variable is recorded as binary, distinguishing whether a formal treaty exists between dyad members in a given year. Alliances are a crucial control because threatening an ally carries substantial political and reputational costs. For example, in the case of NATO, leaders would face severe domestic and international repercussions for undermining treaty commitments by threatening another ally. In the framework of AC, alliances heighten domestic sensitivity because publics generally value the preservation of treaty obligations. Leaders who jeopardize these obligations risk being perceived as unreliable both abroad and at home, thereby increasing the likelihood of punishment if threats are issued but not fulfilled.

Relative capabilities are measured using the Correlates of War *National Material Capabilities* dataset (Singer, 1988), which aggregates indicators such as military personnel, expenditures, iron and steel production, energy consumption, and urban population into CINC. For this study, capabilities are logged and expressed as a relative measure between the threatening and target state, sometimes producing negative values when the initiator is weaker. This control is vital because the balance of capabilities directly affects the credibility of threats: weaker states lack the material means to enforce their rhetoric, reducing both external credibility and domestic expectations. In the AC framework, publics are less likely to punish leaders for backing down when capabilities are clearly insufficient, whereas strong states face more pressure to deliver on threats because their material advantages make inaction appear as weakness.

GDP per capita is included as a logged continuous control. GDP per capita captures economic well-being, which influences the domestic political environment in which leaders operate (Coppedge et al, 2020). In wealthier societies, citizens may be more sensitive to foreign policy blunders because they expect their leaders to preserve international stability and economic prosperity. Conversely, in poorer states, grievances are more widespread and populations may be quicker to mobilize against perceived leadership failures, potentially magnifying audience costs when threats are issued but not carried out. Thus, controlling for GDP per capita ensures that regime type and polarization effects are not confounded by differences in economic development, which independently shape the magnitude of public backlash against foreign policy retreats.

Finally, the regime type of the target state is included as a continuous measure distinguishing democracies from non-democracies. This control reflects the core logic of democratic peace theory, which posits that democracies are less likely to engage in militarized action with one another. When a democracy threatens another democracy, both the target and the domestic audience may perceive the threat as less credible, given the normative and institutional constraints against intra-democratic conflict. Conversely, threats against autocratic or Anocracy targets may be taken more seriously. Including this variable ensures that audience cost dynamics are not confounded by dyadic regime composition, isolating the role of the initiator's regime type and polarization.

Table 2 Summary Statistics of all Variables

VARIABLES	N	Mean	Standard Deviation	Min	Max
Target Reciprocation	1,491	0.323	0.468	0	1
Polity Score	1,491	2.004	6.819	-10	10
Polarization of Society	1,491	4.193	1.979	0	10
Diplomatic Relations	1,491	0.148	0.355	0	1
Formal Alliance	1,491	0.298	0.457	0	1
Contiguity	1,491	0.612	0.488	0	1
Relative Capabilities (logged)	1,491	0.0820	2.032	-8.203	9.031
GDP per capita (logged)	1,491	3.011	1.183	0.138	5.389
Target Regime	1,414	2.888	6.521	-10	10

4.3. Operationalization

The empirical analysis in this thesis relies on the careful integration of several large-N datasets that together allow for the systematic testing of how AC dynamics vary across regime types and are shaped by polarization. The backbone of the data construction is MID dataset, which provides comprehensive coverage of interstate threats and uses of force between 1816 and 2014. MID serves as the temporal anchor because its dyadic structure and wide timespan not only include all relevant militarized interactions but also match seamlessly with the COW series datasets, ensuring consistency across country codes and year identifiers.

To capture the strategic environment of interstate relations, I merged MID with several COW datasets covering trade, diplomatic exchange, direct contiguity, formal alliances, and national material capabilities. These sources are all dyadic in nature and thus align well with the structure of the MID dataset, where the unit of analysis is the directed dyad–year. Each merge was carried out using COW state codes and years as common identifiers. Importantly, the integration process preserved the full set of observations from MID, ensuring that no dyadic disputes were lost during the harmonization of datasets. The final outcome of this stage was the construction of a comprehensive dyadic dataset spanning from 1816 to 2014, which captures both militarized disputes and a range of structural and relational control factors.

While MID and COW datasets provided the dyadic backbone, they needed to be complemented with domestic-level information to capture regime type and polarization—core elements of the theoretical framework. For this purpose, I relied on the V-Dem dataset, which is monadic and thus structured at the country–year level. To reconcile this with the dyadic structure of the COW and MID datasets, I generated two separate V-Dem components: one corresponding to the sender (state A) and the other to the target (state B). By doing so, I was able to integrate regime type, polarization, and economic characteristics into both sides of the dyad without losing alignment with the dyadic dispute data. This step ensured that both challenger and target characteristics could be simultaneously taken into account when analyzing threat credibility and AC likelihood.

Bringing these sources together produced a unique merged dataset that spans nearly two centuries, covers 193 distinct countries, and results in a total of 1,491 directed

dyad–year observations. This dataset allows for the empirical testing of the thesis’s two central inquiries: how regime type influences the likelihood of AC incurrence, including the often-overlooked case of anocracies, and how societal polarization moderates this relationship across regime categories. The operationalization process, therefore, not only consolidates a wide range of established datasets into a coherent structure but also tailors them to address the theoretical and empirical questions at the core of this research.



CHAPTER 5

RESULTS

In this chapter, I present the results of the empirical analyses and assess whether the hypotheses developed in the theoretical section are supported. The results are introduced in two main stages, each designed to capture a different dimension of the relationship between regime type, polarization, and the likelihood of AC incurrence. In the first stage, I estimate models that include only regime type as the independent variable alongside the control variables. This analysis allows me to compare regime types against one another and identify how democracies, anocracies, and autocracies differ in their baseline propensities to generate audience costs when issuing threats. In doing so, the results highlight the relative position of each regime type in the broader spectrum of credibility and domestic vulnerability.

The second stage builds on this baseline by introducing the interaction between regime type and polarization. This step shifts the focus inward, allowing for within-regime comparisons by examining how rising polarization alters the credibility dynamics of each regime type. In other words, whereas the first set of results highlights cross-regime differences in audience cost likelihood, the second analysis reveals how polarization reshapes the internal logic of democracies, anocracies, and autocracies. This design makes it possible to identify how higher or lower polarization alters the expected credibility of threats within each regime type.

Taken together, these two stages provide both a comparative perspective across regimes and a nuanced understanding of regime-specific trajectories under varying polarization conditions. The following sections present the results of each stage in detail, beginning with the baseline models focused on regime type alone, and then moving to the extended models that incorporate the interaction with polarization.

5.1. Empirical Approach: Logistic Regression

To address the research questions of this thesis, I employ logistic regression analyses as the primary empirical strategy. The central dependent variable, audience cost (AC) incurrence, is operationalized through a binary measure drawn from the Militarized Interstate Disputes dataset, which records whether a target state reciprocated or complied in response to a threat. Because this variable takes the form of a dichotomous outcome—coded as either reciprocation or no reciprocation—it is appropriate to use logistic regression, which is specifically designed to model probabilities for binary outcomes. Unlike linear regression, which assumes continuous and normally distributed dependent variables, logistic regression provides more reliable estimates for categorical outcomes and allows the results to be interpreted in terms of predicted probabilities, which is directly relevant for evaluating the likelihood of AC incurrence. By employing this approach, the models generate predicted probabilities that capture the theoretical mechanism under investigation: the comparative likelihood of AC incurrence across regime types, and the moderating role of polarization. Logistic regression serves not only as the most statistically appropriate method given the binary nature of the dependent variable but also as the most substantively meaningful approach for tracing credibility and

domestic vulnerability in international crises. Before going into analyses, it is useful to circle back to hypotheses to show what is aimed to significantly explained:

Table 3 Expected Significant Results

Hypotheses	Variable	Coding	Source
H1: <i>The more democratized a regime gets, meaning having higher polity scores, higher the likelihood of AC incurrence becomes, making their threats to be perceived highly credible by targets.</i>	Target Reciprocation (DV)	Binary variable, coded as 1 if target reciprocates threat, 0 if complies	MID Dataset (Palmer et al, 2022)
H1a: <i>Regimes with high polity scores, meaning Democracies, have a higher likelihood of AC incurrence compared to anocracies and autocracies, making their threats to be perceived highly credible by targets.</i>	Polity Score (IV)	Continuous variable ranging from -10 (autocracy) to +10 (democracy)	V-Dem (Coppedge et al, 2020)
H1b: <i>Regimes with low polity scores, meaning Autocracies, have a lower likelihood of AC incurrence compared to democracies and anocracies, making their threats to be perceived less credible by targets.</i>			
H1c: <i>Regimes with mid-range polity scores, meaning Anocracies, relatively have higher AC incurrence likelihood than Autocracies but lower than Democracies, making their threats to be moderately credible.</i>			

Conditional Hypotheses	Variable	Coding	Source
H2a: <i>The more polarized a democracy is, the lower the propensity to incur AC is perceived compared to a low polarized democracy, making threats more likely to be reciprocated by the targets.</i>	Polity Score (IV) x Polarization (interaction)	Continuous variable ranging from -10 (autocracy) to +10 (democracy) x Continuous variable ranging from 1 (low polarization) to 10 (high polarization)	V-Dem (Coppedge et al, 2020)
H2b: <i>The changes in polarization level in an autocracy would not impact the likelihood of AC incurrence, as targets threat responses will be unaffected by polarization.</i>			
H2c: <i>The more polarized an anocracy is, the higher the propensity to incur AC is perceived compared to a low polarized anocracy, making threats more likely to be complied by the targets.</i>			

5.2. Findings

5.2.1. Logistic Regression for Regime types

This section presents the findings from the logistic regression models that examine the effect of regime type, measured by the Polity scale, on the likelihood of audience cost (AC) incurrence. As outlined earlier, audience costs are proxied through target state responses to threats in the Militarized Interstate Disputes dataset, where reciprocation indicates low perceived credibility and compliance signals high credibility. By modeling these outcomes with logistic regression, we can assess how variation in regime type shapes the probability of reciprocation and, by extension, the likelihood of incurring AC.

The results in Table 4 demonstrate a statistically significant and substantively meaningful relationship between Polity and reciprocation. The coefficient on Polity is negative and significant at the $p < 0.001$ level, showing that higher levels of democracy reduce the probability of reciprocation. Interpreted in odds terms, each one-unit increase in Polity decreases the odds of reciprocation by about 6.7%. This effect accumulates across the scale: moving from anocracy (0) to a consolidated democracy (+10) lowers the odds of reciprocation by roughly 50%, strongly suggesting that democratic institutions strengthen credibility in international crises. Consistent with the figure, predicted probabilities fall from roughly one-third near Polity 0 to the mid-20s.

Table 4 Logistic Regression Results for Polity Score

VARIABLES	Probability of Reciprocation (AC proxy)
Polity Score	-0.069*** (0.011)
Diplomatic Relations	-1.735*** (0.272)
Contiguity	0.069 (0.138)
Formal Alliance	-0.107 (0.139)
Relative Capabilities (logged)	-0.039 (0.030)
GDP per capita (logged)	0.435*** (0.065)
Target Regime	-0.014 (0.009)

Constant	-1.730*** (0.244)
Observations	1,414
Pseudo R2	0.0847

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

When converted into predicted probabilities, the results become even clearer. As illustrated in Figure 3, democracies sit at the low end of reciprocation: a fully democratic initiator shows a predicted probability of about 24–26% that its threat will be met in kind, implying a high perceived AC. This pattern follows directly from the strongly negative democracy slope in the baseline model and is highly statistically significant. In odds language, each one-point increase on the Polity scale reduces the odds of reciprocation by roughly 6–7%. Accumulated across the democratic range, moving from anocracy to consolidated democracy cuts the odds of reciprocation by about half; spanning the full range from –10 to +10 implies roughly a three-quarters reduction. Substantively, this supports the claim that democratic leaders, facing the prospect of domestic punishment for retreat, issue threats that targets treat as more credible on average, and thus reciprocation is less likely.

Anocracies, by contrast, occupy the middle of the curve. Around the anocracy region, the predicted probability of reciprocation is about one-third (~33%), with uncertainty bands that widen modestly but still track the overall downward slope from autocracy to democracy. This “middle ground” reflects institutional inconsistency and fragility: signals from hybrid regimes are neither as binding as those from robust democracies nor as easily dismissed as those from hard autocracies. The figure makes this intuitive—standing at the center of the scale,

anocracies inherit neither the full credibility advantages of established democracies nor the high skepticism that attaches to autocratic threats.

Autocracies anchor the high end. At the strongly autocratic side, the predicted probability of reciprocation is about 42%, low-40s with overlapping intervals. This is substantially higher than the democratic endpoint and visually apparent in the left side of Figure 3, where the line begins near the mid-0.40s and descends steadily as regimes become more democratic. The implication is familiar: authoritarian leaders, comparatively insulated from broad public backlash, send signals that targets are more willing to test, yielding higher reciprocation rates in expectation.

Turning to the controls, several findings reinforce baseline expectations once we map the coefficients into odds changes and compare their magnitudes. Diplomatic relations are associated with a large and precise reduction in reciprocation: the estimated shift on the log-odds scale translates into about an 82% drop in the odds of reciprocation when a diplomatic relationship exists between the dyad, all else equal ($p < .001$). This is the single largest control-variable effect in the model and is stable across specifications. Income per capita (logged) is positively related to reciprocation and also highly significant ($p < .001$): a one-unit increase in logged income—roughly a $2.7\times$ increase in GDP per capita—corresponds to about a 55% increase in the odds of reciprocation. The remaining controls are small and not statistically reliable in this specification. Contiguity raises the odds only slightly and is not significant; formal alliances show a small, non-significant dampening of reciprocation; relative capabilities (logged) are near zero ($\approx 4\%$ lower odds, non-significant); and the target's regime score is small and non-significant as well. Importantly, these

estimates indicate that the polity–reciprocation relationship is not an artifact of omitted dyadic fundamentals: the downward slope with democracy persists once diplomacy, alliances, geography, capabilities, income, and the target’s regime are held constant.

Taken together, the regression results and the predicted-probability trajectory in Figure 3 provide consistent evidence that regime type plays a decisive role in audience-cost dynamics *on average*. Democracies are the most effective at signaling AC (lowest reciprocation), autocracies are the least (highest reciprocation), and anocracies fall in between—reflecting their institutional incoherence. The strength of the baseline democracy effect and the substantive size of the shifts in both odds and probabilities underscore that these regime distinctions are not only theoretically meaningful but also empirically robust in this sample of 1,414 directed dispute observations.

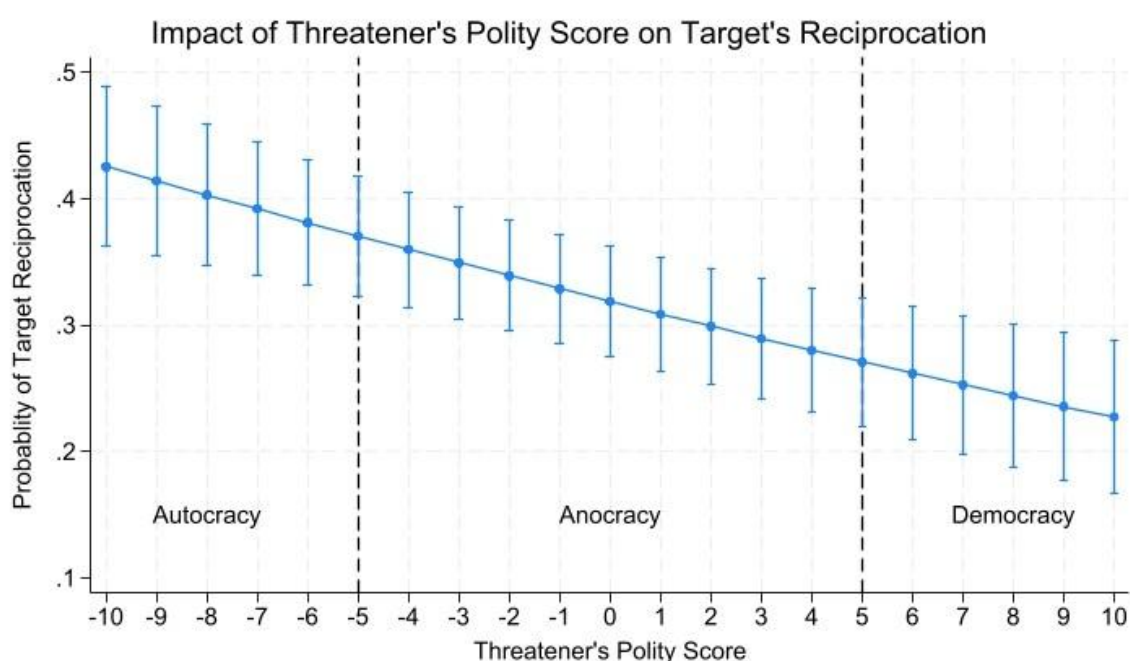


Figure 3 Impact of Threatener’s Polity Score on Target’s Reciprocation

5.2.2. Logistic Regression of Polarization Moderation on Polity Scores

This section turns to the interactive effect of polarization within regime categories, assessing whether societal divisions amplify or dampen the credibility of threats. Table 5 reports the logistic regression models estimated separately for democracies, anocracies, and autocracies, with target reciprocation serving as the dependent variable. Polarization is drawn from the V-Dem dataset and scaled from 0 to 10, allowing us to evaluate how incremental increases in polarization affect the probability that threats are met with resistance. By modeling the interaction within regime types, this analysis isolates how domestic polarization operates in distinct institutional contexts rather than across the regime spectrum as a whole.

Table 5 Logistic Regression Results for Polity Score moderated by Polarization

VARIABLES	(1) Polity Scores' Impact on Reciprocation Likelihood	(2) Polarized Democracies' Reciprocation Likelihood	(3) Polarized Anocracies' Reciprocation Likelihood	(4) Polarized Autocracies' Reciprocation Likelihood
Polity Score	-0.069*** (0.011)			
Societal Polarization		-0.092** (0.036)	0.037 (0.040)	-0.025 (0.038)
Democracy		-1.644*** (0.326)		
Anocracy			0.918*** (0.316)	
Autocracy				0.671** (0.300)

Democracy x Polarization		0.150**		
		(0.070)		
Anocracy x Polarization			-0.172***	
			(0.063)	
Autocracy x Polarization				0.030
				(0.065)
Diplomatic Relations	-1.735***	-1.694***	-1.951***	-1.846***
	(0.272)	(0.273)	(0.270)	(0.271)
Formal Alliance	-0.107	-0.058	-0.115	-0.127
	(0.139)	(0.141)	(0.138)	(0.140)
Contiguity	0.069	0.071	0.088	0.100
	(0.138)	(0.140)	(0.137)	(0.138)
Relative Capability (logged)	-0.039	-0.031	-0.014	-0.029
	(0.030)	(0.030)	(0.029)	(0.030)
GDP per capita (logged)	0.435***	0.451***	0.209***	0.338***
	(0.065)	(0.067)	(0.057)	(0.060)
Target Regime	-0.014	-0.013	-0.007	-0.009
	(0.009)	(0.009)	(0.009)	(0.009)
Constant	-1.730***	-1.055***	-1.340***	-1.726***
	(0.244)	(0.271)	(0.266)	(0.306)
Observations	1,414	1,414	1,414	1,414
Pseudo R2	0.0847	0.0910	0.0656	0.0796

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The results demonstrate important variation across regimes. Within democracies, polarization is associated with a *rise* in reciprocation, meaning a drop in AC incurrence likelihood, that is statistically significant at conventional levels. On the log-odds scale of the logistic model, the within-democracy slope is about +0.06 per one-unit increase in polarization. That corresponds to roughly a 6 % increase in the odds of reciprocation per point of polarization. Aggregated across the full observed

range, this implies on the order of an 80 % increase in the odds. In probability terms (Table 5), the predicted likelihood of reciprocation increases from about 18 % at very low polarization to about 33 % at very high polarization. The uncertainty bands widen at the extremes but the upward pattern holds across most of the scale. Substantively, growing societal division loosens the domestic constraints that usually bolster democratic credibility, so targets are more willing to answer in kind.

In anocracies, the relationship reverses and is statistically precise. The estimated within-anocracy slope on the log-odds scale is about -0.14 per point of polarization. Interpreted as odds, that is roughly a 13 % decrease in the odds of reciprocation for each one-unit increase in polarization. Over the full range, this cumulates to about a three-quarters reduction in the odds. Consistent with this, Table 5 shows predicted probabilities falling from around 50 % at very low polarization to roughly 22 % at very high polarization, a converse surge in AC incurrence likelihood. The magnitude is large and the decline is steady through the midrange. This pattern fits a vulnerability-driven audience-cost logic: when institutions are shaky, polarization concentrates the risks leaders face if they back down, so targets read threats as more credible and are less inclined to reciprocate.

For autocracies, the estimated association between polarization and reciprocation is not statistically distinguishable from zero. The implied per-unit change on the log-odds scale is approximately zero (on the order of five-thousandths per point), which translates to about one-half of one % change in the odds per point—too small to separate from noise. The point predictions in Figure 4 display only a gentle drift (roughly the mid-40s down toward the mid-30s in probability terms across the whole

range), and the uncertainty intervals overlap widely. The safest conclusion is that in authoritarian settings, polarization does not systematically alter targets' responses, consistent with the idea that elite insulation and coercive capacity mute audience-cost mechanisms.

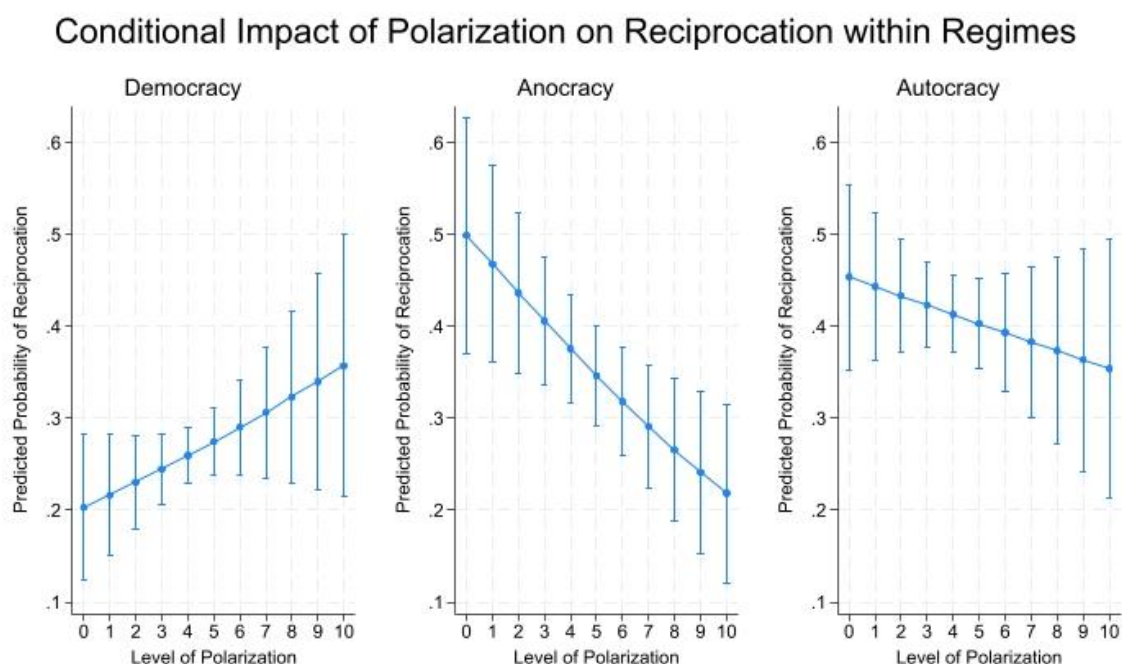


Figure 4 Impact of Threatener's Polarization on Target Reciprocation

The figure visualizing predicted probabilities within each regime type provides an intuitive picture of these dynamics. In democracies, the curve slopes upward: the predicted probability that a target reciprocates rises from roughly 18% at very low polarization to about 33% at very high polarization. In anocracies, the line slopes downward, with predicted probabilities falling from about 50% to near 22% across the same range. In autocracies, the curve is essentially flat with a mild, non-systematic drift—roughly the mid-40s toward the mid-30s—and the uncertainty bands overlap widely. Taken together with the regression estimates in Table 5, these

patterns reinforce the thesis's core claim: polarization moderates audience costs differently by regime. Where democracies normally convey credibility, polarization weakens that effect; where institutionally incoherent anocracies typically lack credibility, polarization heightens it; and where insulated autocracies usually send weak signals, polarization shows no reliable impact.

Turning to the controls, the results are straightforward and consistent with the table. Diplomatic relations are strongly associated with lower reciprocation in every model—a large, precise reduction in the log-odds that holds net of other covariates. Income per capita (logged) is positively related to reciprocation across specifications, indicating higher-income initiators face more frequent reciprocation, all else equal. By contrast, formal alliances, contiguity, and relative capability show no consistent, statistically reliable associations in these models. The target's regime score is small and not significant. These patterns indicate that the polarization effects are not artifacts of omitted dyadic fundamentals; they persist once standard correlates of dispute dynamics are taken into account. Model fit is modest but typical for dyadic outcomes.

Overall, Table 5 and Figure 4 jointly show that polarization's effect is conditional on regime type: it erodes audience costs in democracies, amplifies them in anocracies, and is negligible in autocracies. This divergence underscores the value of analyzing audience costs within regime categories as well as across them, clarifying how institutional structure channels the political consequences of societal polarization in international crises.

CHAPTER 6

CONCLUSION

As the incumbents fundamentally care most about staying in their office to accomplish their ideals or goals, matters like issuing international threat are carried out with one eye on the foreign target and the other on the home arena that decides political survival. Under such dual play, what foreign target weigh, in practice, is not a leader's resolve in the abstract but whether backing down would invite punishment at home—AC that can end a career. This thesis starts from that simple observation and brings ACT back to its core: credibility abroad is an external inference about punishability at home.

At its heart, this thesis asks how the likelihood of AC incurrence differs across regime types—including the often-neglected anocracies—and how social polarization moderates that likelihood within democracies, anocracies, and autocracies. The ACT literature gave strong building blocks on signaling, capabilities, and institutional checks, but it left two blind spots. It, firstly, flattened regime diversity into a democracy–autocracy binary, treating anocracies as residual cases, while, secondly, relegating critical micro-foundations like polarization to the background or to democratic survey settings. Those choices make it hard to see the settings where credibility is most volatile: hybrids where competition and coercion coexist without the buffers of consolidation or the insulation of hardened autocracy, and where partisan identities structure who can punish, how quickly, and at what

scale. Read this way, when accountability pathways are clear, elite insulation is thin, mobilization is easy, and blame cannot be deflected, threats sound different; but where buffers are thick and blame-shifting is credible, they sound hollow. Given these, both factors needed separate vital attention.

To test these claims in a way that matches ACT's externalist logic, I take the target's behavior as the observable window into credibility. The dependent variable is the target's choice in Militarized Interstate Disputes: reciprocation (discounting AC on the issuer) versus compliance (inferring AC are "in play"). The key independent variables are (1) regime type—democracy, anocracy, autocracy—coded using standard Polity/V-Dem classifications; and (2) affective polarization, proxied by V-Dem's ordinal measure of social polarization. The core specification estimates the probability of reciprocation with a regime indicator interacted with polarization, so baseline differences across regimes and polarization's moderating effect are identified jointly rather than assumed. The analysis is conducted at the threat-episode level, aligning the timing of regime and polarization measures with the year of the issuer's threat. To avoid conflating AC with familiar structural drivers of crisis behavior, I include a standard battery of controls: relative capabilities, contiguity, alliances, bilateral trade, GDP per capita. The design is straightforward: if outsiders read high punishability at home, compliance should rise; if they discount punishability, reciprocation should be more likely.

For two sets of hypotheses, I relied on two interrelated theoretical mechanisms. The first is the "normal times" (stability-anchored) mechanism without interaction: regime stability sets the baseline visibility and bite of AC. In democracies,

institutionalized accountability usually converts dissatisfaction into formal penalties, making AC legible; in hardened autocracies, elite insulation and coercion mute AC's visibility; in anocracies, partial competition without full buffers creates lumpy, sometimes existential penalties that are visible but irregular. The second mechanism captures the detrimental impacts of polarization—a risky-shift dynamic layered atop those baselines. Polarization binds in-groups and hardens out-groups, changing how sanctioning coalitions form. In democracies, high polarization can diffuse accountability and enable blame-shifting, lowering inferred AC despite loud conflict. In anocracies, polarization concentrates risk on incumbents: retreat looks like betrayal, mobilization thresholds drop, and AC are inferred to be steep—making threats appear more credible. In autocracies, the same polarization tends to be channeled or suppressed by insulation, so the incremental effect on inferred AC is limited. Together, these mechanisms explain why the same public threat can be discounted in one capital and treated as deadly serious in another: foreign actors are reading different domestic maps tilted by stability and twisted by polarization.

In sum, the thesis asks a comparative question the field has left underexplored and answers it with a design that mirrors how credibility is actually judged. It clarifies who the audience is, what counts as cost, and why anocracies and partisanship sit at the center of contemporary signaling. And it does so in a way that lets a reader see—without turning pages backward—what was done, why it was necessary, and why it matters for connecting domestic political survival to international outcomes.

6.1. Summary of Findings

To test my two sets of hypotheses, I estimated logistic regression models in which the dependent variable is the target state's response in a MID: reciprocation versus compliance. Logit is the appropriate estimator because the outcome is dichotomous and the coefficients can be translated into predicted probabilities and marginal effects that meaningfully summarize crisis behavior. The baseline models compare regimes; the extended models add an interaction between regime and polarization to see how polarization tilts credibility within each regime family. All specifications include the standard controls used in crisis studies (capabilities, contiguity, alliances, trade, etc.).

Two patterns organize the findings. First, the cross-regime baseline behaves as ACT would predict. As initiators move toward the democratic end of the spectrum, targets become less likely to reciprocate and more likely to comply. Read through ACT, that is a straightforward inference: in democracies, accountability pathways are visible and routine, so backing down looks punishable; foreign actors treat public threats as more costly to reverse. At the autocratic end, reciprocation is more common, consistent with thicker elite insulation and a coercive apparatus that blunts the translation of social anger into leadership costs. Anocracies sit between—but not as a mere average. Their imprint is volatility: partial competition without consolidated buffers makes penalties for retreat lumpy. Targets behave as if they see that closeness to the edge.

Second, once polarization is introduced, those baselines tilt, and they do so in regime-contingent ways. In the democratic subsample, higher social polarization is associated with a greater likelihood of reciprocation. Substantively, observers appear to infer that polarized democracies can diffuse responsibility and plausibly reassign

blame after a retreat; the institutional wiring remains, but partisan crossfire muddies who would pay the price. In the anocratic subsample, the sign flips: higher polarization is associated with a lower likelihood of reciprocation. Outsiders read polarized hybrids as more likely to impose steep domestic penalties for reversal—mobilized camps treat compromise as betrayal and can move quickly. In the autocratic subsample, the polarization term is substantively small and statistically indistinct from zero; targets do not update credibility based on polarization alone when buffers remain intact. Together, these interactions say that polarization is not uniformly “bad” or “good” for credibility; its effect depends on how it rewires accountability given the regime’s stability.

These empirical patterns do more than sort regimes among or within themselves; they answer the thesis’s core questions. On the comparative question—how AC incurrence likelihood differs across regime types—the baseline gradients show that targets adjust behavior in line with how punishable leaders appear at home: highest legibility (and thus compliance) in democracies, lowest in autocracies, and a distinct “close-to-the-edge” profile in anocracies. On the moderation question—how polarization conditions those baselines—the interactions demonstrate that partisanship is a mechanism, not mere background: it can blur accountability in democracies (softening deterrence) and concentrate sanction in anocracies (hardening it), while remaining largely inert in autocracies unless insulation cracks. Controls behave sensibly and, importantly, do not overturn the main story. Contiguity raises the feasibility and appetite for pushback; alliances can dampen it when they make testing costlier; material balance and income move outcomes at the margins but do not subsume the regime–polarization effects. This is reassuring: the

results are not an artifact of geography or power. They are the imprint of how outsiders read domestic exposure to AC.

Why does this matter for interpretation beyond the coefficients? Because it gives a diagnostic for crisis reading that aligns with ACT. In democratic settings, the “good sign” for credible restraint is visible elite ownership that tightens responsibility chains before a threat is issued; the “bad sign” is partisan buck-passing that pre-legitimizes retreat. In anocracies, the “good sign” for safe de-escalation is a face-saving exit that spreads ownership across camps (third-party guarantees, elite pacts); the “bad sign” is personalized vows and betrayal rhetoric that make backtracking politically toxic. In autocracies, polarization itself is a weak signal; the meaningful danger signs are cracks in insulation—elite defections, coercive-apparatus splits, or patronage strain—that can suddenly pull AC onto the stage.

What should be taken from this? The models were chosen because the question is binary and behavioral; the variables were selected to let the mechanism compete against familiar alternatives; and the results say that targets are reading domestic exposure to AC, not just institutions on paper. Practically, this yields a compact crisis diagnostic: in democracies, credibility strengthens when ownership of commitments is visibly broad and pre-assigned; in anocracies, credibility strengthens with polarization but safe exits narrow, so de-escalation requires face-saving designs and shared ownership; in autocracies, watch buffers, not noise—credibility reprices only when insulation shows visible cracks. A portable way to interpret crises that treats anocracies as first-class cases and puts polarization inside the mechanism, using logit evidence that translates directly into how often threats are tested or believed.

In short, the results show that the same public threat can sound firm in one capital and flimsy in another because targets are reading different domestic maps—of accountability, insulation, mobilization potential, and deflectibility—fixed by regime stability and tilted by polarization. The logit framework made those readings visible as predicted probabilities and interaction effects that travel cleanly across regime families. What we learn is both conceptual and practical: audiences are coalitions, costs are spectra, anocracies are first-class cases rather than residuals, and polarization’s effect is conditional. Credibility abroad is the mirror of political survival at home; these estimates show how brightly, or dimly, that mirror reflects across today’s regimes.

6.2. Implications

Taken together, the results offer two lessons about where AC and credibility shift. First, anocracies must be treated as crucial cases in ACT. They are now common and often durable, and the empirical patterns show that in hybrid regimes the boundary between a routine policy reversal and a regime crisis is thin—and visible to outsiders (Levitsky & Way, 2002; 2010). That visibility makes public threats more believable because backing down plausibly triggers steep domestic penalties, yet it also narrows safe exits. The lesson is conceptual and practical: theories that collapse hybrids into a democracy–autocracy binary will misread credibility, and crisis management that assumes easy de-escalation in hybrids will overpromise. In practice, do not corner fragile leaders with maximalist public red lines; build exits at the outset—reversible steps, third-party monitoring, and “technical” implementation bodies that let leaders

frame compliance as procedure rather than personal defeat. Use backchannels and face-saving side payments to turn concessions into domestically defensible outcomes. Price in a fragility premium: hybrid-regime threats are often credible, but escalation risks are higher, so design agreements that distribute responsibility and allow partial, staged success.

Second, polarization is not a uniform background condition; it is a regime-contingent moderator of AC. The evidence indicates that polarization reorganizes accountability in different directions within regime families. In democracies, polarization blurs ownership and makes post-hoc blame-shifting credible, softening inferred penalties for retreat and inviting tests unless responsibility is pre-assigned. Practically, rebuild credibility by pre-assigning ownership (legislative authorization, cross-party signs, cabinet co-signatures), setting clear benchmarks you will be judged against, or—if unity is impossible—moving signals off the main stage (quiet demarches, calibrated deployments, conditional assurances). In anocracies, polarization hardens loyalist coalitions and frames compromise as betrayal, sharpening inferred penalties and boosting short-run credibility while raising escalation risk; here, keep pressure mostly private, couple it with domestic insurance (limited security guarantees, economic cushions), and sequence enforcement to avoid involuntary escalation. In hardened autocracies, polarization carries little informational weight for outsiders unless elite insulation cracks; direct signals at the selectorate (ruling coalition, security services, rent access), and be ready to re-price credibility quickly if purges, defections, or fiscal stress appear.

What can be learned for real-world reading of crises is straightforward: trace who clearly owns the commitment, how grievance converts into sanction, how thick the elite and coercive buffers are, and whether blame can be plausibly deflected—then ask how polarization tilts each element in that regime. These lessons update ACT and the wider security literature at once: move anocracies to the forefronts, treat polarization as a conditional mechanism, and anchor inference in target behavior. The payoff is a clearer account of when AC will be inferred and a usable lens for calibrating signals and crafting exits in today’s hybrid, polarized politics.

6.3. Limitations

The inferences in this study are bounded by the data it observes and the way it translates complex signaling into tractable tests. Most fundamentally, the MID corpus captures overt, interstate, militarized crises and, within those, episodes that leaders chose to signal publicly. That scope matters twice over. First, it limits variation in the kinds of signals observed—kinetic-leaning and public rather than private ultimatums, economic coercion, cyber operations, or gray-zone probes where audiences, publicity, and reputational stakes differ. Second, public signaling is itself a strategic choice: leaders go public when they want domestic mobilization, allied reassurance, or international audience validation. The regime–polarization patterns documented here should therefore be read as conditional on those characteristics—overt, interstate, public episodes—not as a claim that all forms of coercion are processed identically by foreign targets.

A second constraint follows from how the outcome is measured. To align with ACT's externalist logic, the analysis models the target's response as a binary decision—reciprocation or compliance—and estimates it with logit so that effects translate cleanly into predicted probabilities. That choice keeps concept, measure, and test coherent, but it is necessarily coarse. Audience costs emerge (or are avoided) through a multi-actor, multi-stage process involving governments, opposition elites, security organs, and external guarantors. Between “push back” and “stand down” lie partial countersignals, quiet concessions, sequencing, third-party mediation, and reputational side-payments. A two-category outcome cannot register those gradations or the cross-pressures that produce them. The proxy is defensible for testing an external-inference claim; it remains a simplified lens on a complex mechanism.

Third, the models treat threats as discrete events, while credibility often evolves in sequences. Repeated exchanges can accumulate and interact: serial reciprocation may embolden a target or radicalize a home audience; serial compliance may create reputational slack—or, conversely, domestic anger at perceived softness. Path dependence, learning, and feedback are partially absorbed by controls and time structure, but the design does not track how earlier moves condition later beliefs about punishability. A dynamic extension—recurrent-event hazards, multi-state (Markov) crisis pathways, distributed lags for prior reciprocation/compliance, or dyadic sequence analysis—would better capture how credibility is made and remade across linked episodes.

Finally, the measurement of domestic context is less granular than the theory's moving parts. Treating autocracies as a single family is parsimonious but conceals

subtype differences—personalist, party, military, monarchic—in elite cohesion, purging norms, and coercive autonomy, precisely the dimensions that map to insulation and mobilization in the argument. Likewise, cross-national indicators of social polarization are ordinal and noisy; near regime thresholds, classification drift can blur genuinely anocratic dynamics. Both features likely attenuate interaction effects toward zero, making the reported patterns conservative. Disaggregating authoritarian subtypes and refining polarization/regime coding at the margins would sharpen the estimates and test whether the “flat” autocratic result masks subtype heterogeneity.

None of these caveats overturn the central finding—that targets behave as if they are reading domestic punishability—but they delineate its scope. The contribution here is most secure for overt, public, militarized crises; it abstracts a complex, multi-actor process into a binary behavioral proxy; it treats episodes as independent; and it aggregates across authoritarian forms. Addressing those four fronts would produce a more finely grained map of when credibility abroad mirrors punishability at home, and how that mirror changes as crises accumulate over time.

6.4. Future Studies

On top of the significant steps taken by now, several avenues can deepen, broaden, and sharpen the claims advanced here by improving how beliefs are measured, how crises are modeled over time, and how scope conditions are specified across regimes.

First, beliefs about punishability can also be observed more directly alongside the behavioral proxy used here. In the face of certain case threat, elite-level studies in target governments—structured interviews and elite surveys with diplomats, defense officials, and legislative foreign-affairs principals—can elicit how they read an opponent’s domestic exposure, which cues they weigh (ownership, sanction pathways, insulation, deflectibility), and how those cues trade off under time pressure. Parallel public surveys in both issuer and target states can map mass inferences about punishability, using experimentally varied vignettes to manipulate regime stability, elite cohesion, and polarization frames. Comparing elite and mass responses will reveal whether “elite polarization” or “mass polarization” better predicts external inference, and whether gaps between them widen or close under crisis.

Second, credibility should be modeled as a sequence rather than a set of isolated shots. The logit framework of this thesis captures whether a particular threat draws reciprocation or compliance; future work can track how earlier exchanges condition later ones. Recurrent-event hazard models, multi-state (Markov) crisis pathways, or sequence analysis at the dyad level can test whether serial reciprocation emboldens targets or radicalizes home audiences, and whether serial compliance creates reputational slack or, conversely, domestic backlash.

Third, the domain of signaling should extend beyond militarized threats. Whether credibility travels similarly in economic coercion or cyber operations is an open question. Future designs can explicitly code threat type (war vs economic vs cyber) then ask whether regime position and polarization interact with those features in

distinctive ways. This expansion matters because different threat ecologies expose different audiences: market actors and bureaucracies in sanctions cases, technical and security communities in cyber incidents, and local social networks in gray-zone contests.

Fourth, treating “autocracy” as a single family is parsimonious but blurs the very dimensions that map to insulation and mobilization in the theory. Disaggregating personalist, party-based, military, and monarchical regimes—and letting polarization interact with these subtypes—would reveal where the “flat” autocratic result hides meaningful variation. In personalist and military regimes, elite polarization (factional splits among insiders and the coercive apparatus) should be the main driver of audience costs because sanctioning capacity is concentrated at the top; public polarization matters chiefly when it couples with cracks in insulation to pull security elites or patrons off the fence. In party-based and monarchic systems, cohesion within the ruling organization or dynasty dampens the effect of mass division until elite discipline erodes. Extending the same logic across regime families also calls for separating elite from public polarization in democracies and anocracies. Mapping these level-specific and feedback effects would clarify when polarization blurs punishability, when it hardens it, and through which actors it travels.

Pursued together, these directions would convert the headline result—targets behave as if reading domestic punishability—into a more granular map of when, how, and for whom that reading changes. They would show whether elites and masses infer costs differently, how credibility accumulates or erodes across linked exchanges, which threat ecologies carry domestic cues most faithfully, and which regime

subtypes truly insulate leaders from AC. The payoff is a theory that travels across instruments of coercion and a toolkit that lets analysts calibrate signals and off-ramps to the specific domestic architecture facing the adversary.



REFERENCES

- Acharya, A., & Grillo, E. (2019). A behavioral foundation for audience costs. *Quarterly Journal of Political Science*, 14(2), 159–190. <https://doi.org/10.1561/100.00017124>.
- Allison, G. T. (1969). Conceptual models and the Cuban missile crisis. *American Political Science Review*, 63(3), 689–718. <https://doi.org/10.2307/1954423>.
- Arbatli, E., & Rosenberg, D. (2021). United we stand, divided we rule: How political polarization erodes democracy. *Democratization*, 28(2), 285–307. <https://doi.org/10.1080/13510347.2020.1818068>
- Barbieri, K., Keshk, O. M., & Pollins, B. M. (2009). Trading data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science*, 26(5), 471–491. <https://doi.org/10.1177/0738894209343887>
- Baum, M. A. (2004). Going private: Public opinion, presidential rhetoric, and the domestic politics of audience costs in U.S. foreign policy crises. *Journal of Conflict Resolution*, 48(5), 603–631. <https://doi.org/10.1177/0022002704267764>
- Bayer, R. (2006). *Diplomatic Exchange Data set, v2006.1*. <https://correlatesofwar.org>
- Bogaards, M. (2009). How to classify hybrid regimes? Defective democracy and electoral authoritarianism. *Democratization*, 16(2), 399–423. <https://doi.org/10.1080/13510340902777800>
- Bowles, S. (2016). *The moral economy: Why good incentives are no substitute for good citizens*. Yale University Press.

- Brewer, M. D. (2005). The rise of partisanship and the expansion of partisan conflict within the American electorate. *Political Research Quarterly*, 58(2), 219–229. <https://doi.org/10.2307/3595624>
- Brown, J. N., & Marcum, A. S. (2011). Avoiding audience costs: Domestic political accountability and concessions in crisis diplomacy. *Security Studies*, 20(2), 141–170. <https://doi.org/10.1080/09636412.2011.572671>
- Brownlee, J., & Miao, K. (2022). Why democracies survive. *Journal of Democracy*, 33(4), 133–149. <https://doi.org/10.1353/jod.2022.0052>
- Brutger, R., & Kertzer, J. D. (2018). A dispositional theory of reputation costs. *International Organization*, 72(3), 693–724. <https://doi.org/10.1017/S0020818318000188>
- Bueno de Mesquita, B., & Siverson, R. M. (1995). War and the survival of political leaders: A comparative study of regime types and political accountability. *American Political Science Review*, 89(4), 841–855. <https://doi.org/10.2307/2082512>
- Cartwright, D. (1971). Risk taking by individuals and groups: An assessment of research employing choice dilemmas. *Journal of Personality and Social Psychology*, 20(3), 361–378. <https://doi.org/10.1037/h0031912>
- Carugati, F. (2020). Democratic stability: A long view. *Annual Review of Political Science*, 23(1), 59–75. <https://doi.org/10.1146/annurev-polisci-052918-012050>
- Chaudoin, S. (2014). Audience features and the strategic timing of trade disputes. *International Organization*, 68(4), 877–911. <https://doi.org/10.1017/S0020818314000174>

- Chiozza, G. (2017). Presidents on the cycle: Elections, audience costs, and coercive diplomacy. *Conflict Management and Peace Science*, 34(1), 3–26.
<https://doi.org/10.1177/0738894215593677>
- Choi, S. W., & James, P. (2007). Media openness, democracy and militarized interstate disputes. *British Journal of Political Science*, 37(1), 23–46.
<https://doi.org/10.1017/S0007123407000026>
- Cilizoglu, M., Tokdemir, E., & Zarpli, O. (2025). To punish or to reward? The effect of sanction threats on public opinion. *International Interactions*, 1–21.
<https://doi.org/10.1080/03050629.2025.2518382>
- Clare, J. (2007). Domestic audiences and strategic interests. *The Journal of Politics*, 69(3), 732–745. <https://doi.org/10.1111/j.1468-2508.2007.00571.x>
- Conrad, C. R., Conrad, J., & Young, J. K. (2014). Tyrants and terrorism: Why some autocrats are terrorized while others are not. *International Studies Quarterly*, 58(3), 539–549. <https://doi.org/10.1111/isqu.12120>
- Coppedge, M. & et al. (2020). *Varieties of democracy: Measuring two centuries of political change*. Cambridge: Cambridge University Press.
- Crisman-Cox, C., & Gibilisco, M. (2018). Audience costs and the dynamics of war and peace. *American Journal of Political Science*, 62(3), 566–580.
<https://doi.org/10.1111/ajps.12347>
- Croco, S. E., Hanmer, M. J., & McDonald, J. A. (2021). At what cost? Reexamining audience costs in realistic settings. *The Journal of Politics*, 83(1), 8–22.
<https://doi.org/10.1086/708912>

- Croco, S. E., Hanmer, M. J., & McDonald, J. A. (2021). At what cost? Reexamining audience costs in realistic settings. *The Journal of Politics*, 83(1), 8–22.
<https://doi.org/10.1086/708912>
- Da Vinha, L. (2021). A tale of two red lines: Managing foreign policy crises in the Obama and Trump administrations. *Comparative Strategy*, 40(1), 63–85.
<https://doi.org/10.1080/01495933.2021.1853441>
- Dahl, R. A. (1971). *Polyarchy: Participation and opposition*. Yale university press.
- De Mesquita, B. & et al. (2005). *The logic of political survival*. MIT press.
- Diamond, L. (2002). Elections without democracy: Thinking about hybrid regimes. *Journal of democracy*, 13(2), 21-35.
<https://doi.org/10.1353/jod.2002.0025>
- Drinot, P. (2017). Remembering Velasco: Contested Memories of the Revolutionary Government of the Armed Forces. In *The Peculiar Revolution: Rethinking the Peruvian Experiment under Military Rule* (pp. 95-120). University of Texas Press.
- Duveroglu, B., & Tokdemir, E. (2025). To Be Experienced or Not? Unpacking the relationship between leader tenure and counterinsurgency efforts. *Terrorism and Political Violence*, 37(3), 386–408.
<https://doi.org/10.1080/09546553.2023.2301553>
- Eckstein, H. (1961). *A theory of stable democracy*. Princeton University Press.
- Eckstein, Harry. (1992). *Regarding politics: Essays on political theory, stability, and change*. Univ of California Press.
- Escribà-Folch, A. (2013). Repression, political threats, and survival under autocracy. *International Political Science Review*, 34(5), 543–560.
<https://doi.org/10.1177/0192512113488259>

- Evers, M. M., Fisher, A., & Schaaf, S. D. (2019). Is there a Trump effect? An experiment on political polarization and audience costs. *Perspectives on Politics*, 17(2), 433–452. <https://doi.org/10.1017/S1537592718003390>
- Eyerman, J., & Hart, R. A., Jr. (1996). An empirical test of the audience cost proposition: Democracy speaks louder than words. *Journal of Conflict Resolution*, 40(4), 597–616. <https://doi.org/10.1177/0022002796040004004>
- Ezrow, N. M., & Frantz, E. (2011). State institutions and the survival of dictatorships. *Journal of International Affairs*, 65(1), 1–13. <http://www.jstor.org/stable/24388178>
- Fearon, J. D. (1994a). Domestic political audiences and the escalation of international disputes. *American Political Science Review*, 88(3), 577–592. <https://doi.org/10.2307/2944796>
- Fearon, J. D. (1994b). Signaling versus the balance of power and interests: An empirical test of a crisis bargaining model. *Journal of Conflict Resolution*, 38(2), 236–269. <https://doi.org/10.1177/0022002794038002004>
- Fearon, J. D. (1997). Signaling foreign policy interests: Tying hands versus sinking costs. *Journal of Conflict Resolution*, 41(1), 68–90. <https://doi.org/10.1177/0022002797041001004>
- Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, 97(1), 75–90. <https://doi.org/10.1017/S0003055403000534>
- Feng, Y. (1997). Democracy, political stability and economic growth. *British Journal of Political Science*, 27(3), 391–418. <https://doi.org/10.1017/S0007123497000197>

- Ferrara, G. M. & Hunt, C. (1979). *Latin America 1978*. Facts on File, Inc.
- Frantz, E. (2003). Breaking down the residual category: Policy stability among dictatorships from a veto players perspective. *Available at SSRN 904263*.
- Frantz, E. (2007a). The International Conflict Behavior of Dictatorships from an Audience Cost Perspective. Available at SSRN: SSRN ID 975158. <https://doi.org/10.2139/ssrn.975158>
- Frantz, E. (2007b). Tying the dictator's hands: Leadership survival in authoritarian regimes. Available at SSRN: SSRN ID 975161. <https://doi.org/10.2139/ssrn.975161>
- Frantz, E. (2018). *Authoritarianism: What everyone needs to know*. Oxford University Press.
- Frantz, E. (2024). Authoritarian survival. In *Research Handbook on Authoritarianism* (pp. 229-243). Edward Elgar Publishing.
- Frantz, E., & Kendall-Taylor, A. (2014). A dictator's toolkit: Understanding how co-optation affects repression in autocracies. *Journal of Peace Research*, 51(3), 332–346. <https://doi.org/10.1177/0022343313519808>
- Fuhrmann, M., & Sechser, T. S. (2014). Signaling alliance commitments: Hand-tying and sunk costs in extended nuclear deterrence. *American Political Science Review*, 58(4), 919–935. <https://doi.org/10.1111/ajps.12082>
- Fuller, C. R. (2015). Regime stability in anocracies: The role of special economic zones. *Jadavpur Journal of International Relations*, 19(2), 85–105. <https://doi.org/10.1177/0973598415627889>

- Gandhi, J., & Vreeland, J. (2004). Political institutions and civil war: unpacking anocracy. *University of Notre Dame. Unpublished manuscript.*
- Gartzke, E., & Lupu, Y. (2012). Still looking for audience costs. *Security Studies*, 21(3), 391–397. <https://doi.org/10.1080/09636412.2012.706486>
- Geddes, B. (1999). What do we know about democratization after twenty years? *Annual Review of Political Science*, 2(1), 115–144. <https://doi.org/10.1146/annurev.polisci.2.1.115>
- Geddes, B., & Zaller, J. (1989). Sources of popular support for authoritarian regimes. *American Journal of Political Science*, 33(2), 319–347. <https://doi.org/10.2307/2111150>
- Geddes, B., Wright, J., & Frantz, E. (2014). Autocratic breakdown and regime transitions: A new data set. *Perspectives on Politics*, 12(2), 313–331. <https://doi.org/10.1017/S1537592714000851>
- Geddes, B., Wright, J., & Frantz, E. (2018). *How dictatorships work: Power, personalization, and collapse*. Cambridge University Press.
- Gelpi, C. F., & Griesdorf, M. (2001). Winners or losers? Democracies in international crisis, 1918–94. *American Political Science Review*, 95(3), 633–647. <https://doi.org/10.1017/S0003055401003148>
- Gibler, D. M. (2009). *International military alliances, 1648-2008*. CQ Press.
- Gilbert, L., & Mohseni, P. (2011). *Beyond authoritarianism: The conceptualization of hybrid regimes*. *Studies in Comparative International Development*, 46(3), 270–297. <https://doi.org/10.1007/s12116-011-9088-x>

- Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., & Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science*, 54(1), 190–208. <https://doi.org/10.1111/j.1540-5907.2009.00426.x>
- Goldstone, J. A., Gurr, T. R., Marshall, M. G., & Ulfelder, J. (2004). *It's all about state structure: New findings on revolutionary origins from global data*. *Homo Oeconomicus*, 21, 429–455. Institute of SocioEconomics. Retrieved from IDEAS (RePEc): <https://ideas.repec.org/a/hom/homoec/v21y2004p429-455.html>
- Guriev, S., & Treisman, D. (2019). Informational autocrats. *Journal of Economic Perspectives*, 33(4), 100–127. <https://doi.org/10.1257/jep.33.4.100>
- Gurr, T. R. (1974). Persistence and change in political systems, 1800–1971. *American Political Science Review*, 68(4), 1482–1504. <https://doi.org/10.2307/1959937>
- Hauenstein, M. (2020). The conditional effect of audiences on credibility. *Journal of Peace Research*, 57(3), 422–436. <https://doi.org/10.1177/0022343319871983>
- Hegre, H. (2001). Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. *American Political Science Review*, 95(1), 33–48. <https://doi.org/10.1017/S0003055401000119>
- Huntington, S. P. (1968). *Political order in changing societies*. Yale university press.
- Huntington, S.P. (1996). Democracy for the Long Haul. *Journal of Democracy* 7(2), 3-13. <https://dx.doi.org/10.1353/jod.1996.0028>

- Hurtado-Torres, S., & Fermandois, J. (2020). The war that didn't break out: Military rule and regional tensions in the Andes in the 1970s. *The International History Review*, 42(5), 967–986. <https://doi.org/10.1080/07075332.2019.1652839>
- Hurwitz, L. (1973). Contemporary approaches to political stability. *Comparative Politics*, 5(3), 449–463. <https://doi.org/10.2307/421273>
- Hutchison, M. L., & Gibler, D. M. (2007). Political tolerance and territorial threat: A cross-national study. *The Journal of Politics*, 69(1), 128–142. <https://doi.org/10.1111/j.1468-2508.2007.00499.x>
- Iyengar, S., Lelkes, Y., Levendusky, M. S., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1), 129–146. <https://doi.org/10.1146/annurev-polisci-051117-073034>
- Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405–431. <https://doi.org/10.1093/poq/nfs038>
- Janusch, H. (2021). Audience, agenda setting, and issue salience in international negotiations. *Cooperation and Conflict*, 56(4), 472–490. <https://doi.org/10.1177/00108367211000784>
- Johnson, J. C. (2017). External threat and alliance formation. *International Studies Quarterly*, 61(3), 736–745. <https://doi.org/10.1093/isq/sqw054>
- Kertzer, J. D., & Brutger, R. (2016). Decomposing audience costs: Bringing the audience back into audience cost theory. *American Journal of Political Science*, 60(1), 234–249. <https://doi.org/10.1111/ajps.12201>

- Kinne, B. J., & Marinov, N. (2013). Electoral authoritarianism and credible signaling in international crises. *Journal of Conflict Resolution*, 57(3), 359–386. <https://doi.org/10.1177/0022002712446124>
- Kohama, S., Quek, K., & Tago, A. (2024). Managing the costs of backing down: A “mirror experiment” on reputations and audience costs in a real-world conflict. *The Journal of Politics*, 86(1), 388–393. <https://doi.org/10.1086/726927>
- Khrushchev, Nikita. (2007). *Memoirs of Nikita Khrushchev. Volume 3, Statesman, 1953-1964*. Pennsylvania State University Press.
- Kurizaki, S. (2007). Efficient secrecy: Public versus private threats in crisis diplomacy. *American Political Science Review*, 101(3), 543–558. <https://doi.org/10.1017/S0003055407070396>
- Kurizaki, S., & Whang, T. (2015). Detecting audience costs in international disputes. *International Organization*, 69(4), 949–980. <https://doi.org/10.1017/S0020818315000211>
- Lebow, R. N. (1983). Miscalculation in the South Atlantic: The origins of the Falkland War. *Journal of Strategic Studies*, 6(1), 5–35. <https://doi.org/10.1080/01402398308437139>
- Leonid, B. (2007). American military presence in Kyrgyzstan: problems and possible repercussions. *Central Asia and the Caucasus*, 4(46), 63-71. <https://www.cac-c.org/index.php/cac/article/view/1098>
- Levendusky, M. S., & Horowitz, M. C. (2012). When backing down is the right decision: Partisanship, new information, and audience costs. *The Journal of Politics*, 74(2), 323–338. <https://doi.org/10.1017/S002238161100154X>

- Levitsky, S., & Way, L. A. (2002). Elections without democracy: The rise of competitive authoritarianism. *Journal of Democracy*, 13(2), 51–65. <https://doi.org/10.1353/jod.2002.0026>
- Levy, J. S. (1988). Domestic politics and war. *The Journal of Interdisciplinary History*, 18(4), 653–673. <https://doi.org/10.2307/204819>
- Levy, J. S. (2012). Coercive threats, audience costs, and case studies. *Security Studies*, 21(3), 383–390. <https://doi.org/10.1080/09636412.2012.706477>
- Levy, J. S., McKoy, M. K., Poast, P., & Wallace, G. P. R. (2015). Backing out or backing in? Commitment and consistency in audience costs theory. *American Journal of Political Science*, 59(4), 988–1001. <https://doi.org/10.1111/ajps.12197>
- Li, X., & Chen, D. (2021). Public opinion, international reputation, and audience costs in an authoritarian regime. *Conflict Management and Peace Science*, 38(5), 543–560. <https://doi.org/10.1177/0738894220906374>
- Lin-Greenberg, E. (2019). Backing up, not backing down: Mitigating audience costs through policy substitution. *Journal of Peace Research*, 56(4), 559–574. <https://doi.org/10.1177/0022343319832641>
- Lipset, S. M., & Rokkan, S. (1967). *Cleavage structures, party systems, and voter alignments: An introduction*. New York Free Press.
- Lisińska, M. (2019). *Argentine Foreign Policy during the Military Dictatorship, 1976–1983: Between a Nationalist and Pragmatic Approach*. Springer.
- Magnani, E., & Barreto, M. L. (2022). Defense policy shaping foreign policy: An alternative interpretation through the study of the Argentine–Chilean relations. *Revista Relaciones Internacionales*, 95(1), 103–126. <https://doi.org/10.15359/ri.95-1.4>

- Mankoff, J. (2013). *The United States and Central Asia after 2014* (CSIS Reports). Center for Strategic and International Studies. Retrieved from <https://www.csis.org/analysis/united-states-and-central-asia-after-2014>
- Maoz, Z., & Russett, B. (1993). Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review*, 87(3), 624–638. <https://www.jstor.org/stable/2938740>
- Marat, E. (2009). Bakiyev wins new geopolitical game over Manas base. *CACI Analyst*. Central Asia-Caucasus Institute & Silk Road Studies Program. Retrieved from <https://www.cacianalyst.org/publications/analytical-articles/item/11868-analytical-articles-caci-analyst-2009-7-1-art-11868.html>
- Marshall, M. G. (2023). Hybrid authority systems and political instability. In *Handbook of Fragile States* (pp. 121-136). Edward Elgar Publishing.
- Masterson, D. (1991). *Militarism and Politics in Latin America*. Greenwood Publishing Group.
- McAdam, D. (1999). *Political process and the development of black insurgency, 1930-1970*. University of Chicago Press.
- McCoy, J., & Somer, M. (2019). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The Annals of the American Academy of Political and Social Science*, 681(1), 234–271. <https://doi.org/10.1177/0002716218818782>
- McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1), 16–42. <https://doi.org/10.1177/0002764218759576>

- McManus, R. W., & Yarhi-Milo, K. (2017). The logic of “offstage” signaling: Domestic politics, regime type, and major power-protégé relations. *International Organization*, 71(4), 701–733. <https://doi.org/10.1017/S0020818317000297>
- Mercau, E. (2019). *The Falklands War: an imperial history*. Cambridge University Press.
- Mercer, J. (2012). Audience costs are toys. *Security Studies*, 21(3), 398–404. <https://doi.org/10.1080/09636412.2012.706488>
- Moon, C., & Souva, M. (2016). Audience costs, information, and credible commitment problems. *Journal of Conflict Resolution*, 60(3), 434–458. <https://doi.org/10.1177/0022002714545222>
- Morrow, J. D. (2000). Alliances: Why write them down? *Annual Review of Political Science*, 3(1), 63–83. <https://doi.org/10.1146/annurev.polisci.3.1.63>
- Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, 12(2), 125–135. <https://doi.org/10.1037/h0027568>
- Mufti, M. (2018). What do we know about hybrid regimes after two decades of scholarship? *Politics and Governance*, 6(2), 112–119. <https://doi.org/10.17645/pag.v6i2.1400>
- Muller, E. N., & Weede, E. (1990). Cross-national variation in political violence: A rational action approach. *Journal of Conflict Resolution*, 34(4), 624–651. <https://doi.org/10.1177/0022002790034004003>

- Myers, D. G. (1978). Polarizing effects of social comparison. *Journal of Experimental Social Psychology*, 14(6), 554-563. [https://doi.org/10.1016/0022-1031\(78\)90049-5](https://doi.org/10.1016/0022-1031(78)90049-5)
- Nanovsky, S., & Knox, C. (2024). *Political stability in authoritarian regimes: The case of Central Asia*. *Central Asian Survey*, 1–24. <https://doi.org/10.1080/02634937.2024.2402067>
- Nomikos, W. G., & Sambanis, N. (2019). What is the mechanism underlying audience costs? Incompetence, belligerence, and inconsistency. *Journal of Peace Research*, 56(4), 575–588. <https://doi.org/10.1177/0022343319839456>
- North, D. C. (1990). A transaction cost theory of politics. *Journal of Theoretical Politics*, 2(4), 355–367. <https://doi.org/10.1177/0951692890002004001>
- Onea, T. A. (2025). How to entrap your protector: Reassessing entrapment in light of the Crimean War crisis. *Review of International Studies*. Advance online publication. <https://doi.org/10.1017/S0260210524000731>
- Palmer, G., McManus, R. W., D’Orazio, V., Kenwick, M. R., Karstens, M., Bloch, C., Dietrich, N., Kahn, K., Ritter, K., & Soules, M. J. (2022). The MID5 dataset, 2011–2014: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 39(4), 470–482. <https://doi.org/10.1177/0738894221995743>
- Park, D. J. (2022). A polarized audience in South Korea and its impact on North Korea policy. *Journal of Indo-Pacific Affairs*. Air University Press. Retrieved from <https://www.airuniversity.af.edu/JIPA/Display/Article/3212588/a-polarized-audience-in-south-korea-and-its-impact-on-north-korea-policy/>

- Partell, P. J., & Palmer, G. (1999). Audience costs and interstate crises: An empirical assessment of Fearon's model of dispute outcomes. *International Studies Quarterly*, 43(2), 389–405. <https://doi.org/10.1111/0020-8833.00125>
- Pate, A. (2020). Trends in democratization: A focus on instability in Anocracies. In *Peace and Conflict 2008* (pp. 27-32). Routledge.
- Piotrowska, B. M. (2020). The price of collaboration: How authoritarian states retain control. *Comparative Political Studies*, 53(13), 2091–2117. <https://doi.org/10.1177/0010414020912277>
- Popkin, S. L. (1991). *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press.
- Potter, P. B. K., & Baum, M. A. (2010). Democratic peace, domestic audience costs, and political communication. *Political Communication*, 27(4), 453–470. <https://doi.org/10.1111/j.1468-2478.2012.00748.x>
- Potter, P. B. K., & Baum, M. A. (2014). Looking for audience costs in all the wrong places: Electoral institutions, media access, and democratic constraint. *The Journal of Politics*, 76(1), 167–181. <https://doi.org/10.1017/S0022381613001230>
- Press, D. G. (2005). *Calculating credibility: How leaders assess military threats*. Cornell University Press.
- Prins, B. C. (2003). Institutional instability and the credibility of audience costs: Political participation and interstate crisis bargaining, 1816–1992. *Journal of Peace Research*, 40(1), 67–84. <https://doi.org/10.1177/0022343303040001206>

- Quek, K., & Johnston, A. I. (2017/2018). Can China back down? Crisis de-escalation in the shadow of popular opposition. *International Security*, 42(3), 7–36. https://doi.org/10.1162/ISEC_a_00303
- Ramsay, K. W. (2004). Politics at the water's edge: Crisis bargaining and electoral competition. *Journal of Conflict Resolution*, 48(4), 459–486. <https://doi.org/10.1177/0022002704266156>
- Rawls, J. (1971). *A Theory of Justice*. Belknap Press.
- Regan, P. M., & Bell, S. R. (2010). Changing lanes or stuck in the middle: Why are anocracies more prone to civil wars? *Political Research Quarterly*, 63(4), 747–759. <https://doi.org/10.1177/1065912909336274>
- Schelling, T. C. (1956). An essay on bargaining. *American Economic Review*, 46(3), 281–306. <https://www.jstor.org/stable/1805498>
- Schelling, T. C. (1970). The diplomacy of violence. In *Theories of Peace and Security: A Reader in Contemporary Strategic Thought* (pp. 64-84). London: Palgrave Macmillan UK.
- Schelling, T. C. (1982). Thinking about nuclear terrorism. *International Security*, 6(4), 61–77. <https://doi.org/10.2307/2538678>
- Schultz, K. A. (1998). Domestic opposition and signaling in international crises. *American Political Science Review*, 92(4), 829–844. <https://doi.org/10.2307/2586306>
- Schultz, K. A. (2001). Looking for audience costs. *Journal of Conflict Resolution*, 45(1), 32–60. <https://doi.org/10.1177/0022002701045001002>
- Schultz, K. A. (2012). Why we needed audience costs and what we need now. *Security Studies*, 21(3), 369–375. <https://doi.org/10.1080/09636412.2012.706475>

- Schwartz, J. A., & Blair, C. W. (2020). Do women make more credible threats? Gender stereotypes, audience costs, and crisis bargaining. *International Organization*, 74(4), 872–895. <https://doi.org/10.1017/S0020818320000223>
- Sechser, T. S. (2011). Militarized compellent threats, 1918–2001. *Conflict Management and Peace Science*, 28(4), 377–401. <https://doi.org/10.1177/0738894211413066>
- Sechser, T. S. (2018). Reputations and signaling in coercive bargaining. *Journal of Conflict Resolution*, 62(2), 318–345. <https://doi.org/10.1177/0022002716652687>
- Singer, J. D. (1988). Reconstructing the Correlates of War dataset on material capabilities of states, 1816–1985. *International Interactions*, 14(2), 115–132. <https://doi.org/10.1080/03050628808434695>
- Slantchev, B. L. (2006). Politicians, the media, and domestic audience costs. *International Studies Quarterly*, 50(2), 445–477. <https://doi.org/10.1111/j.1468-2478.2006.00409.x>
- Slantchev, B. L. (2012). Audience cost theory and its audiences. *Security Studies*, 21(3), 376–382. <https://doi.org/10.1080/09636412.2012.706476>
- Smetana, M. (2025). Microfoundations of domestic audience costs in nondemocratic regimes: Experimental evidence from Putin’s Russia. *Journal of Peace Research*, 62(2), 278–294. <https://doi.org/10.1177/00223433231220252>
- Smith, A. (1998). International crises and domestic politics. *American Political Science Review*, 92(3), 623–638. <https://doi.org/10.2307/2585485>

- Stinnett, D. M., Tir, J., Diehl, P. F., Schafer, P., & Gochman, C. (2002). The Correlates of War (COW) Project direct contiguity data, version 3.0. *Conflict Management and Peace Science*, 19(2), 59–67.
<https://doi.org/10.1177/073889420201900203>
- Stoner, J. A. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology*, 4(4), 442–459.
[https://doi.org/10.1016/0022-1031\(68\)90069-3](https://doi.org/10.1016/0022-1031(68)90069-3)
- Suzuki, A. (2018). Audience costs, domestic economy and coercive diplomacy. *Research & Politics*, 5(3), 1–7. <https://doi.org/10.1177/2053168018787119>
- Svolik, M. W. (2012). *The politics of authoritarian rule*. Cambridge University Press.
- Taddonio, P. (2015). The president blinked”: Why Obama changed course on the “red line” in Syria. *Frontline*. <https://www.pbs.org/wgbh/frontline/article/the-president-blinked-why-obama-changed-course-on-the-red-line-in-syria/>
- Tajfel, H. E. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. Academic press.
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2), 149–178. <https://doi.org/10.1002/ejsp.2420010202>
- Takei, M. (2024). Audience costs and the credibility of public versus private threats in international crises. *International Studies Quarterly*, 68(3), sqae091.
<https://doi.org/10.1093/isq/sqae091>
- Toepfl, F. (2020). Comparing authoritarian publics: The benefits and risks of three types of publics for autocrats. *Communication Theory*, 30(2), 105–125.
<https://doi.org/10.1093/ct/qtz015>

- Tomz, M. (2007). Domestic audience costs in international relations: An experimental approach. *International Organization*, 61(4), 821–840.
<https://doi.org/10.1017/S0020818307070282>
- Tomz, M. R., & Weeks, J. L. P. (2013). Public opinion and the democratic peace. *American Political Science Review*, 107(4), 849–865.
<https://doi.org/10.1017/S0003055413000488>
- Trachtenberg, M. (2012). Audience costs: An historical analysis. *Security Studies*, 21(1), 3–42. <https://doi.org/10.1080/09636412.2012.650590>
- Trager, R. F., & Vavreck, L. (2011). The political costs of crisis bargaining: Presidential rhetoric and the role of party. *American Journal of Political Science*, 55(3), 526–545. <https://doi.org/10.1111/j.1540-5907.2011.00521.x>
- Trinkunas, H. (2024). Regime hybridization and civil–military relations. In *Research Handbook on Civil–Military Relations* (pp. 388–403). Edward Elgar Publishing.
- Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology*, 9(2), 187–204.
<https://doi.org/10.1002/ejsp.2420090207>
- Uzonyi, G., Souva, M., & Golder, S. N. (2012). Domestic institutions and credible signals. *International Studies Quarterly*, 56(4), 765–776.
<https://doi.org/10.1111/j.1468-2478.2012.00746.x>
- Varieties of Democracy*. V-Dem. (2025, August 19). <https://v-dem.net/>
- Villar-Gertner, A. (2014). The Beagle Channel frontier dispute between Argentina and Chile: Converging domestic and international conflicts. *International Relations*, 28(2), 207–227. <https://doi.org/10.1177/0047117814526910>

- Wagner, R. H. (2000). Bargaining and war. *American Journal of Political Science*, 44(3), 469–484. <https://doi.org/10.2307/2669259>
- Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1), 35–64. <https://doi.org/10.1017/S0020818308080028>
- Weeks, J. L. (2012). Strongmen and straw men: Authoritarian regimes and the initiation of international conflict. *American Political Science Review*, 106(2), 326–347. <https://doi.org/10.1017/S0003055412000111>
- Weeks, J. L. (2014). Dictators at war and peace. In *Dictators at War and Peace*. Cornell University Press.
- Weingast, B. R. (1997). The political foundations of democracy and the rule of law. *American Political Science Review*, 91(2), 245–263. <https://www.jstor.org/stable/2952354>
- Weiss, J. C. (2013). Authoritarian signaling, mass audiences, and nationalist protest in China. *International Organization*, 67(1), 1–35. <https://doi.org/10.1017/S0020818312000380>
- Weiss, J. C., & Dafoe, A. (2019). Authoritarian audiences, rhetoric, and propaganda in international crises: Evidence from China. *International Studies Quarterly*, 63(4), 963–973. <https://doi.org/10.1093/isq/sqz059>
- Wigell, M. (2008). Mapping ‘hybrid regimes’: Regime types and concepts in comparative politics. *Democratization*, 15(2), 230–250. <https://doi.org/10.1080/13510340701846319>
- Wong, S. S. (2024). Audience costs, humiliation, and social creativity strategies: How Beijing boosts citizens’ esteem in international conflicts. *European Journal of*

International Relations, 30(3), 569–594.

<https://doi.org/10.1177/13540661241259861>

Yarhi-Milo, K. (2013). In the eye of the beholder: How leaders and intelligence communities assess the intentions of adversaries. *International Security*, 38(1), 7–51. https://doi.org/10.1162/ISEC_a_00128

Yu, C., Whang, T., & Lee, S. (2022). When does audience matter? Challengers' stability and audience costs. *Foreign Policy Analysis*, 18(3), orac011. <https://doi.org/10.1093/fpa/orac011>

Zaller, John. *The nature and origins of mass opinion*. Cambridge university press, 1992.

Zarpli, O. (2023). When do imposed sanctions work? The role of target regime type. *Journal of Conflict Resolution*, 67(7–8), 1482–1509. <https://doi.org/10.1177/00220027221139809>

Zellman, A. (2020). Cheap talk or policy lock? Nationalist frames and sympathetic audience costs in international territorial disputes. *Territory, Politics, Governance*, 8(3), 336–355. <https://doi.org/10.1080/21622671.2018.1561325>

APPENDICES

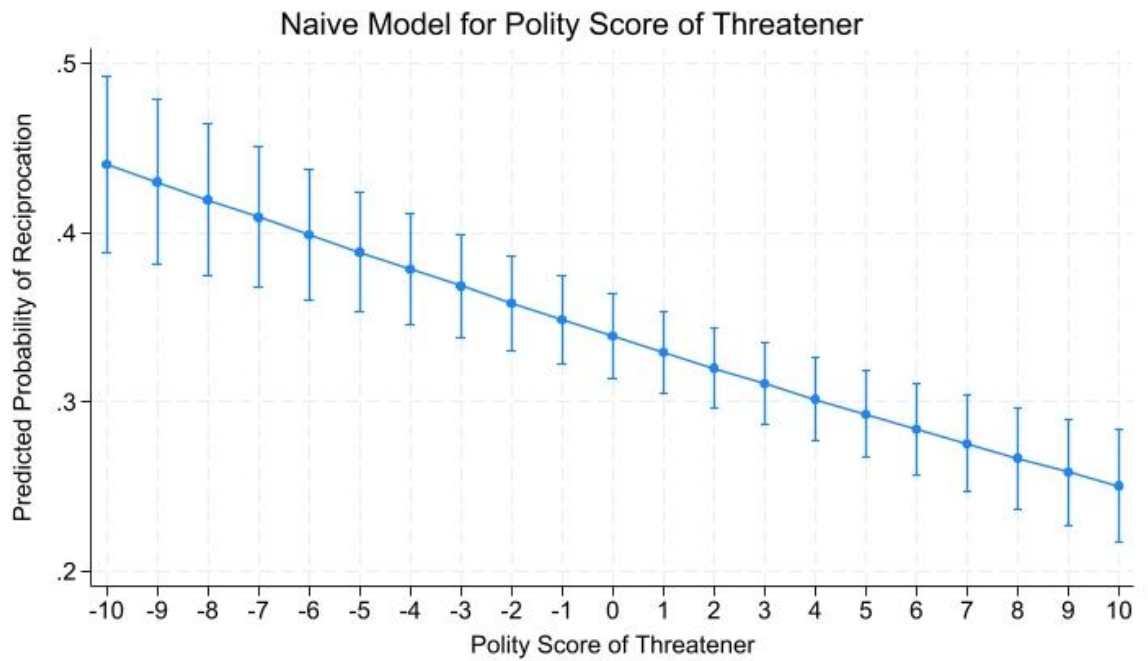


Figure 5 Naïve Model for Polity Score

Naïve Model for Impact of Polarization on Reciprocation within Regimes

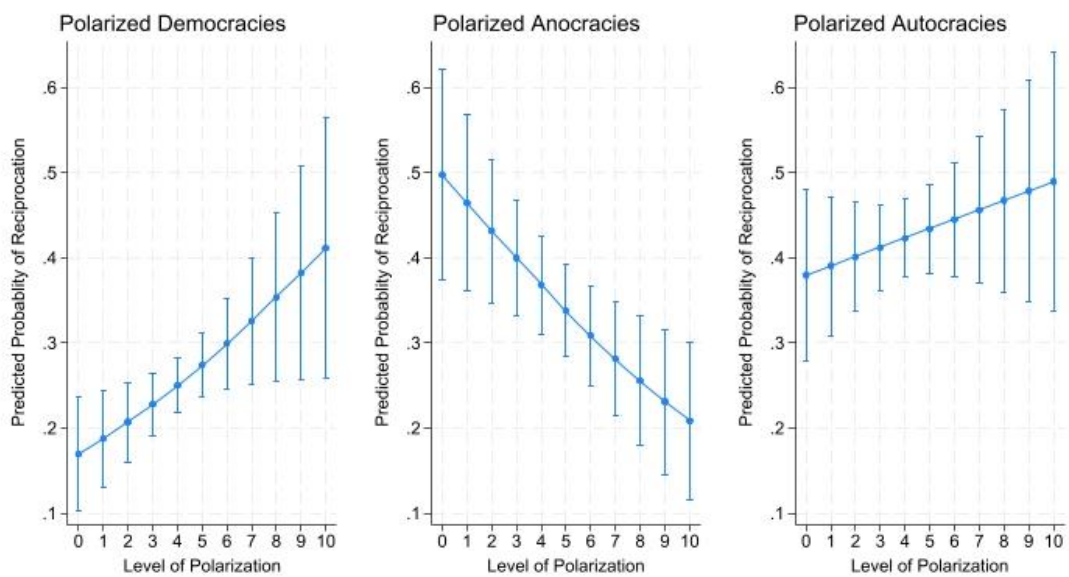


Figure 6 Naïve Model for Polarization's Conditional Impact