Authoritarian States and the Architecture of Global Governance

Ji Hyeon Bae

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Geoffrey Wallace, Chair
Rachel Cichowski
Asli Cansunar
Tom Ginsburg

Program Authorized to Offer Degree:
Political Science

University of Washington

Abstract

Authoritarian States and the Architecture of Global Governance

Ji Hyeon Bae

Chair of the Supervisory Committee:
Geoffrey Wallace
Political Science

How do authoritarian states engage with international institutions and legal norms? While international law and intergovernmental organizations are often associated with liberal democracies, authoritarian regimes increasingly participate in, construct, and invoke them. The purpose of this research is to examine how and why authoritarian states shape international law and organizations, and whether such engagement alters the perceived legitimacy of international rules.

The central claim of this dissertation is that international institutions are not inherently liberal. I argue that two interlocking pillars of global governance, namely, law and organization, can be shaped differently by regime type of states. Authoritarian states do not simply resist international law, they strategically engage with it. They cluster institutionally, reinterpret foundational legal principles, and invoke legal authority to legitimate contentious policies. By examining these dynamics across legal discourse, and public reception, this project develops a unified framework for understanding authoritarian international cooperation.

The first chapter, "Varieties of Intergovernmental Organization: Measuring Member-State Distributions," develops a new measurement framework to capture the regime composition of IGOs. Using a Bayesian approach to model the distribution of member-state regime types,

I show that IGOs exhibit systematic variation in their degree of authoritarian skew rather than central tendency. Organizations skewed towards authoritarian regimes display distinct political characteristics, suggesting that regime composition has meaningful consequences for how international cooperation is structured and maintained.

The second chapter, "The Domestic Sources of International Legal Rhetoric: Contesting Sovereignty Across Regimes," examines how governments interpret the principle of sovereignty in United Nations General Debate speeches. Using text-as-data methods, including word embeddings and semantic similarity analysis, I demonstrate that authoritarian regimes invoke sovereignty more frequently and associate it more closely with non-interference and territorial control. These findings indicate that international legal principles are not fixed in meaning but are strategically reinterpreted in ways that reflect domestic political incentives.

The third chapter, "Content or Cue? Conditional Messenger Effects in the Legitimacy of International Law," investigates whether the source of legal claims shapes public perceptions of legitimacy. Drawing on factorial survey experiments conducted in the Philippines and Australia, I find that citizens continue to view international law as legitimate even when it is endorsed solely by authoritarian states. While message effects condition responses at the margins, the results suggest that legal authority is not fully undermined by its association with authoritarian actors.

Table of Contents

Acknowledgments

Above all, I thank my advisor, Geoff Wallace, for helping me grow as an independent political scientist over the past several years. Geoff has been everything I could have wished for in an advisor, and I cannot thank him enough for sitting through so many drafts while always seeing the best version of them. He always saw the broader contribution of my work before I did, pushing me to think bigger than I thought I could. Early on, as a second-year PhD student, Geoff reminded me that no one is a better expert on my research topic than I am, at a moment when I was starting to black out from nervousness during my M.A. defense. That was one of countless moments when Geoff gave me much needed support and confidence, and in doing so, helped me grow into the scholar I am today. Beyond his intellectual guidance, Geoff's extraordinary organizational clarity made every milestone feel reachable. I always left our meetings with a clear sense of direction, which is no small gift in the fog of a dissertation.

I also thank Rachel Cichowski, who has supported me since my very first days at UW, when I served as her teaching assistant. Insights from her seminars contributed to the shaping of this dissertation, and her invitations to public law conferences in Bergen and Atlanta opened doors I would not have found on my own. Those experiences were among the highlights of my graduate school years.

Asli Cansunar has been one of my biggest supporters and most trusted sounding boards. I thank her for always being there when I needed her, whether for methodological guidance or professionalization advice. Her door was always open, and I walked through it often.

I met Tom Ginsburg at a conference in Stockholm, where he served as my discussant. Having him read my manuscript and offer feedback was already more than I could have hoped for; having him join my committee was an invaluable privilege. His comments on Chapter 2 were transformative, and the chapter is substantially sharper for them. Beyond the dissertation, he generously read related manuscripts and offered feedback that continues to shape my research, a kindness that I will not forget.

I thank Saadia Pekkanen for serving as my Graduate School Representative. Her comments at the UWISC seminars were always a highlight, not only remarkably insightful but delivered with such thoughtfulness and kindness. I truly don't know what I appreciated more, the sharpness of her observations or the generosity behind them. I am so grateful that our connection extended to my dissertation committee.

I also thank the faculty of the Department of Political Science for their generous feedback and mentorship throughout the program. Jon Mercer and Beth Kier offered invaluable comments at various stages of this project, and I am grateful for the intellectual community they helped create. I learned enormously and genuinely enjoyed the seminars led by Tony Gill. I learned how to ask better questions by watching James Long ask brilliant ones. I have cherished the departmental intellectual life, including the International Security Colloquium, the Severyns Ravenholt Seminar, and the Political Economy Forum.

This brings me to Jonas Tallberg, whom I first met at ISA San Francisco and who later invited me to a workshop at Stockholm University. Being included in that gathering of scholars was one of the most meaningful moments of my graduate career, the first time I truly felt I belonged among researchers working on questions I cared about. Simply knowing that people wanted to hear more about my work served as a quiet reassurance during moments when I was not sure I was on the right track.

I owe a great deal of gratitude to the whole team of data scientists at the University of Washington eScience Institute. I benefited greatly from *all* of their programs. Naomi Alterman was always kind enough to answer any question I had, and I would not have felt like I belonged in the data science community had it not been for her. I also thank Bryna Hazelton for helpful feedback during weekly meetings, and Jenny Holcomb, whose quiet generosity I sincerely appreciate. The eScience floor closed at four, but Jenny would let me stay longer, and those extra hours mattered.

Working with Valentina Staneva during the Data Incubator program was a turning point for my research and career. She was kind enough to explain complex language models in so many different ways until I grasped both the intuition and the rigorous math behind them. Valentina spent more than fifteen hours every week working through code with me side by side, and I thank her for that extraordinary generosity of time and patience.

I am so fortunate to have met Curtis Atkisson through the Data Science for Social Good program in summer 2024. Curtis has been a generous collaborator and a genuine intellectual partner, and working with him has made this dissertation stronger in ways I am still discovering. The Bayesian project we built together was the most challenging, most fun, and most rewarding learning experience of my graduate years, the kind of collaboration and mentorship only Curtis could magically pull off. I could not have asked for a better partner in that work.

I am also incredibly lucky to have shared this journey with amazing friends in the political science department. My cohort, in particular, Jana Foxe, Jessica Sciarone, have been the greatest support throughout the program. We have powered through pandemic, comprehensive exams, and countless challenges together. I thank Christianna Parr and Jeff Grove for long walks and delicious dinners. Our conversations always started with research but ended with food and baseball. To Jintong Han, I will always miss our afternoon walks to get boba; you made coming to campus to work fun. I thank Wendy He for countless library hours and the work ethic I tried to match. Thanks also to Anna Nguyen for sweet treats and emotional support, and to Ian Callison for invaluable feedback and commiserating on our surveys - More people should have warned us how difficult it is to field them. To Brian Leung for support especially as we neared the finish line; to Ji Hyeon Chung, Minji Jeong, Chanwoong Yoon, and Eunsik Yoo for being such a wonderful group of friends; to Megan Erickson and Bree Bang-Jensen for generous advice as I prepared for the market; and to Jon Beck, whose guidance as lead TA made my first experience in the classroom far less daunting than it might have been.

Studying at the University of Washington has always been a source of pride. I made sure to take advantage of everything the campus offered, the spectacular libraries, access to any book or article I could want, and the computing resources at the lab. I already miss reading at the Elisabeth Miller Library, playing racquetball with my good friend Jintong at the IMA, and occasionally escaping to the library at the Bothell campus. I loved the quiet spaces in Odegaard and my assigned carrel at Suzzallo. I even completed my yoga teacher training at the UW, and being able to guide my own meditation has been a lasting gift. I also met a wonderful group of friends outside the political science department, Priyam, Ashwini, and Shweta. We shared so much fun and laughter over many Sunday mornings, and I am grateful for every one of them.

None of this would have been possible without my undergraduate advisors. Don Moon was my advisor for three years, during which I worked as his research assistant writing Stata code for projects on the rational design of institutions, middle power diplomacy, and WTO judicial rulings, topics that have since become my own. I am grateful to him for assigning me tasks that required reading foundational journal articles and for trusting me with substantive intellectual work. At the same time, I was also discovering my own intellectual passions. I absolutely loved Sunil Kim's comparative politics class as a freshman at Kyung Hee University. I found everything about dictators puzzling, and Kim helped me reshape that curiosity into an actionable research agenda. It was only natural that I would eventually delve into the topic of authoritarian international cooperation. I also thank Jiyoun An and Won-Tak Hong for believing in my potential as a researcher.

I am deeply grateful for my family in South Korea. My grandfather, Joong-woong Baek, was my biggest fan before he passed away in 2024, just as I was completing the first draft of this dissertation. We never said it, but we both knew for sure I was his favorite grandchild. He used to tell me stories about what it was like to live under an authoritarian government when he was young, stories that, in hindsight, shaped my curiosity about political science from an early age. When I was working at home in South Korea over the summers, he used

to slice fruit for me and shush everyone to turn the volume down, because I was doing very serious research in English.

It was usually my grandmother, Sool-yi Kim, who had to turn down the volume of the singing competition on the television. She is the strongest, smartest, and most thoughtful woman I know. Had she lived in different times, I am certain she would have earned three doctoral degrees. She is the only person who tells me to work less and enjoy more, to go out with friends and practice singing. From my grandparents, with whom I grew up from the age of five, I learned to be generous with myself and to dare to dream without limits.

My parents, Minho Bae and Jeungim Baek, have given me unconditional belief in everything I do. I could not have crossed an ocean and built a life there without knowing they were behind me every step of the way. I told them I would go to the United States to pursue a PhD for five years (which became six) in 2020, when the pandemic was at its peak. They supported my decision without hesitation, even as the world was shutting down around us, and trusted that I would find my way. My younger brother, Junhan Bae, took extra care to make sure I could settle into a new country. He is now the coolest brother, running a successful reptile retail business, traveling to Indonesia to shake hands with lizard farmers and close deals. I find myself thinking of him whenever international commerce comes up in my research.

Lastly, I thank my husband Junhyuk Oh. He made coffee every single day with different latte arts, a small routine that made starting to write each morning feel enjoyable. Junhyuk really made sure *nothing* stood in the way of my academic work. We even postponed our own wedding just so I could participate in a research workshop. Junhyuk has always wanted me to have everything I want in the world, and he has been more committed to my academic career than I sometimes was myself. I talked about my research so often that he can now explain it to his own friends, and does so with greater pride than I do. I am lucky that we get to carry it all together.

Authoritarian States and the Architecture of Global Governance

1 Introduction

Global governance rests on international law and intergovernmental organizations (IGOs), by helping states circumscribe cooperation problems (Abbott & Snidal, 1998; Keohane & Martin, 1995; Martin, 1994; Stein, 1982). International law binds states to rules, formal organizations facilitate monitoring and enforcement, and legalization increases accountability. Institutions reshape the payoff structure by making cooperation more attractive and deviation more costly.

Based on the insight that states can harvest cooperation gains through international institutions, existing work has developed an argument that the variation in institutional design is attributable to the type of problem that states face (Barsoom, 2022; Chayes & Chayes, 1993; Lyall, 2010). Some states are more likely than others to sign and legalize treaties, covering a wide range of issue areas like riparian management, humanitarian treatment, and trade liberalization (Abbott et al., 2000; Allee & Peinhardt, 2010; Moravcsik, 2000; Simmons & Danner, 2010; Tir & Stinnett, 2011). These strands of literature share the idea that membership composition of IGOs determines the cooperation problem type, and institutional designs that we observe.

In much of the literature on international cooperation, however, these institutional features are associated with the preferences of liberal democracies. International institutions are often treated as components of the liberal international order (Börzel & Zürn, 2021). Even for questions like why autocrats ratify seemingly counter-intuitive treaties like the Convention Against Torture, the underlying setting has been that international law enshrines liberal values (Vreeland, 2008).

Yet this sits uneasily with contemporary global politics. Authoritarian states are not antagonizing international institutions. It is quite the contrary. A growing number of authoritarian states are creating their own international organizations with codified rules. Not only do they use these organizations as anchor points for cooperation, but also rely on legal principles for legitimating contentious policies (Libman & Obydenkova, 2018; Obydenkova & Libman, 2019).

Because regime type structures domestic political incentives, informational environments, and vulnerability to external scrutiny, it plausibly shapes how governments design and use international institutions. Unlike issue-area or power-based explanations, regime type directly affects how leaders perceive legal constraint and the political costs of delegation.

This observation generates a central puzzle. If international law and organizations are expected to constrain state power, why would authoritarian regimes actively establish them? More broadly, *what happens when international law is written, interpreted, and defended by authoritarian states?* This dissertation comprises three essays, moving from conceptualization of authoritarian formal international cooperation, rhetorical strategies of autocrats' international law, and to empirical testing of whether the rhetoric actually changes how the public think about contentious policies.

In the big picture, this dissertation advances a single proposition that authoritarian states actively structure global governance. They operate within the liberal international order while simultaneously reshaping it. Each chapter stands alone in its theoretical and methodological contribution, yet together they demonstrate that authoritarian engagement with international institutions is not episodic or superficial. The following section first contextualizes the dissertation within the broader intellectual landscape.

2 Authoritarianism and Global Governance

To situate this dissertation within the broader literature, it is necessary to engage foundational debates in international relations about the nature of global governance and the role of international institutions. Much of the classic scholarship on international cooperation emerged from attempts to explain how states overcome collective action problems in an anarchic system (Keohane & Martin, 1995; Stein, 1982). Institutions were theorized as devices that reduce transaction costs, provide information, facilitate monitoring, and lengthen the shadow of the future (Abbott & Snidal, 1998; Kahler, 2000; Koremenos et al., 2001; Martin, 1994). In this tradition, international organizations are functional responses to cooperation dilemmas.

The rise of institutionalist theory coincided historically and intellectually with the consolidation of what later came to be called the liberal international order. Established and emerging liberal democracies were seen as the primary architects and beneficiaries of multilateral institutions (Hyde, 2011; Moravcsik, 2000; Pevehouse, 2002). Multilateralism itself became closely associated with liberalism as a political project, emphasizing rule-based cooperation, legal constraint, and collective decision-making (Ruggie, 1982). International law and organization were not only instruments of cooperation but also components of a liberal international order.

The implicit tie between international institutions and liberalism is prevalent. Studies of legalization emphasize precision, obligation, and delegation as mechanisms that constrain power, ultimately restricting state's autonomy (Abbott et al., 2000). Research on international courts highlights their independence and authority (Alter et al., 2019). Work on human rights treaties examines how legal commitments empower domestic actors and promote accountability (Simmons, 2009). Even critiques of international law often presuppose a liberal baseline, questioning whether states comply with rules that enshrine liberal values (Goldsmith & Posner, 2005). Across the literature, international institutions are typically treated as extensions of a liberal governance model.

Within this dominant paradigm, authoritarian states appear as either reluctant participants or strategic defectors. A substantial body of work asks why autocrats sign treaties that appear to constrain them (Hollyer, 2011; Vreeland, 2008), arguing that participation provides informational or signaling benefits others emphasize domestic audience costs or elite competition. Similarly, studies of membership in international organizations examine whether institutions socialize states toward liberal practices or promote democratic consolidation (Bush, 2011; Pevehouse, 2002; Poast & Urpelainen, 2015). In these accounts, the analysis is limited to domestic political factors within the bracketed scope of democratic IGOs.

What remains underexplored is the possibility that authoritarian states do not merely enter liberal institutions but actively lead and reshape them. Recent scholarship has begun to document the rise of authoritarian-led regional organizations and alternative multilateral forums (Bae & Moon, 2025; Bae & Salvati, 2026; Cottiero & Schneider, 2026; Cottiero et al., 2025; Debre, 2022; Libman & Obydenkova, 2018). This literature challenges the notion that international cooperation is inherently liberal by showing that autocracies also invest in formalized, rule-based structures. My work lends support to this burgeoning field of authoritarian cooperation and contributes to the field by offering tighter theory for international/domestic linkage as well as empirical grounds.

Parallel developments appear in the literature on authoritarian international law. Before anything, the term "authoritarian international law" needs clarification. Rather than treating international law as a liberal baseline which autocracies enter into (or deflect from), I shift attention to legal principles that are designed and shaped under authoritarian leadership. This move reorients the inquiry from participation to active authorship, from how autocrats use liberal law to how they structure legal order itself.

Ginsburg (2020) sets the ground for defining Authoritarian International Law (AIL) as norms and principles advanced by autocratic states and offers empirical grounds to build ex-ante expectation on why democracies and autocracies would codify a distinct set of norms. While Ginsburg (2020) advances the scholarship by treating authoritarian states' role seriously in the shaping of global legal order, I enrich the concept by investigating how states recast old norms. For example, sovereignty and non-interference principles were not introduced by autocrats but they are used by contemporary autocrats to resist external scrutiny. We know comparatively less about how authoritarian governments reinterpret existing foundational principles embedded in the liberal order and how such reinterpretations interact with audience perceptions.

Then the question arises, whether such authoritarian legal rhetoric will make any difference in the eyes of the ordinary citizens. A related debate concerns legitimacy of international law. Classical legal theory attributes legitimacy to procedural fairness, clarity, and coherence (Franck, 1990; Fuller, 1969). Empirical research shows that citizens are more likely to support policies framed as legally mandated (Chilton & Tingley, 2013; Wallace, 2013). This literature often assumes that the authority of international law derives from its institutional features or moral content. However, endorsement environments in these studies are typically embedded in liberal contexts, where the credibility of law is rarely attributed to normatively suspect actors.

Throughout the existing strands of work, what remains insufficiently theorized is how regime type shapes the architecture, meaning, and perceived legitimacy of international institutions themselves. This dissertation intervenes at this juncture. Rather than treating authoritarian states as peripheral actors within a liberal order, this dissertation shows that they participate in shaping its structure, meaning, and authority. Authoritarian regimes cluster within institutional structures, reinterpret foundational legal norms, and invoke legal authority to legitimate contested policies. In doing so, they challenge the assumption that international law and multilateral organization are inherently liberal artifacts.

International law is not a fixed constraint imposed by liberal democracies. This project reconceptualizes international cooperation as a neutral site, an arena open to all regime types. Once we detach the ideas of human rights, democracies, and liberal agendas from the concept of international institution, a myriad of exciting questions emerge. I hope to address at least some of the initial questions in this dissertation.

3 Outline of the Dissertation

Building on the theoretical framework developed in the previous section, the dissertation proceeds in three steps. It first examines how authoritarian regimes shape the organizational architecture of international cooperation, then how they reinterpret foundational legal principles within those institutions, and finally whether such strategies influence public perceptions of legitimacy. Each chapter represents a standalone study with distinct theoretical and methodological contributions, while collectively advancing a unified angle on how authoritarian states structure global governance.

Chapter 1. Authoritarian International Cooperation and the Varieties of Intergovernmental Organizations

Chapter 1 "Varieties of Intergovernmental Organization: Measuring Member-State Distributions" opens the broader theme through conceptualization and operationalization of *what it means* for an Inter-Governmental Organizations (IGOs) to be authoritarian centered. Authoritarian Intergovernmental Organizations (AIGOs), IGOs in which a majority of member states are autocracies, are not merely ad hoc alliances but formalized entities with codified treaties. In parallel, I conceptualize Authoritarian International Law (AIL) as the legal norms and principles advanced by autocratic states. Together, these concepts provide a framework for analyzing how autocracies shape global governance through institutional design and legal discourse.

How can we accurately capture whether an international organization's membership leans more toward democracies or autocracies? Existing approaches often rely on mean-based measures of regime composition, which obscure important variation. To address this, the first essay develops a new methodological framework that measures the full distribution of member states' regime types in lieu of averaging democracy scores. Using a Bayesian beta-distribution model, my approach identifies whether an IGO is skewed toward autocracies or democracies and reveals the extent of heterogeneity in its membership. This methodological innovation forms the basis of a standalone paper and underpins a new dataset.

Substantively, it show how autocracies cluster through intergovernmental organizations and show these allow them to survive longer in incumbency, even in the existence of "black knights" of democratic intergovernmental organizations' counterforce. While doing so, I introduce a novel measurement of member state distribution within IGOs, and compare how this recovers fruitful information that is previously missed from sample averages.

Upon unpacking theories of how benefits of IGOs can be extended to authoritarian states, I re-anchor theoretical discussion with the new measurement of IGOs' regime distribution. It further shows that a country's membership in authoritarian-skewed IGOs can dampen qualities of democracy. The remaining parts of the dissertation rests on Chapter 1 as it shows the significance of authoritarian states' international institutions in domestic political spheres.

Chapter 2. Sovereignty Reframed: Rhetorical Contestation Across Regime Types

Whereas the first essay examines how autocracies embed their preferences into the formal design of international institutions, the second focuses on how they invoke and reinterpret existing norms in discourse. Chapter 2, titled "The Domestic Sources of International Legal Rhetoric: Contesting Sovereignty Across Regimes" shifts attention to the legal architecture, addressing the question of whether authoritarian states can not only introduce "new" norms but also recast the meaning of already-existing norm, the sovereignty principle.

What remains striking is that, when confronted with international scrutiny, autocrats do not reject international law but instead appropriate it as a means of defense. I hypothesize that certain legal principles, most notably sovereignty, serve as the cornerstone of authoritarian rhetoric.

To test this claim, Chapter 2 trace how authoritarian states deploy and reinterpret sovereignty in international fora, particularly in the UN General Debate since 1945. Autocrats often recast sovereignty as a rhetorical shield, foregrounding its state-centered dimensions: non-intervention, territorial integrity, and regime autonomy, contrary to its competing interpretations emphasizing responsibility to protect populations and government accountability towards their constituents. Using computational text analysis, I show that authoritarian states not only invoke sovereignty more frequently than democracies, but also do so in ways that emphasize its restrictive, state-centric interpretation. By measuring the semantic distance between sovereignty and its surrounding narratives - territory, intervene, interfere - this paper demonstrates that autocrats rhetorically invoke sovereignty principle in a way that helps them deflect criticism from actors from liberal democracies.

Methodologically, this essay employs pre-trained word embedding models to analyze patterns of legal rhetoric at scale. Rather than counting the frequency of individual words, I measure the semantic distance between sovereignty and conceptually related terms such as territory, intervention, and interference. Word embeddings represent words as vectors in high-dimensional space, allowing me to capture relational meaning based on contextual usage. This approach moves beyond traditional bag-of-words methods, which are limited in detecting latent semantic associations. By examining how sovereignty is positioned relative to competing concepts across regime types, I identify systematic differences in how authoritarian and democratic states construct its meaning in international discourse.

Theoretically, this essay pushes AIL beyond a catalog of doctrinal inventions. It shows that authoritarian engagement with international law is not only a matter of creating new norms, but also of reinterpreting existing ones to serve regime legitimation. Drawing on the historical evolution of sovereignty as a political construct, I argue that sovereignty is not a fixed legal object but a site of continual contestation, where authoritarians strategically re-anchor its meaning, particularly when their regimes are under threat.

Chapter 3. Impacts of Authoritarian International Law

Chapter 3, entitled "Content or Cue? Conditional Messenger Effects in the Legitimacy of International Law," closes the dissertation by showing authoritarian international law *matters*. Even if autocrats can build institutions and reframe norms, a deeper question remains: Do these strategies actually persuade citizens? The third essay turns to this puzzle. I investigate whether and why citizens find international law less legitimate when it is endorsed by authoritarian states. Using online factorial survey experiments in the Philippines and Australia, I test if the compliance pull associated with international legal endorsement holds even when the messenger is seen as normatively suspect.

Existing scholarship shows that international law increases public support for state actions when framed as legally compliant. Yet it remains unclear what creates that authority in the first place. I ask whether that pull still operates when the law is endorsed by a bloc of authoritarian states. In particular, I focus on refugee policy, where I find that participants update their policy preferences when informed that failure to repatriate violates international law, if that law is backed only by authoritarian governments.

The third essay tests impacts of authoritarian international law through online factorial survey experiments in the Philippines and Australia. Participants increased their support for refugee acceptance when policies were framed as legally required, even when given authoritarian endorsements. The same pattern held in the opposite direction. When international law was said to require repatriation, support for refugee acceptance decreased regardless of the endorsing states' regime types. These results suggest that authoritarian international law has significant persuasive power.

This paper offers several key theoretical contributions. First, it empirically evaluates the significance of authoritarian international law by examining how its invocation affects the liberal international order. The public does not become more supportive of an authoritarian state's contentious foreign policy, even when such actions are framed as consistent with international law. Second, the paper advances the experimental literature of international law by testing whether legal legitimacy is driven more by the content of the message or by the identity of the messenger. Taken together, these three essays advance a unified research agenda on how authoritarian states shape international institutions and norms through both formal design and strategic discourse.

Beyond these theoretical contributions, the research also resonates with the experiences of practitioners on the ground. With support from the University of Washington's Human Rights Grant in 2022, I conducted exploratory interviews with NGO workers engaged with Rohingya refugee issues. These interviews suggest that governments invoke regional charters emphasizing sovereignty and non-interference to resist external criticism of their refugee policies. While anecdotal, it illustrates how authoritarian interpretations of international law can filter into concrete policy arenas, constraining humanitarian responses and shaping the lived experience of displaced populations.

4 Innovations and Contributions

The three essays demonstrate that international law is not the exclusive domain of liberal democracies. By moving from institutional structure to legal discourse to audience reception, the dissertation provides a comprehensive account of how authoritarian states operate within the architecture of international law. I argue that authoritarian states cluster through formal organizations, reinterpret foundational legal principles to protect regime autonomy, and invoke legal authority to legitimize contentious policies. International law thus emerges not as a fixed liberal artifact, but as a site of political contestation.

This dissertation has broader implications for the study of international cooperation and authoritarian studies. First, it expands the understanding of domestic sources of international cooperation by investigating under-explored areas of authoritarian cooperation. While existing work explored autocrats' participation in already-existing treaties and organizations, I look at new pathways pioneered by authoritarian leadership. My new data reveals that there are growing number of such international institutions designed and led by authoritarian state leaders.

Second, I make methodological contribution to the field of international organizations by reviewing conventional measurement of sample-mean based democratic density and expanding the method frontier by introducing a new measurement. Instead of taking a specific solution to the given question in hand, I offer a general solution by estimating the full distribution shape parameters of all IGOs. While the dissertation relied on a subset of dataset with a specific focus on authoritarian- and democratic-skewed IGOs, the new dataset is full of new opportunities that future research can take.

Previously, the theoretical concept of heterogeneity was used interchangeably with concepts like asymmetry, power disparity, and dispersion (Allee & Elsig, 2016; Bae & Moon, 2025), due to lack of detailed measures. My new approach unlocks empirical testings of hypotheses that distinguish distributional shapes.

Third, I show that international law differs between regime types, not just based on new ideas they propose, but also how they re-frame existing principles. This insight bridges international law scholarship with research on authoritarian politics. Rather than asking whether authoritarian states comply with or defect from liberal norms, I demonstrate that they participate in shaping the meaning and reception of those norms. International law emerges not as a static constraint on power, but as a strategic and interpretive arena in which regime types compete to define legitimate governance.

Fourth, this dissertation reveals that the legitimacy of international law depends on both its substantive content and the identity of those who invoke it. More broadly, it contributes to the study of international law by reframing legitimacy as a politically contingent outcome rather than an inherent property of legal rules. Much of the existing literature treats international law as possessing built-in authority derived from procedural fairness, legalization, or moral content. I test whether the same legal content enjoys limited legitimacy when delivered by illiberal messengers. Findings from survey experiments demonstrate that international law can retain persuasive force even when endorsed by authoritarian states. This dissertation therefore advances a relational concept of legitimacy, treating it as contingent upon the relationship between legal claims and their audiences. The same legal rule (identical in content and form) may be perceived as appropriate or binding depending on who invokes it and how it is received. Legitimacy therefore emerges through interaction between institutional authority and public interpretation.

Fifth, it demonstrates that regime type is not merely a background variable in international cooperation. It structures institutional composition, influences the semantic evolution of legal principles, and conditions how legal authority is received by publics. These contributions suggest that understanding contemporary global governance requires integrating institutional design, legal rhetoric, and audience perception within a unified analytical framework.

More broadly, by integrating computational text methods and survey experiments, this dissertation advances empirical ground for understanding authoritarian international cooperation. It opens new avenues for research on how regime type influences the evolution of global governance and invites scholars to reconsider the domestic sources of international institutions.

5 Organization of the Remainder of the Dissertation

The dissertation proceeds sequentially across these three dimensions. Chapter 1 conceptualizes and measures authoritarian-skewed intergovernmental organizations (AIGOs), establishing how regime composition within IGOs can be systematically identified. By introducing a distributional approach to membership composition, the chapter provides the conceptual and empirical foundation for distinguishing democratic- and authoritarian-skewed institutions. Given the distinctive distributional shape of IGO's member states, I argue that existing theories of IGOs might not perfectly map onto behaviors of AIGOs. Therefore, this part justifies why we need a separate framework for understanding *authoritarian international cooperation.*

Chapter 2 examines how they contest and reinterpret legal meaning. I situate this piece within the broader literature of authoritarian international law. Unlike previous work that focuses on original legal principles that treat such principles as allegedly 'illiberal' in and of itself, I suggest that investigation of how autocrats not only write new principles but also reboot existing ones with new framing can help us understand the full picture of authoritarian engagement of international law.

Lastly, I ask whether such authoritarian legal strategies influence public perceptions of legitimacy. This piece is pivotal not only in continuing the discussion by addressing actual ramifications of authoritarian use of international law, but also by posing a puzzling question in and of itself. I bring attention to a unique aspect of authoritarian international law in that we have little empirical ground to predict ex ante whether illiberal messengers can successfully present their messages in the form of law. These essays form a unified research agenda on the political uses of international law beyond liberal democracies.

The concluding chapter synthesizes the findings of the three essays and reflects on their broader implications for the study of international law and global governance. It revisits the central argument that international institutions are not inherently liberal, but are instead shaped by the political characteristics of participating states. By integrating insights from institutional design, legal discourse, and audience reception, the conclusion highlights how authoritarian engagement with international law reshapes both the structure and meaning of global governance. It also discusses the limitations of the current study and outlines avenues for future research on authoritarian international cooperation and the evolving international legal order.

References

Abbott, K. W., Keohane, R. O., Moravcsik, A., Slaughter, A.-M., & Snidal, D. (2000). The Concept of Legalization. *International Organization*, *54*(3), 401–419. https://doi.org/10.1162/002081800551271

Abbott, K. W., & Snidal, D. (1998). Why States Act through Formal International Organizations. *Journal of Conflict Resolution*, *42*(1), 3–32. https://doi.org/10.1177/0022002798042001001

Allee, T., & Elsig, M. (2016). Why do some international institutions contain strong dispute settlement provisions? New evidence from preferential trade agreements. *The Review of International Organizations*, *11*(1), 89–120. https://doi.org/10.1007/s11558-015-9223-y

Allee, T., & Peinhardt, C. (2010). Delegating Differences: Bilateral Investment Treaties and Bargaining Over Dispute Resolution Provisions. *International Studies Quarterly*, *54*(1), 1–26. https://doi.org/10.1111/j.1468-2478.2009.00575.x

Alter, K. J., Hafner-Burton, E. M., & Helfer, L. R. (2019). Theorizing the Judicialization of International Relations. *International Studies Quarterly*, *63*(3), 449–463. https://doi.org/10.1093/isq/sqz019

Bae, J., & Moon, D. (2025). Regime-type composition of an intergovernmental organisation and dispute settlement mechanisms. *The British Journal of Politics and International Relations*, 13691481251329774. https://doi.org/10.1177/13691481251329774

Bae, J., & Salvati, F. (2026). United Nations Discourse on Liberalism across Regime Types. *Global Studies Quarterly*, *6*(1), ksag001. https://doi.org/10.1093/isagsq/ksag001

Barsoom, P. N. (2022). Is the Good News about Compliance Good News about Cooperation?, 29.

Börzel, T. A., & Zürn, M. (2021). Contestations of the Liberal International Order: From Liberal Multilateralism to Postnational Liberalism. *International Organization*, *75*(2), 282–305. https://doi.org/10.1017/S0020818320000570

Bush, S. S. (2011). International Politics and the Spread of Quotas for Women in Legislatures. *International Organization*, *65*(1), 103–137. https://doi.org/10.1017/S0020818310000287

Chayes, A., & Chayes, A. H. (1993). On Compliance. *International Organization*, *47*(2), 175–205. http://www.jstor.org/stable/2706888

Chilton, A., & Tingley, D. (2013). Why the Study of International Law Needs Experiments. *Columbia Journal of Transnational Law*, *52*, 64.

Cottiero, C., Hafner-Burton, E. M., Haggard, S., Prather, L., & Schneider, C. J. (2025). Illiberal regimes and international organizations. *The Review of International Organizations*, *20*(2), 231–259. https://doi.org/10.1007/s11558-024-09556-1

Cottiero, C., & Schneider, C. J. (2026). International Financial Institutions and the Promotion of Autocratic Resilience. *International Organization*, 1–38. https://doi.org/10.1017/S0020818325101276

Debre, M. J. (2022). Clubs of autocrats: Regional organizations and authoritarian survival. *The Review of International Organizations*, *17*(3), 485–511. https://doi.org/10.1007/s11558-021-09428-y

Franck, T. M. (1990). *The Power of Legitimacy among Nations*. Oxford University Press.

Fuller, L. L. (1969). Human Interaction and the Law. *The American Journal of Jurisprudence*, *14*(1), 1–36. https://doi.org/10.1093/ajj/14.1.1

Ginsburg, T. (2020). Authoritarian International Law? *American Journal of International Law*, *114*(2), 221–260. https://doi.org/10.1017/ajil.2020.3

Goldsmith, J. L., & Posner, E. A. (2005). *The limits of international law*. Oxford University Press.

Hollyer, J. R. (2011). Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance. *Quarterly Journal of Political Science*, *6*(3-4), 275–327. https://doi.org/10.1561/100.00010059

Hyde, S. D. (2011). Catch Us If You Can: Election Monitoring and International Norm Diffusion: ELECTION MONITORING AND NORM DIFFUSION. *American Journal of Political Science*, *55*(2), 356–369. https://doi.org/10.1111/j.1540-5907.2011.00508.x

Kahler, M. (2000). Conclusion: The Causes and Consequences of Legalization. *International Organization*, *54*(3), 661–683. https://doi.org/10.1162/002081800551244

Keohane, R. O., & Martin, L. L. (1995). The Promise of Institutionalist Theory. *International Security*, *20*(1), 39. https://doi.org/10.2307/2539214

Koremenos, B., Lipson, C., & Snidal, D. (2001). The Rational Design of International Institutions. *International Organization*, *55*(4,), 761–799. http://www.jstor.org/stable/3078615

Libman, A., & Obydenkova, A. V. (2018). Understanding Authoritarian Regionalism. *Journal of Democracy*, *29*(4), 151–165. https://doi.org/10.1353/jod.2018.0070

Lyall, J. (2010). Are Coethnics More Effective Counterinsurgents? Evidence from the Second Chechen War. *American Political Science Review*, *104*(1), 1–20. https://doi.org/10.1017/S0003055409990323

Martin, L. (1994). Heterogeneity, Linkage and Commons Problems. *Journal of Theoretical Politics*, *6*(4), 473–493. https://doi.org/10.1177/0951692894006004004

Moravcsik, A. (2000). The Origins of Human Rights Regimes: Democratic Delegation in Postwar Europe. *International Organization*, *54*(2), 217–252.

Obydenkova, A. V., & Libman, A. (2019). *Authoritarian Regionalism in the World of International Organizations: Global Perspective and the Eurasian Enigma* (1st ed.). Oxford University Press.

Pevehouse, J. C. (2002). Democracy from the Outside-In? International Organizations and Democratization. *International Organization*, *56*(3), 515–549. https://doi.org/10.1162/002081802760199872

Poast, P., & Urpelainen, J. (2015). How International Organizations Support Democratization: Preventing Authoritarian Reversals or Promoting consolidation? *World Politics*, *67*(1), 72–113. https://doi.org/10.1017/S0043887114000343

Ruggie, J. G. (1982). International regimes, transactions, and change: Embedded liberalism in the postwar economic order. *International Organization*, *36*(2), 379–415. https://doi.org/10.1017/S0020818300018993

Simmons, B. A. (2009, October). *Mobilizing for Human Rights: International Law in Domestic Politics* [Google-Books-ID: qmDxDMruTMMC]. Cambridge University Press.

Simmons, B. A., & Danner, A. (2010). Credible Commitments and the International Criminal Court. *International Organization*, *64*(2), 225–256. https://doi.org/10.1017/S0020818310000044

Stein, A. A. (1982). Coordination and Collaboration: Regimes in an Anarchic World. *International Organization*, *36*(2,), 299–324. http://www.jstor.org/stable/2706524

Tir, J., & Stinnett, D. M. (2011). The Institutional Design of Riparian Treaties: The Role of River Issues. *Journal of Conflict Resolution*, *55*(4), 606–631. https://doi.org/10.1177/0022002710393917

Vreeland, J. R. (2008). Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture. *International Organization*, *62*(01). https://doi.org/10.1017/S002081830808003X

Wallace, G. P. (2013). International Law and Public Attitudes Toward Torture: An Experimental Study. *International Organization*, *67*(1), 105–140. http://www.jstor.org/stable/43282154

Varieties of Intergovernmental Organization: Measuring Member-State Distributions

Jihyeon Bae

Curtis Atkisson*

Abstract

How do we characterize the political composition of intergovernmental organizations (IGOs)? Existing work relies on sample means of member-state democracy scores and deterministic thresholds to distinguish "autocratic" from "democratic" IGOs. We show that this practice is conceptually incomplete and statistically fragile. First, sample means omit information about the shape of the underlying membership distribution, conflating unimodal, skewed, and polarized organizations. Second, mean-based summaries are sensitive to sampling variability especially for IGOs with small number of member states. Third, deterministic statements based on such means encourage overconfident categorical claims that obscure genuine uncertainty in organizational type. We propose an alternative that models member-state democracy scores as Beta-distributed and estimates the corresponding shape parameters in a Bayesian framework. For each IGO–year, we obtain posterior distributions of the Beta shape parameters, from which we derive measures of mean location, precision, and probabilistic assignment (unimodal, U-shaped, democracy-skewed, autocracy-skewed). Using simulations and empirical illustrations, we demonstrate that beta means are more stable than arithmetic sample means and that probabilistic shape assignments avoid the perils of overconfident cutoffs. Finally, we show how these measures can be used to study whether exposure to autocracy-skewed versus democracy-skewed IGOs is associated with subsequent changes in state democracy scores.

*eScience Institute, University of Washington

1 Introduction

How should we characterize the political composition of intergovernmental organizations (IGOs)? A large literature in international relations treats the regime types of member states as central to explaining variation in institutional design, authority, and cooperation outcomes (Jo & Namgung, 2012; Koremenos et al., 2001; Mansfield & Pevehouse, 2008). Whether IGOs are dominated by democracies or autocracies is thought to shape delegation, pooling, amendment rules, and the scope of institutional constraints (Allee & Peinhardt, 2010; Bae & Moon, 2025; Hyde & Saunders, 2020; Tallberg & Vikberg, 2025). Despite the theoretical importance of membership composition, empirical work has relied on relatively blunt summaries to operationalize what an IGO's "regime dominance" is.

International organizations, often associated with democratic leadership, have been understood to undergird the liberal international order system that reflects broadly shared norms (Keohane & Martin, 1995; Krasner, 2001; Lake et al., 2021). However, as the institutional ecosystem of global governance splinters into partly competing institutional orders (Bartels et al., 2023; Börzel & Zürn, 2021; Leeds et al., 2025; Stephen, 2025), there is growing discussion on accumulated backlash against the liberal international order. In this context, unpacking *why* and *how* institutions are dominated by which types of states, and how to measure that composition accurately, becomes increasingly consequential. Yet existing measures of IGO composition are not sharp enough to track these shifts with the precision this debate demands.

In this paper, we argue that existing approaches suffer from what we term *compositional collapse*, the reduction of a full membership distribution to a single scalar summary. [1] When structural information is absorbed into a point estimate, the sample mean retains (although we later show that even this comes with limitation) only the center of gravity of an IGO's membership while discarding the shape, dispersion, and polarization that theory invokes but measurement has not

[1] Whether that summary is a sample mean, a percentage of democratic members, or a standard deviation, the underlying distributional structure, we argue, is discarded. We focus primarily on the sample mean as our point of comparison because it is the most widely used measure in the literature and because it provides the clearest baseline against which to demonstrate the costs of compositional collapse. The logic, however, applies to any scalar reduction of a membership distribution.

captured.

This collapse is not merely a technical inconvenience. We aim at persuading the readers that how regime composition is measured has substantive consequences for inference about international institutions. IGOs are political coalitions, not random samples of states, and their memberships often exhibit skewness, clustering, or polarization across regime types.

We propose a distributional framework for characterizing IGO membership composition that moves beyond stand-alone point estimates. Rather than collapsing member-state democracy scores into a single summary statistic, we model the full distribution of scores within each IGO–year. Using a Bayesian Beta framework, we estimate shape parameters that jointly govern central tendency, dispersion, and distributional form. We also introduce theoretical Beta mean, based on shape parameters, instead of an arithmetic mean. [2] While the Beta mean remains a useful summary, it is no longer interpreted in isolation. It is anchored to a specific distributional shape and accompanied by posterior uncertainty. This approach yields interpretable quantities, including concentration, skewness, and probabilistic distributional types, that align more closely with theoretical claims about heterogeneity and coalition structure.

This paper makes three contributions to the study of international institutions. The first is methodological. Existing work has made significant progress in characterizing IGO membership composition through arithmetic means of member-state democracy scores, generating important insights about how organizational character shapes institutional design and cooperation outcomes (Cottiero & Haggard, 2023; Debre, 2022; Pevehouse, 2002a). We build on this foundation by asking what additional information lies beyond the mean. Following recent calls for more precise measurement of IGO heterogeneity (Debre & Sommerer, 2025), we model member-state democracy scores as draws from a Beta distribution and estimate the corresponding shape parameters within a Bayesian framework. This yields posterior distributions over the full shape of each IGO's membership, not merely its central tendency, and supports probabilistic rather than deterministic classification. The resulting measures complement existing approaches while offering greater

[2] The Beta mean, $\frac{\alpha}{\alpha+\beta}$, is derived directly from the shape parameters α and β, rather than computed as the arithmetic average of observed values.

stability across small samples, transparency about uncertainty through credible intervals, and sensitivity to the distributional features of membership that theory often invokes but measurement has not yet captured.

The second contribution is empirical. A large and consequential literature examines how international institutions shape domestic political outcomes. On the democratizing side, scholarship has shown that membership in democracy-dense IGOs raises the costs of anti-democratic behavior, socializes norms of electoral competition, and locks in governance commitments that constrain future incumbents (Moravcsik, 2000; Pevehouse, 2002a, 2005; Poast & Urpelainen, 2015). On the authoritarian side, research has demonstrated that participation in autocracy-dense institutions stabilizes incumbent survival, facilitates resource transfers among regime elites, and diffuses norms of non-interference that insulate leaders from external democratic pressure (Bush et al., 2025; Cottiero & Schneider, 2026; Debre, 2022; Libman & Obydenkova, 2018; Obydenkova & Libman, 2019).

Yet these two strands of literature have developed largely in isolation, treating institutional environments as if states experience only one type at a time. We show that for most states, democratic and authoritarian institutional environments operate simultaneously, and that modeling portfolio exposure probabilistically, weighting each organization by the posterior probability of its regime composition, recovers systematic relationships that mean-based and deterministic approaches obscure. In doing so, we offer a more complete empirical accounting of how the shifting political composition of global governance translates into domestic political consequences.

The third contribution is a dataset. We release IGO–year estimates of posterior means and credible intervals for the Beta shape parameters (α, β), the Beta mean and precision, and posterior probabilities for four distributional regime types: unimodal, U-shaped, democracy-skewed, and autocracy-skewed. These data are designed to integrate with existing empirical strategies while enabling more theoretically precise operationalization of heterogeneity, coalition structure, and regime dominance within international organizations. Although we illustrate the framework using democracy scores, it applies equally to any member-state attribute scaled to the unit interval, including trade openness, human rights performance, or economic development.

The remainder of the paper proceeds as follows. We first situate our approach within existing theories and measurement practices in the literature on IGOs. We then formalize the limitations of sample-mean-based and deterministic approaches in a dedicated problem statement. Next, we introduce the Beta-based Bayesian framework and describe the estimation strategy. Finally, we illustrate the analytical advantages of the distributional approach by applying our measures to a substantive question in the IGO literature.

2 IGO Membership Composition and Institutional Design

The composition of IGOs' member states has occupied a special place in the theory of international institutions. A core insight of the rational design literature is that variation in institutional design reflects underlying disparities among members, including differences in power, preferences, and domestic political constraints (Abbott & Snidal, 2000; Koremenos et al., 2001). States delegate authority to IGOs in order to manage cooperation problems, such as monitoring, enforcement, and information provision, but the form and extent of that delegation depend critically on who the members are and how they perceive particular cooperation problems as especially salient (Abbott & Snidal, 1998). A. Thompson (2006) and Voeten (2005) further contend that endorsement from a diverse membership sends credible information to international observers that a decision is politically unbiased, thereby generating a legitimation effect.

Member heterogeneity mediates these mechanisms by shaping bargaining leverage, coalition formation, and the distribution of gains from cooperation. Recent work provides microfoundations for these dynamics. Knapp (2025) shows that states sharing similarities along regime type, wealth, region, and affiliation tend to form coalitions within IGOs. Martin (1994) conceptualizes heterogeneity as the distribution of interest intensities and capacities among states, arguing that diversity can facilitate cooperation through issue linkage. By contrast, other accounts emphasize that asymmetric interests and capacities generate distributive conflict, complicating institutional design when the benefits and burdens of cooperation are unevenly shared (Libecap, 2025).

Mansfield and Pevehouse (2008) look at the early stage of IGO's life-cycle and show that democratizing states are more likely to *join* IGOs with economic goals, rather than those embedded in political issue areas. Membership composition further shapes institutional trajectory after the design stage, influencing both voting behavior and access rules within organizations. For example, Meyerrose and Nooruddin (2025) show that backsliding states (from democracies) are more likely to vote against the United Nations Human Rights Council's Resolution, aligning closely with autocracies. Tallberg et al. (2016) shows that IGOs with more democratic members are also more likely to be open to non-state actors.

Given the significance of membership composition in the understanding of IGOs, empirical work has focused on two broad tasks when operationalizing regime composition. One emphasizes the central tendency of members, often focusing on the characteristics of the modal or majority state, and the other focuses on asymmetry, highlighting differences among members and the implications of uneven distributions of power or regime type. A substantial body of research has sought to measure the average level, using sample mean, of democracy among member states for the main analysis or for robustness check (Cottiero & Haggard, 2023; Debre, 2022, 2025; Hafner-Burton et al., 2025; Pevehouse, 2002a). This strand of literature is particularly important for highlighting interactions between authoritarian states' leadership and their unique preferences and ensuing institutional designs. In essence, classifying IGOs based on the theoretical lens of whether they are *predominantly* composed of authoritarian states as members is at the heart of this group of research.

Beyond measures of central tendency, a growing body of scholarship has sought to operationalize distributional features more directly. Theories of heterogeneity have often used asymmetry and imbalance almost interchangeably, without accompanying measurement of the underlying distribution (Abbott & Snidal, 1998). In practice, operationalization has taken several forms. Allee and Peinhardt (2010) capture power asymmetry within preferential trade agreements using a dummy variable indicating whether the agreement involves at least one country from the global North and one from the global South — a binary reduction of a continuous and multidimensional concept.

Regime-type heterogeneity is often measured using the standard deviation of member-state democracy scores (Bae & Moon, 2025; Tallberg & Vikberg, 2025), which summarizes dispersion but is uninformative about skewness or polarization. Smith (2000) adopts the Herfindahl-Hirschman index to capture concentration among members, borrowing a measure designed for market share data that assumes no meaningful ordering among units. Taken together, these approaches reflect genuine need across the field in moving beyond the mean, but they stop short of modeling the distributional shape that theories of coalition formation, regime dominance, and institutional design ultimately require and benefit from.

Existing work supplements the sample mean with alternative measures such as the median democracy score or the proportion of democratic members. While these indicators mitigate some concerns about outliers, they introduce limitations of their own. The median discards information about the shape of the underlying distribution, treating sharply polarized organizations the same as those clustered around the middle of the regime spectrum. Proportion-based measures, meanwhile, discard information contained in continuous democracy scores by imposing binary classifications at the country level in order to characterize an IGO as, for instance, 60% democratic. However, both approaches still fail to capture the full distributional dynamics that shape formation, contestation, and institutional design within IGOs. By relying on thresholded or location-based summaries, existing scholarship risks assuming that dominance follows directly from simple majorities, thereby obscuring the coalition dynamics most relevant to institutional design.

While the sample mean and standard deviation of members' democracy scores provide intuitive summaries of an IGO's democratic character, they pose important limitations. For instance, an organization evenly split between autocracies and democracies yields the same mean score as one composed entirely of hybrid regimes, even though the distributional dynamics, and thus the potential for coalition building, contestation, and institutional design, may be markedly different. Similarly, the presence of a few outlier democracies in an otherwise authoritarian IGO can disproportionately shift the mean upward, masking its fundamentally autocratic character. These examples reflect intuitive questions that scholars of international institutions routinely confront. In the problem statement that follows, we examine these limitations in a systematic and data-driven manner.

Against this background, we introduce a generalizable framework for characterizing the attributes of IGO member states, focusing on both their central tendency and their dispersion across key features. Although we illustrate it with democracy scores, the method applies equally to other features of interest, enabling richer and more precise comparisons across IGOs. As a working example, we estimate the parameters of the Beta distribution (α and β) and derive the beta mean and variance. These estimates provide a compact and flexible summary of the distributional properties of member-state characteristics within an IGO.

To be clear, earlier empirical work based on sample point estimates remains valuable. The sample mean is interpretable, facilitates comparisons across studies, and has generated important insights about IGO composition and design. Our aim is not to displace these contributions but to extend them by modeling the full distribution of member-state characteristics in ways that provide a tighter link between theoretical claims about heterogeneity and their empirical measurement. We hope the distributional information recovered from our framework allows researchers to revisit well-established findings, and ask whether those effects hold uniformly across organizational types or concentrate among IGOs with particular distributional profiles.

3 Problem Statement

In the big picture, we make three claims. First, the sample mean can be an incomplete and sometimes misleading measure of regime-type dominance within IGOs. Second, sample mean is vulnerable to sampling variability. Third, deterministic approach can lead to overconfidence in the estimated value. We provide data-driven evidence showing why and to what extent these conventional approaches misstate organizational composition.

What, then, is the appropriate approach? Before turning to the three empirical problems, we begin by examining the underlying distribution of democracy scores within IGOs. If the mean is to serve as a sufficient summary of IGO composition, the underlying distribution would need to be approximately symmetric *and* unimodal. We first check whether our data meets such condition.

Our main variable of interest, the democracy score, is measured using V-Dem's *v2x_polyarchy* index, which is defined on the open interval $(0, 1)$ (Coppedge et al., 2025). In practice, scores can approach 0 or 1 but never take these boundary values exactly. We conduct normality checks for each IGO across member polyarchy scores at the organization's inception year. The results indicate that member-state democracy scores within IGOs frequently depart from normality. Using the Shapiro–Wilk test, approximately 41% of IGOs reject the null hypothesis of normality.

These features of the data suggest that member-state democracy scores within IGOs are more naturally understood as draws from a bounded and potentially asymmetric distribution rather than from a normal distribution. Democracy scores are constrained to the unit interval, often exhibit skewness toward one pole of the regime spectrum, and may cluster near the boundaries in organizations dominated by either autocracies or democracies. The Beta distribution is well suited to capturing these properties. Defined on the open interval $(0, 1)$, it accommodates a wide range of distributional shapes, including unimodal, skewed, and polarized configurations, through its two shape parameters.

Moreover, compared to other candidate families of bounded distributions, the Beta distribution is particularly attractive because its parameters directly control both location and shape, allowing skewness and polarization to be represented parsimoniously with interpretable parameters. Treating IGO membership composition as Beta-distributed therefore provides a flexible and theoretically coherent framework for modeling regime composition while respecting the empirical properties of the data. We therefore quantify how sample mean deviates from quantities generated by a Beta distribution, in order to assess the magnitude of the problems introduced by mean-based summaries.

Claim 1. Omission and Misrepresentation from Sample Mean

Substantively, using the arithmetic sample mean for data that are distributed as a Beta distribution leads to omitting information. Two IGOs may share the same average democracy score even if one is composed of states clustered around the middle of the spectrum and the other is sharply polarized between autocracies and democracies. For example, both the World Intellectual Property Organization (WIPO) and the Organization of American States (OAS) have mean democracy scores near 0.5, yet the latter is composed of a mixture of autocracies and democracies that produce a U-shaped distribution. The sample mean alone cannot distinguish these cases.

Instead, these distributions are well characterized by the Beta family. Modeling IGO membership with a Beta specification also provides a principled basis for classification. Because the two shape parameters jointly govern the center and form of the distribution, they naturally generate four quadrants: unimodal, U-shaped (polarized), democracy-skewed, and autocracy-skewed.

Conditional on its two shape parameters, α and β, the Beta distribution spans a wide range of functional forms that map directly onto substantively meaningful types of IGO membership composition (see Figure 1). When both shape parameters exceed 1, the distribution is unimodal, indicating a relatively homogeneous membership. When both parameters are below 1, the distribution is U-shaped, corresponding to a polarized organization composed of both autocracies and democracies. When $\alpha > 1$ and $\beta < 1$, the distribution is skewed toward democracies, whereas $\alpha < 1$ and $\beta > 1$ produces the opposite pattern.

One prominent example of U-shaped IGOs would be the Southeast Asia Treaty Organization (SEATO). The SEATO was a Cold War-era mutual defense organization formed in 1954 to prevent the spread of communism in Southeast Asia. Created by the United States, France, Great Britain, Australia, New Zealand, the Philippines, Thailand, and Pakistan, it dissolved in 1977 after failing to gain widespread regional acceptance. It is one of the many examples of U–shaped IGOs (in terms of regime types) during the Cold War. For unimodal IGOs, one example is the Asia-Pacific Economic Cooperation (APEC). For Authoritarian-skewed IGOS, there is the Gulf Cooperation Council (GCC), and for democratic-skewed IGOs, there is Benelux (Luxembourg, Netherlands, Belgium), where certain regime types form a clear dominant mode. Figure 1 illustrates the four distributional types Beta framework (unimodal, U-shaped, right-skewed, and left-skewed), demonstrating how organizations with similar mean democracy scores can differ substantially in their underlying distributional shape.

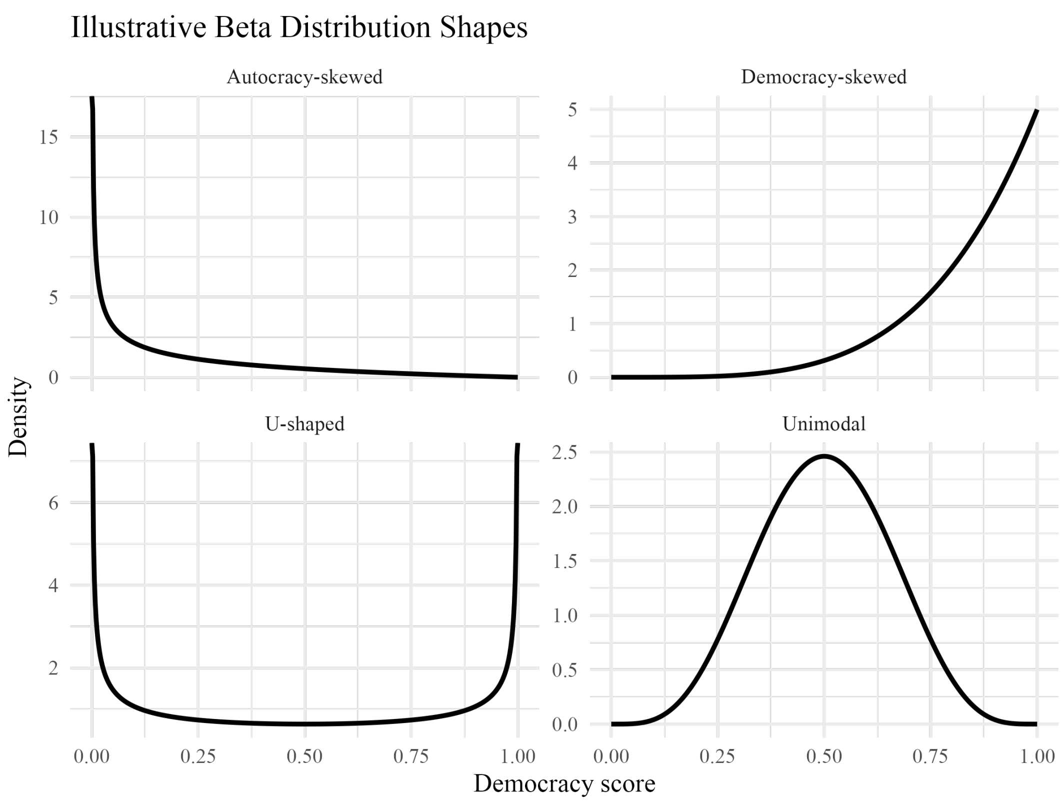


Figure 1: Illustrative shapes of the Beta distribution.

To further quantify how the sample mean can misrepresent the underlying distribution, we specify a Beta-distributed data-generation process and repeatedly draw samples of size $N = 1,000$ from it. We then compare the arithmetic mean of each simulated sample (i.e., the sum divided by N) to the theoretical mean of the generating distribution. The main text presents the intuition and setup of this comparison, while Appendix A reports the full simulation design, estimation procedure, and implementation details.

For example, assume we focus on cases where both α and β are less than 1, which produce U-shaped distributions. In these settings, the underlying distribution places most probability mass near 0 and 1, with relatively few observations near the center. As a result, the sample (arithmetic) mean becomes highly unstable. Depending on whether a small sample happens to draw more observations near one extreme or the other, the mean can deviate substantially from the theoretical mean.

This instability is evident in simulation results. Across 1,000 simulated samples drawn from the same U-shaped Beta distribution, sample means vary widely around the theoretical mean, yielding a total sum of squared deviations of 3.51 (MSE = 0.00351), substantially larger than in unimodal or symmetric distributions. These discrepancies are especially pronounced when IGO membership is small, where a handful of extreme draws can dramatically shift the mean.

Our first claim is that the sample mean is prone to drifting away from the true theoretical mean as a function of distributional shape and sample size. These failures are present even in skewed and unimodal cases, but are most pronounced in IGOs with small numbers of member states.

3.1 Claim 2. Sampling Variability

A second problem is that sample statistics fail to account for sampling variability. By sampling variability, we refer to the inherent randomness in how democracy scores are measured and observed. This uncertainty has at least two layers. At the micro level, a country's electoral democracy score ($v2xpolyarchy$ variable) can be affected by idiosyncratic coding judgments, such as how observers interpret the data on any given day. Although the V-Dem electoral democracy index achieves high inter-coder reliability, it cannot eliminate the role of such random noises. At the macro level, the democracy score observed in a single year is only one realization of a broader, underlying political trajectory. A protest or electoral crisis that shaped a country's score in a given year might easily not have occurred under marginally different circumstances. The political events we observe, in these regards, are just one manifestation of alternative results we could have observed.

To illustrate an unobserved counterfactual scenario in which one thing went differently, we conduct a simulation in the following steps.[3] For each IGO-year, we construct a pool of candidate polyarchy scores for each member state by drawing from a ±2-year window around the observed year. This windowed pool represents plausible alternative realizations of member states' democracy levels and provides a practical approximation of sampling variability. From this pool, we repeatedly sampled alternative realizations of member scores (where year t's democracy could have been the one from near future from ±2) and recomputed the IGO's *means* (sample mean and theoretical beta-mean) democracy level in each trial. This procedure shows how easily the sample mean can shift depending on which year or coding noise is realized.

Figure 2 compares the standard deviation of sample means for each IGO across 100 simulations with the standard deviation of simulated beta means. In other words, it compares the variability of simulated sample means with that of Beta means across IGOs of different sizes in year 2010. The y-axis compares the spread of sample mean and beta mean. Positive values indicate the sample mean is more volatile than the Beta mean. Negative values, on the other hand, mean that the Beta mean is more vulnerable to sampling variability. The plot shows there are more points above 0, meaning that beta means are generally more robust (less volatile) from sampling variability compared to sample means. The Beta mean's stability advantage becomes consistent for larger organizations. For IGOs with fewer than 30 member states, the two estimators perform similarly, with neither showing a clear advantage. Beyond 30 members, however, the Beta mean is more

[3]For the simulation exercises, we rely on maximum likelihood estimation (MLE) of the Beta parameters rather than the Bayesian framework used in the main analysis. Fits are computed using `fitdist()` from the `fitdistrplus` package (Delignette-Muller & Dutang, 2015). The purpose of this section is to compare the sensitivity of sample means and beta means under repeated resampling. MLE provides a fast and consistent way to obtain point estimates across thousands of simulated draws, which makes large-scale simulation computationally feasible. By contrast, the Bayesian procedure, while central to our substantive analysis, would be prohibitively costly at this scale.

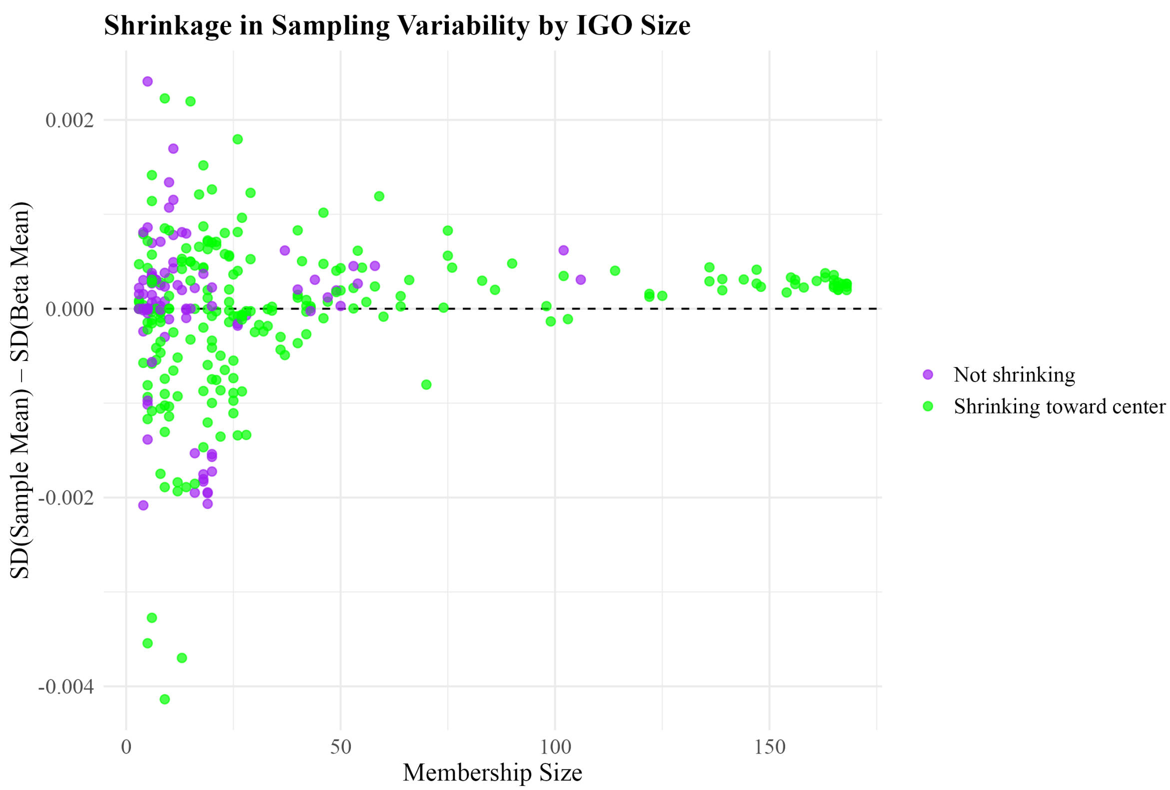


Figure 2: Comparison of simulated sample means and beta means across IGOs in 2010.

stable in the majority of cases, and for IGOs with 60 or more members, 94% of cases show lower volatility in the Beta mean than the sample mean. [4]

Two bonus secondary implications follow from modeling sampling variability explicitly, shrinkage toward the center of the democracy scale and increased precision. First, we find that Beta-mean offers a more conservative estimate of democracy score as it shrinks to the center of the distribution, 0.5, as its sample size increases. Points are colored by whether the beta estimate pulls values closer to the center (0.5) of the democracy spectrum. [5] Points in Fig. 2 are colored green when the distance from the beta mean to 0.5 is smaller than the distance from the sample mean to 0.5; in other words, when the beta mean is closer to the center (0.5). Since, in our data, the beta mean is closer to 0.5 than the sample mean in many cases, the remaining difference in distances is small. Formally, let $d_s = |m_s - 0.5|$ denote the distance of the sample mean m_s from 0.5, and

[4] See Appendix C2 for group-size contingent comparison.

[5] Note that while the y-axis is the difference in variability between sample mean and beta mean, shrinkage is a directional check on the mean.

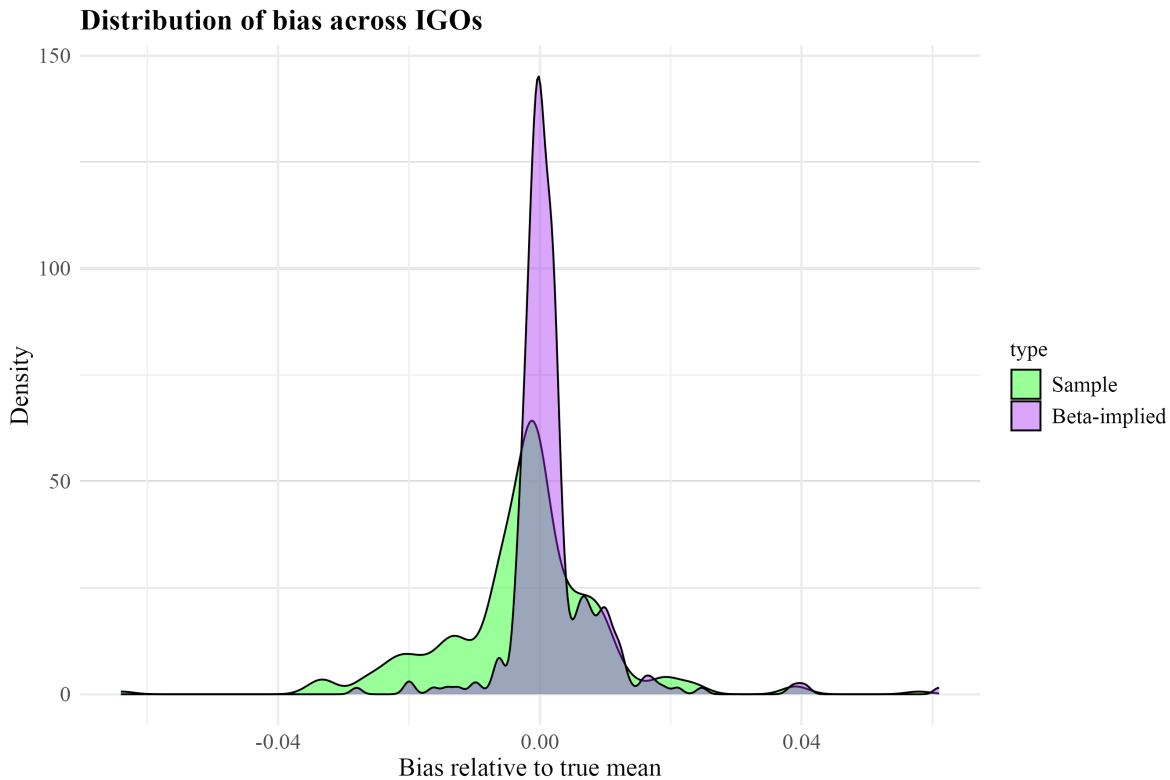


Figure 3: Comparison of simulated sample means and beta means across IGOs in 2010.

let $d_b = |m_b - 0.5|$ denote the distance of the beta mean m_b from 0.5. In a large share of cases, $d_b < d_s$. A practical implication is that regressions using the beta mean as a predictor will often produce smaller coefficient magnitudes than those using the sample mean, serving as a conservative measure.

Second, the Beta mean performs better in terms of precision around the observed mean. Figure 3 plots the empirical distribution of bias in 2010 for two estimators of an IGO's average democracy level: the arithmetic sample mean and the Beta mean obtained by fitting a Beta distribution to resampled membership scores. The horizontal axis reports bias relative to the observed 2010 mean, while the vertical axis shows the corresponding density across simulated draws.

Both estimators are centered near 0, indicating little systematic bias. However, the distribution of bias for the Beta mean is markedly more concentrated, with a sharper peak and thinner tails than that of the sample mean. This tighter concentration indicates lower variability and smaller absolute deviations from the observed mean across resamples. Substantively, this pattern reflects shrinkage toward the center. When membership is modest or the underlying distribution is skewed or bimodal, the Beta mean pulls extreme realizations back toward 0.5.

3.2 Claim 3. Perils of Deterministic Statements

We zoom out to make a broader claim. A third problem arises not from the sample mean itself, but from the practice of converting any noisy point estimate into a deterministic claim. [6] Even small variability can push an observation just above or below a chosen threshold, producing a category that appears definitive but is in fact highly unstable. Because membership composition changes incrementally over time, deterministic cutoffs can produce abrupt "category jumps" in time-series data that do not reflect substantive organizational change.

The result is that organizations with nearly identical member compositions may be placed into different bins, while the same organization could fall into different categories across repeated samples. For example, IEXB (bureau of international expositions that oversees world's fairs) in 1972 is one outlier in its IGO history. Its estimated α parameter was only 0.09 short of exceeding the threshold of 1. A deterministic rule would rigidly classify the organization into a different regime type. Such deterministic statements therefore not only exaggerate confidence but also mask the continuous nature of the underlying distribution.

On a theoretical level, there is over-confidence in putting an IGO into a category with a deterministic approach. It has a hidden assumption that measurement uncertainty is negligible and that latent political attributes have clear-cut boundaries, when in fact both are continuous. As a result, deterministic labeling can amplify apparent discontinuities between IGOs and produce spurious inferences about institutional divergence, when in reality such differences may stem from sampling variability rather than substantive change. A probabilistic framework, in contrast, retains this uncertainty and represents the continuous nature of institutional composition with more

[6] This problem also applies to a common robustness-check measure of IGO authoritarian proportion, which classifies IGOs based on the share of member states whose democracy scores fall below a 0.5 cutoff. Although widely used, this percentage-based indicator relies on a hard threshold on a continuous index and therefore inherits the same instability and overconfidence problems described in the text.

transparency.

4 Bayesian Approach to Distributional Estimates

Here, we introduce our Bayesian measurement strategy for the Beta shape parameters and, equally importantly, show why this framework can mitigate each of the three problems identified above. Our goal is not only to describe the method, but also to demonstrate why it provides a more coherent and transparent representation of IGO membership distributions than conventional approaches.

Recall that the two Beta shape parameters, α and β, jointly determine both the location and shape of an IGO's regime-type distribution. The Beta mean, $\hat{\mu} = \frac{\alpha}{\alpha+\beta}$, captures where the distribution is centered on the democracy–autocracy spectrum, while the relative magnitudes of α and β govern the overall shape, whether the distribution is unimodal, U-shaped, or skewed toward one pole.

Now with the goal of estimating two parameters, we adopt Bayesian approach. This assumes that the data we observe is drawn from a population, which itself is one random manifestation among a larger space of possibilities. Hence, parameter estimates generated from maximum likelihood estimation (MLE) represent a single point estimate, whereas Bayesian inference treats parameters themselves as probability distributions. Instead of estimating one fixed set of parameter values, we draw from the posterior distribution. Posterior draws in this context refer to simulated samples from the joint probability distribution of the two Beta shape parameters, conditional on the observed democracy scores of IGO member states. Each draw represents a plausible combination of α and β given the data, and the collection of draws characterizes our full uncertainty about the true distributional shape.

Against this background, we employ Bayesian estimation of the Beta shape parameters for each IGO–year,[7] which yields full posterior distributions for α and β. This feature directly addresses the first problem. The Bayesian framework models the entire distribution of democracy

[7] Bayesian models are estimated using the `brms` package (Bürkner, 2017), with post-processing via the `posterior` package (Bürkner et al., 2025).

scores and hence preserves the shape information that the sample mean systematically discards. Because the theoretical mean and precision are derived from the full posteriors, our estimates incorporate uncertainty about the underlying membership distribution rather than collapsing it into a single statistic. For instance, based on the posterior draws, we can state that an organization has an 80% probability of being classified as an "authoritarian-skewed IGO" (AIGO) under our definition with a mean of X and a 95% credible interval Y-Z. [8] That is, out of all posterior samples $s = 1, \dots, S$, we count the number of draws satisfying $(\alpha_s < 1, \beta_s > 1)$ and divide by S. This probabilistic assessment reflects the inherent uncertainty of small or heterogeneous memberships while systematically leveraging distributional information to improve classification.

We rely on weakly informative priors for both parameters, a Student-t distribution with 3 degrees of freedom, location 0, and scale 2.5, placed on the log-odds of μ and the log of ϕ respectively.[9] These priors are weakly informative in the sense that they do not strongly favor any particular shape of the membership distribution, but gently regularize estimates away from degenerate boundary values, which would be theoretically implausible. On the log-odds scale, a scale of 2.5 is diffuse enough to place substantial prior mass across the full $[0, 1]$ range of μ, while the heavy tails of the Student-t allow for occasional extreme values without unduly penalizing them. These priors are centered at substantive neutrality, yet wide enough to let the data dominate in typical IGO-years.[10]

Importantly, researchers are not limited to taking the modal assignment at face value. Because the posterior percentages are retained, one can apply alternative thresholds for confidence in classification. For example, if no category exceeds 0.75 posterior probability, the case may be left unclassified, flagged as uncertain, or analyzed separately. This flexibility allows for more cautious and transparent use of categorical labels compared to deterministic approaches, while still making full use of the underlying Bayesian information.

The Bayesian approach also mitigates the second problem, sampling variability. Rather than

[8] The Gulf Cooperation Council (GCC) in 1986 has a posterior mean of 0.115 with a 95% credible interval of [0.0893, 0.153] and a 0.998 posterior probability of being uni-modal, reflecting not merely that its members are autocratic on average, but that they are *homogeneously and certainly* so.

[9] This is default setting from the *brms* package.

[10] On the log-odds scale, location 0 corresponds to $\mu = 0.5$, implying equal probability mass on either side of the democracy–autocracy spectrum. A scale of 2.5 on this scale covers roughly $\mu \in (0.07, 0.93)$.

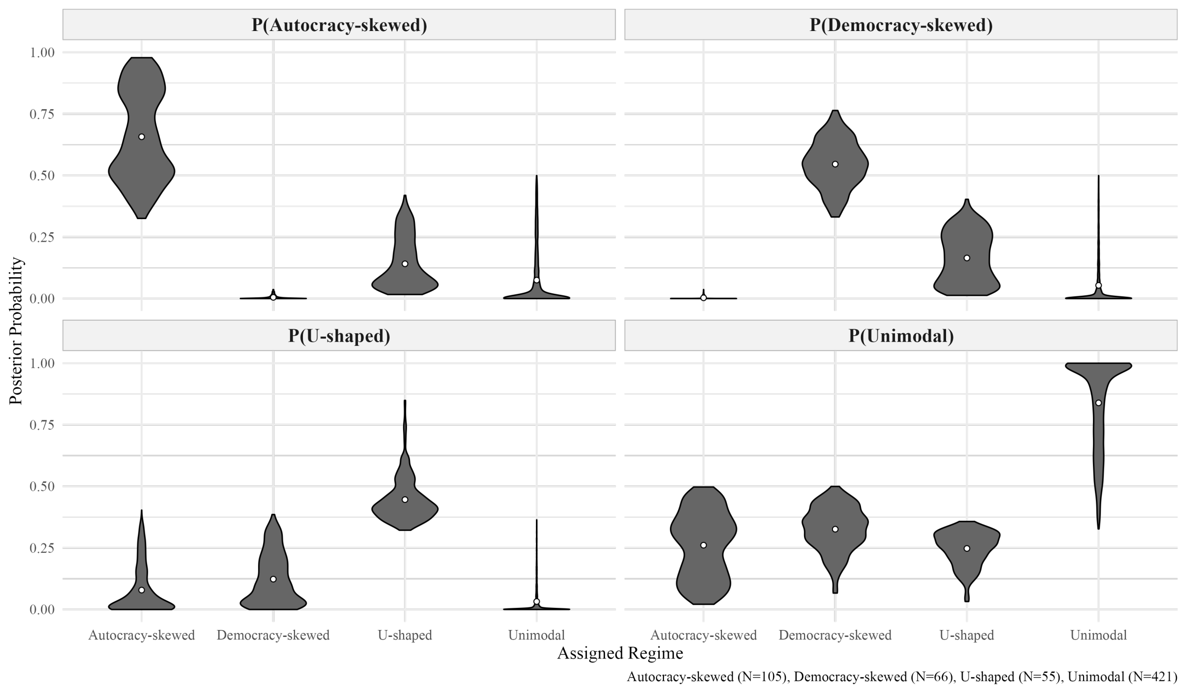


Figure 4: Bayesian probabilistic regime assignment.

treating each country's observed democracy score as fixed, the posterior distribution incorporates uncertainty arising from small memberships, skewed shapes, or noisy realizations of political events. This induces a natural shrinkage toward plausible central values when information is limited. As a result, the posterior-based Beta mean is substantially less sensitive to random errors than the raw sample mean from its MLE counterpart.

Finally, Bayesian estimation addresses the third problem by eliminating the need for deterministic cutoffs. Instead of assigning each IGO to a single regime-shape category, we compute the posterior probability that a given IGO–year falls into each of the four Beta quadrants. For example, an organization might have an 80% probability of being autocracy-skewed and a 15% probability of being unimodal (four quadrants naturally add up to 1). This probabilistic characterization acknowledges uncertainty. Small fluctuations in the data no longer force abrupt category changes, and observations near boundaries are explicitly represented as such. In this way, the Bayesian framework replaces fragile deterministic labels with transparent probability statements grounded in the underlying distribution.

Each panel in Figure 4 displays the posterior probability of assignment to a given shape category, conditional on that category being chosen based on the highest probability of assignment. The violin densities capture the dispersion of probabilities, with medians and inter-quartile ranges indicated by the white dots and black bars. Uni-modal cases show the highest concentration near one, reflecting strong certainty in this classification. By contrast, democracy-skewed and U-shaped cases exhibit broader spreads, highlighting greater uncertainty. The figure thus conveys not only which shape is most likely for each IGO–year, but also the degree of credibility in that assignment, information that would be obscured by deterministic cutoffs alone.

Assignment to distribution regime was based on the estimated shape parameters. When both α and β estimates are greater than 1, an IGO for a given year is classified as a uni-modal regime. If both are less than 1, it is U-shaped, or polarized in a substantive sense. If $\alpha > 1, \beta < 1$, then it is skewed toward 1, democracy. If it is skewed toward 0, with $\alpha < 1, \beta > 1$, it is skewed towards autocracy. Each IGO–year has a posterior probability of belonging to each regime shape, and the modal assignment reflects the category with the highest probability. For example, Shanghai Cooperation Organization (SCO) in year 2004 has a 0.96 posterior probability of being unimodal and a 0.03 probability of being autocracy-skewed, making the unimodal label highly reliable while still acknowledging classification uncertainty. In year 1996, the SCO showed 0.84 in unimodal and 0.13 in autocracy-skewed.

Importantly, there was no case in which an autocracy-skewed distribution coincided with a Beta mean centered on the democratic side of the spectrum, and vice versa. This indicates that while the shape and location capture distinct features of IGO composition, they remain internally coherent. Skewed distributions align with corresponding underlying mean rather than contradicting them.

We release a harmonized dataset at the level of IGO year that reports distributional estimates of member state regime composition derived from the Bayesian Beta framework. Each row in the dataset corresponds to a unique IGO-year. For every IGO-year, the dataset includes posterior summaries of the two Beta shape parameters.[11] Each of these variables records the posterior

[11] To make classification uncertainty explicit and usable for future research, the dataset includes four variables with the suffix “perc”.

probability that a given IGO year belongs to one of the four regime shape categories. These probabilities are computed as the proportion of posterior draws, out of 5000 Bayesian samples, in which the estimated alpha and beta parameters fall within the corresponding region of the shape space.

For an illustrative example, the variable capturing autocracy skewed composition records the share of posterior draws in which α is less than 1 and β is greater than 1, while the unimodal variable records the share of draws in which both parameters exceed one. In addition to these probabilistic measures, we provide a modal regime shape label for each IGO year, defined as the category with the highest posterior probability. In the example above, the SCO in year 1996 will thus be labeled as ”unimodal” shape. These measures allow researchers to characterize IGO membership composition flexibly, whether by using continuous probabilities, applying confidence thresholds, or relying on categorical assignments, depending on theoretical interests.

Figure 4 traces the distribution of IGO regime compositions over time. In the 1980 cross-section, each line represents the prevalence of a different regime-shape category (unimodal, autocracy-skewed, U-shaped (polarized), and democracy-skewed) from top to bottom.

Unimodal IGOs are the most common throughout the period, followed by democracy-skewed organizations. Notably, the decline in unimodal IGOs overlaps with a sharp increase in democracy-skewed IGOs beginning in the late 1970s, coinciding with the onset of the third wave of democratization. At the same time, U-shaped IGOs composed of both democracies and autocracies gradually decline, particularly toward the end of the Cold War. As suggested earlier through the example of SEATO, many of these organizations were created during the Cold War to counter the spread of communism, often bringing together states with sharply different regime types.

Having introduced our distributional framework and showcased its measurement properties, we now turn to a substantive application. We ask whether the political composition of a country's IGO portfolio, characterized not merely by the average regime type of its members, but by the full distributional structure of those memberships, shapes subsequent domestic democratic outcomes. Specifically, we examine whether states simultaneously embedded in both democracy-skewed

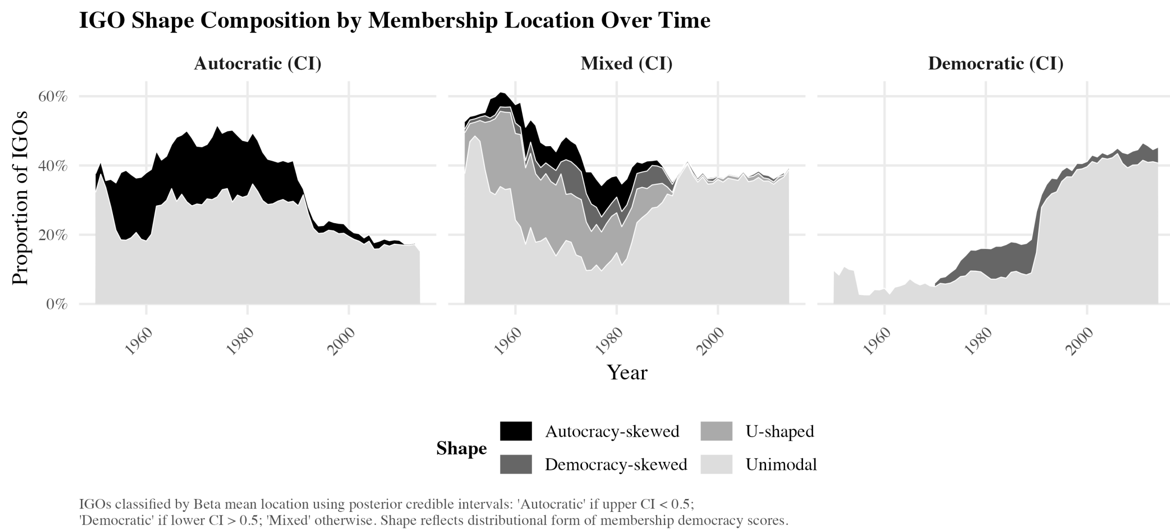


Figure 5: Stacked areas show the distributional shape composition within each location category.

and autocracy-skewed institutional environments experience systematically different democratic trajectories than those whose portfolios are more homogeneous.

Figure 5 presents the joint distribution of IGO shape and membership location over time. A notable asymmetry emerges across panels. Democratically-located IGOs are overwhelmingly unimodal in distributional form, whereas autocratically-located IGOs display a more heterogeneous mix of shapes, including a substantial Autocracy-skewed component. This asymmetry suggests that autocratic skewness constitutes a distinct distributional phenomenon not reducible to membership location alone, lending further justification to our focus on the skewed components of a country's IGO portfolio as the primary quantity of interest. No Autocracy-skewed IGO has a posterior credible interval lying entirely above the 0.5 democracy threshold, and no Democracy-skewed IGO has a credible interval lying entirely below it. This perfect separation between shape classification and CI-based location provides additional evidence of measurement validity, confirming that our distributional shape captures substantively meaningful variation in IGO composition without contradictory implication.

5 Application: Does Exposure to IGOs affect democracy?

5.1 Introduction

Most states belong to both democratic and authoritarian intergovernmental organizations simultaneously. Yet existing scholarship has examined these institutional environments as if states experience only one at a time. A large literature establishes that membership in democracy-dense IGOs promotes democratization (Pevehouse, 2002a, 2002b). On the other hand, a growing literature establishes that membership in autocracy-dense IGOs stabilizes authoritarian rule (Cottiero & Haggard, 2023; Debre, 2022). What happens when states are embedded in both simultaneously? And which institutional environment wins in terms of net effect?

The question sits at the intersection of a long tradition in international relations that examines how external environments penetrate domestic political life (Gourevitch, 1978; Putnam, 1988; Whitehead, 2021). Within this tradition, international organizations occupy a distinctive position. Unlike diffuse systemic pressures, they create structured membership environments that supply normative models and political constraints in a more sustained way (W. R. Thompson, 1996).

Prior work typically treats these institutional environments in isolation, examining either the democratizing effects of democracy-dense IGOs or the authoritarian-stabilizing effects of autocracy-dense IGOs, but rarely both simultaneously. This separation is both empirically and theoretically incomplete. Empirically, most states belong to IGOs of both types simultaneously, a fact that existing research designs have not fully accommodated. Theoretically, the net effect of IGO exposure on domestic democracy depends on the relative weight and character of competing institutional pressures, not on any single organizational affiliation considered in isolation.

We therefore shift the focus from single-organization effects to portfolio exposure. Rather than asking whether a state belongs to an authoritarian or democratic IGO, we examine the composition of its entire IGO membership environment. Using the posterior probabilities generated by our Bayesian Beta framework, we construct certainty-weighted measures of exposure to autocracy-skewed and democracy-skewed IGOs. This allows us to evaluate the net institutional environment to which a state is exposed, while preserving uncertainty about each organization's regime composition.

We extend the current understanding in three ways. First, we apply a distributional measurement framework that moves beyond sample means and percentage-based classifications, allowing us to characterize the full shape of each IGO's membership composition rather than just its central tendency.

Second, we bring together the democratization and authoritarian stability literatures within a single empirical framework, estimating the simultaneous effects of exposure to democracy-skewed and autocracy-skewed IGOs on subsequent regime trajectories.

Third, our probabilistic exposure measures enable a more precise test of the causal mechanisms at stake. Rather than asking simply whether a country belongs to an autocratic IGO, we can ask how confidently that IGO is classified, how skewed its membership is, and how that uncertainty propagates through to estimated effects, allowing us to distinguish between organizations where autocratic dominance is clear and concentrated versus diffuse and contested.

5.2 Literature Review

Existing scholarship has established that membership in democracy-dense IGOs is associated with democratization from autocracies and consolidation of nascent democracies (Pevehouse, 2002a, 2002b). Poast and Urpelainen (2015) take a closer look at the processes. IGOs consolidate democracies through sharing norms, solidifying electoral institutions, and offering practical solutions to governance problems. [12] Relatedly, Moravcsik (2000) presents the "lock-in" theory where new democracies can lock-in their preferred set of democratic governance norms so that even future incumbents cannot reverse to authoritarian practices. Democratic IGOs increase the cost of anti-democratic behavior by threatening exclusion and publicizing information on electoral misconduct, especially when the country holds geopolitical importance (Donno, 2010; Pevehouse, 2002a).

[12] Admittedly, they point to limited impact of IOs on preventing transitional democracies' reversal to authoritarianism. The reason being, IOs cannot directly enforce or initiate military intervention.

In addition, IGOs socialize member states by transmitting norms, institutional templates, and governance practices that shape how domestic actors understand legitimate political arrangements. Finnemore (1993) frames IGOs as teachers of norms, and provides examples of socialization of science policy practices. Greenhill (2010) shows that IGO's membership composition can be particularly important in socialization process since human rights norms are transmitted among states.

The mirror logic applies for autocracies. First, autocrats are capable of pooling resources and redistribute to members at times of crisis. Libman and Obydenkova (2018) and Obydenkova and Libman (2019), in turn, show that authoritarian institutional cooperation can generate tangible material benefits, such as oil subsidies, labor mobility, and financial transfers, which strengthen regime resources and bolster elite survival. Cottiero and Schneider (2026) additionally offers an empirical evidence that autocrats benefit from financial support through authoritarian financial institutions and boost their regime survival.

Second, norms that emphasize sovereignty principle against competing values of human rights can also be shared and normalized among members of AIGOs. The same logic of normative diffusion can be applied to norms of non-interference (Bae & Salvati, 2026; Ginsburg, 2020). For example, the Shanghai Cooperation Organization Charter famously includes three principles, including anti-terrorism, anti-extremism, and anti-intervention. Pauselli and Urzúa (2024) further shows that autocrats exploit traditional values like the sovereignty principle to resist the expansion of international liberal norms and oppose LGBTQ+ rights. These are often invoked at the time of diplomatic pressure from liberal democracies to justify military drills on the border, and acquiescence to human rights violation.

Third, IGOs can pool and redistribute resources among authoritarian members, providing material support that can either bolster incumbent survival in autocratic regimes (Cottiero & Schneider, 2026; Libman & Obydenkova, 2018). The resource channel operates not only through financial transfers but also through organizational programs. Bush et al. (2025) demonstrates that authoritarian states can mitigate the political risks associated with international scrutiny by supplying low-quality election observers, thereby neutralizing monitoring without incurring reputation costs.

The closest existing work to our question is the "black knights" literature, which contends that the democratizing influence of international institutions (or liberalization effort, in general) is restrained when autocratic states provide alternative sources of legitimacy and material support to incumbent regimes (Chou, 2017; Tolstrup, 2014, 2015). This work acknowledges that states face competing international pressures simultaneously.

Yet even this literature frames the problem as a binary, either a black knight is present or it is not. It does not model the full portfolio of institutional environments a state inhabits, nor does it weight each organization by the confidence with which it can be classified as democratizing or authoritarian. The result is that the *degree* of competing institutional pressure remains unmeasured.

Our work also speaks to the literature on regime complexity, which examines why states participate in multiple, functionally overlapping IGOs (Panke & Stapel, 2023). Existing accounts emphasize how institutional overlap shapes strategic choice among forums. For example, Busch (2007) shows that states anticipate future disputes and gravitate toward institutions whose decision rules are likely to produce favorable outcomes, while Clark (2022) demonstrates that overlapping regional financial arrangements can serve as exit options that enhance bargaining leverage in negotiations within competing organizations.

Indeed, there has been research that answers the puzzle of why autocracies and new democracies would join international institutions that are mainly dominated by democracies (Moravcsik, 2000; Vreeland, 2008). Although less explored, the opposite direction (why -or even whether- democracies join autocratic-skewed IGOs) is also plausible. Libman and Obydenkova (2018) shows that authoritarian states do not attach any human rights conditionalities and can draw more diverse (in terms of both political and economic conditions) set of countries, contrary to democracies' selective accession to autocracies.

More broadly, the increasing density of international institutions means that states rarely operate within a single organizational environment. Instead, they are embedded in complex institutional ecologies characterized by overlapping memberships, competing rules, and partially redundant governance structures (Alter & Raustiala, 2018). In such environments, the political composition of any single IGO provides only a partial view of the institutional pressures states face. Our portfolio-based framework therefore aligns with recent efforts to conceptualize international cooperation as a network of overlapping institutions rather than a collection of isolated organizations.

To establish empirical ground for this claim, we test the assumption that states join both democracy-skewed and autocracy-skewed IGOs, regardless of domestic regime types. Figure 6 shows density of country-year exposure to authoritarian-dominated IGOs by regime type. This addresses, "of all the IGOs this country belongs to, what share are classified as authoritarian-skewed?" It therefore reflects the institutional environments states face. Exposure, the horizontal axis, is the share of an individual country's IGO memberships that are classified as autocracy-skewed, weighting each organization by the posterior probability of that classification. There is substantial overlap in exposure across autocracies, hybrid regimes, and democracies. No regime type is insulated from authoritarian-dominated institutional environments. The implication follows: Most states, regardless of their domestic regime type, are exposed to both authoritarian-skewed and democratic-skewed IGOs at the same time. Hence, the effect of IGOs on democracy should be understood as a net-effect of both "democracy from above," and "autocracy from above."

Figure 6 further motivates our decision to model IGO exposure probabilistically and along multiple dimensions. Because states are simultaneously embedded in democracy-skewed and autocracy-skewed institutional environments, analyses that rely on mean-based summaries or deterministic classifications risk obscuring the conditions under which institutional influence is most consequential. The distributional overlap documented here provides a baseline empirical justification for treating institutional exposure as continuous, uncertain, and multidimensional in the outcome analyses that follow.

5.3 Theory and Observable Implication

Why would we anticipate AIGOs and DIGOs to cancel out each other? The mechanisms identified by existing scholarship suggest that IGO influence operates through three channels,

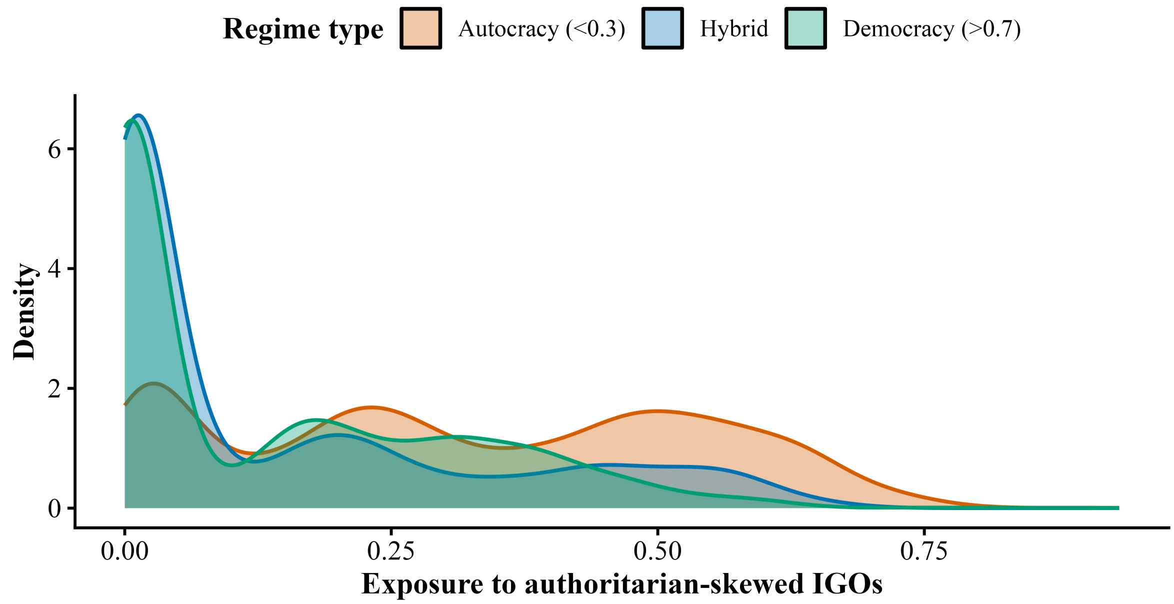


Figure 6: Distribution of country exposure to autocracy-skewed IGOs by regime type.

namely, cost manipulation, norm socialization, and resource provision. Each of these factors can be activated by both democratic and authoritarian institutional environments simultaneously and in opposite directions.

When AIGOs are also present, each channel faces a countervailing force. The costs of anti-democratic behavior are reduced by the availability of alternative legitimacy sources and exit options. Autocratic IGOs supply recognition and diplomatic cover that insulate incumbents from the reputational consequences that democratic IGOs threaten (Tolstrup, 2015).

Normative socialization is contested by competing institutional templates. Where democratic IGOs transmit norms of electoral accountability and human rights, authoritarian IGOs diffuse principles of sovereignty and non-interference that provide ideological justification for resisting external democratic pressure (Finnemore, 1993; Ginsburg, 2020).

Likewise, material conditionality is weakened when authoritarian IGOs supply resources unconditionally. The threat of democratic IGO exclusion loses credibility when alternative patrons are available, reducing the leverage that conditionality depends on (Libman & Obydenkova, 2018). The reactive nature of authoritarian IGO channels is illustrated by the cases of Cambodia and the Philippines. When the United States withdrew promised electoral assistance and the European Union threatened rice tariffs following Hun Sen's dissolution of the opposition party, China pledged $588 million in aid within days, directly replacing the conditional democratic pressure with unconditional authoritarian patronage (Bae, 2022). The cost-manipulation channel of democratic IGOs was neutralized not by the absence of pressure, but by the availability of an alternative. Similarly, when the United Nations Human Rights Council investigated Duterte's drug war and liberal democracies threatened sanctions, China's representative publicly endorsed the Philippines' human rights record at the same forum, contesting the normative claim on its own terrain. These cases illustrate that authoritarian IGO channels do not operate independently of democratic pressure. They substitute for conditional resources when liberal pressure escalates.

This naturally places the AIGO effectiveness finding under a more demanding test.[13] Does authoritarian IGO exposure retain its autocratizing influence even when states are simultaneously embedded in democracy-skewed institutional environments? If AIGO channels are partly reactive, activated in response to democratic pressure, then controlling for simultaneous DIGO exposure should attenuate the estimated AIGO effect. If autonomous channels dominate, the effect should persist. The answer has direct implications for how we understand the robustness of authoritarian international influence in an era of institutional overlap. These considerations lead to the following empirical expectation.

Hypothesis (Portfolio balance): The net democratic trajectory of a state reflects the overall balance of its IGO portfolio. States with higher democracy-skewed exposure show stronger democratic trajectories, partially offsetting the negative effect of autocracy-skewed exposure.

5.4 Data and Model Fitting

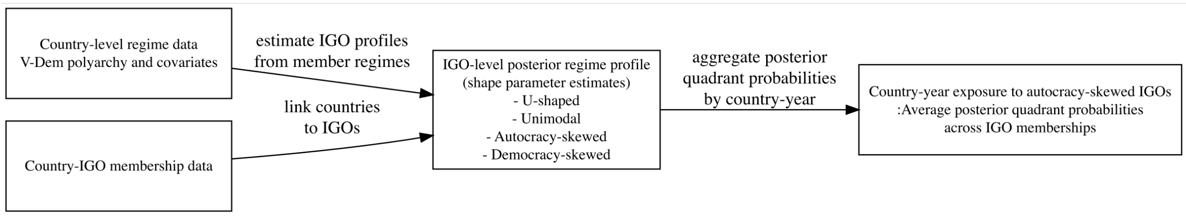
The analysis estimates a series of Bayesian regressions in which a country's electoral democracy score, *v2x_polyarchy*, is modeled as a function of its lagged IGO portfolio, capturing the extent to

[13] The test is symmetric of course. Depending on theoretical interest, one might equally ask whether DIGO effectiveness persists in the presence of authoritarian counterforces. We return to this question in the results.

which membership is concentrated in autocracy-skewed versus democracy-skewed organizations. By linking variation in IGO exposure to changes in domestic democratic outcomes, the regression framework allows us to evaluate whether greater engagement with authoritarian IGOs is associated with lower subsequent democracy scores, and whether exposure to democracy-skewed IGOs offsets or amplifies these effects.

To capture how countries are embedded within different types of IGOs, we construct a country's measure of certainty-weighted exposure to autocracy-skewed and democracy-skewed IGOs.[14] Rather than assigning each IGO to a single category, we use the full posterior distribution of member-regime types estimated for every IGO. For each country–year, we sum the posterior probability that each of its IGO memberships belongs to the autocracy-skewed (or democracy-skewed) quadrant of the regime-composition space, and then normalize this quantity by the total posterior mass across all four quadrants, which is simply the number of IGOs a country is in. The resulting measures represent the expected share of a country's IGO portfolio that is autocracy-skewed or democracy-skewed, weighting each organization by the model's uncertainty about its regime composition. Conceptually, these indicators capture not only how many autocratic or democratic IGOs a country belongs to, but also how confidently those IGOs can be classified based on their membership profiles. Figure 5.4 illustrates these steps in a flowchart.

[14] In addition to the quadrant-based exposure measures, our data can be used to construct a pair of mean-based exposure indicators that summarize how centrally autocratic or democratic a country's IGO portfolio is. These indicators reflect the average leaning of a country's IGO memberships, weighting each organization by the model's uncertainty regarding the position of its posterior means.



Our framework also partially addresses a well-known methodological concern in this literature. Cottiero and Haggard (2023), echoing Pevehouse (2002a), exclude the focal country from the IGO-level democracy average to avoid mechanical correlation between the predictor and the outcome. [15] Our approach mitigates this concern. First, the Beta distribution aggregates member scores into shape parameters rather than a direct mean, attenuating the mechanical relationship between any single country's score and the resulting distributional estimate. Second, our exposure measure averages posterior probabilities across a country's entire IGO portfolio, further diluting the influence of any single country-IGO observation. The attenuation is also largest where it matters most. In small IGOs where the focal country's score would otherwise carry disproportionate weight, the posterior distribution is already more diffuse, naturally reducing the precision of the classification and hence its leverage on the exposure measure.

To address potential confounding factors that may jointly influence countries' IGO membership profiles and their domestic democratic trajectories, we include a set of standard political, economic, and demographic covariates drawn primarily from the V-Dem dataset and complemen-

[15] Consider an example. Suppose we wish to test whether membership in a democratic IGO promotes democratization. A natural approach is to measure IGO democracy as the average polyarchy score of all member states and regress a country's own democracy score on that average. The problem is that the focal country's score is itself included in the average used as the predictor where the level of analysis is at country-year. The regression is therefore partly asking whether a country's democracy score predicts its own democracy score, a relationship that is trivially true by construction.

tary international data sources. First, we control for domestic and international policy autonomy using V-Dem's indicators of autonomy in domestic policymaking (*v2svdomaut*) and insulation from external influence in foreign policymaking (*v2svinlaut*). These variables capture the extent to which a state retains independent authority over internal and external policy decisions, net of constraints imposed by other states, and help account for baseline sovereignty conditions that may shape both IGO participation and regime outcomes. We also include freedom of foreign movement (*v2clfmove*), which proxies for openness and mobility and is commonly associated with broader civil liberties and regime type.

To capture economic capacity and international integration, we control for logged GDP, population size, and trade exposure, measured by exports and imports. These variables account for differences in material resources and global embeddedness that may independently affect both institutional affiliations and democratic stability. Finally, we include the number of coup attempts as an indicator of recent political instability, which is strongly associated with short term regime change and democratic erosion.

We then estimate a series of Bayesian regression models in which a country's democracy score, $v2x\_polyarchy_{it}$, is modeled as a function of its certainty-weighted IGO portfolio and a set of standard covariates. The main predictors are exposure to autocracy-skewed IGOs and exposure to democracy-skewed IGOs, entered simultaneously to capture the relative weight of authoritarian versus democratic memberships in a country's overall IGO portfolio. Both IGO exposure predictors are lagged by one year to reduce endogeneity concerns.

We also include country-level random intercepts, $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$,to account for unobserved, time-invariant heterogeneity across states, such as historical legacies and institutional features not captured by observed covariates. The empirical strategy follows a Bayesian ANCOVA-style specification in which these country-specific intercepts are treated as group-level effects and all key predictors are lagged by one year. Bayesian approach enables partial pooling across countries while still absorbing persistent cross-national differences in baseline democracy levels. Year-specific indicators capture global shocks, functioning analogously to year fixed effects but estimated with full posterior uncertainty. The Bayesian framework yields posterior distributions for all parameters, allowing direct probabilistic interpretation of effect sizes and facilitating the counterfactual simulations discussed below.

Formally, our estimate of interest is the expected level of democracy at time $t+1$, conditional on prior democracy at time t, given differential exposure to autocracy and democracy-skewed IGOs.The ANCOVA structure includes the one-year lagged democracy score, $y_{i,t-1} = v2x\_polyarchy_{i,t-1}$, as a predictor. Substantively, this design allows us to assess whether greater exposure to autocracy- or democracy-skewed IGOs is associated with higher or lower subsequent democracy after conditioning on a country's previous democracy level and time-invariant characteristics. Although our substantive interest concerns changes in democracy, we model democracy levels rather than differentials. This needs justification. [16] A change-score specification implicitly assumes that last year's democracy level carries over perfectly into the current year and that all countries follow the same adjustment dynamicsVan Breukelen (2013). This assumption is difficult to justify in a setting where democratic persistence varies across regime types and outcomes are bounded. We therefore adopt an ANCOVA-style specification that conditions on prior democracy while allowing the data to determine how strongly past regime levels shape current outcomes.

We model $y_{it} = v2x\_polyarchy_{it} \in (0, 1)$ with a Beta likelihood,

$$y_{it} \sim \text{Beta}(\mu_{it}\phi,\ (1 - \mu_{it})\phi),$$

where $\mu_{it} \in (0, 1)$ is the conditional mean and $\phi > 0$ is a precision parameter. The mean is linked to predictors via

$$\text{logit}(\mu_{it}) = \alpha_i + \lambda_t + \rho\,\text{logit}(y_{i,t-1}) + \beta_1\,\text{AutoIGO}_{i,t-1} + \beta_2\,\text{DemoIGO}_{i,t-1} + X_{i,t-1}\boldsymbol{\gamma}.$$

We include country-specific random intercepts $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ and year effects λ_t, where $X_{i,t-1}$ denotes the vector of lagged covariates and Z_{it} denotes the vector of year indicators.

[16]See Pearl (2016) for further discussion on ANCOVA versus CHANGE model.

5.5 Results and Discussion

Our models tell a consistent story that a country would experience a decline in democracy index if it were connected to more authoritarian-skewed IGOs. In addition, it also supports the "democracy from above" theory that stronger connection to democratic-skewed IGOs are also related to higher democracy score. The estimated effect of exposure to democracy-skewed IGOs is larger in magnitude than the effect of exposure to autocracy-skewed IGOs. Substantively, this implies that, for comparable marginal changes in portfolio composition, democratic institutional environments are associated with stronger gains in democracy than the losses associated with authoritarian environments.

Figure 7 presents counterfactual changes in democracy scores when each country's exposure to autocracy-skewed IGOs is shifted from its observed value to the minimum and maximum observed levels.[17] Consistent with theoretical expectations, countries would show lower democracy scores if they were linked to authoritarian-skewed IGOs to the highest possible degree. For example, if Nepal were exposed to authoritarian-skewed IGOs at the maximum observed level, its electoral democracy score would decrease by an estimated 0.025. While the magnitude of this change may appear modest, it is important to note that the electoral democracy index aggregates ten subcomponents and is therefore inherently a composite measure. In policy reports, changes of 0.05 are often used as a threshold to classify countries as undergoing democratization or autocratization (Coppedge et al., 2025; Nord et al., 2025).

For counterfactuals, we set Authoritarian Skewed $\text{IGOs}_{i,t-1}$ to the global minimum and maximum observed in the sample while holding other covariates fixed. The plot illustrates how the predicted democracy score responds to these changes in the autocracy-share of a country's IGO portfolio while holding all other covariates, including the lagged democracy score, constant. In other words, each line shows how much higher or lower a country's democracy level would be expected to be if it were embedded in a more authoritarian or less authoritarian IGO environment,

[17] For ease of interpretation, I report predicted values based on the model estimates. Table Appendix D in the Appendix reports the raw coefficient estimates.

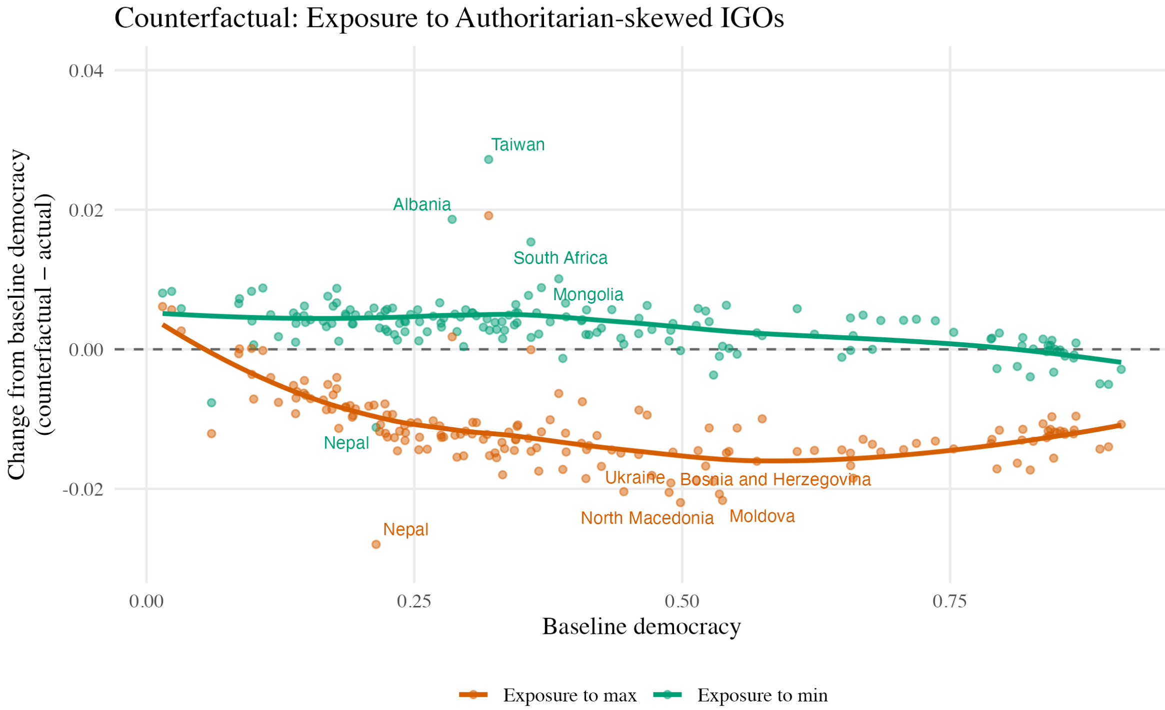


Figure 7: Counterfactual when exposure to AIGO changes.

conditional on its prior regime status. For example, countries with a democracy score of 0.5 at time t are predicted to experience an average decline of about 0.02 in democracy when exposure to autocracy-skewed IGOs is set to its maximum.

Countries positioned toward the left side of each panel (autocratic states) show relatively modest changes in predicted democracy when exposure to autocracy-skewed IGOs increases. In contrast, countries with mid-range democracy scores exhibit noticeably larger predicted declines under the high-exposure scenario. This pattern is consistent with the theoretical expectation that autocracy-skewed IGOs exert stronger influence where institutions are more open and thus more vulnerable to shifts in external constraints and incentives.

The relatively modest predicted changes among fully autocratic states warrant careful interpretation. On one reading, this pattern reflects a theoretical saturation effect in that consolidated autocracies have already activated the survival mechanisms that authoritarian IGOs supply, including resource networks, legitimacy cover, and elite coalition support. In this sense, additional AIGO exposure would be largely redundant. On another reading, however, the flatness near the lower end of the democracy spectrum may be partly mechanical. Because the democracy index is bounded, states already near the floor have limited room to decline further regardless of their institutional environment. Our current design cannot cleanly separate these two interpretations. Future work using unbounded outcome measures, such as leadership tenure, would help disentangle whether the null marginal effect of AIGO exposure on fully autocratic states reflects genuine theoretical saturation or the constraints of a bounded outcome scale. For now, we flag this ambiguity as a limitation for full transparency and refrain from overclaiming the flat counterfactual pattern as definitive evidence of either mechanism.

A parallel pattern arises at the upper end of the democracy spectrum, though with an important asymmetry that provides slightly more theoretical purchase. For consolidated democracies, we might attribute the flat counterfactual lines under a mechanical ceiling effect, that states already near the top of the bounded democracy scale have little room to score higher regardless of their institutional environment. However, consolidated democracies also show limited predicted decline under the high Autocratic-skewed IGOs exposure scenario, and here the ceiling argument offers no relief. (These states have ample room to fall on the scale, yet do not.) The flatness in both directions therefore cannot be dismissed as purely mechanical. It instead suggests that strong domestic institutions, dense civil societies, and deeply embedded democratic norms act as genuine buffers against external organizational pressure regardless of its direction. When democratic consolidation has already occurred domestically, the marginal influence of any single IGO environment (whether autocracy-skewed or democracy-skewed) is limited. This finding aligns with accounts emphasizing that IGO influence is most consequential for states where domestic political outcomes remain genuinely contested, and least consequential where they have already been settled, whether in an authoritarian or democratic direction (Moravcsik, 2000; Pevehouse, 2005).

Hybrid regimes occupy a qualitatively different position. Neither the constraints of consolidated autocracy nor the buffers of consolidated democracy are firmly in place. Domestic institutions remain contested, incumbent survival is less certain, and the costs and benefits supplied by external organizational environments carry greater weight in the calculations of both elites and opposition actors. This openness operates symmetrically. Hybrid regimes are not only most vulnerable to the autocratizing influence of authoritarian IGOs but also most responsive to the democratizing pressure of democracy-skewed ones. The finding that intermediate regime positions show the largest counterfactual swings in both directions is therefore consistent with a broader literature suggesting that external influence (whether democratizing or autocratizing) is particularly important precisely where domestic political outcomes remain genuinely uncertain.

Being embedded in an IGO environment dominated by authoritarian members is associated with lower subsequent democracy, even after controlling for a country's prior democracy level and domestic covariates. Including each country's lagged democracy score is a conservative test, as it may absorb some of the variation attributable to IGO embeddedness over time, understating the true effect. Nevertheless, the association persists, lending confidence to our findings.

Moreover, the heterogeneous pattern across the democracy spectrum suggests these effects are not uniform. Exposure to IGOs dominated by authoritarian members appears most consequential for countries occupying intermediate regime positions, neither fully autocratic nor consolidated democracies, where external organizational environments may more readily shape domestic political trajectories.

Figure 8 similarly shows a counterfactual estimates when all countries were tied to democratic-skewed IGOs to the strongest possible degree. In this scenario, countries are assumed to be embedded in institutional environments dominated by democratic member states. Consistent with the positive coefficient on democracy-skewed IGO exposure, the model predicts higher democracy scores under this counterfactual relative to observed levels. Substantively, this pattern supports the "democracy from above" theory that institutional environments dominated by democratic states are associated with improvements in domestic democratic outcomes (Pevehouse, 2002a, 2005).

The asymmetric pattern across the democracy spectrum in both 7 and 8 is substantively meaningful. Autocratic regimes (those on the left-side on the x axis) appear relatively insulated from further decline because many institutional constraints associated with democratic accountability are already absent. Highly democratic regimes, by contrast, possess institutional and societal buffers

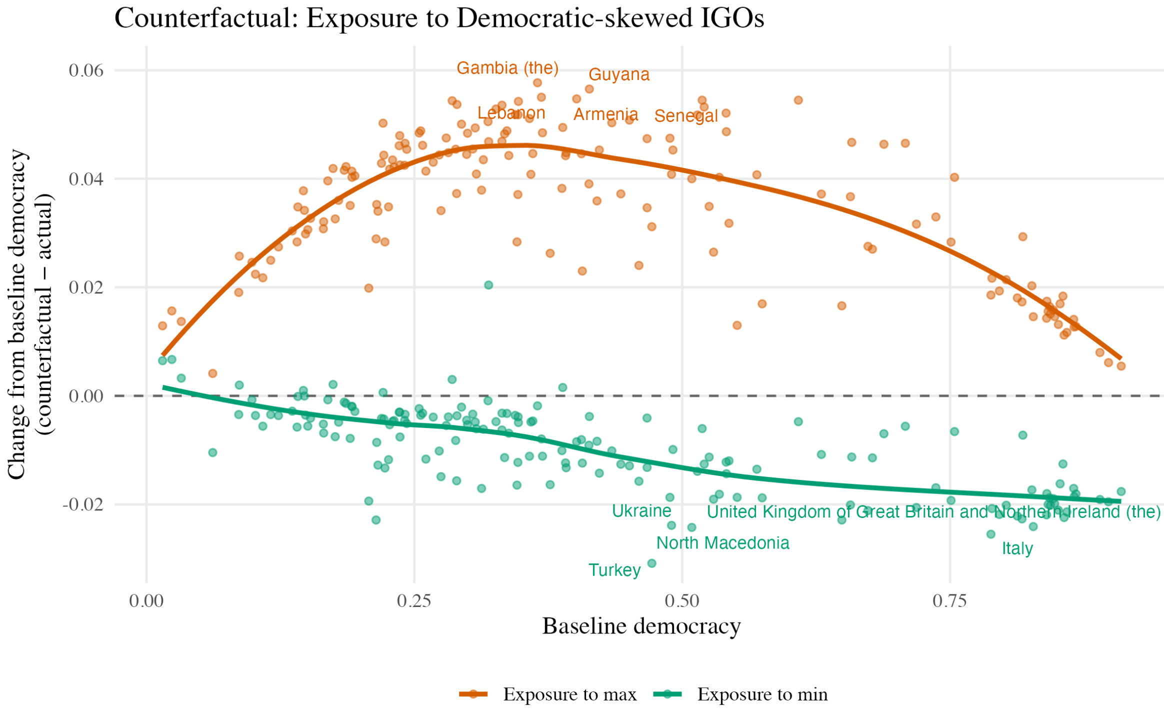


Figure 8: Counterfactual when exposure to DIGO changes.

that limit the influence of external organizational environments. Hybrid regimes exhibit the greatest volatility in predicted democracy under the counterfactual scenarios, showing the largest responses to changes in exposure to autocracy-skewed IGOs.

These patterns underscore the analytical value of modeling IGO composition probabilistically rather than relying on mean based or deterministic classifications. Because the exposure measures weight each IGO by the posterior probability that it is autocracy skewed or democracy skewed, the counterfactuals reflect graded differences in institutional environments rather than sharp category boundaries. As a result, the estimated effects are not driven by a small number of borderline cases but instead capture systematic variation in the authoritarian or democratic character of countries' broader IGO portfolios.

5.6 Model Evaluation

To assess whether our framework and measures improve upon conventional approaches, we estimate a set of alternative models that differ only in how they operationalize IGO exposure. All models use the same outcome, covariates, and country random effects, but for the measurement of a country's IGO portfolio. Instead of averaging posterior probabilities directly, we construct an alternative, more conservative exposure measure by converting the full posterior distribution of each IGO–year into a categorical indicator. Specifically, each IGO is assigned to the regime-shape category with the highest posterior probability, and these binary classifications are then aggregated to the country–year level. [18]

We evaluate model performance using Pareto-smoothed importance sampling (PSIS) leave-one-out (LOO) cross-validation(Vehtari et al., 2017). LOO successively refits the model while holding out one observation at a time and assesses how well the model predicts the omitted data point. The resulting expected log predictive density (ELPD) summarizes out-of-sample predictive accuracy. Higher ELPD values indicate better out-of-sample predictive performance. We compare our main model against a set of more parsimonious rival specifications. Table 1 reports the predictive performance (ELPD) for five competing models. The full specification serves as our baseline and includes both exposure variables (Democracy-skewed and Autocracy-skewed) along with the full set of covariates. The subsequent rows progressively remove key predictors. Model 3 excludes country-level share of the democracy-skewed and autocracy-skewed IGO measures. Model 4 uses mean-based measurement instead of shape-based measurement, using a country's share of IGO portfolio whose posterior probability of mean is less than 0.5. Model 2 excludes political and economic covariates like trade volume and GDP, and Model 1 is the restricted model where I only include lagged democracy score as the only predictor.

Overall, the model tracks the observed democracy scores closely, indicating good predictive fit. We report posterior predictive plot in the Appendix E to further illustrate that the model

[18] As we established earlier, this binary assignment is derived from the posterior distribution of shape parameters rather than from sample means or deterministic cutoffs, thereby avoiding the instability of mean-based classifications. This sets a hard test for our main model.

Model	ΔELPD	SE(ΔELPD)
M5: Full Model (Autocracy-Skewed Shape Measure)	0.0	0.0
M3: Full Controls, No IGO Predictors	-44.1	10.2
M4: Full Model (Autocracy-centered Mean Measure)	-44.6	9.8
M2: IGO Shape Exposure + Controls Excluded	-502.7	54.1
M1: Baseline (Lagged Democracy Only)	-664.3	59.3

All models fit on identical observations. ΔELPD relative to best-fitting model.

Table 1: Model Comparison

reproduces the distribution of observed values well. The full model using the autocracy-skewed shape-based IGO exposure measure performs best, while the model substituting the location-based measure performs nearly identically to a model that excludes IGO predictors entirely. This near-equivalence suggests that the location-based operationalization, capturing the share of IGO memberships whose posterior mean falls below 0.5, carries little predictive information beyond what the controls already provide. By contrast, the shape-based measure, which captures the probability that an IGO's membership distribution is skewed toward autocracy, does meaningfully improve fit. The two restricted baseline models, one including only lagged democracy and one adding shape-based IGO exposure without full controls, fit substantially worse, confirming that both the control variables and the IGO exposure measure contribute independently to predictive performance.

This result is not merely a robustness check but also useful evidence for the paper's central methodological claim. What matters for democratic erosion is the ideological coherence and homogeneity of an IGO's membership, in terms of a tight, like-minded autocratic bloc that a location-based measure will systematically miss. An IGO whose members average 0.3 on the democracy scale looks identical to one whose members cluster tightly near zero, yet the latter represents a qualitatively different institutional environment. The fact that the location-based operationalization adds nothing beyond the controls suggests that what the controls already absorb is sufficient to account for the general tendency of low-democracy states to join low-democracy IGOs. The residual signal recovered by the shape-based measure therefore reflects something the controls cannot capture, the distributional structure of institutional environments, not merely their average character.

5.7 Replication and Extension: Debre (2022)

To situate our framework relative to existing work, we replicate and extend discussion in comparison with Debre (2022), which provides the most direct prior test of authoritarian IGO effects on regime survival. This exercise serves two purposes. First, it allows us to assess whether differences in results are attributable to measurement, to sample scope, or to the conceptualization of IGO exposure. Second, it demonstrates that our distributional framework recovers and extends substantive findings beyond our own modeling strategy, lending external validity to the approach. Our framework differs from Debre's along three dimensions, each of which carries theoretical and empirical implications. This means that our results do not directly contradict or replace Debre's finding, but rather we uncover information that's otherwise unavailable without our distributional estimates.

Country sample. Debre restricts her analysis to autocratic regimes, following the GWF dataset (Geddes et al., 2014). The outcome of interest is regime survival, modeled as time-to-event using Cox proportional hazard models. Our analysis includes all country-years regardless of regime type, modeling democracy levels as a continuous outcome. This difference reflects not merely a technical choice but a substantive one. Debre asks whether authoritarian IGO membership prolongs autocratic rule, conditional on already being autocratic. We ask whether IGO portfolio composition shapes democratic trajectories across the full regime spectrum.

IGO exposure measure. Debre (2022) operationalizes IGO exposure through *Authoritarian Density*, defined as the average democracy score of the members of the most autocratic regional organization a country belongs to in a given year. This "highest autocratic density" approach captures the intensity of the single most extreme organizational affiliation. Our approach instead constructs a certainty-weighted portfolio measure that aggregates exposure across all IGO memberships, weighting each organization by the posterior probability that it belongs to the autocracy-skewed or democracy-skewed quadrant of the membership distribution space. These two approaches encode different theories of how IGO influence operates. Debre's measure implicitly assumes that the most extreme organizational affiliation dominates, consistent with a threshold or tipping-point logic. Our portfolio measure assumes that the net balance of institutional environments matters, consistent with a cumulative exposure logic. Testing both within the same robustness framework therefore speaks directly to which mechanism better characterizes IGO influence.

Distributional measurement. Beyond sample and scope, our measure also differs in how it characterizes each IGO's regime composition. Debre (2022) relies on the arithmetic mean of member democracy scores, which, as we demonstrated in Section 3, is subject to compositional collapse, sampling variability, and overconfident classification. Our Bayesian Beta framework models the full distribution of member scores and yields posterior probabilities over shape categories rather than a single point estimate. This difference is consequential for classification. Several well-known authoritarian clubs, including the Shanghai Cooperation Organization, the African Union, and the Commonwealth of Independent States, are unimodally distributed rather than J-shaped. A quadrant-based or mean-based measure would classify these organizations differently than our location-based posterior probability, producing different exposure measures and different downstream estimates.

Debre (2022) finds that authoritarian IGO membership reduces the hazard of democratization but shows no significant association with autocratic regime replacement. Our replication and extension reveals that both findings are sensitive to measurement and estimation choices in ways that carry theoretical implications.

On democratization, substituting the arithmetic mean-based *AutDensity* measure with our probabilistic location-based measure and estimating with a frequentist Cox model attenuates the effect to nonsignificance. The signal re-emerges only when country-level heterogeneity is modeled explicitly through Bayesian partial pooling, suggesting that the frequentist specification lacks the statistical power to detect the effect with a more precise measure. This is not a contradiction of Debre's finding but a clarification of the conditions under which it holds.

On autocratic replacement, the null result dissolves entirely when we shift from a location-based to a shape-based measure. Our probabilistic shape measure, which captures exposure to IGOs with a J-shaped autocracy-skewed membership distribution rather than merely autocratically centered ones, produces a large and credible negative association with autocratic replacement.

We assess out-of-sample predictive performance using PSIS-LOO cross-validation for all three model specifications. The results, reported in Appendix G, show comparable predictive accuracy across measures, suggesting that differences in coefficient significance reflect measurement precision rather than model fit.

Table 2 reports coefficient estimates from Bayesian Cox models using the autocratically centered exposure measure. For democratization, the estimated effect is negative and meaningful. This indicates that a country's greater exposure to autocratically centered IGOs reduces the hazard of democratization. For autocratic replacement, the estimate is positive and the credible interval crosses zero, yielding a null result.

Table 3 reports results using the autocracy-skewed exposure measure, which captures exposure to IGOs with a J-shaped membership distribution. Here the pattern changes substantially. The democratization effect remains negative and credible, consistent with the location-based results. For autocratic replacement, however, the shape-based measure produces a large negative and credible estimate, indicating that exposure to autocracy-skewed IGOs also significantly reduces the hazard of autocratic regime breakdown.

The two measures therefore diverge precisely on the autocratic replacement outcome. The location-based measure finds no effect while the shape-based measure finds a strong one, suggesting that what drives incumbent protection is not merely membership in autocratically centered organizations but specifically exposure to the most extreme and homogeneous autocratic clubs.

Table 2: Bayesian Cox Model: Location-Based Exposure Measure

	Democratization	Aut. Replacement
Autocratically centered IGO exposure	-2.402	2.731
	[-4.854, -0.067]	[-0.707, 6.132]
Authority	-0.420	-0.178
	[-1.578, 0.679]	[-1.634, 1.052]
Hegemon	0.357	0.221
	[-0.449, 1.172]	[-0.514, 0.926]
Cold War	0.861	-0.851
	[-0.217, 1.921]	[-2.120, 0.262]
Contagion (dem.)	-0.603	
	[-1.786, 0.339]	
Contagion (aut.)		-1.004
		[-2.408, 0.135]
Regime type (1)	6.298	0.795
	[4.853, 7.978]	[-0.424, 1.912]
Regime type (2)	-2.264	-0.624
	[-6.185, 0.509]	[-2.331, 0.892]
Regime type (4)	2.680	1.605
	[1.288, 4.090]	[0.598, 2.565]
GDP growth	-4.766	-2.768
	[-8.156, -1.256]	[-4.988, -0.226]
Ln resources	-0.223	-0.161
	[-0.417, -0.045]	[-0.307, -0.014]

Table 3: Bayesian Cox Model: Shape-Based Exposure Measure

	Democratization	Aut. Replacement
Autocracy-skewed IGO exposure	-2.537	-4.569
	[-5.098, -0.007]	[-6.973, -2.201]
Authority	-0.568	-0.108
	[-1.772, 0.535]	[-1.527, 1.118]
Hegemon	0.581	0.516
	[-0.311, 1.470]	[-0.242, 1.279]
Cold War	0.942	-3.239
	[-0.157, 2.040]	[-4.502, -2.062]
Contagion (dem.)	-0.653	
	[-1.880, 0.297]	
Contagion (aut.)		-1.201
		[-2.616, -0.033]
Regime type (1)	6.337	1.320
	[4.893, 8.012]	[0.044, 2.469]
Regime type (2)	-1.841	-0.085
	[-5.711, 0.916]	[-1.695, 1.357]
Regime type (4)	2.985	1.706
	[1.694, 4.446]	[0.725, 2.633]
GDP growth	-4.779	-4.495
	[-8.243, -1.356]	[-6.884, -1.718]
Ln resources	-0.154	-0.122
	[-0.335, 0.028]	[-0.271, 0.043]

6 Conclusion

This paper advances the study of intergovernmental organizations by moving beyond mean-based summaries of member-state attributes to a distributional approach that is both flexible and interpretable. We show that common practices (using the sample mean to characterize "average" membership and then imposing deterministic thresholds) can misstate organizational composition in three ways. They omit shape information that distinguishes unimodal, skewed, and polarized memberships. Second, they are highly sensitive to sampling variability and measurement noise. Third, they are prone to making overconfident categorical claims that obscure genuine uncertainty. Modeling membership with a Beta distribution and estimating its shape parameters within a Bayesian framework addresses these problems directly. The resulting estimates capture both the location and dispersion of regime types within organizations while allowing researchers to make probabilistic, rather than deterministic, statements about regime composition.

Recovering distributional characteristics matters because a large literature in international relations argues that the regime composition of international organizations shapes institutional design, authority, and cooperation outcomes (Bae & Moon, 2025; Koremenos & Betz, 2012; Koremenos et al., 2001). Whether IGOs are dominated by democracies or autocracies affects delegation, enforcement mechanisms, and the diffusion of political norms. If regime composition is mischaracterized, inferences about the political foundations of institutional design may also be misleading. By treating membership composition as a *distribution* rather than a point estimate, our framework aligns measurement more closely with theoretical claims about heterogeneity, coalition structure, and bargaining within international institutions.

Methodologically, our results offer practical guidance. First, when characterizing IGO composition, researchers should report both location ($\hat{\mu}$) and precision ($\hat{\phi}$), and, when relevant, posterior probabilities over shape categories. Second, where downstream analyses require a scalar predictor, the beta mean provides a more conservative and stable estimator than the raw sample mean, particularly with small membership and non-normal shapes. Third, when categorical bins are analytically useful, we recommend retaining posterior probabilities (and optionally applying confidence thresholds) rather than imposing hard cutoffs that ignore classification uncertainty.

More broadly, the framework provides a principled way to connect theories of international institutional design with the empirical realities of heterogeneous member-state attributes, enabling IR scholars to more accurately assess how the composition of IGOs shapes their behavior, rules, and long-term political consequences.

The approach is general. Although we illustrate with regime type, the same logic applies to other member attributes: economic size, trade openness, conflict exposure, or human-rights performance. Researchers can fit Beta (or alternative bounded distributions) to any $(0, 1)$-scaled indicator, estimate (α, β) over time. [19] Our model can be used to investigate distributional features as institutional outcomes. In outcome modeling, posterior summaries can enter as covariates, or full posterior draws can be propagated through hierarchical models to account for measurement uncertainty.

We accompany this paper with a dataset of IGO–year estimates reporting posterior summaries of (α, β), the implied mean and precision, and posterior probabilities for unimodal, U-shaped, democracy-skewed, and autocracy-skewed distributions. By making distributional information about membership composition directly available to researchers, these data facilitate more transparent and theoretically grounded empirical analyses of international institutions.

An important avenue for future research concerns the dynamics of distributional change within IGOs. In this paper, we focus primarily on measuring the shape of membership distributions in a given IGO–year. Yet IGOs are not static coalitions. Membership expansions, withdrawals, and regime changes within member states can gradually alter the underlying distribution of political regimes. For instance, under what conditions do previously unimodal organizations become autocracy-skewed, or polarized distributions converge toward a unimodal structure? Understanding when and why IGOs transition across distributional regimes remains an open empirical question. Such shifts may occur during periods of geopolitical realignment, regional democratization waves, or institutional expansion. Tracking shifts across distributional regimes therefore opens new op-

[19] For example, Voeten (2021) emphasizes the significance of member state ideology distribution.

portunities to examine how evolving membership structures shape the trajectory of international organizations and the political influence they exert on member states.

References

Abbott, K. W., & Snidal, D. (1998). Why States Act through Formal International Organizations. *Journal of Conflict Resolution*, *42*(1), 3–32. https://doi.org/10.1177/0022002798042001001

Abbott, K. W., & Snidal, D. (2000). Hard and Soft Law in International Governance. *International Organization*, *54*(3), 421–456. https://doi.org/10.1162/002081800551280

Allee, T., & Peinhardt, C. (2010). Delegating Differences: Bilateral Investment Treaties and Bargaining Over Dispute Resolution Provisions. *International Studies Quarterly*, *54*(1), 1–26. https://doi.org/10.1111/j.1468-2478.2009.00575.x

Alter, K. J., & Raustiala, K. (2018). The Rise of International Regime Complexity. *Annual Review of Law and Social Science*, *14*(1), 329–349. https://doi.org/10.1146/annurev-lawsocsci-101317-030830

Bae, J. (2022). Chinese Patronage and Southeast Asian Democracy in Distress. *Asia-Pacific Social Science Review*, *22*(4). https://doi.org/10.59588/2350-8329.1471

Bae, J., & Moon, D. (2025). Regime-type composition of an intergovernmental organisation and dispute settlement mechanisms. *The British Journal of Politics and International Relations*, 13691481251329774. https://doi.org/10.1177/13691481251329774

Bae, J., & Salvati, F. (2026). United Nations Discourse on Liberalism across Regime Types. *Global Studies Quarterly*, *6*(1), ksag001. https://doi.org/10.1093/isagsq/ksag001

Bartels, L. M., Daxecker, U. E., Hyde, S. D., Lindberg, S. I., & Nooruddin, I. (2023). The Forum: Global Challenges to Democracy? Perspectives on Democratic Backsliding. *International Studies Review*, *25*(2), viad019. https://doi.org/10.1093/isr/viad019

Börzel, T. A., & Zürn, M. (2021). Contestations of the Liberal International Order: From Liberal Multilateralism to Postnational Liberalism. *International Organization*, *75*(2), 282–305. https://doi.org/10.1017/S0020818320000570

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Gabry, J., Kay, M., Vehtari, A., Magnusson, M., Češnovar, R., Lambert, B., Adıgüzel, O., Socolar, J., & Kallioinen, N. (2025, February). Posterior: Tools for Working with Posterior Distributions. Retrieved September 11, 2025, from https://cran.r-project.org/web/packages/posterior/index.html

Busch, M. L. (2007). Overlapping Institutions, Forum Shopping, and Dispute Settlement in International Trade. *International Organization*, *61*(04). https://doi.org/10.1017/S0020818307070257

Bush, S. S., Cottiero, C., & Prather, L. (2025). Zombies ahead: Explaining the rise of low-quality election monitoring. *The Review of International Organizations*, *20*(2), 261–291. https://doi.org/10.1007/s11558-024-09554-3

Chou, M. (2017). Have the black knights arisen? China's and Russia's support of autocratic regimes. *Democratization*, *24*(1), 175–184. https://doi.org/10.1080/13510347.2015.1124089

Clark, R. (2022). Bargain Down or Shop Around? Outside Options and IMF Conditionality. *The Journal of Politics*, *84*(3), 1791–1805. https://doi.org/10.1086/719269

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A., Fish, M. S., Fox, L., Gastaldi, L., Gjerløw, H., Glynn, A., Good God, A., Grahn, S., Hicken, A., Kinzelbach, K., Marquardt, K. L., . . . Ziblatt, D. (2025). V-dem codebook v15 [Version 15, March 2025]. Retrieved October 6, 2025, from https://www.v-dem.net/documents/55/codebook.pdf

Cottiero, C., & Haggard, S. (2023). Stabilizing Authoritarian Rule: The Role of International Organizations. *International Studies Quarterly*, *67*(2), sqad031. https://doi.org/10.1093/isq/sqad031

Cottiero, C., & Schneider, C. J. (2026). International Financial Institutions and the Promotion of Autocratic Resilience. *International Organization*, 1–38. https://doi.org/10.1017/S0020818325101276

Debre, M. J. (2022). Clubs of autocrats: Regional organizations and authoritarian survival. *The Review of International Organizations*, *17*(3), 485–511. https://doi.org/10.1007/s11558-021-09428-y

Debre, M. J. (2025, March). The Dictators' Club: Theorizing Regional Organization Membership and Regime Survival. In *How Regional Organizations Sustain Authoritarian Rule* (1st ed., pp. 19–41). Oxford University PressOxford. https://doi.org/10.1093/9780198903635.003.0002

Debre, M. J., & Sommerer, T. (2025). Are non-democracies taking over? Autocratization and international organization membership. *Cooperation and Conflict*, *61*(1). https://doi.org/https://doi.org/10.1177/00108367251364170

Delignette-Muller, M. L., & Dutang, C. (2015). Fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*, 1–34. https://doi.org/10.18637/jss.v064.i04

Donno, D. (2010). Who Is Punished? Regional Intergovernmental Organizations and the Enforcement of Democratic Norms. *International Organization*, *64*(4), 593–625. https://doi.org/10.1017/S0020818310000202

Finnemore, M. (1993). International organizations as teachers of norms: The United Nations Educational, Scientific, and Cutural Organization and science policy. *International Organization*, *47*(4), 565–597. https://doi.org/10.1017/S0020818300028101

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

Ginsburg, T. (2020). Authoritarian International Law? *American Journal of International Law*, *114*(2), 221–260. https://doi.org/10.1017/ajil.2020.3

Gourevitch, P. (1978). The second image reversed: The international sources of domestic politics. *International Organization*, *32*(4), 881–912. https://doi.org/10.1017/S002081830003201X

Greenhill, B. (2010). The Company You Keep: International Socialization and the Diffusion of Human Rights Norms. *International Studies Quarterly*, *54*(1), 127–145. https://doi.org/10.1111/j.1468-2478.2009.00580.x

Hafner-Burton, E., Pevehouse, J. C. W., & Schneider, C. J. (2025). Good governance in autocratic international organizations. *The Review of International Organizations*, *20*(2), 323–356. https://doi.org/10.1007/s11558-024-09550-7

Hyde, S. D., & Saunders, E. N. (2020). Recapturing Regime Type in International Relations: Leaders, Institutions, and Agency Space. *International Organization*, *74*(2), 363–395. https://doi.org/10.1017/S0020818319000365

Jo, H., & Namgung, H. (2012). Dispute Settlement Mechanisms in Preferential Trade Agreements: Democracy, Boilerplates, and the Multilateral Trade Regime. *The Journal of Conflict Resolution*, *56*(6), 1041–1068. Retrieved December 6, 2024, from https://www.jstor.org/stable/23414731

Keohane, R. O., & Martin, L. L. (1995). The Promise of Institutionalist Theory. *International Security*, *20*(1), 39. https://doi.org/10.2307/2539214

Knapp, A. (2025). Twin diplomacy: Strategic co-author selection and resolution outcomes at the UNSC. *The Review of International Organizations*. https://doi.org/10.1007/s11558-025-09610-6

Koremenos, B., & Betz, T. (2012, December). The Design of Dispute Settlement Procedures in International Agreements. In J. L. Dunoff & M. A. Pollack (Eds.), *Interdisciplinary Perspectives on International Law and International Relations* (1st ed., pp. 371–393). Cambridge University Press. https://doi.org/10.1017/CBO9781139107310.019

Koremenos, B., Lipson, C., & Snidal, D. (2001). The Rational Design of International Institutions. *International Organization*, *55*(4,), 761–799. http://www.jstor.org/stable/3078615

Krasner, S. D. (2001, December). 1. Problematic Sovereignty. In S. D. Krasner (Ed.), *Problematic Sovereignty* (pp. 1–23). Columbia University Press. https://doi.org/10.7312/kras12178-003

Lake, D. A., Martin, L. L., & Risse, T. (2021). Challenges to the Liberal Order: Reflections on *International Organization*. *International Organization*, *75*(2), 225–257. https://doi.org/10.1017/S0020818320000636

Leeds, B. A., Mosley, L., Rosendorff, B. P., & Zarakol, A. (2025). The Future of Global Governance and World Order. *International Organization*, *79*(S1), S1–S11. https://doi.org/10.1017/S0020818325101185

Libecap, G. D. (2025). Distributional Issues in Contracting for Property Rights.

Libman, A., & Obydenkova, A. V. (2018). Understanding Authoritarian Regionalism. *Journal of Democracy*, *29*(4), 151–165. https://doi.org/10.1353/jod.2018.0070

Mansfield, E. D., & Pevehouse, J. C. (2008). Democratization and the Varieties of International Organizations. *Journal of Conflict Resolution*, *52*(2), 269–294. https://doi.org/10.1177/0022002707313691

Martin, L. (1994). Heterogeneity, Linkage and Commons Problems. *Journal of Theoretical Politics*, *6*(4), 473–493. https://doi.org/10.1177/0951692894006004004

Meyerrose, A. M., & Nooruddin, I. (2025). Trojan horses in liberal international organizations? How democratic backsliders undermine the UNHRC: Contribution to the Special Issue "Autocratic Regimes, Democratic Backsliding, and International Organizations". *The Review of International Organizations*, *20*(1), 125–156. https://doi.org/10.1007/s11558-023-09511-6

Moravcsik, A. (2000). The Origins of Human Rights Regimes: Democratic Delegation in Postwar Europe. *International Organization*, *54*(2), 217–252.

Nord, M., Altman, D., Angiolillo, F., Fernandes, T., Good God, A., & Lindberg, S. I. (2025). *Democracy report 2025: 25 years of autocratization – democracy trumped?* (Democracy Report 2025). University of Gothenburg. Retrieved March 5, 2026, from https://www.v-dem.net

Obydenkova, A. V., & Libman, A. (2019). *Authoritarian Regionalism in the World of International Organizations: Global Perspective and the Eurasian Enigma* (1st ed.). Oxford University Press.

Panke, D., & Stapel, S. (2023). Towards increasing regime complexity? Why member states drive overlaps between international organisations. *The British Journal of Politics and International Relations*, *25*(4), 633–654. https://doi.org/10.1177/13691481221115937

Pauselli, G., & Urzúa, M.-J. (2024). Why Autocracies Fear LGBTQ+ Rights. *Journal of Democracy*, *35*(2), 80–92. https://doi.org/10.1353/jod.2024.a922835

Pearl, J. (2016). Lord's Paradox Revisited – (Oh Lord! Kumbaya!) *Journal of Causal Inference*, *4*(2), 20160021. https://doi.org/10.1515/jci-2016-0021

Pevehouse, J. C. (2002a). Democracy from the Outside-In? International Organizations and Democratization. *International Organization*, *56*(3), 515–549. https://doi.org/10.1162/002081802760199872

Pevehouse, J. C. (2002b). With a Little Help from My Friends? Regional Organizations and the Consolidation of Democracy. *American Journal of Political Science*, *46*(3), 611. https://doi.org/10.2307/3088403

Pevehouse, J. C. (2005, January). *Democracy from Above: Regional Organizations and Democratization* [Google-Books-ID: Buu4BsTyQccC]. Cambridge University Press.

Poast, P., & Urpelainen, J. (2015). How International Organizations Support Democratization: Preventing Authoritarian Reversals or Promoting consolidation? *World Politics*, *67*(1), 72–113. https://doi.org/10.1017/S0043887114000343

Putnam, R. D. (1988). Diplomacy and Domestic Politics: The Logic of Two-Level Games. *International Organization*, *42*(3), 427–460. http://www.jstor.org/stable/2706785

Smith, J. M. (2000). The Politics of Dispute Settlement Design: Explaining Legalism in Regional Trade Pacts. *International Organization*, *54*(1), 137–180. https://doi.org/10.1162/002081800551145

Stephen, M. D. (2025). The Diffusion of Global Power and the Decline of *Global* Governance. *Ethics & International Affairs*, *39*(2), 145–158. https://doi.org/10.1017/S0892679425100129

Tallberg, J., Sommerer, T., & Squatrito, T. (2016). Democratic memberships in international organizations: Sources of institutional design. *The Review of International Organizations*, *11*(1), 59–87. https://doi.org/10.1007/s11558-015-9227-7

Tallberg, J., & Vikberg, C. (2025). Democracy, Autocracy, and the Design of International Organizations. *International Studies Quarterly*, *69*(2), sqaf034. https://doi.org/10.1093/isq/sqaf034

Thompson, A. (2006). Coercion Through IOs: The Security Council and the Logic of Information Transmission. *International Organization*, *60*(01). https://doi.org/10.1017/S0020818306060012

Thompson, W. R. (1996). Democracy and peace: Putting the cart before the horse? *International Organization*, *50*(1), 141–174. https://doi.org/10.1017/S0020818300001697

Tolstrup, J. (2014). External Influence and Democratization: Gatekeepers and Linkages. *Journal of Democracy*, *25*(4), 126–138. https://doi.org/10.1353/jod.2014.0076

Tolstrup, J. (2015). Black knights and elections in authoritarian regimes: Why and how ¡span style=”font-variant:small-caps;”¿R¡/span¿ ussia supports authoritarian incumbents in post-¡span style=”font-variant:small-caps;”¿S¡/span¿ oviet states. *European Journal of Political Research*, *54*(4), 673–690. https://doi.org/10.1111/1475-6765.12079

Van Breukelen, G. J. P. (2013). ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. *Multivariate Behavioral Research*, *48*(6), 895–922. https://doi.org/10.1080/00273171.2013.831743

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC [arXiv:1507.04544 [stat]]. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Voeten, E. (2005). The Political Origins of the UN Security Council's Ability to Legitimize the Use of Force. *International Organization*, *59*(03). https://doi.org/10.1017/S0020818305050198

Voeten, E. (2021). *Ideology and International Institutions*. Princeton University Press.

Vreeland, J. R. (2008). Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture. *International Organization*, *62*(01). https://doi.org/10.1017/S002081830808003X

Whitehead, L. (2021). Regional organizations and democratic conditionality: Family resemblances and shaming. *International Political Science Review*, *42*(4), 546–560. https://doi.org/10.1177/0192512120937474

Contribution

Jihyeon Bae: Conceptualization; Methodology; Data curation; Formal analysis; Software; Writing – original draft preparation; Writing – review and editing.

Curtis Atkisson: Methodology; Investigation; Writing – review and editing.

Appendix A Deviation of Sample Mean from Theoretical Beta Mean

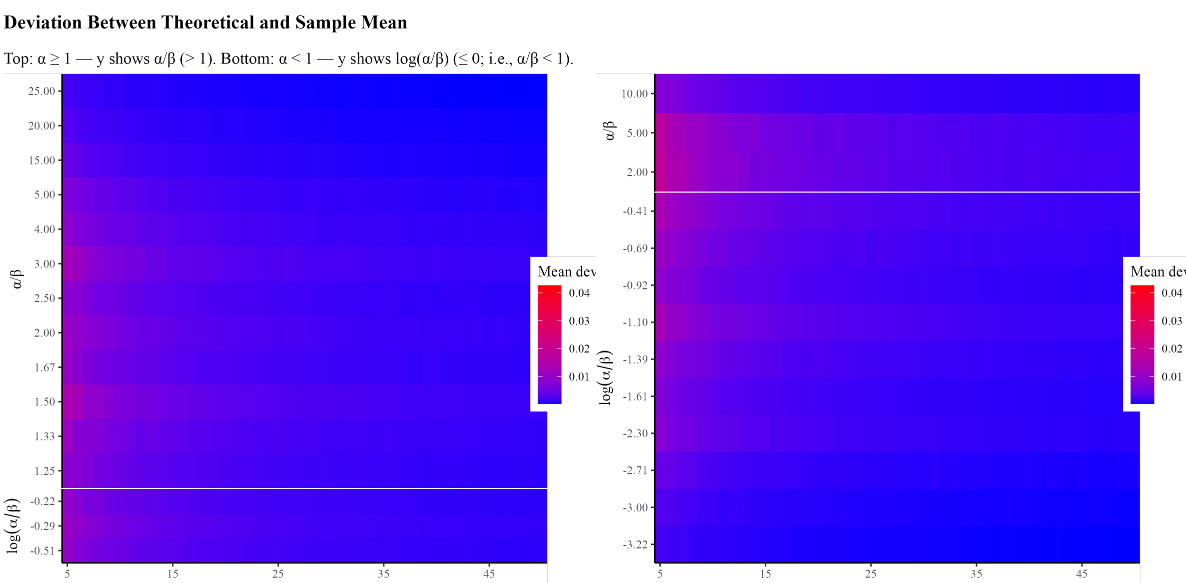


Figure A1: Deviation of the sample mean from the theoretical Beta mean across simulated draws.

Sample means can deviate substantially from the theoretical means implied by the Beta distribution, especially when both α and β are less than 1. Because U-shaped Beta distributions place most of their mass near the boundaries, small samples tend to draw disproportionate numbers of extreme values, causing the sample mean to vary widely across realizations.

Deviations also spike when the α/β ratio departs from 1, producing skewed distributions. When one of the shape parameters falls below 1, the distribution becomes highly asymmetric, concentrating probability near a boundary. In these cases, the sample mean is pulled strongly by a few extreme observations, creating large discrepancies between the empirical mean and the beta theoretical mean.

These deviations are not merely statistical curiosities. Substantively, a right-skewed distribution may contain a handful of highly democratic members that inflate the mean, while a left-skewed distribution may contain a small number of autocratic outliers that depress it. In both scenarios, the sample mean offers a misleading picture of organizational composition. The Beta framework, by contrast, models these asymmetries directly rather than obscuring them.

Appendix B Parameterization of the Beta Distribution

We use the mean–precision parameterization:

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \phi = \alpha + \beta, \qquad \alpha = \mu\phi, \qquad \beta = (1-\mu)\phi,$$

where $Y \sim \mathrm{Beta}(\alpha, \beta)$.

Appendix C Group-size contingent Sampling Variability

Figure C2 disaggregates the sampling variability comparison by IGO size group. Each boxplot shows the distribution of SD(Sample Mean) – SD(Beta Mean) across IGOs within that size category for the year 2010. Positive values indicate greater volatility in the sample mean relative to the Beta

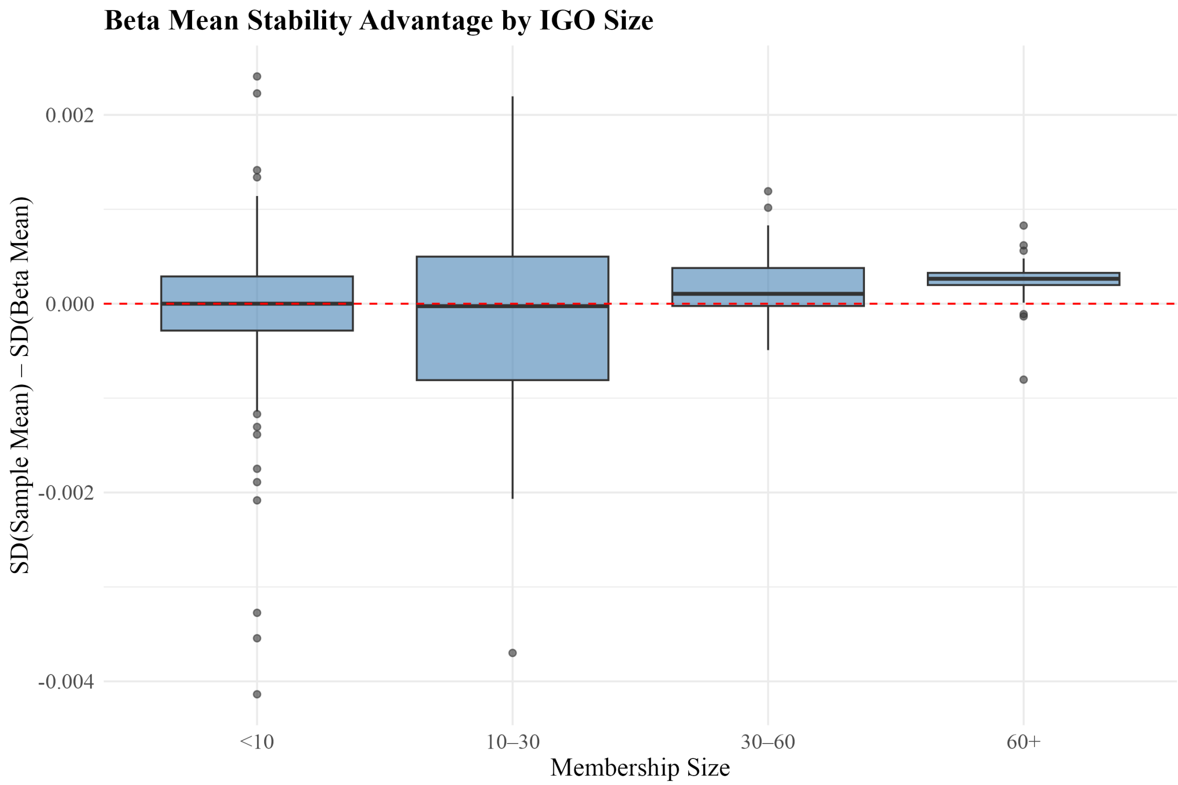


Figure C2: Difference in sampling variability between sample mean and Beta mean by IGO size group, 2010.

mean. The results show that the Beta mean's stability advantage is concentrated among larger IGOs, with the clearest gains appearing in organizations with 30 or more member states.

Appendix D Main Model Coefficient Estimates

Table 4: Main Model

Variable	Estimate [95\% Credible Interval]
v2xpolyarchy_Intercept	-2.311 [-2.420, -2.199]
Lagged Democracy	4.020 [3.963, 4.075]
Lagged Autocracy-skewed IGO Share	-0.094 [-0.123, -0.064]
Lagged Democracy-skewed IGO Share	0.870 [0.649, 1.094]
Domestic Autonomy	-0.045 [-0.064, -0.026]
International Autonomy	-0.005 [-0.023, 0.012]
Freedom of Movement	0.129 [0.119, 0.138]
Logged GDP	0.020 [0.003, 0.036]
Number of Coup Attempts	-0.179 [-0.201, -0.156]
Interpolated Log Exports	-0.018 [-0.029, -0.006]
Interpolated Log Imports	0.016 [0.004, 0.029]

Appendix E Posterior Predictive Checks

Posterior predictive (PP) plots provide visual evidence of model fit by comparing predicted values with observed data. Figure Appendix E displays the PP check for our main model.

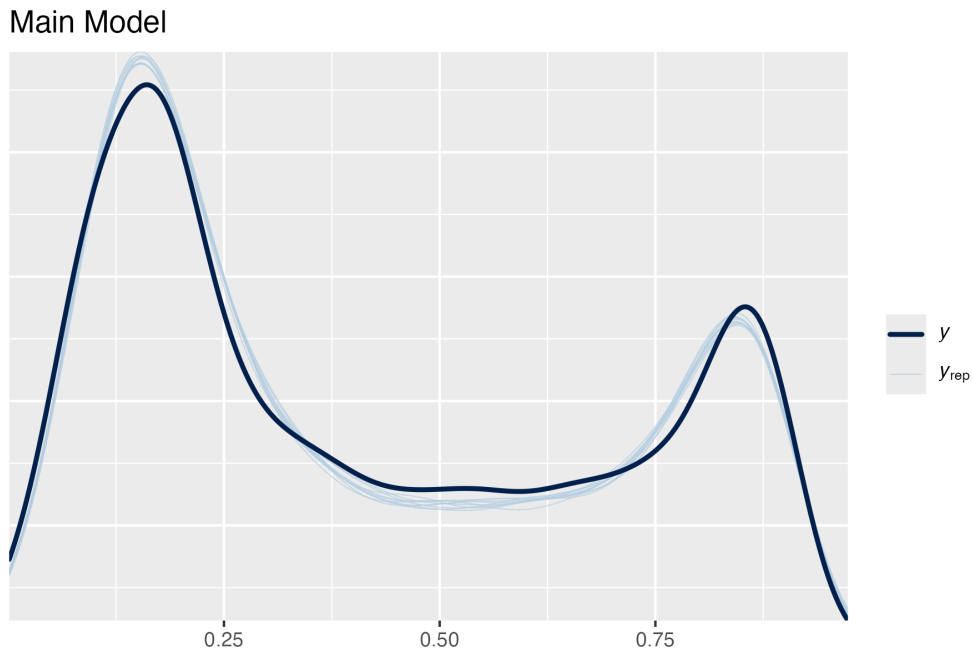


Figure E3: Posterior predictive check for the main model.

Appendix F Summary of Exposure Measures

	Probabilistic	**Deterministic**
Location	`AIGO_pct_post_lag`	`pct_location_aut_lag`
Shape	`autoc_pct_post_lag`	`pct_quad_askew_lag`
Shape + Location	—	`pct_sl_askew_aut_lag`

Appendix G LOO Comparison for Extension Models

Below tables report PSIS-LOO cross-validation results for autocratic replacement and democratization models respectively. All three measures achieve comparable out-of-sample predictive accuracy, with differences well within one standard error.

Model	ΔELPD	SE(ΔELPD)
Autocracy-skewed exposure	0.0	0.0
Autocratically centered exposure	-1.8	4.2
Baseline (AutDensity)	-4.4	3.6

Loo for Authoritarian Regime Changes.

Model	ΔELPD	SE(ΔELPD)
Baseline (AutDensity)	0.0	0.0
Autocratically centered exposure	-0.2	3.3
Autocracy-skewed exposure	-1.6	2.8

Loo for Democratization.

The Domestic Sources of International Legal Rhetoric: Contesting Sovereignty Across Regimes

Abstract

This paper introduces a theory of *regime-contingent legal appropriation*, showing that authoritarian and democratic states not only invoke international norms at different frequencies, but reinterpret their meanings to serve divergent legitimation strategies. This reframes international legal principles as rhetorical resources whose functions depend on domestic political institutions. It further contributes to the study of authoritarian international institutions by advancing the concept of "recasting" rather than originating legal norms, and by offering a systematic empirical framework for analysis. Drawing on a corpus of over 10,000 United Nations General Debate speeches (1946–2022), I argue that sovereignty is not a neutral or uniformly interpreted legal principle, but one that is strategically adapted according to regime type. Authoritarian regimes, particularly military and monarchic ones, invoke sovereignty more frequently than democracies and tie it more explicitly with Westphalian concepts of territoriality and non-interference. These rhetorical strategies are especially pronounced during periods of perceived political risk. To test these claims, I integrate computational text analysis using pre-trained word embeddings and sentiment measures with panel regression models. I construct measures of semantic distance between "sovereignty" and conceptually-related terms and show that authoritarian states consistently

interpret sovereignty in more restrictive, traditionalist terms. Furthermore, using sentiment cues to capture perceived risk, I find that authoritarian leaders are more likely to invoke sovereignty as a rhetorical shield when facing risk. By tracing how domestic political institutions shape international legal rhetoric, this paper contributes to debates on norm contestation, authoritarian legitimation, and the political use of international law.

1. Introduction

In a 2018 speech before the United Nations (UN) General Assembly, Myanmar threw down a resounding question when accused of genocide against the Rohingya Muslim minority: "We request the States Members of the United Nations to differentiate between the motives that are behind the actions of the General Assembly. Is there a genuine will to protect human rights, or is the aim to hijack human rights issues for political purposes? (UN General Debate, 2018)" [1] One year earlier, at the UN Security Council, Myanmar's then-leader Aung San Suu Kyi invoked the principle of non-interference. The regime expressed gratitude to members "who upheld the principle of non-interference in the internal affairs of sovereign countries," positioning liberal democracies' human rights concerns as violations of its sovereignty. [2] Years later, after a military coup, the junta continued to invoke sovereignty to justify its repression and silence critics at home and abroad.

There are unusual aspects about this case. Most notably, Myanmar sought to build and deliver a legal claim as strongly as possible, rather than rejecting the legitimacy of international law. This

[1] This 2018 speech ranks within the top 5th percentile in terms of the semantic distance between sovereignty and interfere, indicating one of the strongest rhetorical associations between the two terms across all observations in the corpus. I provide more technical details in the later empirical section.

[2] HTUSAN Associated Press, E. (2017, November 13). Myanmar's Suu Kyi now benefits from Southeast Asia's silence. *Associated Press: Worldstream*.

is counter-intuitive in that Myanmar could have opted out of the international system that does not serve its domestic political goal, especially when international institutions are functioning as watchdogs. Plus, Myanmar is implicitly accusing international critics of violating its sovereignty, and thus themselves acting in contravention of international law. It invoked the principle of sovereignty, in specific, which is considered one of the main pillars of contemporary international politics and order (Krasner 1999, Fowler & Bunck, 1995; Franck, 1990). Myanmar's case makes us wonder whether some aspects of sovereignty principle serve the authoritarian leaders' handbook. These general insights lead to the research question of this paper: *Do democratic and authoritarian states use prominent international legal principles, sovereignty, differently? And in particular, why and how would an authoritarian regime use international law in global forums instead of denying its authority?*

This paper argues that domestic political institutions shape how international legal principles - especially sovereignty - are invoked, interpreted, and timed. First, authoritarian states tend to invoke the sovereignty principle more frequently than democratic counterparts. They proactively use international law for their own advantage, as opposed to questioning the validity of liberal international order altogether (Lake et al., 2021; Tallberg et al., 2020). How is it possible that autocrats can benefit from the language of international law? This leads to my next argument: Authoritarian states frame sovereignty in a way that aligns well with their political interests by attaching state-centric notions like "non-intervention" and "territorial integrity" to sovereignty more closely than their democratic counterparts. Of the many historical interpretations of sovereignty, authoritarian states disproportionately favor the traditional Westphalian model which emphasizes territorial integrity without any external interruption, as opposed to more progressive interpretations. Third, I argue authoritarian states are especially likely to invoke the principle of sovereignty in a defensive term when they feel menaced by liberal actors. If this claim were right, then we should observe a more sensitive relationship between the sense of threat and changes in frequency of invocation conditional on the speaker's domestic political institutions.

Among many key principles of international law, I focus on the principle of sovereignty given its political saliency in contemporary international politics as well as in the shaping of international legal governance. While sovereignty is one of the oldest and most debated concepts in international relations and international law (Krasner 1999; Barkin & Cronin 1994; Philpott 1995; Osiander 2001), existing work has not systematically explored how political regime characteristics shape the way states deploy this principle in global fora. This study extends the theoretical discussion on sovereignty with a text-as-data approach to test three hypotheses about authoritarian rhetoric on sovereignty within a corpus of the United Nations General Debate speeches from 1946 to 2022.

By integrating computational text analysis with theory-driven hypotheses, I offer a new empirical lens for understanding how legal principles are adapted, repurposed, and invoked across different political regimes. I test my argument using a text embeddings model that operationalizes semantic distance between keywords related to the term "sovereignty." The text corpus comes from the United Nations General Debate from 1945 until 2022, which offers a unique opportunity to analyze state leaders' stances on a public forum. The text analysis shows two interesting findings. First, autocratic regimes tend to associate state-centric values of "territory," "non-interfere," "non-intervene," and their umbrella idea of "Westphalia" more closely than their democratic counterparts. Second, authoritarian states invoke the sovereignty principle more frequently, *especially* times of heightened risk.

This article makes several contributions. First, this paper develops a theory of regime-contingent legal appropriation that encompasses not just democracies but autocracies' usage of international law. It has been often intuitive to think that authoritarian states would skirt international legal norms, since it is part of the bedrock of liberal international order (Lake et al., 2021). This is not to argue that they are active supporters of international law. Rather, I complicate the current understanding of authoritarian states' relationship with international legal principle.

Second, this theoretical framework contributes to the standing debate on whether we need a separate category of "authoritarian" international law (Ginsburg 2020 a; Mallat 2020). While previous work has been focused on the innate nature of the sovereignty principle and its historical background to gauge whether it should fall into the category of liberalism or illiberalism, I switch attention to how it is deployed. I argue that it is crucial to address the question of when it is used in specific context and timing. By looking at not just original illiberal principles introduced by authoritarian states, but also how they "reboot" old principles like sovereignty, we can better formulate an understanding of *authoritarian* international law. This adds a new dimension to the growing literature on authoritarian international law (Debre 2021; Weiner 2020), shifting the focus from formal institutional design to discursive strategies that leverage foundational legal principles for legitimation.

Third, this paper offers a methodological contribution by moving beyond anecdotal evidence and explicitly testing a widely asserted assumption in the literature on authoritarian international cooperation (Bae and Moon, 2025; Tallberg and Vikberg, 2025; Cottiero et al., 2025). Scholars often contend that autocracies invoke sovereignty and non-intervention as shields against liberal scrutiny (Debre 2025, p.12), yet this claim has rarely been subjected to systematic empirical evaluation. These dynamics are conventionally difficult to test because legal rhetoric is ambiguous, context-dependent, and not easily reducible to observable variables. To address this challenge, I use a text-embedding model to measure the frequency, meaning, and timing of sovereignty invocations across regime types. This approach makes it possible to capture not only how often states refer to sovereignty but also the semantic neighborhoods in which they embed it, providing a rigorous way to assess whether autocracies mobilize sovereignty differently. In doing so, I turn a long-standing argument into a tractable empirical proposition and demonstrate how computational text analysis can bring new precision to the study of international law and organizations.

Finally, this paper offers interesting insights to broader fields of political science; authoritarian studies, international law, and norm contestation. I extend foundational work on sovereignty as a norm (Krasner 1999, Glanville 2016, Lake 2003) by introducing domestic political institution as a systematic source of variation in its rhetorical deployment. Moreover, I advance research on norm contestation in global governance (Bae & Salvati 2026, Risse et al. 1999, Simmons & Shaffer 2023), showing that sovereignty is not merely invoked differently across states but is imbued with diverging semantic content that maps onto political regime characteristics. This contributes to the literature on how different states shape and reinterpret international legal norms and institutions (Ambrosio 2008, Tallberg et al. 2020).

I trace one of the key principles, "the principle of sovereignty." This principle is especially well-suited for my research for several reasons. First, it has been invoked consistently across time and countries, offering a strong empirical basis for cross-national comparison. Second, its meaning and implications have been the subject of persistent scholarly debate; the principle of sovereignty has evolved dramatically, at times even taking on opposing interpretations. Third, it is one of the most frequently cited principles alleged (but not empirically tested) to support autocracy, particularly for the way it has been used to shield states from external scrutiny over human rights violations (Cottiero et al. 2024; Debre 2022).

This paper proceeds in the following order. I begin by reviewing a broad body of literature on the authoritarian use of international institutions, the principle of sovereignty, and authoritarian engagement with international law. A comprehensive overview of how states have used the legal language requires thorough review from various angles. I then describe the anticipated relationship between the regime type and observable implications on the frequency, interpretation, and timing of invocation of the sovereignty principle.

2. Literature on Rhetoric and International Law

Much of the existing literature on international law and domestic political institutions has centered on democracies, emphasizing how rule-based cooperation advances liberal values such as transparency, accountability, and human rights. From this perspective, international law is often viewed as both a reflection of democratic norms and a mechanism for sustaining them. Scholars have shown that democratic governments are more likely to participate in, comply with, and internalize international legal commitments (Keohane 1984; Simmons 2009; Dai 2007). International human rights law has been theorized as an extension of domestic democratic institutions, designed to lock in liberal rights protections and to reassure external accountability of compliance (Moravcsik 2000; Hathaway 2002; Hafner-Burton & Tsutsui 2005). In this view, the international legal order functions as a reinforcing layer of liberal governance, providing democracies with both normative legitimacy and credible commitment mechanisms.

Although early theories largely portrayed autocracies as peripheral or resistant to international law, a growing body of research shows that they, too, use international institutions instrumentally. Recent research shows that authoritarian states share a distinct set of expectations about intergovernmental cooperation with fellow autocrats, and that membership in such arrangements often contributes to regime survival (Debre 2021; Obydenkova & Libman 2019). One mechanism through which authoritarian cooperation can enhance incumbents' regime survival is the legitimation of otherwise contentious policies (Cottiero & Haggard 2023). Legal language embedded in an organization's charter also reinforces the contours of international law in ways that favor incumbent regimes by reducing the perceived responsibility of other states to intervene in cases of human rights violation (Weiner, 2020). Even approaches generally skeptical of international law acknowledge the importance of rhetoric in sustaining the international legal order. State leaders need legalistic and moralistic rhetoric to signal their preferences to domestic audience and international citizens (Goldsmith & Posner 2005).

Yet, existing debate on authoritarian states' engagement with international law has been focused on whether they introduce "new" illiberal principles. When Ginsburg (2020 b) introduces *authoritarian international law,* it is conceptualized as a political tool for diffusing authoritarian rule across borders (221). Some of the key principles, such as separatism, extremism, and sovereignty, are highlighted as central to the distinct legal repertoire of authoritarian regimes. While Ginsburg's contribution has opened an important line of inquiry, subsequent scholars have raised questions about the origins and normative characteristics of these so-called "illiberal" principles. Notably, Mallat (2020) argues that norms like non-intervention and Westphalian sovereignty have long histories as global, even liberal, doctrines, and that counterterrorism initiatives have often been championed by democracies themselves. The justificatory pattern can appear in democracies, as illustrated by the United States' reliance on security discourse to legitimize counter-terrorism measures during the War on Terror.

By shifting attention from *who creates* legal principles to *how* different actors use existing ones, this paper addresses the apparent tension between authoritarian reliance on international law and the liberal origins of many of the principles they invoke. Rather than focusing solely on legal texts or institutional design (Bae & Moon 2025; Tallberg & Vikberg 2025; Koremenos et al., 2001), I examine how international law is *employed* in practice. This approach helps reconcile a central paradox; authoritarian states use international law to legitimize and extend incumbency, yet the very principles they invoke, most notably sovereignty, remain foundational to the liberal international order embraced by democracies.

Recent studies have identified similar examples of non-intervention and sovereignty (often used with similar meanings) when characterizing authoritarian *uses* of international law (Flonk & Debre 2025, Hall et al. 2025). Despite growing work, there has been no large-scale, cross-regime, empirical analysis testing whether autocrats, in practice, emphasize sovereignty more often and strongly than democracies, or whether they invoke it in ways consistent with the notion of "unlimited sovereignty" (Cottiero et al.,2025).

3. Theory and Hypotheses

3.1. Principle of Sovereignty as Authoritarian Playbook

Sovereignty is a deeply contested concept, both in theory and in practice. This paper does not seek to definitively reconcile its many meanings or endorse any singular definition. Instead, it traces the empirical patterns of how states invoke and deploy the concept of sovereignty in contemporary international discourse. In this section, I introduce mechanisms through which the sovereignty principle can benefit authoritarian survival. This part builds up to the first hypothesis that sovereignty will be invoked *more frequently* by authoritarian states than democratic counterparts, all else equal.

What is sovereignty? Scholarly and diplomatic definitions typically center on two core dimensions; the internal authority of a state over its territory, and protection from external interference (Lake, 2003; Philpott, 2010; Hinsley, 1966). Existing interpretations vary widely, particularly in terms of how much authority a state is presumed to exercise domestically and how strongly it is insulated from outside interference.

In line with this broader concept, Krasner (1999) identifies the logic of Westphalian sovereignty as a constant balance between nonintervention and effort to alter other states under the justification of human rights or capitalism[3]. In this framework, Westphalian sovereignty is defined by two core elements; territoriality and autonomy, grounded in the absence of external authority over domestic affairs. While Krasner (1999) argues that the logic of consequences routinely leads powerful states to violate the model, this paper departs from his emphasis on imposition. Instead, it contends that international legal norms are not merely enforced upon weaker states but are also appropriated and purposefully reshaped by them. Nonetheless, this paper shares the general view that the evolution of sovereignty is shaped by instrumental calculations. My contribution lies in demonstrating that these calculations are not only about power politics but also domestic political institutions.

Several characteristics of the sovereignty principle make it appealing to autocrats. First, it is multifaceted. To address transnational issues, states grant some degree of authority to supranational entities by signing onto binding treaties. To improve domestic policy outcomes and maximize gains from cooperation, many state leaders forgo absolute sovereignty. This creates a constant tension between enhancing collective action and preserving national control. Second, the significant reinterpretation of sovereignty by democracies, especially in the context of

[3] Krasner (1999) stresses that there are multiple types of sovereignty.

Responsibility to Protect (R2P) debates and interventions, has made it easier for autocrats to shape the discussion (Bellamy et al. 2016, Wheeler 2000). The boundary between respect for national autonomy and the enforcement of human rights has been ceaselessly contested, especially in the Non-Governmental Organization (NGO) sphere (Hafner-Burton 2013, 64-86). Owing to the unique development history of the sovereignty principle, it is one of the few authoritative legal principles that allows autocrats to reshape it with relatively greater flexibility.

Second, unlike any other social norms, the principle of sovereignty holds a privileged status as both a codified and customary norm in international law. As a result, sovereignty is invoked with sufficient frequency to generate observable empirical implications. For example, during the United Nations Human Rights Council's Universal Periodic Review (UPR) sessions, China has repeatedly invoked the principle of sovereignty and non-interference in response to criticism of its policies in Xinjiang, Tibet, and Hong Kong. In its 2018 national report, China stated that it "firmly opposes interference in its sovereignty and internal affairs under any pretext" and emphasized that "the request to release those who are under compulsory measures … is an interference in China's judicial sovereignty." (Human Rights Council, 2019) This episode is representative of a broader pattern in which authoritarian states invoke sovereignty when confronted by international legal bodies over alleged human rights violations.

hetorical conflict over how to interpret and apply international legal principles is not a new phenomenon. State leaders often contest the validity of a legal question by disputing facts, challenging the legitimacy of the law itself, or arguing that the legal provisions invoked are inadmissible to the issue at hand (Morse & Pratt 2025). Similarly, Búzás (2018) demonstrates how actors exploit the gap between law and norm, adhering to the legal commitments on the book, while evading their underlying spirit. These studies suggest that rhetorical evasion is a broader political practice rooted in the interplay between domestic legitimation needs and international scrutiny.

While existing studies have primarily documented rhetorical evasion among democratic states, authoritarian regimes are by no means exempt from this tendency. Indeed, the strategic invocation of the sovereignty principle tends to be, this paper argues, *more* pronounced in autocracies, where leaders face stronger incentives to use international law defensively. *If* sovereignty were a tool in the authoritarian leader's playbook, we would expect to observe a higher frequency of its usage by authoritarian states. This is because authoritarian regimes tend to face international scrutiny, particularly from liberal democracies, due to rampant violations of civic, political rights and electoral norms (Hyde 2011; Simmons 2009; Bush et al., 2024). Koliev (2020) demonstrates that democracies are more likely to care about international norms and condemn norm violators, given their sensitivity to the demands of domestic groups and commitment to international standards. It is also shown that non-democracies are more likely to be criticized for violating international norms than democracies (Hendrix & Wong 2013). Given that authoritarian states are more likely to be in the position of defending themselves from such criticism, it follows that defensive rhetoric should be considered worthy of devoting speech time at global forum like the United Nations General Assembly.

As such, authoritarian leaders frequently frame sovereignty as a counterweight to human rights. It serves authoritarian interest to frame the principle of sovereignty is a competing value against human rights and rule of law (Cooley, 2015). Risse et al. (1999) similarly point out that "repressive state denies validity of human rights norms as subject to international jurisdiction and claims non-intervention norm" (703).

For instance, in 2000, Myanmar delivered a speech that heavily borrows the cloak of international law: “The principles enshrined in the Charter have served the world community well for more than five decades. Sovereignty, sovereign equality, respect for territorial integrity and non-intervention in internal affairs are cardinal principles that remain vital for the peace and security of all nations… [T]he justification of interference in internal affairs under certain conditions is a concept easily vulnerable to abuse by the powerful for their narrow national interests.”

In such moments, the principle of sovereignty becomes a particularly potent defense, offering a legal and normative shield for policies that might otherwise draw international condemnation. If authoritarian leaders systematically invoke sovereignty to resist external scrutiny, and if they are disproportionately targeted by international shaming, we should therefore expect them to deploy sovereignty-based rhetoric more frequently than their democratic counterparts.

This argument does not assume that democracies are inherently more principled on human rights. Rather, authoritarian regimes are more likely to adopt a defensive posture because they face greater international scrutiny. In contrast, democracies can more easily justify interventionist behavior under alternative norms. The Bush Doctrine, for instance, framed sovereignty as secondary to spreading democracy, legitimizing breaches of non-intervention when dealing with authoritarian regimes (Russett, 2005). Thus, while democracies may invoke other values to justify external action, authoritarian regimes are more likely to fall back on sovereignty. This descriptive statement is a baseline proposition that is maintained throughout the paper.

H1 (frequency): Authoritarian states invoke the principle of sovereignty more frequently than democracies in international fora.

3.2. Contested Sovereignty: Diverging Authoritarian and Democratic Interpretations

I further contend that authoritarian states attach different meanings to the sovereignty principle than democratic states. This extends the baseline proposition that focuses on the sheer frequency of invocation by looking at even greater disparity attached to the same principle. The goal of this section is to briefly sketch the different versions of sovereignty have existed in contemporary world politics and theorize that autocracies have their favorite version and try to revert to it. Authoritarian states tend to prefer reverting back to what is known as “Westphalian sovereignty,” which puts domestic political autonomy before anything else. Authoritarian states are more likely than democracies to present sovereignty within the context of a strict non-intervention principle.

Earlier discussion on sovereignty from the 16th Century to the Peace of Westphalia in 1648 reflects a general idea that states have *unrestricted* authority over the territory. Paris (2020) explains that earlier theorists like Jean Bodin in the 16th Century envisioned sovereignty as a centralized authority within a bounded territory, *unrestrained* by law. As early as then, it has been continuously debated if sovereignty can be conflated with unlimited freedom of action in the domestic realm and with complete absence of accountability (Glanville, 2016). Early political philosophy focused on the relationship between state and society when defining sovereignty.

In the literature of international relations, the principle of sovereignty began to be understood as a central concept after the Peace of Westphalia in 1648, often presented as the start of contemporary international state system. After the Thirty Years war, the very idea that each state is equally sovereign within its territory and should mutually respect each other’s domestic affairs. As many have noted, the Peace of Westphalia - consisting of Treaty of Osnabruck, and Treaty of Munster- introduced the notion that rulers can set the religion of their own state without external meddling. It is often understood as the idea that "basic attributes such as the existence of a government with control of its territory were now, along with Christianity, the criteria for becoming a state… [T]he Treaty removed all legitimate restrictions on a state's activities within its territory (Philpott, 1995, p. 364)." The United Nations Charter Article 2 (4) later codified a general ban on the threat or use of force against states, consolidating a non-intervention expectation that pre-dated 1945 in customary practice.

Norms around what constitute statehood *after* the Peace of Westphalia attained a status as revolutionary idea of sovereignty (Philpott, 2010). Westphalian sovereignty certainly held an important position in the shaping of contemporary international politics, but was not sacrosanct. European countries delegated policy autonomy to supranational authority through a series of regional integration, and non-interference principle was collectively and formally violated when the United Nations sanctioned intervention with the goal of upholding standards of human rights practices.

The balancing between preservation of state's domestic autonomy and the goal of achieving human rights protections has become more salient especially after the Second World War, departing from Westphalian sovereignty (Krueder-Sonnet et al., 2014). The conflict between desire to regulate human rights violation and securing domestic authority has continued to grow and manifested itself into the tangible form as the debate around military intervention has sparked after the Cold War (Cohen and Deng 2016). Scholars have started emphasizing the contested nature of sovereignty by pointing out states' right to rule should not be unlimited (Barkin and Cronin 1994). With growing discourse on the responsibility to protect (R2P), the degree of authority granted to each state by the legal principle of sovereignty dwindled over time. Looking at the brief history of long-standing balance (or conflict) between sovereignty and competing values brings a skeptical lens to the idea that the non-intervention principle in and of itself contains an "illiberal" intention to mask human rights violations.[4]

During decolonization, sovereignty was invoked as a language of emancipation. Liberation movements and newly independent states framed sovereignty through self-determination, territorial integrity, and freedom from external domination, using the norm of non-intervention to shield anti-colonial projects and state formation. This anticolonial repertoire positioned sovereignty as a claim against empire, not as a license for domestic impunity. It is for this reason, that sovereignty's uses often diverge from authoritarian survival, that this paper's approach is necessary to test how meanings vary across regime types.

So far, I have outlined various interpretations of the sovereignty principle. The historical overview serves to contextualize the menu of available meanings that states may draw from when invoking sovereignty. Among different meanings of sovereignty that have existed in history, authoritarian states favor a reversion to the Westphalian model, emphasizing rigid non-interference and domestic political autonomy. In other words, authoritarian states envision the world view where the idea of responsibility to protect, and restriction of policy autonomy for the sake of human rights were absent. As the earlier interpretation of sovereignty grants more domestic autonomy vis-a-vis competing liberalization values, it aligns more closely with the political aims of autocrats. Simmons and Gorman's (2021) point out that the "older" sovereignty-based order conflicts with the liberal international order, as the former emphasizes territoriality. In the border-based sovereignty narrative, the state and society function along physical borders, which diminishes the

[4] Sovereignty is contested even within the single legal document. Section 1 in the preamble of the Final Act of the Helsinki Conference (1975) emphasizes states' responsibility to respect other's sovereign equality, including right to freely choose domestic systems and its own laws and regulations. Section 4 further emphasizes the principle of non-intervention by mandating states to respect territorial integrity. On the other hand, section 7 obliges states to respect the universal significance of human right, pleading for "endeavor jointly and separately,…, to promote universal and effective respect."

influence of liberal agendas promoted by the international community. An understanding of sovereignty anchored in territorial boundaries and state autonomy within those borders can serve to justify autocrats' domestic policies, no matter how repressive they may be. This leads to the second argument: Authoritarian states are more likely to associate sovereignty with the core ideas of the Westphalian model.

Considering the tension around the sovereignty principle, it can be unsurprising that state leaders with divergent political goals show contrasting usages of it. It is less of the inherently malicious intention of authoritarian state leaders to diffuse their novel ideology, but more of a theoretical ambiguity on a set framework of sovereignty that leads to "authoritarian" international law, different from "democratic" international law. It is less the case that authoritarian states invented the new *illiberal* principle of non-intervention.

While this section referenced the historical evolution of the sovereignty principle, the core claim is not about temporal change *per se*, but rather about cross-regime variation in how states interpret sovereignty at a given time. Specifically, authoritarian regimes disproportionately favor a more restrictive conception of sovereignty that emphasizes non-intervention and domestic autonomy within borders.

Examples abound. In its 2005 General Debate speech, Myanmar declared: "Efforts and strategies to tackle transnational issues need to be coordinated at all levels and carried out with the necessary respect for the sovereignty and territorial integrity of States … The strength of our Organization lies in its universality and legitimacy, anchored in the Charter of the United Nations ... Among these cardinal principles are respect for the sovereignty, territorial integrity and equality of States, non-interference in their internal affairs and the non-use or threat of use of force.

(UNGD, Myanmar, 2005)" [5] This statement foregrounds sovereignty as a defensive norm, placing it before, and in contrast to, broader transnational agendas, thereby reinforcing the regime's emphasis on autonomy and resistance to external scrutiny.

Cambodia voiced a similar position more than a decade later. In 2018, its representative told the General Assembly:[6] "I would like to draw the attention of all States Members of the United Nations to the vitality of the Charter of the United Nations. We must all give the proper respect to that crucial document by avoiding interference in or damage or disruption to the sovereignty of independent States. We regretfully have to highlight the fact that the issue of human rights nowadays has become the way some powerful nations strive to impose civilization on other nations, or, according to their operating standards, as a pretext for interfering in the name of protecting political rights."

These anecdotal examples reveal how governments reassert sovereignty as a barrier against external criticism, portraying human rights as a competing value for sovereignty principle. Such rhetoric illustrates the broader claim of this paper: Authoritarian regimes reinterpret sovereignty as a defensive instrument, anchored in the logic of non-intervention, to preserve regime autonomy and legitimacy. This brings us to the second hypothesis.

H2 (Emphasis on non-intervention): Authoritarian states attach state-centered values to the principle of sovereignty to a greater extent than democracies.

Before moving on to the next section, two major critiques might arise regarding the current depiction of Westphalian sovereignty. Historians debate whether the Peace of Westphalia should

---

[5] Myanmar's 2005 speech ranks in the top 7th percentile in terms of the semantic distance between *sovereignty* and *interfere*. More discussion on the semantic distance follows in the methods section.
[6] Cambodia's 2018 speech ranks in the top 2nd percentile in terms of the semantic distance between *sovereignty* and *interfere.*

be treated as a watershed moment for statehood, and whether it encompasses the idea of territorial autonomy at all. Osiander (2001) questions the historical foundations of the Westphalian narrative altogether, pointing out that the treaties of Münster and Osnabrück make no explicit reference to sovereignty, territoriality, or legal equality. Relatedly, Krasner (1999) argues that Westphalian sovereignty has never been an empirical reality: Great powers routinely violate norms of non-intervention when it suits their interests.

To engage with each of these claims, I make two points. First, in conjunction with Krasner's critique, I do not argue that the Westphalian model accurately captures historical reality, but rather that it persists as a powerful rhetorical and conceptual resource, especially for autocrats seeking to deflect external scrutiny and resist evolving interpretations of sovereignty that prioritize human rights and democratic norms.

Second, showing whether it is historically accurate that Westphalian system includes territoriality and non-intervention as core elements is not critical to the overall argument. It is the gradual detachment from the most restrictive version where states have *unlimited* control over their societies that build up to the claim that autocrats would try to revert back to more restrictive definition. I echo Philpott's (2010) defense against a claim that Westphalia cannot be conflated with non-interference. While acknowledging that the principle of non-intervention is not explicitly mentioned in the statement, he points out that non-intervention became standard and expected state practices as intervention from the Habsburg dynasty lost any legitimacy after the Westphalia settlement.

3. 3. Timing of invoking the sovereignty principle

The core argument so far is that authoritarian states lean on the sovereignty principle, particularly its emphasis on non-interference and territorial integrity. Under conditions of heightened international risk, invoking sovereignty offers a way to assert regime legitimacy, reframe external scrutiny as unjustified interference, and hence deflect blame. While the sovereignty principle is a central feature of international legal discourse, we still know relatively little about when states choose to invoke it. If states use legal principles strategically, we should expect variation in the timing of invocation based on political conditions.

I hypothesize that perceived political risk increases the likelihood of sovereignty invocation, especially for authoritarian regimes. This argument builds on scholarship linking domestic insecurity to repressive and defensive behavior. For instance, states are more likely to violate personal integrity rights during wartime or under conditions of internal political conflict (Poe and Tate 1994; Davenport 2007). The logic here is similar to the multiplier effect of hypothesis 1: When leaders face external threats and engage in state violence, their heightened vulnerability to international criticism increases their incentives to foreground sovereignty in political discourse.

Periods of crisis heighten the tension between international obligations and domestic imperatives, making legal justifications especially attractive to authoritarian state leaders. By relying on legal instruments, governments can frame controversial measures as lawful responses to exceptional circumstances, thereby avoiding the appearance of outright lawlessness. Experimental evidence shows that the noncompliance costs declines when governments frame violations as applicability challenges; that is, when they argue that the legal principle in question does not apply to the specific circumstances as it conflicts with other competing values (Morse &

Pratt 2025). Even the most repressive leaders benefit from invoking international law in ways that appear consistent with prevailing global norms, using legality itself as a form of rhetorical legitimacy. The appeal to law functions as a low-cost signaling mechanism for the "denial of self-interest" (Goldsmith and Posner 2005, 157), allowing states to convey to both domestic and international audiences that their motives behind contested behavior are neutral. By contrast, failing to do so rhetorically risks signaling the opposite.

I assume that the same external shocks render different degrees of perceived risk, depending on the domestic structure. There are several reasons for prioritizing a subjective measure of risk perception. First, it serves the purpose of the research to unravel ambiguity around legal principles even though they might be spoken the same at face value. To accurately understand the source of variation in legal rhetoric, it is crucial to take a consistent approach that tackles the assumption that the observed phenomenon implies the same underlying preferences.

Second, there are numerous confounders between observable exogenous shocks and risk perception when conducting cross-national analysis. States experience different numbers and magnitudes of civil conflicts, economic setbacks, and political changes, on top of global shocks like pandemic or climate change. While it might be possible to include these observable factors as covariates, such shocks often bring about drastically distinct domestic political repercussions depending on institutional setting. For example, insular countries are more susceptible to climate-related shocks, while those with weak domestic legal systems can suffer more from civil war than others with rigid institutions.

H3 (Risk Amplification): Authoritarian regimes increase their invocation of sovereignty under heightened perceived risk.

Thus far, the argument has treated authoritarian regimes as a unified category. Yet existing scholarship on authoritarian politics cautions against this aggregation. There is another key channel through which risk translates into political behavior, variation in institutional constraints across authoritarian subtypes. I argue that differences in institutional structure moderate how authoritarian leaders respond to external risk, including their reliance on the sovereignty principle as a rhetorical resource. Because institutional constraints shape how regimes absorb, deflect, or redistribute pressure, they also shape the extent to which leaders seek to mitigate risk through legal rhetoric rather than material or procedural responses.

Importantly, this argument does not claim that all autocracies face risk in the same way, nor that they possess identical tools for managing it. Empirical work shows that authoritarian regimes differ systematically in their infrastructural power and internal constraints, shaping how they absorb and respond to pressure (Lai and Slater 2006). Gerschewski (2013) conceptualizes authoritarian survival as resting on a combination of repression, co-optation, and legitimation, emphasizing that these pillars are unevenly available across regimes. While his framework applies to authoritarian rule broadly, it also explicitly allows for within-authoritarian variation in how regimes manage vulnerability.

In a similar vein, studies propose that authoritarian regimes adjust the balance among material, institutional, and symbolic strategies in response to external pressure, depending on their constraints and available resources (Svolik 2012; Levitsky and Way 2010). As a result, even when confronted with similar levels of international scrutiny, dictators may differ in their reliance on sovereignty rhetoric as a legitimating strategy, depending on the institutional tools at their disposal.

While there has been no direct test of within-authoritarian variation in rhetorical strategies, two strands of literature offer suggestive insights. First, work on authoritarian regime types emphasizes differences in how military, monarchical, and civilian dictatorships organize authority and coordinate elite support (Cheibub et al. 2010; Gandhi et al. 2007). Military and royal regimes often possess well-defined coercive or dynastic institutions but face weaker international legitimacy, particularly when power is obtained through non-electoral means. Civilian dictatorships, by contrast, are more likely to rely on nominal party systems or electoral institutions that provide alternative forms of procedural legitimation (Ezrow and Frantz, 2011; Boix and Svolik, 2013).

These differences suggest that authoritarian subtypes may not rely equally on sovereignty rhetoric when international scrutiny rises. For military and monarchical dictatorships, invoking sovereignty offers a low-cost, norm-compliant signal that can counteract diplomatic marginalization without requiring substantive concessions. Civilian dictatorships, especially those with party or electoral façades, may already engage in legitimacy mimicry that reduces the need for overt reliance on sovereignty language.

Second, research on authoritarian foreign policy highlights differences in risk tolerance and responsiveness across regime types, particularly between military and civilian dictatorships (Weeks 2012; Snyder 1984). This strand of research theorizes that civilian dictators are not that different from democracies in their aversion to initiating costly conflicts, whereas military dictators (particularly personalistic ones) are more prone to belligerence due to the costs of inaction. Although developed to explain variation in conflict behavior rather than discourse, this work underscores a broader point: Institutional subtype conditions how leaders evaluate responses to external pressure.

What remains unclear is how such institutional differences translate into rhetorical behavior in low-cost international settings. Unlike diplomacy or military action, speechmaking in forums such as the UN General Debate involves minimal material stakes, and existing theories do not specify how authoritarian subtypes should differ in their use of legal rhetoric. While theories of authoritarian foreign policy focus on material behavior, this study evaluates whether their core implication that institutional subtype conditions responses to external pressure extends to rhetorical reliance on international legal principles. The following hypothesis therefore examines whether institutional differences among autocracies are associated with systematic variation in their reliance on sovereignty rhetoric under risk.

H3b (conditional responsiveness): *Among autocracies, military and monarchical regimes exhibit a stronger increase in sovereignty invocation under heightened perceived risk than democracies.*

4. Data and Method

UNGA Speech Data: What do the leaders say and how do they speak?

Every year, more than 200 state representatives open up the United Nations General Assembly meeting with statements during the General Debate (UNGD) over several days. Unlike other units of the UN, the UNGA comprises all of the members and decides on key issues like the UN budget approval, appointment of the Secretary General, and election on the non-permanent members of the UN Security Council. As one of the first parts of the UNGA session, UNGD statements are known as the "barometer of international opinions" (Smith 2006, p.155) because these are rare opportunities for small states to have their voices heard to a wide range of audiences. All speaker states are expected to keep a voluntary 15-minute time limit for speech. Often times, countries who speak in languages other than the UN official language provide their own translators. If not, the UN staff provide a post-speech translation for documentation. In various aspects, UNGD statements offer standardized procedures for all countries with consistent format. These also play a crucial role in setting the agenda for the remaining sessions and incentivize states to put efforts in condensing their viewpoints in a concise way. Since sovereignty and human rights are one of the most important and widely contested topics across space and time, UNGD can offer hints on which one states prioritize given the limited amount of time allocated equally to all. Hence, it clarifies how states carefully balance competing values.

I use the corpus of 10568 speech transcripts, collected by Baturo et al. (2017), covering more than 200 countries from 1946 until 2022. The UNGD corpus has recently gained popularity in the studies of political science owing to its substantive significance.[7] While I join burgeoning efforts to incorporate computational text analysis into the field of international relations, I contribute to the current computational work on the UNGD corpus data in many ways. Most notably, I connect pre-trained language model back to theory-deduced hypotheses testing. While still taking advantage of vastness of training data already used to build pre-trained model, I test specific hypotheses that allows me to conduct conventional statistical tests on a fine-grained level.

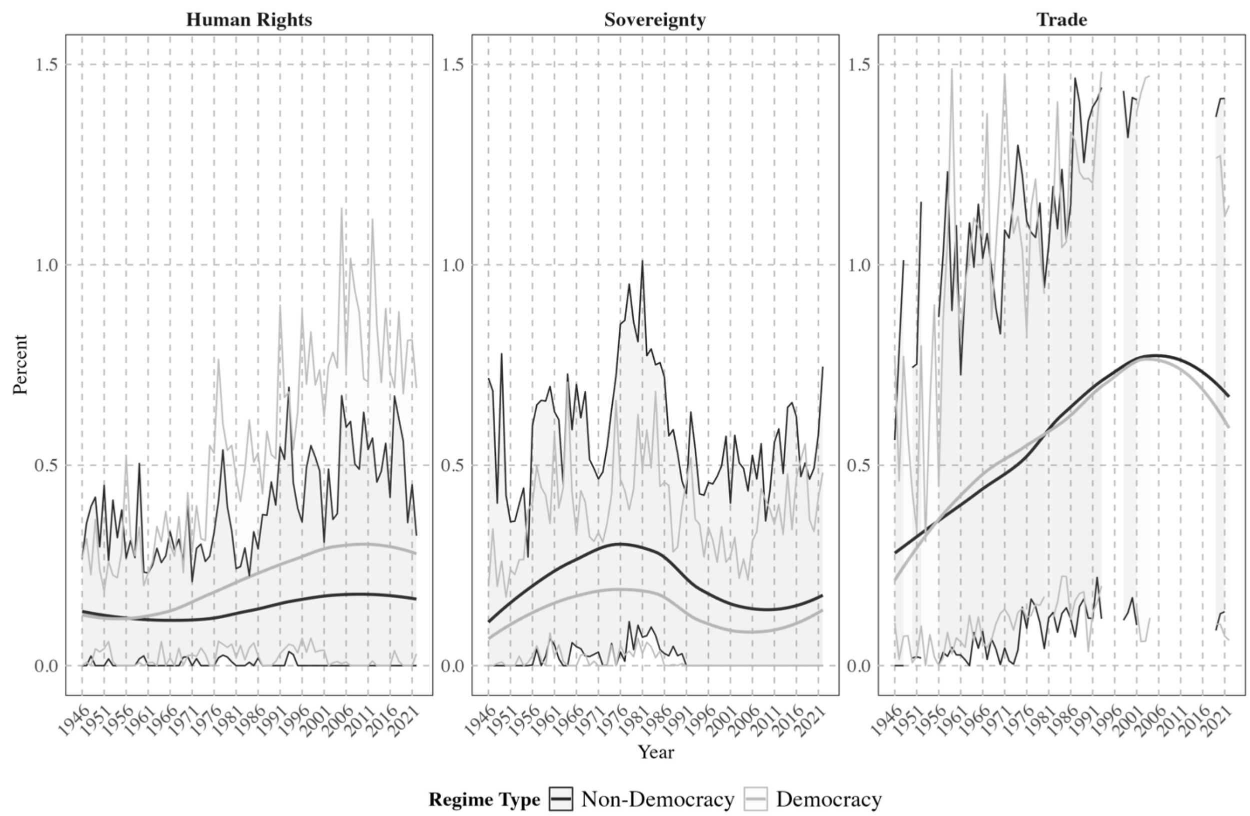
[7] For example, Simmons and Shaffer (2023) trace the international discourse around border anxiety, expressed as a tension between domestic sovereignty and global governance. Turco(2024) used an unsupervised approach by showing structural topic models around global climate governance. Hosli and Kantorowicz(2024) has used pre-trained WordScore to trace policy convergence among the EU member states. Similarly, Chelotti et al., (2022) leveraged WordScore to measure state preference convergence with the EU positions and showed socialization effect through international organizations.

4.1. Frequency of International Legal Rhetoric

If the sovereignty principle were agnostic, we would not expect any systematic variation in its usage. The most intuitive way to detect such variation is to examine how frequently the term appears. To account for variation in speech length, I calculated the proportion of mentions from three keyword-families, *sovereignty*, *rights*, and *trade*, relative to the total number of words in each UN General Debate statement. Each keyword-family represents a small dictionary of related expressions that capture the broader semantic field of each concept (for instance, *sovereignty* includes "non-intervention," "non-interference," and "territorial integrity," while *rights* includes "human rights," "civil rights," and "political rights").[8] This normalization ensures that the measure reflects rhetorical emphasis rather than the sheer volume of speech text. These keyword-families were selected because they appear consistently across years and countries, providing a stable basis for temporal comparison.

In [Figure 1], the darker line represents observation points of autocracies, operationalized by the V-dem dataset's electoral democracy score smaller than or equal to 0.5 (Coppedge et al., 2025).[9] Electoral emphasis of the polyarchy index matters for this paper's argument that sovereignty invocations function as defensive rhetoric among authoritarian states. The strength of this response varies with institutional capacity to transmit constituent demands, where that conduit is weaker, the reaction is stronger.

[8] [Appendix 9] provides the full list of dictionary terms. The procedure for computing dictionary-based proportions replicates the mftPercent() function from Kraft et al. (2024), with minor adjustments.
[9] [Appendix 1] shows the proportion of keyword appearance across six-fold regimes as a robustness check (Gandhi et al., 2010).



[Figure 1] Proportion of keyword appearance per speech

Compared to relatively neutral terms like *trade*, the difference in the frequency of keyword-families across regime types was pronounced. Autocracies consistently invoked the *sovereignty* family more often than democracies across all years. In contrast, the *human rights* keyword families were invoked more often by autocracies than democracies. The relative prominence of *human rights* references among democracies became increasingly pronounced after 1960, whereas autocracies showed little change over time. By contrast, references to *sovereignty* remained persistently higher in authoritarian speeches. Because all countries face the same time constraints in the General Debate, these proportions capture the relative rhetorical priority that states allocate to each concept. Hypothesis 1 is well supported with descriptive statistics.

4.2. Semantic Meanings around the Sovereignty Principle

Descriptive data shows that authoritarian states, especially with individualistic leaders, tend to invoke sovereignty more often than democracies. While raw term frequencies offer a useful starting point for understanding the prominence of sovereignty and human rights in international discourse, they fall short of capturing different meanings on what democracies and autocracies attach to the same term. The main analysis focuses on disentangling ambiguity around the principle.

To compare rhetorical patterns between regime types in a more rigorous way, I used the Global Vectors for Word Representation (GloVe) (Pennington et al. 2014) model, trained based on Wikipedia sources. It is a static language model that is developed by features of the entire corpus and co-occurrence matrix of terms (Jurafsky et al. 2024, 23). Embedding refers to vector representation of words and whose representation is trained by testing how likely it is for one word to be a neighbor with the target word. Word embedding models add a distinct methodological contribution to the study of international legal rhetoric because they enable systematic measurement of meaning rather than mere frequency (Smith 2020).

Unlike bag-of-words or dictionary counts that treat "sovereignty" as the same token, embeddings recover the *meaning* a term acquires from its contexts. I project UN speeches into this vector space and quantify how the semantic neighborhood of "sovereignty" shifts across regimes and over time. This yields three advances over bag-of-word approach. First, it measures contestation as semantic re-signification rather than mere frequency. Second, it provides commensurable, continuous distances that support panel designs with fixed effects. Third, it enables theory-driven tests of how autocracies versus democracies mobilize the same vocabulary. In short, the approach turns debates about "what they mean by sovereignty" into observable, falsifiable quantities.

I assess whether the way sovereignty is invoked by different regime types reflects general semantic proximity to concepts like territory, colonial, or Westphalia. While GloVe is trained on a general corpus, its semantic space offers a consistent reference point for comparing rhetorical strategies across states. Pre-trained models like GloVe leverage the volume of billions of tokens of Wikipedia that are used for training, capturing numerical representation of our natural language. The main advantage of the GloVe model is that it captures the underlying "rhetorical neighborhood" of a term by analyzing the constellation of concepts and frames that tend to appear alongside it. The GloVe model allows us to systematically identify and compare these semantic patterns across regime types and over time, providing a contextualized understanding of how key norms are invoked in international settings. This approach helps illuminate not just *how often* states talk about sovereignty or human rights, but *how* they talk about them, and whether those rhetorical strategies diverge by political system. The theory argues that autocracies employ the sovereignty principle by shifting meanings rather than denying them. By measuring semantic proximity, the method captures exactly that mechanism and shows when and for whom it occurs.

After constructing semantic similarity measures using GloVe, I run time-series cross-sectional regression models to test whether different regime types are systematically associated with distinct rhetorical uses of sovereignty. This modeling strategy allows me to assess how meaning shifts across political systems, while controlling for temporal shocks.

There are five dependent variables, each representing semantic distance generated by the GloVe model. These are pairs of semantic distance index between sovereignty and 1) "colonial," 2) "territory," 3) "Westphalia," 4) "intervene" and 5) "interfere." To examine how different regime types invoke sovereignty in relation to core normative frames, I compute the semantic proximity between each UN speech and keyword pairs related to sovereignty. For each speech, I calculate the average semantic distance between the speech's embedding and the embeddings of both keywords in the pair.[10] This indicates how conceptually intertwined the two keywords are within that speech. Word embeddings are real-valued vectors, and cosine similarity between a speech and a couple of keywords ranges from -1 (exact opposite) to 1 (identical). I convert this to cosine *distance* as 1 - cosine similarity, so higher values indicate greater semantic separation. Because cosine distances not easily interpretable in isolation, I standardize them (using z-scores) to enable cross-model comparisons and interpret the results in terms of deviations from the mean value. Before generating embeddings, all words in the speeches were lemmatized to their base forms.

The first pair, 'sovereignty' and 'colonial', reflects postcolonial interpretations of sovereignty. The second pair, with 'territory,' captures sovereignty's association with centralized territorial control and domestic autonomy within the physical border. The third, 'Westphalia,' is an explicit use of the term that is often understood to encompass traditional, non-interventionist understanding rooted in the Westphalian tradition.

Inclusion of pair *sovereignty* and *colonial* needs an explicit justification. Glanville (2016) points out that sovereignty as non-interference was states' last resort to fight colonial power. Not only do I heed to Glanville's (2016) call for caution, I empirically decouple non-intervention as a mask for human rights violations from non-intervention as a shield against colonial repression. While this paper's central claim is that authoritarian states invoke the sovereignty principle, with

[10] To be clear with the method, this is not a distance between two keywords. I calculate the distance between each speech and the pair of keywords as a group. This approach captures how closely a given document rhetorically aligns with both concepts together. Conceptually, it tells us the degree to which a given speech rhetorically links 'sovereignty' and 'territory.' This is opposed to measuring only the inter-keyword distance that ignores how speakers position themselves with respect to those meanings.

the goal of justifying human rights violation, its pattern can be over-estimated without denominating its usage during the colonial period. Hence, I aim at separating out cases where authoritarian states that are often post-colonial states used the sovereignty principle to guard themselves from imperialism.

For an illustrative example, Egypt advocated for its right over the Suez Canal, challenging British control. The UK Parliament provides the entire debate text corpus including a full transcript from the House of Lords on December 13th, 1956. One of the house members states: “I am well aware of the temptation there is to think one could obtain a solution by force. We could not do it in the case of Persia, because we were bound by our obligations under the United Nations.” Amidst growing tension over the Suez Canal, Egypt's representatives at the United Nations reinforced its defense. In a statement to the UN Security Council on October 5, 1956, Egyptian Foreign Minister Mahmoud Fawzi stated: "The nationalization of the Suez Canal Company is an exercise of Egypt's sovereign rights and is in full conformity with international law." This has little to do with an effort to justify contentious policy. V-Dem’s additive polyarchy scores for Egypt and the UK in 1956 are 0.15 and 0.92, hitting the near-end of the spectrum of full autocracy and full democracy. This is one of the many examples where anti-colonial use of sovereignty principle could grossly distort the analysis.

In sum, this paper differentiates between sovereignty rhetoric grounded in anti-colonial self-determination and that which may be used by authoritarian states to obscure human rights issues. By distinguishing between sovereignty as a right to self-determination and as a possible cover for human rights abuses, this research reveals its dual roles—as both protective and potentially manipulative.

Lexical similarity between a pair of words is operationalized by the cosine similarity of word embedding. This is a standard measurement of similarity that increases with a greater semantic overlap (Manning et al. 2008; Kraft & Klemmensen 2024). Alternative to this is simple Euclidean distance, but since the GloVe has high-dimensional vectors, I adopted cosine similarity which is appropriate for handling vectors with varying lengths. In other words, cosine similarity focuses more on lexical distance than the magnitude of vectors, which otherwise could be inflated for common terms.

The main explanatory variable of interest is a binary variable of democracy, based on electoral democracy score. I report the binary specification for two reasons. First, it offers an intuitive and theory-congruent contrast making effect sizes easy to interpret and comparable to prior work. Second, it avoids strong functional-form assumptions about how incremental changes in polyarchy map onto rhetoric. Year-to-year measurement noise and potential nonlinearity can obscure patterns when treated as continuous.

I rely on the regime classification developed by Cheibub, Gandhi, and Vreeland (2010) because it links regime type directly to institutional constraints on executive authority**,** which is central to theorizing how leaders respond to risk perception. As a robustness check, I include continuous democracy scores, from V-Dem's polyarchy index, as well as Polity score. Robustness check against alternative measures of democracy is included in [Appendix 8], which show consistent result.

4.3. Covariates

To control for confounders between the relationship between regime type and semantic concepts around sovereignty, I consider following covariates. Document-level variables are word counts (WC) and word per sentences (WPS). While the UNGA itself limits the total speech time for all countries, it is still crucial to mitigate the spurious relationship by over-weighting keywords that happened to appear more frequently simply due to the length of the remarks. Similarly, average word counts per sentence within each statement have to be considered. This can also be a good proxy measure of statements that were delivered in languages that are not one of the 6 UN official languages.

For country-level controls, I include a measure of country–year participation in militarized interstate disputes, using the Militarized Interstate Disputes (MID) dataset (Palmer et al. 2020). States involved in interstate disputes are more likely to emphasize territorial integrity and related concepts in their international rhetoric. I also include a binary indicator of colonial history, based on the Authoritarian Regime Dataset (Teorell et al. 2018). While the original dataset distinguishes among ten former colonial rulers, I collapse this information into a binary variable indicating whether a country experienced any colonial rule.

Given countless observable and unobservable factors that can potentially complicate the relationship between legal invocation and risk, I try to measure "perceived" risk by directly operationalizing displayed sense of risk in the same given speech. I rely on a computational linguistics model, the "Linguistic Inquiry and Word Count (LIWC)", to operationalize the expressed sense of insecurity (Pennebaker et al., 2011; Pennebaker et al., 2001; Boyd et al., 2022). While not widely used as of yet in the political science domain, it is a common approach adopted by psychologists to detect often-used words that signal a speaker's emotional status. The LIWC-22 dictionary, often used in the computational linguistics field, allows researchers to quantify text data by categorizing words into psychologically meaningful clusters, including cognitive processes and sentiments. Unlike other conventional sentiment dictionaries like AFINN, or NRC

Emotion Lexicon, LIWC dictionary is maintained by empirical psychologists and validated against numerous datasets from various domains (Donohue et al., 2013). It is shown to be powerful enough to even predict authoritarian state leaders' survival by analyzing speeches and estimating sentiments based on the dictionary (Windsor et al., 2018).

Among many sentiment constructs, I focus on four key variables. First, I capture the degree of overall emotional tone through "negative tone" sentiment score. Second, "focus past" variable is expected to capture the extent to which the statement emphasizes past events, based on inclusion of past tense verbs and retrospective words. Third, "risk" measures the frequency of words that are related to uncertainty, caution, or physical threat. In 2020, for example, the highest risk score of 2.58 in my sample appears in Tunisia's address, which features risk-laden phrasing such as "very critical international situation," "serious repercussions," "deep concern," and "the threat persists."[11] Last, "negation" score increases when the speaker uses words like "no," "never," or "nothing" to negate neighboring terms. The risk measure captures leaders' expressed sense of insecurity rather than exogenous threat. This is appropriate for the present analysis, which examines how sovereignty rhetoric co-varies with contemporaneous rhetorical signals of vulnerability, rather than attempting to model objective crisis onset.

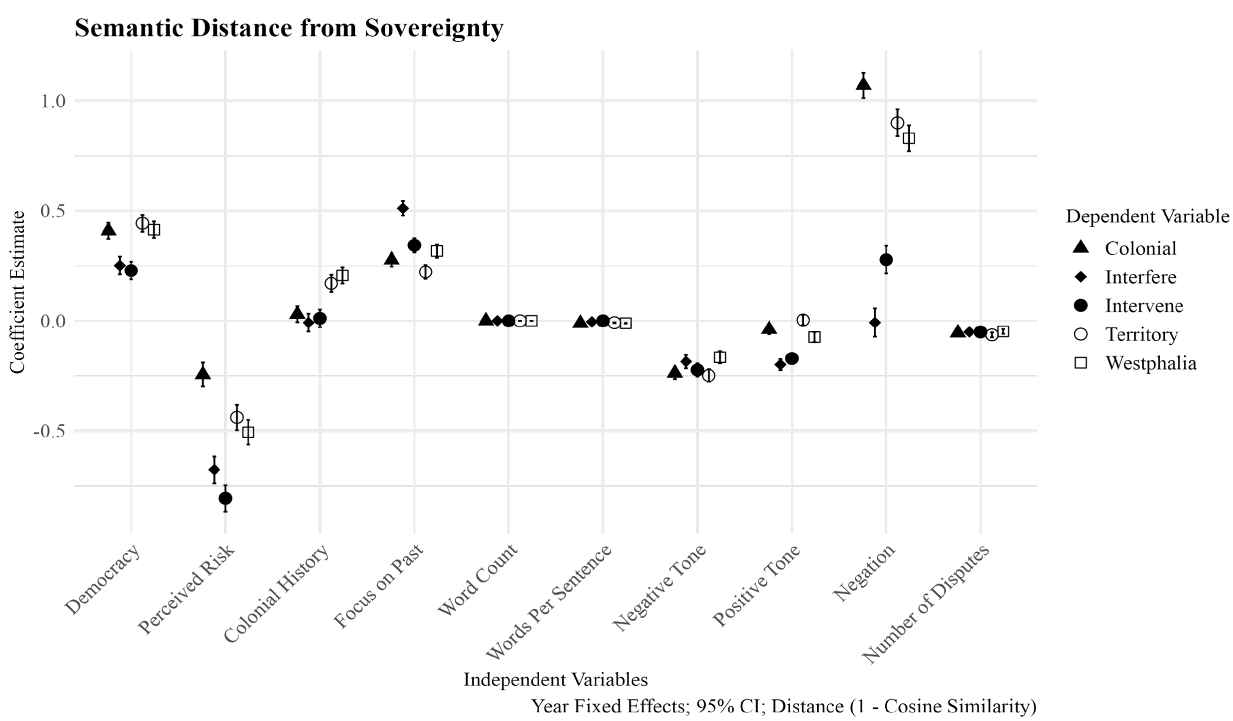
5. Empirical Evidence

The results provide strong support for Hypothesis 2, which posits that authoritarian states are more likely than democracies to invoke *sovereignty* in its traditional Westphalian sense.[12] As

[11] The risk scores are well-dispersed and right-skewed, ranging from near 0 up to 3, with most speeches between 0.5 and 1.0. In 2020, the lowest score is the United States (0.2), followed by Oman, indicating meaningful variation across speeches rather than a few outliers.
[12] [Appendix 2] shows that a group of authoritarian states tend to show narrower semantic distance between the target term "sovereignty" and "Westphalia."

shown in [Plot 1], higher levels of democracy are consistently associated with greater semantic distance between *sovereignty* and traditionally loaded terms such as *territory*, *Westphalia*, and *colonial*.

To reiterate, lower values of semantic distance indicate closer rhetorical association. After controlling for potential confounders, change from autocracy to democracy is associated with a 0.44 standard deviation increase in the distance between *sovereignty* and *territory*, which is a meaningful divergence corresponding over half the total variation in rhetorical usage. The effect is even greater for *Westphalia* (0.41), and similarly large for *colonial* (0.41). These results suggest that authoritarian states are more likely to rhetorically anchor their sovereignty claims in historically defensive or non-interventionist terms. Democracies, by contrast, invoke sovereignty in ways that are semantically more distant from these traditional reference points. The pattern is consistent for subcategories of sixfold regime, as reported in [Appendix 3].



Plot 1. Estimated Semantic Distance

Among the country-level covariates, the number of militarized disputes each year is negatively associated with the semantic distance across all keyword pairs (*colonial, interfere, intervene, territory, and Westphalia)*. During periods of heightened conflict, state leaders are more likely to rhetorically link sovereignty with territorial defense. This is in line with conventional understanding that during the belligerent times, state leaders find it important to emphasize the connection between territory and justification based on sovereignty. An increased overlap between sovereignty and "territory", as well as "Westphalia" is observed across all regime types.

More disputes are associated with "sovereignty" moving closer to the colonial/anticolonial semantic neighborhood, consistent with states narrating coercion, border claims, or external pressure as legacies of colonial interference. The effect is stronger after 1960, suggesting that decolonization institutionalized an anticolonial frame as the default language of contestation, so later disputes are increasingly voiced through that repertoire.[13]

Countries with colonial history did not show any meaningful correlation with the distance between sovereignty and the word "colonial." They also attached weak connection between the other two pairs - sovereignty and territory, and sovereignty and Westphalia. Instead of aligning sovereignty with traditional notions such as territorial integrity or the Westphalian model of absolute state autonomy, post-colonial states prioritize alternative dimensions of sovereignty that reflect their unique historical and geopolitical contexts. For instance, these states may associate sovereignty more strongly with self-determination, resistance to external interference, or solidarity with other post-colonial nations, thereby de-emphasizing classical Western interpretations of sovereignty rooted in territoriality and the Westphalian tradition.

[13] [Appendix 4] reports estimates using only observations from 1961 onward, a window motivated by the 1960 decolonization surge and UNGA Resolution 1514.

To take a closer look into this result, India's first prime minister, Jawaharlal Nehru's speech at the Bandung Conference in 1955 can be useful. It was a meeting of newly independent countries in Asia and Africa. In his speech, he stated: "We meet here to find ways and means to safeguard peace and freedom... The sovereignty we claim is not a sovereignty of aggression or isolation. It is the sovereignty to shape our own destiny, to cooperate freely with others, and to strive for a world where colonialism and imperialism have no place." By no means did they invoke the sovereignty to mask any human rights violation. This aligns with the findings that countries with colonial histories de-emphasize traditional constructs like territory and Westphalia, instead redefining sovereignty in ways that reflect their unique historical and political experiences.

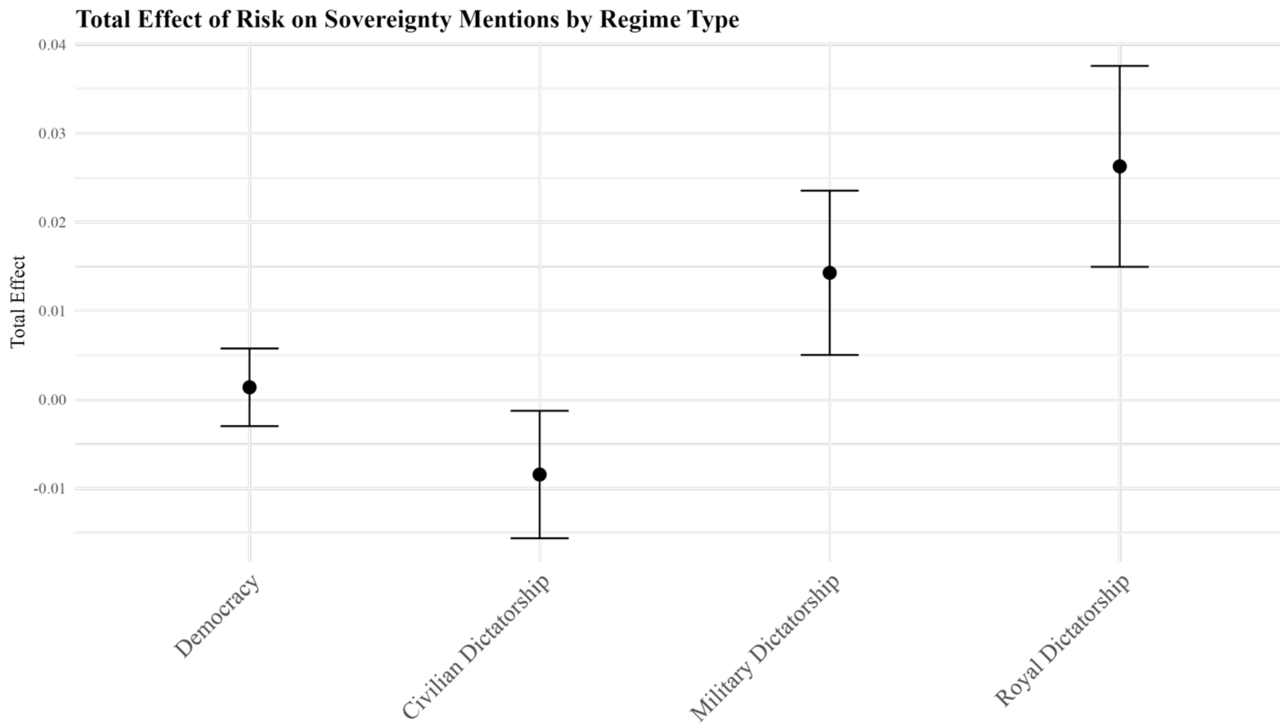
Sentiment analysis using the LIWC-22 dictionary offers additional insight. While general "negative tone" was weakly associated with shifts in semantic meaning, the *risk* sentiment score and *negation* linguistic style consistently showed significant explanatory power. This suggests that targeted psychological indicators, such as references to uncertainty or threat, are more predictive than aggregate sentiment measurements. Estimated coefficient of risk score for each dependent variable shows that countries across all regime types tend to find closer relationship for pairs. Overall, the risk score consistently reduced the distance between all pairs of keywords. Risk score had the biggest effect in reducing the distance between the meanings of "sovereignty" and "intervene," but it had the smallest effect on "sovereignty" and "colonial."

5.1. Risk and the Rhetorical Salience of Sovereignty

Hypothesis3 states a positive relationship between perceived risk level and the frequency of sovereignty invocation from autocracies than democracies. The unit of analysis remains the country–year. Sovereignty invocation is measured as the frequency of sovereignty-related terms, defined as the number of such terms divided by the total word count of each speech. Consistent with this expectation, perceived risk is positively associated with sovereignty invocation across regime types. However, the absence of a statistically significant regime-level interaction suggests that risk sensitivity alone does not distinguish democracies from autocracies. Instead, meaningful variation emerges once we account for institutional heterogeneity within authoritarian regimes. [Appendix 6] reports coefficient estimates for full transparency.

The absence of a statistically significant interaction at the aggregate regime level warrants careful explanation. As discussed in the theoretical framework, aggregation across regime types may obscure meaningful differences in how authoritarian states respond to risk. Similar rhetorical responses by democracies and civilian dictatorships may dampen aggregate differences, even if other authoritarian subtypes respond more defensively. I now turn to the auxiliary hypothesis that examines whether the relationship between perceived risk and sovereignty invocation varies across authoritarian regime types.

To test *H3b, Military and royal authoritarian states will invoke the principle of sovereignty when they feel at risk, compared to democracies-*, I regress the frequency of sovereignty invocation on the sixfold regime type, risk score, and their interaction. Since I do not theorize any intra-democracy subtypes variation, I pooled three different types of democracies (parliamentary, presidential, and mixed) into one single bin of democracy. [Appendix 7] reports coefficient estimates based on full six-category based classification. The results are consistent.



Plot 2. Estimated Total Effect

Plot 2 shows that at baseline levels of perceived risk, civilian dictatorships invoke sovereignty more frequently than democracies, as reflected in the positive main effect of the civilian regime coefficient (0.017).[14] This suggests that, holding other factors constant and under ordinary conditions, civilian autocracies are more likely than democracies to emphasize sovereignty in their international rhetoric. In contrast, the baseline coefficients for military (0.002) and royal dictatorships (−0.001) are statistically insignificant, implying no meaningful difference from democracies in their average invocation of sovereignty.

Next, I examine the *total marginal effect of* perceived risk by regime type, calculated by summing the main effect of risk and the regime-specific interaction terms. For democracies, the estimated marginal effect is small and statistically insignificant (0.0014, 95% CI: −0.0030 to

[14] Please refer to the [Appendix 5] for full coefficient estimates.

0.0058), suggesting that changes in perceived risk do not meaningfully affect the frequency of sovereignty invocation. By contrast, the effect of risk is positive and statistically significant for both military (0.0143, 95% CI: 0.0050 to 0.0235) and royal dictatorships (0.0263, 95% CI: 0.0150 to 0.0376), indicating that these regimes invoke sovereignty more frequently as perceived risk increases. This divergence underscores the heterogeneity within authoritarian regimes. While royal and military autocrats turn to sovereignty rhetoric defensively under pressure, civilian autocrats may already rely on institutionalized or partisan legitimacy that makes further rhetorical escalation unnecessary.

While my argument does not explicitly distinguish between personalist and non-personalist civilian leaders, the empirical findings are consistent with Weeks's (2012) logic of domestic institutional constraint. In her theory, non-personalist civilian leaders are not that different from democracies in their lower propensity to initiate conflict, as other elites within the regime participate in foreign policy decisions and are often reluctant to resort to brute force. Relatedly, scholarship that follows Cheibub et al.'s (2010) regime classification reaches a similar conclusion. Because civilian dictators possess established institutional (often partisan) resources, they are not fundamentally different from democracies in their incentives to provide public goods and maintain engagement with international law during ordinary times (Ginsburg 2021; Ch. 1). In this sense, civilian dictatorships begin from a higher baseline of sovereignty rhetoric than other autocratic subtypes, reflecting their need to engage in international discourse as part of their legitimation strategy. By contrast, military and royal dictatorships rely less on such rhetoric under normal conditions but turn to sovereignty language more aggressively when they perceive heightened threats. This heterogeneity underscores that authoritarian international legal behavior cannot be reduced to a uniform logic of deflection or resistance; it is contingent on the type of domestic institutions.

Among the auxiliary variables, several results align with theoretical expectations. Countries experiencing a higher number of militarized disputes are more likely to invoke sovereignty, consistent with the association between external conflict and territorial defense rhetoric. Longer speeches (WC) and greater sentence complexity (WPS) also increase the likelihood of sovereignty invocation, reflecting more elaborate discursive framing in extended statements. Negative emotional tone (tone_neg) is positively associated with sovereignty mentions, suggesting that leaders draw on sovereignty language under more defensive or confrontational rhetorical conditions. Conversely, positive tone (tone_pos) and negation are negatively related to sovereignty frequency, indicating that more optimistic or conciliatory speeches place less emphasis on sovereignty. Colonial history does not show significant effects, implying that contemporary rhetorical variation is better explained by political regime and situational risk than by historical or structural factors.

6. Conclusion

When contentious political events arise, especially those involving international security or human rights, both democracies and autocracies turn to the most authoritative language available, international law. What is striking, however, is that they often invoke the same legal principle, sovereignty, to advance opposing claims. On the one hand, democracies argue that autocracies have violated the principle of sovereignty, by shedding light on the domestic repression. On the other hand, autocracies also argue that it was their sovereignty that were infringed on. Autocracies contend that democracies' diplomatic pressure, criticism, often coupled with financial sanctions and embargo violate *their* right to non-interference and *their* sovereignty. This paper examines this puzzle and situates it within a broader question of how domestic political institutions shape the use of international law. I develop a theory of regime-contingent legal appropriation, which posits that variation in domestic institutions leads states to frame and deploy the sovereignty principle in systematically different ways.

Findings from text analysis demonstrate that there are substantial differences in how democracies and autocracies invoke and interpret the principle of sovereignty. Authoritarian regimes, particularly royal and military dictatorships, tend to anchor sovereignty more closely to traditional ideas of non-interference and territorial integrity than democracies do. This reliance on sovereignty as a tool of rhetorical legitimation intensifies during periods of risk, when autocratic leaders use legal language to deflect external criticism.

This study contributes to the broader literature on international law and regime behavior in several ways. First, it expands the theory of how authoritarian states can strategically use international institutions. By highlighting that the sovereignty principle, while ostensibly universal, can take on distinct meanings depending on the political context of the state invoking it, this study moves beyond textual comparisons of "what is on the books." It instead reveals how international law operates as a malleable political resource, one that regimes appropriate and reinterpret to serve their own domestic political needs.

Second, the results underscore the importance of accounting for variation in domestic political institutions when analyzing international legal discourse. Regime type not only shapes how states interpret legal principles but also how they deploy them as diplomatic tools. Rhetorical engagement with international law, how leaders frame, justify, or contest legal norms, should therefore be incorporated into broader discussions of authoritarian international institutions and their modes of legitimation.

Furthermore, this study calls for future research on how other legal principles, particularly emerging or evolving norms, are adopted and adapted across different political regimes, shaping the broader contours of international governance. While this paper focuses on the long-standing principle of sovereignty, its analytical framework can be extended to newer domains where meaning remains fluid and contestation continues to unfold. The empirical strategy presented here, which compares latent and context-dependent interpretations of legal norms, provides a way to capture what is often difficult to observe directly, how states use shared vocabularies to advance different political objectives.[15] Future research can apply this approach to hard-to-measure norms such as terrorism, or liberalism, examining how democracies and autocracies reinterpret these norms in ways that reflect their institutional structures and policy priorities.

In the broader picture, this paper complicates the idea that legal texts carry the same meaning for all states that endorse them, using the sovereignty principle as a critical test case. This is a hard test because sovereignty principle has existed for a long time, providing ample time for states and courts to discuss, apply, interpret, and adjudicate it. If there is an ambiguity around such a high-profile principle, there should be greater discrepancies around newly emerged principles.

This article also speaks to the work of non-governmental organizations that stand at the forefront of resistance to authoritarian policies. In an era of expanding transnational repression, where despotic governments increasingly exploit international institutions, rhetoric grounded in the sovereignty principle can further embolden authoritarian claims before international audiences.

[15] For example, Finnemore and Sikkink (1998) point out the empirical challenge of observing norms that are often inferred from behavior or language rather than directly measured.

When state leaders invoke the moral and legal gravity of sovereignty, they raise the political cost for NGOs to monitor, document, and challenge abuses, effectively narrowing the discursive space for accountability.

Recent scholarship on transnational advocacy networks highlights similar dynamics, showing that not only states but also illiberal nonstate actors increasingly contest liberal norms while emulating the organizational and rhetorical strategies of their democratic counterparts (Reiners 2022; Hall et al., 2025). Building on these insights, this study offers an analytical framework for understanding how sovereignty-based language operates as a rhetorical strategy to legitimize or shield illiberal practices. By tracing how such discourse manifests across regimes and actors, it helps illuminate why global civil society now faces a more complex and constrained environment for monitoring, advocacy, and norm promotion.

References

Acharya, Amitav. "Who are the norm makers? The Asian-African conference in Bandung and the evolution of norms." *Global Governance* (2014): 405-417.

Ambrosia, A. (2008). *The Shanghai Cooperation Organization: Norms and challenges*. Journal of International Organizations Studies, 6(2), 34-50.

Bae, J., & Salvati, F. 2026. “United Nations Discourse on Liberalism across Regime Types.” *Global Studies Quarterly*. https://doi.org/10.1093/isagsq/ksag001.

Bae, J., & Moon, D. (2026). Regime-type composition of an intergovernmental organisation and dispute settlement mechanisms. *The British Journal of Politics and International Relations*, *28*(1), 235-258.

Barkin, J. S., & Cronin, B. (1994). The state and the nation: Changing norms and the rules of sovereignty in international relations. *International Organization*, 48(1), 107–130.

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2), 1–9.

Bellamy, Alex J., and Tim Dunne. "R2P in Theory and Practice." *The Oxford handbook of the responsibility to protect*(2016): 3-17.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Behavior Research Methods*, 54(2), 526–543.

Boix, C., & Svolik, M. W. (2013). The Foundations of Limited Authoritarian Government: Institutions, Commitment, and Power-Sharing in Dictatorships. *The Journal of Politics*, *75*(2), 300–316.

Bush, S. S., Cottiero, C., & Prather, L. (2024). Zombies ahead: Explaining the rise of low-quality election monitoring. *The Review of International Organizations*, 1-31.

Chelotti, N., Huhe, N., & Woll, C. (2022). Preferences, convergence, and the role of EU institutions: Evidence from WordScore analysis. *Journal of European Public Policy*, 29(6), 921–943.

Cheibub, J. A., Gandhi, J., & Vreeland, J. R. (2010). Democracy and dictatorship revisited. *Public Choice*, *143*(1-2), 67–101.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Kyle L. Marquardt, Juraj Medzihorsky, Daniel Pemstein, Linnea Fox, Lisa Gastaldi, Eitan Tzelgov, Yi-ting Wang, and Steven Wilson. 2025. "V-Dem Methodology v15" Varieties of Democracy (V-Dem) Project.

Cohen, R., & Deng, F. M. (2016). *Sovereignty as responsibility: Conflict management in Africa*. Brookings Institution Press.

Cottiero, C., & Haggard, S. (2023). Authoritarian cooperation and legitimation: The role of international organizations. *International Studies Quarterly*, 67(2), 234–248.

Cottiero, C., Hafner-Burton, E. M., Haggard, S., Prather, L., & Schneider, C. J. (2025). Illiberal regimes and international organizations. *The Review of International Organizations*, *20*(2), 231-259.

Cooley, A. (2015). Authoritarianism goes global: Countering democratic norms. *Journal of Democracy*, *26*(3), 49-63.

Debre, M. J. (2025). *How Regional Organizations Sustain Authoritarian Rule: The Dictators' Club*. Oxford University Press.

Debre, M. J. (2021). Authoritarian regimes and international organizations: Examining participation patterns. *Review of International Organizations*, 16(4), 685–713.

Donohue, S. E., McDaniel, M. A., & Richards, J. B. (2013). Evaluating LIWC measures of psychological constructs: A meta-analysis. *Behavior Research Methods*, 45(1), 1–20.

Ezrow, Natasha M., and Erica Frantz. "STATE INSTITUTIONS AND THE SURVIVAL OF DICTATORSHIPS." *Journal of International Affairs*, vol. 65, no. 1, 2011, pp. 1–13.

Flonk, D., & Debre, M. J. (2025). Hollow multilateralism: how autocracies contest the norms and procedures of international organizations. *International Affairs*, *101*(4), 1463-1482.

Fowler, Michael Ross, and Julie Marie Bunck. *Law, power, and the sovereign state: the evolution and application of the concept of sovereignty*. Penn State Press, 2010.

Franck, T. M. (1990). *The power of legitimacy among nations*. Oxford University Press.

Goldsmith, Jack L., and Eric A. Posner. *The limits of international law*. Oxford University Press, 2005.

Gandhi, J., & Przeworski, A. (2007). Authoritarian institutions and the survival of autocrats. *Comparative Political Studies*, 40(11), 1279–1301.

Gerschewski, J. (2013). The three pillars of stability: legitimation, repression, and co-optation in autocratic regimes. *Democratization*, *20*(1), 13–38.

Ginsburg, T. *Democracies and International Law*. 2021. Cambridge: University of Cambridge.

Hafner-Burton, Emilie M. 2013. "The International Human Rights Legal System." In *Making Human Rights a Reality*, [p.64-86]. Princeton University Press.

Hafner-Burton, Emilie M., Laurence R. Helfer, and Christopher J. Fariss. 2011. "Emergency and Escape: Explaining Derogations From Human Rights Treaties". *International Organization* 65:673-707.

Poe, Steven C., and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis". *American Political Science Review* 88:853-72.

Ginsburg, T. (2020) a. Authoritarian international law? *American Journal of International Law*, 114(2), 221–260.

Ginsburg, T. (2020) b. How authoritarian regimes use international law. *Journal of Democracy*, 31(4), 44–58.

Glanville, L. (2016). *Sovereignty in the international legal order: Norms, practices, and challenges*. Cambridge University Press.

Hendrix CS, Wong WH. When Is the Pen Truly Mighty? Regime Type and the Efficacy of Naming and Shaming in Curbing Human Rights Abuses. *British Journal of Political Science*. 2013;43(3):651-672.

Hosli, M. O., & Kantorowicz, J. (2024). Policy convergence in the EU: A WordScore analysis of member states' speeches. *European Union Politics*, 25(1), 45–68.

Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing* (3rd ed.). Prentice Hall.

Koremenos, B., Lipson, C., & Snidal, D. (2001). The rational design of international institutions. *International organization*, *55*(4), 761-799.

Krasner, S. D. (1999). *Sovereignty: Organized hypocrisy*. Princeton University Press.

Kraft, P. W., & Klemmensen, R. (2024). Lexical ambiguity in political rhetoric: why morality doesn't fit in a bag of words. *British Journal of Political Science*, *54*(1), 201-219.

Koliev, F. (2020). Shaming and democracy: Explaining inter-state shaming in international organizations. *International Political Science Review*, *41*(4), 538-553.

Lake, D. A. (2003). The new sovereignty in international relations. *International studies review*, *5*(3), 303-323.

Lake, D. A., Martin, L. L., & Risse, T. (2021). Challenges to the liberal order: Reflections on international organization. *International organization*, *75*(2), 225-257.

Mallat, C. (2020). The paradox of illiberal principles: Sovereignty, non-intervention, and human rights. *Global Constitutionalism*, 9(3), 301–323.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Naderi, B. (2024, October 1). *UNGA: China's policies grounded in legitimate arguments.* China's Diplomacy in the New Era. Retrieved from http://en.chinadiplomacy.org.cn/2024-10/01/content_117462992.shtml.

Hall, N., Reiners, N., Basu, S., Chaudhry, S., Henry, L., Medie, P. A., Sundstrom, L. M., Vilán, A. & Zvobgo, K. (2025). Challenges and Opportunities for Transnational Advocacy. *International Studies Review*, *27*(3).

Osiander, A. (2001). Sovereignty, international relations, and the Westphalian myth. *International Organization*, 55(2), 251–287.

Palmer, G., D'Orazio, V., Kenwick, M., & Lane, M. (2020). The MID4 dataset: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 37(2), 253–275.

Paris, R. (2020). *The practice of sovereignty and its limits in modern politics*. Cornell University Press.

Reiners, N. (2022). *Transnational Lawmaking Coalitions for Human Rights.* Cambridge: Cambridge University press.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Philpott, D. (1995). Sovereignty: An introduction and brief history. *Journal of International Affairs*, 48(2), 353–368.

Risse, T., Ropp, S. C., & Sikkink, K. (1999). *The power of human rights: International norms and domestic change*. Cambridge University Press.

Russett, B. (2005). Bushwhacking the Democratic Peace. *International Studies Perspectives,* 6, 395-408.

Smith, A. D. (2006). *Speech at the United Nations: A historical perspective*. Routledge.

Simmons, B. A. (2009). *Mobilizing for human rights: international law in domestic politics*. Cambridge University Press.

Simmons, B. A., & Goemans, H. E. (2021). Built on borders: Tensions with the institution liberalism (thought it) left behind. *International Organization*, *75*(2), 387-410.

Simmons, B. A., & Shaffer, G. (2023). The discourse of borders: Sovereignty, global governance, and state anxiety. *International Organization*, 77(2), 401–427.

Tallberg, J., & Vikberg, C. (2025). Democracy, Autocracy, and the Design of International Organizations. *International Studies Quarterly*, *69*(2).

Tallberg, J., Lundgren, M., Sommerer, T., & Squatrito, T. (2020). Why international organizations commit to liberal norms. *International studies quarterly*, *64*(3), 626-640.

Teorell, J., & Wahman, M. (2018). Institutional stepping stones for democracy: How and why multipartyism enhances democratic change. *Democratization*, *25*(1), 78–97.

Turco, M. (2024). Climate governance in international fora: A structural topic model approach. *Global Environmental Politics*, 24(1), 72–95.

United Nations Office of Legal Affairs. (2009). *Repertory of practice of United Nations organs: Supplement No. 10 (2000–2009), Article 2(7)*. United Nations. Available at https://legal.un.org/repertory/art2/english/rep_supp10_vol1_art2_7.pdf.

Weiner, M. (2020). Authoritarian regimes, legitimacy, and international cooperation. *International Organization*, 74(3), 555–578.

Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1), 35–64.

Windsor, L. C., Dowell, E. M., & Graetz, N. (2018). Speech sentiment analysis and authoritarian survival. *Political Analysis*, 26(3), 334–351.

Wheeler, N. J. (2000). *Saving strangers: Humanitarian intervention in international society*. OUP Oxford.

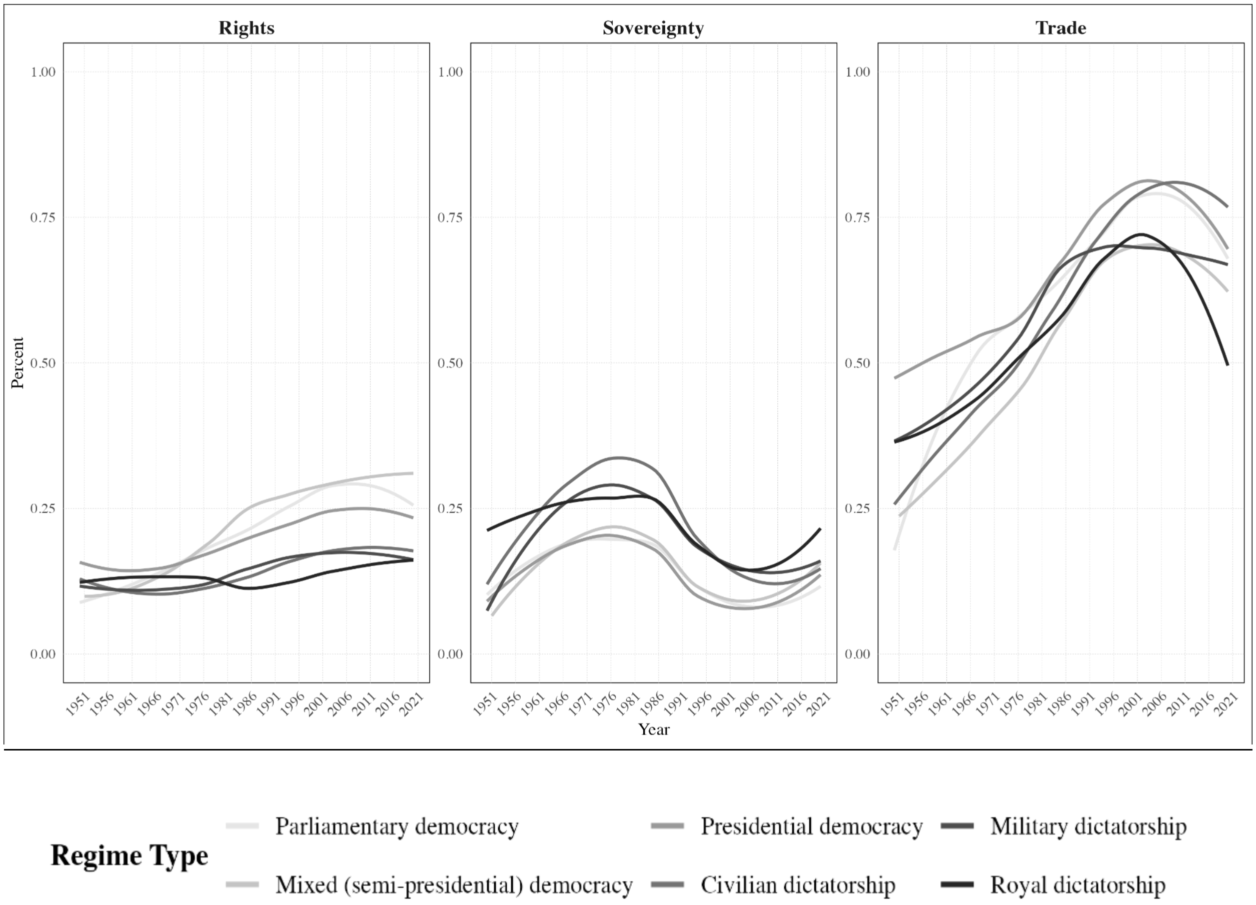
Morse, Julia C., and Tyler Pratt. forthcoming. "Smoke and Mirrors: Strategic Messaging and the Politics of Noncompliance." *American Political Science Review*: 1–19.

Búzás, Zoltán I. 2018. "Is the Good News About Law Compliance Good News About Norm Compliance? The Case of Racial Equality." *International Organization* 72 (2): 351–385.

UN Human Rights Council. 2019. *Report of the Working Group on the Universal Periodic Review: China:appendum*. A/HRC/40/6/Add.1. 31st session. Geneva: United Nations.

Appendices

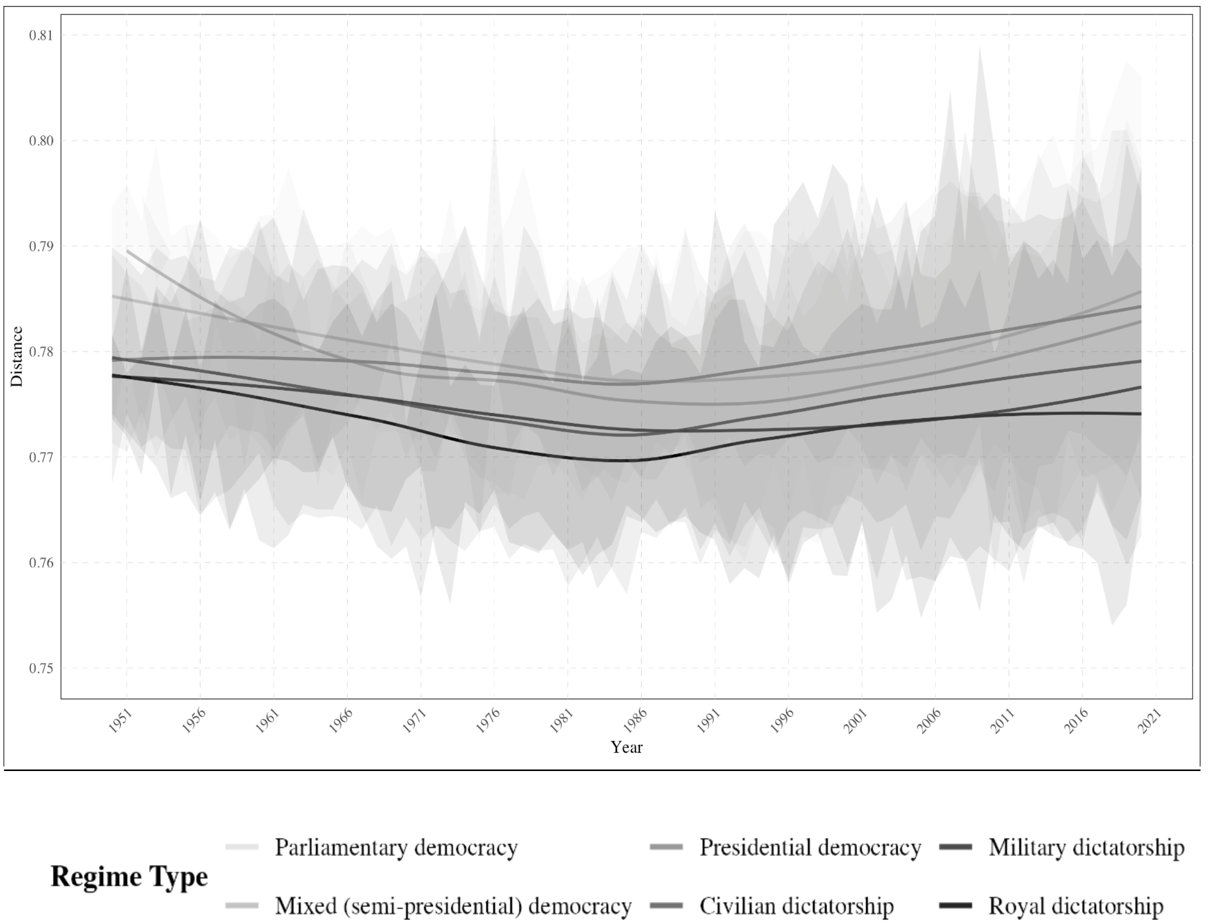
[Appendix 1] Proportion of keyword appearance per speech using six-fold regime categories



I used Gandhi et al.(2010)'s six-fold regime categories to further understand variation within a broader category of dictatorship. The darker the shade is, the closer the regime is to the royal dictatorship. If the principle of sovereignty were a normatively fair and neutral term that was applied to all states, then we should have observed a consistent pattern across all types of regimes like "trade."

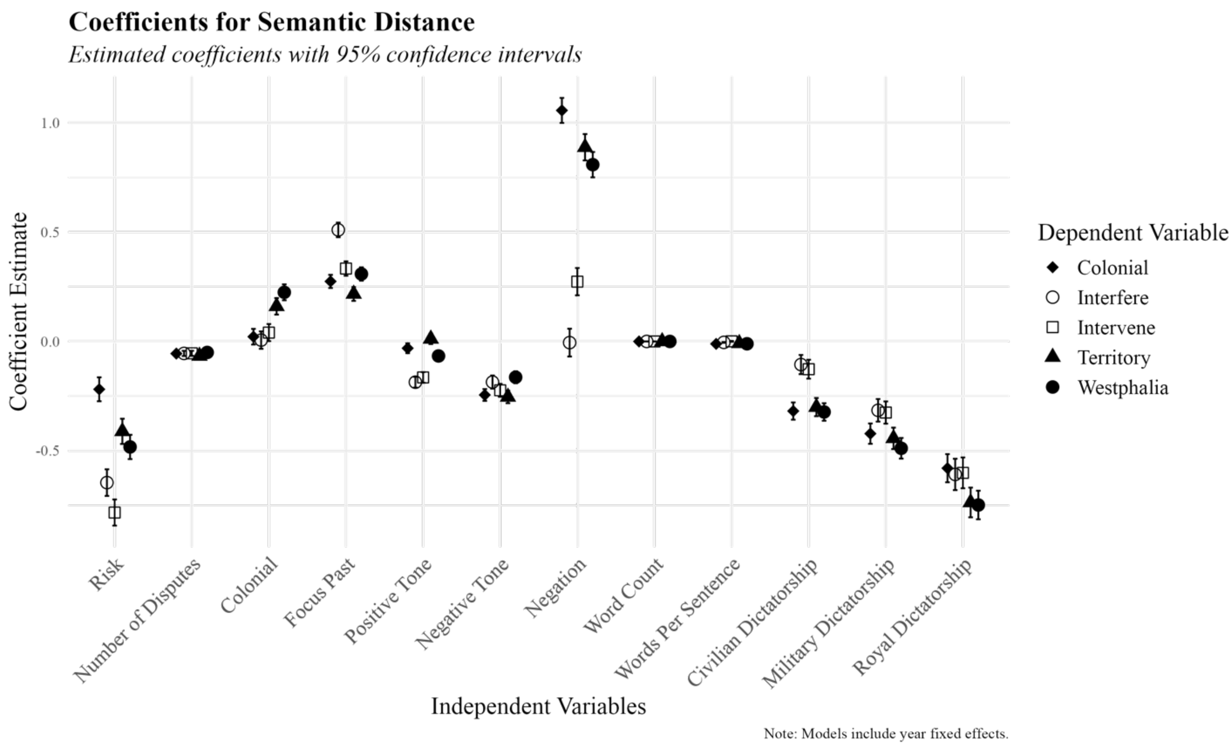
[Appendix 2] Semantic Distance Between Sovereignty and Westphalia

The shade of the line gets darker as the regime type moves along the spectrum of democracy to royal dictatorship. The y-axis is lexical similarity (or semantic distance) between "Sovereignty" and "Westphalia." With little temporal change on the concurrence of these two terms, it is generally noticeable that authoritarian states tend to associate Westphalia more closely with the term "sovereignty" than democracies do.



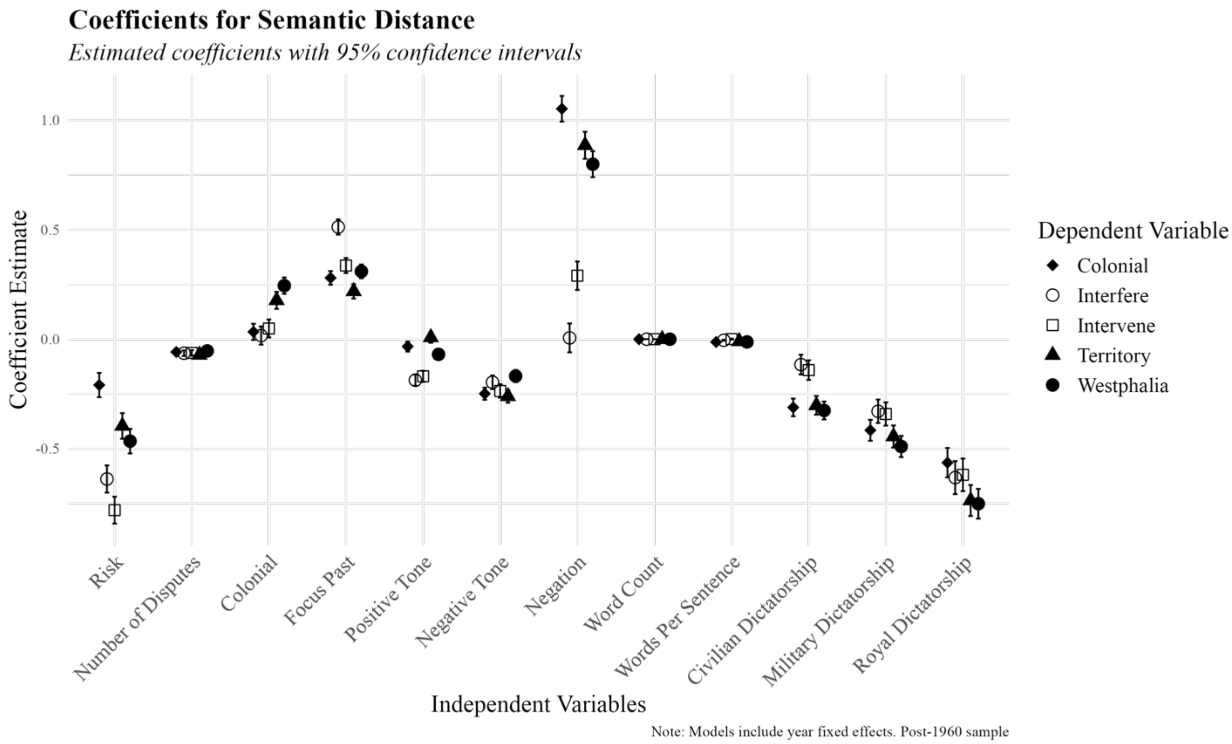
[Appendix 3] Using sub-authoritarian regime classification as Robustness Check to H2

I further use six-fold regime category to further test the logic to variants of democracy and autocracy. I used parliamentary democracy as a baseline of comparison vis-a-vis mixed democracy, presidential democracy, civilian dictatorship, military dictatorship, and royal (or monarchic) dictatorship. As shown in [Plot 2], the coefficient for royal dictatorship is significantly negative, indicating a smaller semantic distance between *sovereignty* and *Westphalia* relative to the baseline of parliamentary democracy. In other words, royal dictatorships tend to associate sovereignty more closely with Westphalian principles than other regime types do.

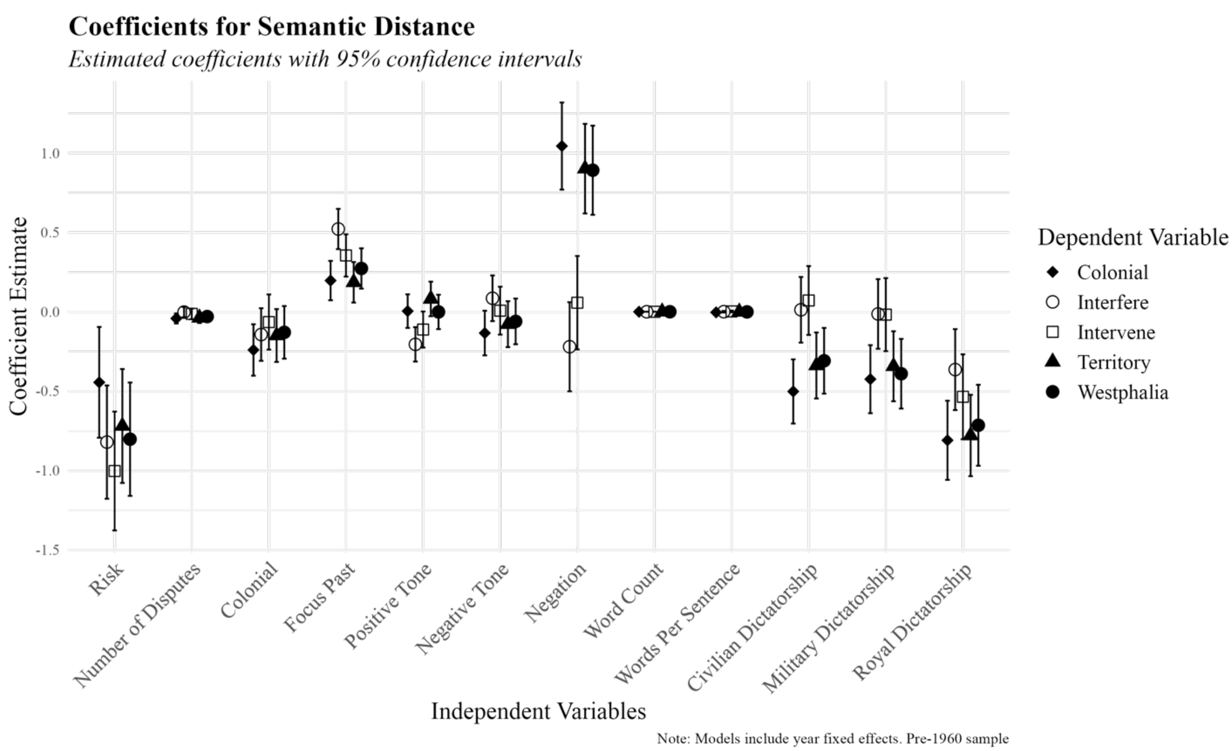


[Appendix 4]

4.1. Post 1960 sample



4.2. Pre 1960 sample

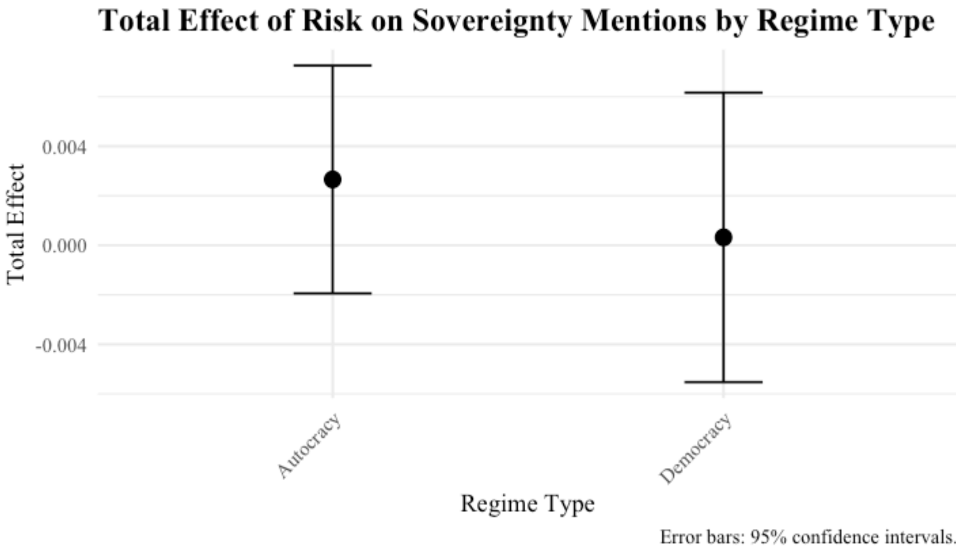


[Appendix 5] Four Coefficient Estimate for Interaction term with four-regime types

Note that frequency rate is measured by the number of term appearance over total word counts *100.

	Dependent variable: frequency
avg_risk	0.001
	(0.002)
as.factor.dd_regime_pooled.Civilian dictatorship	0.017***
	(0.003)
as.factor.dd_regime_pooled.Military dictatorship	0.002
	(0.004)
as.factor.dd_regime_pooled.Royal dictatorship	−0.001
	(0.005)
avg_mid_num_dispute	0.002***
	(0.000)
colonial	0.001
	(0.001)
WC	0.000**
	(0.000)
WPS	0.001***
	(0.000)
tone_neg	0.006***
	(0.001)
tone_pos	−0.002***
	(0.001)
negation	−0.004*
	(0.002)
avg_risk × as.factor.dd_regime_pooled.Civilian dictatorship	−0.010**
	(0.004)
avg_risk × as.factor.dd_regime_pooled.Military dictatorship	0.013**
	(0.005)
avg_risk × as.factor.dd_regime_pooled.Royal dictatorship	0.025***
	(0.006)
Num.Obs.	8905
R2	0.059
R2 Adj.	0.051
AIC	−28636.3
BIC	−28529.9
RMSE	0.05

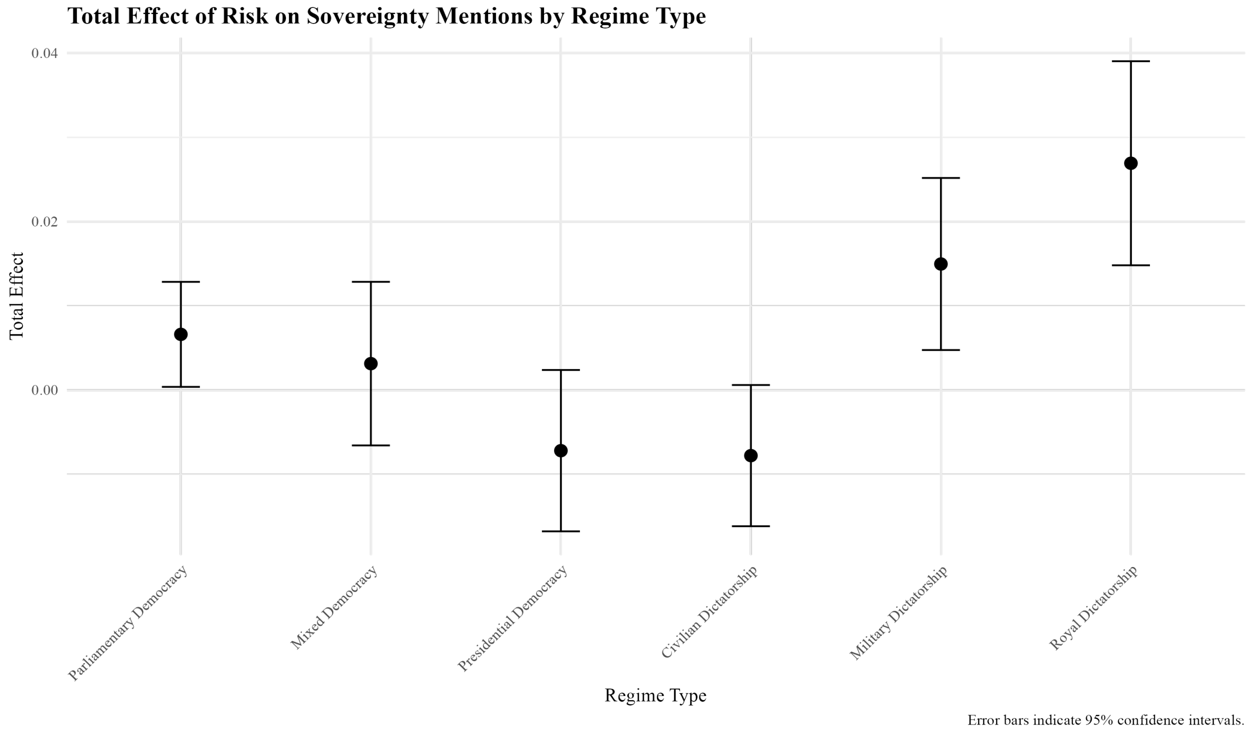
[Appendix 6] Model summary of interaction model with binary variable



Year FE Models with Binary Democracy * Risk Interaction

	(1)
risk	0.003
	(0.002)
as.factor(democracy)1	−0.010***
	(0.003)
WC	0.000***
	(0.000)
WPS	0.001***
	(0.000)
tone_neg	0.007***
	(0.001)
tone_pos	−0.003***
	(0.001)
negation	−0.005*
	(0.002)
focuspast	−0.004***
	(0.001)
colonial	0.001
	(0.001)
mid_num_dispute	0.002***
	(0.000)
risk × as.factor(democracy)1	−0.002
	(0.003)
Num.Obs.	8936
R2	0.056
R2 Adj.	0.047
AIC	−28792.9
BIC	−28707.7
RMSE	0.05

[Appendix 7] Model summary of interaction term model with sixfold categories



	(1)
avg risk	0.007*
	(0.003)
as.factor.dd regime.Mixed (semi-presidential) democracy	0.003
	(0.005)
as.factor.dd regime.Presidential democracy	0.014***
	(0.004)
as.factor.dd regime.Civilian dictatorship	0.023***
	(0.004)
as.factor.dd regime.Military dictatorship	0.008+
	(0.004)
as.factor.dd regime.Royal dictatorship	0.004
	(0.005)
avg mid num dispute	0.002***
	(0.000)
colonial	0.001
	(0.001)
WC	0.000**
	(0.000)
WPS	0.001***
	(0.000)
tone neg	0.006***
	(0.001)
tone pos	−0.003***
	(0.001)

negation	−0.005**
	(0.002)
avg risk × as.factor.dd regime.Mixed (semi-presidential) democracy	−0.003
	(0.005)
avg risk × as.factor.dd regime.Presidential democracy	−0.015**
	(0.005)
avg risk × as.factor.dd regime.Civilian dictatorship	−0.015***
	(0.004)
avg risk × as.factor.dd regime.Military dictatorship	0.008
	(0.005)
avg risk × as.factor.dd regime.Royal dictatorship	0.020**
	(0.006)
Num.Obs.	8905
R2	0.061
R2 Adj.	0.051
AIC	−28639.9
BIC	−28505.1
RMSE	0.05

[Appendix 8] Alternative Measures of Democracy
8.1. Polity

Year Fixed Effects Models: Polity2 and Sovereignty Rhetoric

	territory	**Colonial**	**Westphalia**	**Interfere**	**Intervene**
Polity2	0.034***	0.033***	0.034***	0.019***	0.016***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Perceived Risk	−0.477***	−0.272***	−0.542***	−0.687***	−0.821***
	(0.030)	(0.029)	(0.029)	(0.033)	(0.032)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Word Count	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Words per Sentence	−0.006***	−0.009***	−0.008***	−0.003***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.279)
Negative Tone	−0.219***	−0.214***	−0.131***	−0.157***	−0.199***
	(0.015)	(0.014)	(0.014)	(0.016)	(0.016)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Positive Tone	0.014	−0.025*	−0.061***	−0.184***	−0.157***
	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
	(0.247)	(0.031)	(<0.001)	(<0.001)	(<0.001)
Negation	0.883***	1.051***	0.809***	−0.014	0.275***
	(0.032)	(0.030)	(0.031)	(0.034)	(0.033)
	(<0.001)	(<0.001)	(<0.001)	(0.672)	(<0.001)
Focus on Past	0.212***	0.261***	0.302***	0.500***	0.326***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.017)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Disputes	−0.063***	−0.053***	−0.047***	−0.050***	−0.050***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Colonial	0.148***	0.022	0.195***	−0.013	0.006
	(0.020)	(0.019)	(0.019)	(0.021)	(0.021)
	(<0.001)	(0.230)	(<0.001)	(0.524)	(0.774)
Num.Obs.	8202	8202	8202	8202	8202
R2	0.302	0.361	0.341	0.253	0.235
R2 Adj.	0.294	0.354	0.334	0.245	0.227
AIC	18873.4	17948.0	18281.2	19971.4	19703.0
BIC	18950.6	18025.1	18358.3	20048.6	19780.1
RMSE	0.76	0.72	0.74	0.82	0.80

[Appendix 8]
8.2. Continuous Polyarchy Score

Year Fixed Effects Models: Polyarchy and Sovereignty Rhetoric

	territory	**Colonial**	**Westphalia**	**Interfere**	**Intervene**
Perceived Risk	−0.477***	−0.272***	−0.542***	−0.687***	−0.821***
	(0.030)	(0.029)	(0.029)	(0.033)	(0.032)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Word Count	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Words per Sentence	−0.006***	−0.009***	−0.008***	−0.003***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.279)
Negative Tone	−0.219***	−0.214***	−0.131***	−0.157***	−0.199***
	(0.015)	(0.014)	(0.014)	(0.016)	(0.016)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Positive Tone	0.014	−0.025*	−0.061***	−0.184***	−0.157***
	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
	(0.247)	(0.031)	(<0.001)	(<0.001)	(<0.001)
Negation	0.883***	1.051***	0.809***	−0.014	0.275***
	(0.032)	(0.030)	(0.031)	(0.034)	(0.033)
	(<0.001)	(<0.001)	(<0.001)	(0.672)	(<0.001)
Focus on Past	0.212***	0.261***	0.302***	0.500***	0.326***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.017)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Disputes	−0.063***	−0.053***	−0.047***	−0.050***	−0.050***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Colonial	0.148***	0.022	0.195***	−0.013	0.006
	(0.020)	(0.019)	(0.019)	(0.021)	(0.021)
	(<0.001)	(0.230)	(<0.001)	(0.524)	(0.774)
Num.Obs.	8202	8202	8202	8202	8202
R2	0.302	0.361	0.341	0.253	0.235
R2 Adj.	0.294	0.354	0.334	0.245	0.227
AIC	18873.4	17948.0	18281.2	19971.4	19703.0
BIC	18950.6	18025.1	18358.3	20048.6	19780.1
RMSE	0.76	0.72	0.74	0.82	0.80

[Appendix 9] Dictionary used for descriptive analysis

sovereignty:
- territory
- territorial
- integrity
- border
- interfere
- interference
- independent
- independence
- self
- meddle
- westphalia
- sovereign

trade:
- tariff
- world trade organization
- economic
- market
- investment
- development
- growth
- financial
- trade

rights:
- human rights
- fundamental
- human-rights
- civic rights
- civic-rights
- liberal
- responsibility
- protect
- responsible

Content or Cue? Conditional Messenger Effects in the Legitimacy of International Law

Abstract

Can international law shape public preferences when it is endorsed by authoritarian states? Existing research shows that international law can influence public opinion, but we know little about whether this holds when non-democracies are the primary endorsers. This paper examines whether international law retains persuasive force when it is endorsed only by non-democratic states. I expand existing understandings of compliance pull by theorizing conditional legitimacy, distinguishing between baseline legitimacy rooted in legal content and additional cues derived from the political identity of endorsing actors. I test this argument using a survey experiment in Australia and the Philippines that varies both the legal frame (message) and the regime type of endorsers (messenger) in the context of refugee policy. The results show that international law can influence policy preferences and perceived legality even when endorsed by authoritarian states. Messenger effects also interact with legal content: Sovereignty-based claims are influential regardless of messenger, whereas human-rights claims depend on endorsement context. These findings contribute to the empirical literature on the microfoundations of legal legitimacy and speak to the growing research on the topic of authoritarian international cooperation.

1 Introduction

In response to international criticism over its military operations in Rakhine State, the government of Myanmar appealed to international legal principles of sovereignty and non-interference. At the United Nations General Debate in 2017, Myanmar's representatives argued that the situation constituted an internal matter and that diplomatic opprobrium is in violation of established norms of state sovereignty. International law was mobilized as an authoritative justification for contested state conduct. Whether such reliance on international law makes audiences more accepting of contested state conduct remains unanswered. If legal legitimacy is inseparable from the credibility of its endorsers, authoritarian sponsorship should undermine its persuasive force. If compliance pull operates independently of messenger credibility, endorsement should not matter, enabling Myanmar to cloak contested actions in the authority of international law.

Literature suggests that international law exerts a "compliance pull," that individuals are more likely to support policies when they are framed as legally mandated. Because legal obligation carries moral and procedural authority, leaders invoke international law to legitimize contested actions. Consistent with this logic, international agreements can shape domestic and foreign policy attitudes by signaling global norms (Chayes & Chayes, 1993; Simmons, 2009).

Yet we know much less about whether these persuasive effects extend to international law when it is endorsed by authoritarian governments alone. This paper addresses this question directly: Can international law retain its persuasive and legitimating force *even* when endorsed solely by authoritarian states? By isolating the regime type of endorsing states, the study adjudicates messenger-based accounts of legal legitimacy. This study shifts attention from audience-level moderators to the regime identity of the messenger, testing whether the persuasive authority of international law survives authoritarian endorsement.

The topic matters for three reasons. First, it speaks directly to the foundations of legitimacy in international law, specifically, whether legal authority can persuade when the messenger itself is viewed as politically illegitimate.[1] While previous studies have traced how narratives and

[1] As will be discussed in theory section, political legitimacy is understood as the public's belief that an institution's

peer opinion shape legitimacy for international organizations (Dellmuth & Tallberg, 2015; Ecker-Ehrhardt et al., 2025; Tallberg & Zürn, 2019), we have limited understanding on whether the same legal narratives and content have varying persuasive effects when the messenger differs. This paper connects legal theories of legitimacy with empirical research on international law to develop clearer microfoundations for how legal obligation influences public opinion.

Second, this paper contributes to research on authoritarian international institutions by testing the public-facing consequences of authoritarian uses of international law. Scholars argue that authoritarian governments increasingly deploy international legal rhetoric, organizations, and procedures to legitimize contentious policies and reinforce regime stability (Bae & Moon, 2025; Cottiero & Schneider, 2026; Cottiero et al., 2025; Ginsburg, 2020; Tallberg & Vikberg, 2025). Existing work emphasizes how authoritarian-led institutions can provide a veneer of legality or collective approval for policies that might otherwise face normative resistance (Cottiero & Haggard, 2023; Debre, 2022). Yet this literature largely assumes that such legal strategies are effective. We lack direct evidence on whether international law, when used by authoritarian states, actually shapes public opinion in the intended direction. By isolating the persuasive effects of authoritarian endorsement, this paper provides a microfoundational test of a central mechanism underlying authoritarian international law.

Third, this paper expands the experimental literature on international law by broadening the scope of inquiry beyond policies aligned with liberal norms. Existing studies primarily examine how international law increases support for humanitarian or rights-protecting policies (Chaudoin & Lim, 2026; Strezhnev et al., 2019; Wallace, 2013). I instead examine whether international law retains persuasive force across the full range of policy applications, including contexts where legal principles are invoked to justify more restrictive or contested measures like deportation of refugees.

This paper's focus on refugee policy has timely implications as well. Refugee acceptance is one of the few policy domains where legal obligation is simultaneously embedded in constitutional law, and international treaties, most notably the 1951 Refugee Convention and its 1967 Protocol.

exercise of authority is justified and appropriate (Dellmuth & Tallberg, 2015; Tallberg & Zürn, 2019).

As a result, refugee law has long been understood to exert a strong normative pull, shaping political behavior and public expectations even in the absence of robust enforcement mechanisms (Cope & Crabtree, 2022). At the same time, refugee policy is susceptible to competing political framings, variously constructed as a humanitarian duty, a legal obligation, or a threat to sovereignty, making it a strategic site for rhetorical contestation (Strezhnev et al., 2019). Since refugee policy combines high legalization with high political salience, it provides a strong empirical setting to test for theories of international law's persuasive power.

This paper proceeds as follows. In the next section, I review the theoretical background of legitimacy of international law and identify limitation of applying existing theory to explain the question of whether the endorser states matter. I then introduce the main argument connecting international law and varying degree of legitimacy conditional on the type of messengers that endorse it. The second section then describes the research design, including the experimental component of online surveys. The third section discusses the main findings, followed by the fourth section of conclusion and implications.

2 Theory and Hypotheses

Public opinion takes a special place in the studies of international law. States comply with international human rights law, even in the absence of a centralized enforcement mechanism (Barsoom, 2022; Hollyer, 2011), and one of the causal chain is through domestic actors (Finnemore, 1993; Simmons, 2009). The core insight of compliance theory is that perceived legitimacy towards international law enhances the compliance with its content (Chayes & Chayes, 1993). Domestic actors advocate for compliance with international law that their states have previously ratified, serving as a mechanism that connects formal treaty ratification to meaningful international commitment (A. Chilton & Linos, 2021). Simmons (2009) states "human rights treaties matter most when they have domestic political and legal traction (p.12)." People *want* their government to comply with international law and incentivize compliance through voting (Dai, 2005). International law shapes behavior not simply because it exists, but because it is understood by the public as carrying normative weight.

Crucially, legitimacy does not arise independently from legal texts or institutional design. The concept of *opinio juris*, the belief that certain practices are legally obligatory, underscores that international law is sustained by shared social and cognitive understandings. In this sense, public opinion is not peripheral to law's authority: It is constitutive of it. Without understanding when and why citizens form beliefs about legal obligation, explanations of international law's legitimacy remain incomplete. If citizens vary in how they perceive weight of law, then the degree of the law's effectiveness in changing state policies should vary accordingly.

Given the central role of public beliefs in the making of international law, a substantial body of work on public opinion examines whether individuals change their beliefs about a policy when they learn it is prohibited by international law. Across a range of issue areas, the evidence shows a consistent pattern that exposure to international law meaningfully shifts public opinion (A. Chilton & Tingley, 2013). For instance, when informed about the existence and content of international humanitarian law, the public reduces support for torture, drone strikes, and wartime violence (Fariss, 2014; Kreps & Wallace, 2016; Lupu & Wallace, 2019; Wallace, 2013).

On the enforcement side, Putnam and Shapiro (2017) show that citizens of the United States are more in favor of imposing sanctions on foreign countries that violated international human rights law. This pattern of compliance pull and consequent political actions appears at the elite decision making level as well (Simmons & Danner, 2010).

These findings establish that international law can enjoy legitimacy and generate compliance pull. Policies that align with international law tend to receive greater public support. In much of this literature, legal authority is assumed to operate independently of the actor invoking it. If legal legitimacy derives solely from the content of legal obligation, then endorsement should not matter. If, however, credibility and regime identity shape perceptions of legality, then the legitimacy of international law may be conditional. The next section develops these competing expectations.

2.1 Sources of legitimacy behind international law

Legitimacy refers to the perception that a rule or authority is normatively appropriate and therefore deserving of voluntary compliance (Ecker-Ehrhardt et al., 2024; Levi et al., 2009; Tyler, 2006). Individuals comply not because of coercion, but because they view a rule as rightful. Legitimacy entails not only a personal belief in rightful authority but also a perception that this claim to authority is socially recognized and generally accepted by others within the relevant political community (Bicchieri, 2006; Brennan et al., 2013). [2]

What exactly is it about international law that makes people accept its authority? Once we identify the source of international law's legitimacy, or the "active ingredient," then we can readily answer whether that same mechanism would operate if international law was presented by authoritarian states. Existing research implies two distinct pathways through which international law may shape public opinion. First, individuals may respond to legal claims because they perceive law itself as carrying normative authority. Second, individuals may rely on cues about who endorses a legal claim to assess its credibility and neutrality.

Classical legal theory locates the source of legitimacy in the intrinsic properties of law itself. From this perspective, compliance stems from the perceived moral appropriateness of legal rules and the procedures through which they are generated, rather than from enforcement or coercion (Tyler, 2006). Franck (1990) similarly argues that legitimacy arises from distinctive features of law, such as clarity, coherence, and procedural fairness, which signal that a rule deserves obedience. Fuller (1969) further contends that law carries an internal morality, grounded in traits such as generality and nonretroactivity, that generates an independent obligation to comply. Under this view, legal claims exert persuasive force because they signal a morally relevant duty, not because of the identity of the actor who invokes them. This leads to the baseline hypothesis.

[2] Importantly, legitimacy is conceptually distinct from policy preference. Individuals may support or oppose a policy based on material interests or ideological agreement without attributing legitimacy to the authority advancing it. Some empirical literature uses shifts in policy support as evidence of legitimacy effects, yet theoretical discussions treat legitimacy as a belief about rightful authority rather than agreement with a specific outcome. This paper therefore distinguishes between policy evaluation and legitimacy beliefs, while recognizing that the former may serve as an observable behavioral implication of the latter.

Hypothesis 1 (Baseline)**.** *Exposure to international law affects public preferences relative to no legal framing.*

2.2 Regime type as source cues

Studies complicate the baseline account by emphasizing the conditional nature of international law's legitimacy. Institutional design explanations highlight the role of legalization (Wallace, 2013), procedural fairness (Ghassim, 2024; Ghassim et al., 2022; Tyler, 2001), and ideological alignment between legal content and individual moral beliefs (Ecker-Ehrhardt et al., 2024) in shaping perceptions of rightful authority. These accounts focus primarily on features of the legal message, what is written on the paper (both legal substance and institutional rules).

Framing theory similarly suggests that international law operates not only through its substantive content, but also through *how* that content is presented and communicated (Chong & Druckman, 2007, 2013). Political actors strategically shape public reactions by selecting and emphasizing particular legal frames. For example, Morse and Pratt (2022, 2025) show that governments can mitigate the reputational costs of non-compliance by reframing the meaning and applicability of legal rules, even when the underlying obligation remains unchanged. This body of work underscores that legal influence is communicative: perceptions of legitimacy depend not only on the rule itself, but the narrative it is articulated.

Granting that international endorsement can generate legitimacy, Chapman (2007) argues that citizens "consider the source"(p.741) of legal approval. On messenger, prior research shows that endorsement by heterogeneous actors enhances perceived impartiality. This suggests that audiences attend to characteristics of endorsers when evaluating the validity of legal claims (Claude, 1966; Grieco et al., 2011; Hurd, 2005; Thompson, 2006; Voeten, 2008). When legal outcomes emerge through multi-stage processes such as proposal, drafting, revision, and negotiation, audiences infer that these layered procedures signal neutrality and appropriate decision-making structures, which in turn shape evaluations of whether outcomes are impartial and justified (Thibaut & Walker, 1975).

I then need to answer: *How might regime type operate as a heuristic cue?* In contexts where citizens lack detailed knowledge about how international legal claims are generated or applied, regime type provides a readily available heuristic. Regime identity may serve as a shortcut for evaluating whether a legal claim reflects impartial application of rules or instrumental manipulation of them. Democracies are often associated with transparency, accountability, and rule-bound decision-making, whereas authoritarian regimes are frequently perceived as strategically selective or self-interested in their legal commitments (Simmons, 2009; Weeks, 2008). These perceptions reflect a broader association between international institutions and liberal democratic leadership. [3]

If such heuristics operate in evaluations of institutional leadership, they may extend to evaluations of regime-based endorsements. Coalitions composed exclusively of authoritarian states may be perceived as biased or self-interested. As a result, international law endorsed solely by authoritarian states may be viewed as less impartial or broadly representative, weakening its legitimacy (Gangl, 2003). This leads to a set of two hypotheses.

Hypothesis 2 (Conditional Messenger Effect)**.** *The perceived legitimacy of international law will vary by the regime type of the endorsing coalition.*

Hypothesis 3 (Directional Messenger Effect)**.** *Democratic and mixed coalitions are expected to generate greater perceived legitimacy than authoritarian coalitions.*

So far, the literature points to pathways through which legitimacy judgments may arise. International law may carry a baseline persuasive force rooted in the intrinsic properties of legal norms. At the same time, the magnitude of that persuasive effect may be conditioned by political cues about who invokes the law. Rather than posing mutually exclusive explanations, these accounts imply that legal content and messenger identity operate jointly, with messenger characteristics shaping how legal claims are interpreted and evaluated. In sum, international law exerts influence not only through its substantive content, but also through political heuristics about who invokes it, making regime type a critical, yet underexplored, dimension of legitimacy.

[3]For example, Arias and Hulvey (2025) show that although global publics generally hold favorable views of the United Nations, institutional legitimacy declines when leadership shifts to actors perceived as normatively incongruent with the institution's liberal orientation.

2.3 Heterogeneous Treatment Effects

Effect of international law on public opinion varies substantially across individuals. To deepen understanding of the conditions under which authoritarian international law is most likely to persuade, I examine treatment heterogeneity across several preregistered moderators. Prior preferences, political orientation, and dispositional traits all shape whether legal information is incorporated or resisted.

First, prior perceptions and policy predispositions can matter. Individuals enter the survey with different baseline attitudes toward international law and refugee policy. Prior work shows that attitude strength constrains persuasion. People with strong pre-existing views are less likely to revise their preferences in response to legal information (Chaudoin, 2014; Chong & Druckman, 2013; Cope & Crabtree, 2022). If international law moves preferences even among respondents with firm priors, then the compliance-pull effect operates under especially demanding conditions.

A second set of moderators captures political interest, participation, and awareness. Respondents who are more politically engaged or informed may be more likely to notice, interpret, and update based on legal cues. Likewise, political ideology, particularly the conservative–liberal divide, has been shown to structure reactions to international law (Wallace, 2013). Individuals who place low importance on living in a democracy may place little value on procedural legitimacy and thus may not differentiate among democratic, mixed, or authoritarian messengers. Conversely, respondents who feel threatened by liberal legal obligations may find authoritarian legal claims more attractive than democratic ones, even when they recognize the normative tension.

A third group of moderators concerns nationalism, cosmopolitanism, and international trust. Dispositional orientations shape how people evaluate obligations toward outsiders and how they interpret the international system (Bayram, 2017; Kertzer & Brutger, 2016). More cosmopolitan respondents may be generally more receptive to human-rights–based legal arguments, while more nationalist individuals may view sovereignty-based claims as more legitimate, potentially amplifying messenger effects when the endorsing coalition signals a particular geopolitical identity.

3 Experimental Design

The survey was fielded online in Australia (December 2024) and the Philippines (February–March 2025) using a proprietary online panel of adult respondents. [4]

Two country cases are selected to balance comparability and theoretical leverage. Australia and the Philippines are similar along several dimensions that make them suitable for cross-national replication: both are embedded in the Asia–Pacific regional order and regularly exposed to international legal claims advanced by major powers, including the United States and China. China, in particular, has become a central legal actor in the region, frequently emphasizing sovereignty, non-intervention, and deference to domestic authority, often in tension with liberal human-rights–based interpretations of international law (Ambrosio, 2008). This shared regional and geopolitical context reduces confounding from differences in the salience or content of international legal discourse.

At the same time, the cases differ in ways that are theoretically relevant to the study's hypotheses. Importantly, the theory implies that messenger effects should not be uniform across contexts. If persuasive effects were driven solely by the global image of democracies and autocracies, the same legal endorsement should exert similar influence across countries. By contrast, a genuine messenger effect predicts that the impact of international legal claims depends on how endorsing regime types are perceived by different publics. Cross-national variation therefore provides a test of whether observed effects reflect contextual interpretation of the messenger rather than a fixed reputational advantage.

Australia is characterized by strong rule-of-law institutions and a public discourse in which international law is closely associated with liberal democratic norms. As a long-standing participant in, and contributor to, the liberal international legal order, Australia is more often positioned as an agenda-setter and enforcer of international legal rules. The Philippines, by contrast, exhibits weaker rule-of-law enforcement and greater exposure to illiberal governance practices. In addition,

[4] Respondents were selected via stratified quota sampling based on key demographics (age, gender, and region).I worked with TGM research, which recruited a population-representative respondents via opt-in panels, email lists, and mobile networks. As quota targets were met, additional respondents were re-sampled from the panel to fill remaining demographic cells. Table 1 summarizes the resulting 12 treatment conditions and the pure control group. I doubled the control-group size to enable clearer comparisons across aggregated treatment categories.

its experience with international law is more frequently shaped by external scrutiny and constraint, as international legal norms are often invoked against the state in areas such as human rights and governance. This asymmetric position within the international legal order shapes how international law is experienced, as a source of authority and voice in Australia, and more often as an external constraint in the Philippines. These differing experiences may condition how international legal arguments are interpreted when they are invoked by different messengers.

These cases also strengthen the study's external validity. First, much of the experimental literature on international law focuses on Western democracies; examining similar mechanisms in a Southeast Asian context expands the scope of existing findings. Second, cross-national replication allows for testing the hypothesized effects beyond the context of consolidated democracy. Finally, messenger effects are expected to vary across countries. Even when exposed to the same legal content and endorsing states, audiences may evaluate the credibility of legal claims differently depending on their prior interactions with, and perceptions of, those states.

3.1 Vignette and Issue Area

To test how international law conditions public preferences, I focus on refugee policy. This issue area offers several advantages for identifying the persuasive effects of international law. First, refugee policy is a highly salient and polarizing topic on which citizens typically hold well-formed prior opinions (Bansak et al., 2016). Strong priors reduce the risk of overstating treatment effects, which can occur when respondents are asked to evaluate unfamiliar or low-salience issues (Mutz, 2011). By selecting an issue where attitudes are already shaped, the design provides a conservative test of whether international law can meaningfully shift preferences.

Second, refugee policy is deeply embedded in contemporary international law through decades of codification and judicial rulings. The 1951 Refugee Convention, for example, established the principle of non-refoulement, which prohibits returning refugees to countries where they face threats to life or freedom. Even non-signatories are not fully exempt, as the protection of refugees has evolved into a norm of customary international law. These legal obligations have fueled enduring tensions between states' international commitments and domestic political concerns, making refugee policy an ideal issue area to examine how international law conditions public opinion. To echo Cope and Crabtree (2022), international law on refugee policy concerns legal obligations owed to non-citizens. This incongruence between legal accountability and responsibility bearer gives the refugee issue particular political salience.

Third, it presents a hard test for the compliance-pull mechanism since attitudes toward refugees are often deeply rooted and resistant to persuasion. Strength of prior attitudes, conditional on the topic, is indeed one of the key determinants of persuasion effectiveness (Druckman, 2022a). If public preferences over a issue area where people already have strong opinion on is influenced by international law, then the same effects are likely present across other issue areas.

Sovereignty and human rights offer sharply contrasting legal frames for refugee policy. For many authoritarian leaders, the principle of sovereignty implies that states retain the exclusive right to determine how to handle refugee admissions and may prioritize non-interference over individual protection. By contrast, the human rights principle stipulates that states have an affirmative obligation to safeguard refugees from harm, including a prohibition on refoulement. Because these principles can point toward opposite policy conclusions, refugee policy provides an opportunity to examine how people respond when international law is framed around competing legal norms.

To clarify, the choice of refugee policy as the policy domain does not require that publics in Australia and the Philippines hold similar baseline attitudes toward refugees. Rather, refugee governance is an issue area in which international legal obligations are well defined, politically contested, and frequently invoked by both states with divergent regime types. This makes it a particularly suitable domain for testing whether the persuasive force of international law depends on the regime characteristics of its endorsers, even across contexts with different underlying preferences.

The survey vignette begins with a hypothetical scenario: "In one country, members of the opposition political party are enduring severe repression under the ruling government. As a result, these individuals are seeking refuge in [respondent's country]. They claim to be political refugees fleeing persecution. Their home country has officially requested our administration to repatriate them, intending to subject the political dissidents to the domestic judicial system. The current administration of [respondent's country] is faced with the choice of whether to let the refugees stay or send them back to their home country. The following questions are about your perspective on this complex issue."

The vignette was intentionally phrased in abstract and neutral terms to minimize unintended priming effects. The scenario avoids naming specific countries, leaders, or real-world events, and instead describes a generic political dispute using institutionally neutral language. This is to prevent country-specific confounding variables to influence the treatment (Brutger et al., 2023; Tomz & Weeks, 2013). Plus, it is hard to vary a specific country's regime type even in the hypothetical setting and hurts experimental realism (Druckman, 2022b). As Brutger et al. (2023) notes, salient cue-givers show larger treatment effect, which makes my choice of unspecified country name a harder test than directly mentioning well-known country examples of autocracies or democracies. Respondents were explicitly informed that the scenario was hypothetical, though designed to resemble real-world dilemmas, in order to balance realism with experimental control.

Next, the treatment groups receive additional prompts that have varying treatments. Pure control group does not receive any additional prompts. There are two treatments in my experiment. First, I vary the information about the types of state leaders that supported international law. I mention that international law relevant to refugee policies is supported by a group of countries [that are non-democracies / that are democracies / that includes both democracies and non-democracies]. For the treatment of "both democracies and non-democracies," I randomize the order of each topic to prevent the priming effect. [5] I also adopted the term "non-democracies" instead of autocracy or dictatorship. This follows Tomz and Weeks (2013)'s note that tyrants, dictators can come across as being evocative language, while non-democracy offers more neutrality. The terminology adopted here is therefore intended to cue regime type while minimizing emotional priming. [6]

Second, I vary the information about the contents of law. For those treated with the "sovereignty"

[5] Similar approach by A. S. Chilton and Versteeg (2016)

[6] Dafoe et al. (2018) further warns that even simply labeling a country as a democracy can send implicit cues about the country's geographic location. To test whether my vignette passes the information equivalence assumption, I ask respondents at the very end of the survey if this hypothetical scenario reminded them of any countries in particular.

message, I give the prompt that states international law says that "Countries have to respect national sovereignty. Some legal experts argue that if our administration refuses to send the refugees back to their home country despite the repatriation request, it may be seen as a violation of the sovereignty of the refugees' home country."

For the "human rights" message treatment, the prompt states " law says that countries have to respect human rights. Some legal experts argue that if our administration decides to send the refugees back to their home country despite the refugees' potential exposure to violence, it may be seen as a violation of the obligation to protect human rights." This vignette closely resembles actual rhetoric of authoritarian state leaders.

The factorial design is well suited to the core question of whether messenger identity shapes public reactions to international law. Without varying the legal content, a null messenger effect would be ambiguous: respondents might reject the statement because they distrust the messenger, or simply because they disagree with the substantive content of the legal claim. By introducing two substantively distinct legal frames (human rights and sovereignty), the design separates disagreement with the *message* from reactions to the *messenger*. Messenger effects can therefore be evaluated across substantively different legal principles, ensuring that responses reflect source-based judgments rather than simple policy disagreement. In this way, the factorial structure isolates the persuasive consequences of messenger legitimacy.

Table 1: Survey Experiment Design: Modified 2×3 Factorial

	Message Type		
Messenger Type	Sovereignty	Human Rights	Control
No Messenger	T1	T2	Control
Democracies	T3	T4	
Non-democracies	T5	T6	
Both	T7	T8	

Notes: The design crosses 2 message types (Sovereignty, Human Rights, Control) with 3 messenger conditions (No Messenger, Democracies, Non-democracies, Both). The Control column spans all four messenger conditions, but messenger assignment is not meaningful when no message is provided.

I do not include a fully crossed condition in which respondents receive both legal messages (sovereignty and human rights) and a specific messenger type. Presenting respondents with two competing legal claims in a single treatment introduces conceptual and interpretive ambiguity. It becomes unclear whether respondents are reacting to the relative strength of the frames, the perceived inconsistency of the law, or the identity of the endorsing states. Research in political communication shows that competing frames do not simply cancel out. Instead, higher-quality or more resonant frames dominate (Chong & Druckman, 2007). A treatment containing both messages would therefore test which legal norm is stronger, not the effect of messenger legitimacy. Because this would confound the two mechanisms, the design limits the "both messages" condition to the no-messenger baseline.

I also included Factual Manipulation Checks (FMCs) targeting both the experimental component and the general scenario. FMCs serve to validate whether respondents accurately perceived the randomized information about the messenger and the legal content (Mutz, 2021). Following Kane and Barabas (2019), placement of FMCs immediately after the treatment is not expected to bias treatment effects, so I included them directly after the vignette to reduce cognitive fatigue. Importantly, FMCs were used only for validation, and respondents were *not* dropped based on their FMC responses.

Separate from FMCs, I implemented an instructional attention check at the start of the survey, prior to the vignette. This item required respondents to follow a simple directive (e.g., selecting a specified option) and was designed to screen out individuals who were not reading instructions carefully. Then, I excluded respondents who failed this instructional check at the beginning of the survey. These cases were treated as *non-participants* rather than as survey dropouts. Because the firm continued sampling until all demographic quotas were met, the exclusion of these inattentive respondents did not alter the demographic balance of the final sample.

All analyses use the full set of respondents who passed the instructional attention check. In Australia, 1,227 respondents were randomly assigned to either the control group or one of the eight treatment cells. The mean approval of refugee admission in the pure control group was 3.74 on a five-point scale. In the Philippines, 1,134 respondents were similarly randomized, with a control-group mean approval of 3.21. Across both countries, respondents exposed to the human-rights frame, which emphasizes that returning the refugees would violate international human-rights obligations, expressed substantially higher support for accepting refugees than both the control group and those assigned to the sovereignty frame.

3.2 Outcome Variables

Respondents reported whether the government should admit the refugees or repatriate them to their home country. This item measures support for the refugee-admission policy referenced in the vignette. Responses are recorded on a five-point scale (1 = strongly oppose refugee repatriation; 5 = strongly support admission).

I also ask whether the respondent thinks it is legally appropriate to send the refugees back to their home country or to let them stay. Even though the respondents might not approve of one policy, they might still think it is legally appropriate. Strezhnev et al. (2019), for instance, shows that conservatives often failed to recognize that rejecting refugee admissions violated international law when elites provided contradictory cues. Building on this insight of confirmation bias, I distinguish between perceived legitimacy and perceptions of legality. The distinction is important in practice. Even when individuals do not support a particular foreign policy, they may still accept the authority of an international judicial ruling once the validity of refugee policy is adjudicated by a third party.

In other words, this measure captures impacts that extend beyond personal policy preferences to judgments about the legal validity of each option under international law as endorsed by different messengers. Because these judgments are categorical, I operationalize legality differently from approval. Unlike approval (measured on a five-point scale), legality is coded as a binary outcome, reflecting the binary nature of judicial rulings as a simple "yes" or "no." Respondents were asked to complete the response to, "If this question were brought to a UN legal body, it would say...: (1) "[Respondent country] should send political refugees back to their home country," or (2) "[Respondent country] should let political refugees stay." Response is not based on personal policy preference, but what is *expected* decision from the UN.

4 Results

Legitimacy of international law is conditioned by both the legal content and the type of endorsers. Across two countries, respondents who were exposed to international law that mandates states to repatriate political refugees were more likely to support policy of sending them back. Figure 1 reports the percentage change in refugee approval relative to the country-specific control mean, with 95% confidence intervals.[7]

Given the complexity of the factorial design, varying message type, messenger type, and country, the result section proceeds in three steps. First, I examine the message effect, the average impact of introducing an international legal obligation, irrespective of who endorses it. Second, I analyze messenger effects within each message type. Because the sovereignty and human-rights messages push preferences in opposite substantive directions, evaluating messenger variation within each frame provides cleaner interpretation. Finally, I compare results across Australia and the Philippines to assess whether these patterns generalize across two countries' political contexts. I repeat this three stages for both dependent variables, approval and legality.

4.1 Approval Rate

Figure 1 illustrates how international law reshapes support for refugee acceptance across message types and messengers. On average, legal information can shift public preferences, but the magnitude and direction of these effects vary sharply by the two conditioning variables. Human-rights message, on average, do not generate meaningful change in approval rate in Australia. In the Philippines, however, human-rights messages increase support by 7.49 percentage points (CI: 3.76,

[7] Percentage change is calculated as the difference between the treatment group's mean approval and the country-specific control mean, divided by the control mean and multiplied by 100. Because Australia and the Philippines differ in their baseline mean approval values, expressing outcomes as percentage change relative to each country's own control mean centers the scale at zero for both samples. This normalization allows the plot to highlight how much each treatment increased or decreased approval within a country, independent of its overall level of support, making differences in treatment effects visually comparable across countries. See Appendix A for raw points.

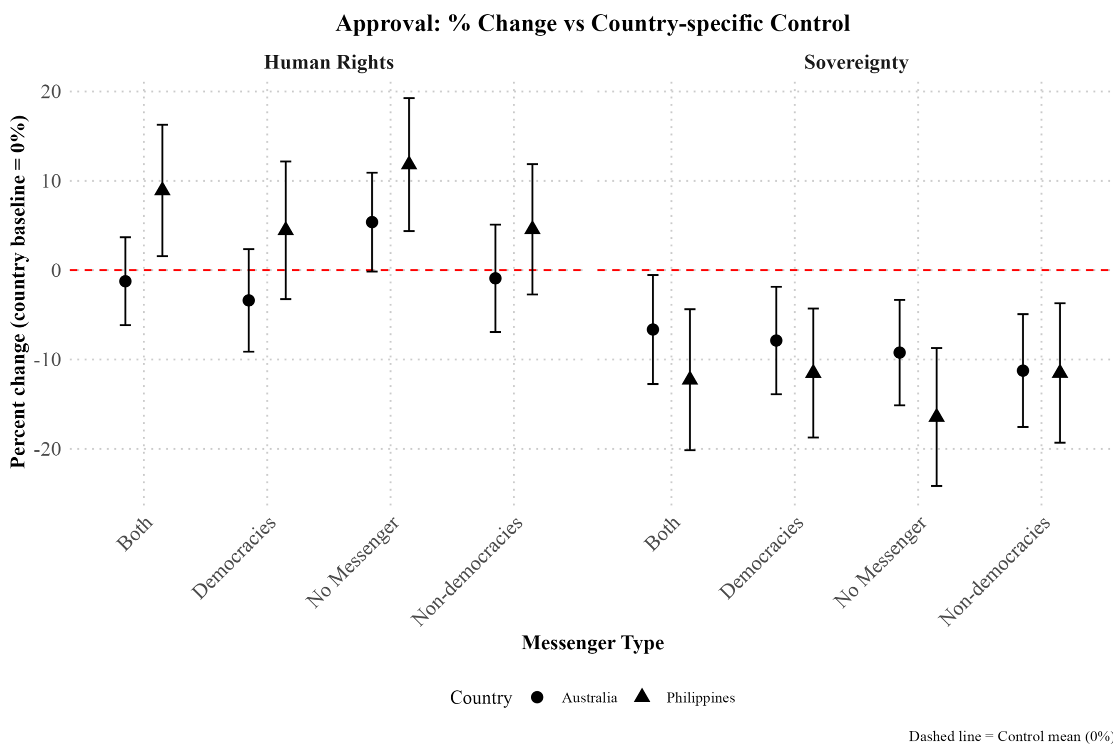


Figure 1: Percentage change in approval rates by treatment

11.21), on average across messenger conditions, indicating a substantively meaningful positive response to the human-rights obligation.

Sovereignty message, which frames repatriation as required under international law, generates much consistent effects. Approval for refugee acceptance declines by 8.75 percentage points(CI: 11.79, 5.70) compared to the control group in Australia and by 12.94 percentage points(CI: 16.77, 9.12) in the Philippines, on average. Substantively, this means that the average Filipino respondent moves from a baseline approval of 3.21 on a five-point scale in the control group to approximately 2.8 following exposure to the sovereignty treatment. This represents a shift of nearly half a scale point, moving respondents from mild approval toward ambivalence or mild disapproval.

The patterns become more nuanced once messenger condition is introduced. Under the human-rights message, messenger identity affects responses in asymmetric ways. In Australia, approval increases only when the legal obligation is presented without any messenger. Estimated differences between messenger conditions are small and statistically indistinguishable from one another. A similar pattern appears in the Philippines. Filipino respondents react positively when the human-rights obligation is delivered without a messenger or by a mixed coalition, but do not show meaningful shift in opinion when it is attributed to a group of countries with one regime type.

Under the sovereignty message, messenger variation matters much less. In both countries, sovereignty-based obligations reduce approval in every messenger condition, compared to the control group. Although the magnitude varies somewhat across messengers, the direction of the effect is consistently negative. Filipino respondents react most strongly in the no-messenger condition, with a decline of 16.44 percentage points, whereas Australian respondents react most strongly when the sovereignty obligation is endorsed by non-democracies, with a decline of 11.24 percentage points.

4.2 Legality

Legality is measured as a binary assessment of whether the policy is viewed as legally permissible under international law. For analysis, responses are coded as 1 if the respondent selected the "let refugees stay" option and 0 if they selected the "send refugees back" option. Group means therefore represent the share of respondents who believe international law favors admitting refugees. Then, treatment effects thus reflect percentage-point changes in the proportion of respondents who think that letting refugees stay is the legally appropriate ruling relative to the Control group.

Across both samples, the legality results show a clear asymmetry between human-rights and sovereignty messages. Human-rights messages increase perceived legality only in the Philippines, while sovereignty messages reliably reduce legality judgments in both countries. In Australia, the Control group reports a high baseline legality rate of 0.75. Human-rights messages generate only a small and statistically indistinguishable change in legality (–3.07%, CI: –7.09, 0.95), and messenger identity does not meaningfully condition this response. Australian respondents make nearly identical legality judgments across all messenger conditions, remaining clustered around the Control group's baseline.[8]

[8] Indeed, under a 90% confidence interval, the democracies–human-rights condition in Australia spans from −14.89

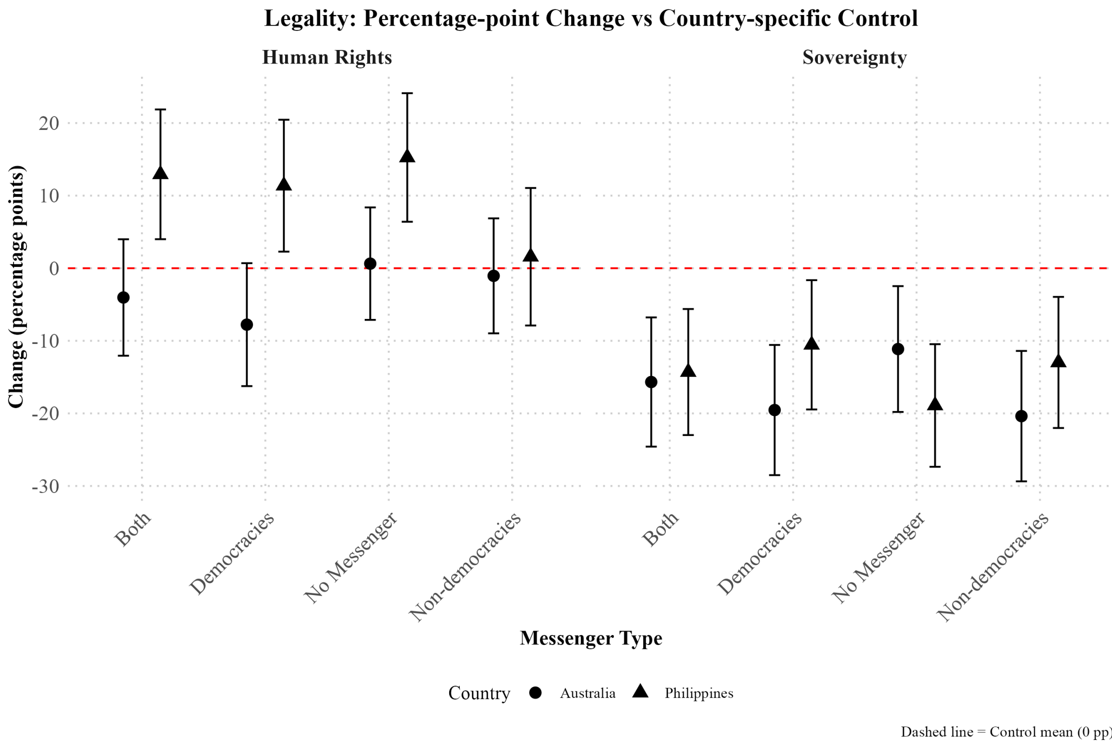


Figure 2: Percentage change in legality rates by treatment

The patterns diverge sharply in the Philippines. The control group in the Filipino respondent pool shows a much lower baseline legality rate (0.493), leaving considerably more room for persuasion. Here, the human-rights message meaningfully increases the share of respondents judging acceptance as lawful, but only under certain messenger conditions. When the message is presented anonymously, legality rises by 15.26 percentage points (CI: 6.40, 24.11), and when both democratic and non-democratic states jointly endorse the message, legality increases by 12.94 percentage points (CI: 3.99, 21.87). These are the only messenger conditions under which human-rights claims meaningfully shift legality judgments.

Sovereignty messages tell a more consistent story across both countries. Regardless of messenger identity, sovereignty-framed obligations sharply reduce the proportion of respondents who view admission as legally permissible. In Australia, sovereignty lowers legality by 16.68 pp (CI: –21.12, –12.25) compared to the control group. The largest drop occurs when democracies deliver

to −0.66 percentage points, indicating that legality judgments may even decline slightly when the obligation is attributed specifically to democratic states.

the message (–19.54 pp, CI: –28.51, –10.57), though all estimated effects lie well below the Control group's confidence interval.

The Philippine sample shows similarly strong declines. Legality falls by 18.91 pp (CI: –27.36, –10.46) under the no-messenger condition and by -14.30 pp (CI: –22.98, –5.62) under the both-messenger condition. In every messenger condition, respondents become less likely to judge refugee acceptance as legally permissible, indicating that messenger credibility plays little to no moderating role. Rather, sovereignty claims exert a strong, directional influence on legality assessments irrespective of who delivers them.

4.3 Discussion

Experiment results reveal two interesting findings. First, Filipino respondents do not become more receptive of refugees when exposed to human rights law, but do believe that the UN legal body will think it is legally appropriate to accept refugees. This effect is null in Australia case. Australian respondents do not change their expectation on how the UN community will respond, even when they are given with additional information on human rights law. Second, sovereignty message that mandates repatriation of refugees is always effective in making people become less receptive of refugees. In other words, messenger does *not* condition the perceived legitimacy when the message is "illiberal." These two findings are worth further explanation.

4.3.1 Filipino experience with international law

The divergence between approval and legality judgments among Filipino respondents likely reflects the distinction between personal policy preferences and expectations about external legal authority. Unlike Australian respondents, Filipinos may be more accustomed to experiencing international law as an external evaluative force directed at their own government rather than as a domestically internalized normative commitment. In Australia, international law is more often embedded within domestic legal and political discourse as part of a liberal normative order. In the Philippines, by contrast, international law is frequently encountered through criticism and external scrutiny, particularly on human-rights issues.

As a result, exposure to human-rights–based legal arguments increases Filipino respondents' expectations about how a UN legal body would rule, even when their own policy preferences remain unchanged. In other words, respondents update their beliefs about what international law requires without necessarily internalizing that requirement as a personal policy preference. This pattern is consistent with a dual-process account of legitimacy, in which individuals can recognize the authoritative status of international legal norms while maintaining independent policy attitudes shaped by domestic political considerations.

This finding also expands the social cue theory (Chu, 2025; Pauselli, 2023). While existing work emphasizes in-group/out-group identification between sender and recipient, my results suggest that persuasion may also depend on how recipients have historically encountered international legal authority. In contexts where international law is experienced as external scrutiny or constraint, legal cues may function primarily as signals about global judgment rather than as normative commitments to be internalized. Messenger effects are therefore shaped not only by identity proximity, but also by the institutional and historical context through which the public experience international law. International law may persuade differently in contexts where it is perceived as voice and participation than where it is perceived as constraint and evaluation.

4.3.2 Why sovereignty, but not human rights?

I now turn to the second question that arises from the finding: Why would citizens in both countries be more consistently persuaded by the sovereignty than human rights message? Across both Australia and the Philippines, the data show a consistent pattern that respondents are most persuaded by the human-rights message only when the message is presented anonymously, and specifying any messenger (democratic, non-democratic, or mixed) substantially attenuates the effect. This was not the case for sovereignty message treatment. The messenger condition appears largely muted across both countries and across both outcome measures in the sovereignty treatment.

While this is beyond the preregistered hypotheses, the results reveal an interaction between message type and messenger identity. I offer a few explanations based on existing literature. First, human-rights rhetoric is highly politicized, and messenger attribution activates skepticism. When a human-rights claim is tied to a particular state, respondents may interpret it as strategic, self-interested, or hypocritical, especially if the messenger's own rights record is salient. The same obligation that appears principled in the abstract becomes more contestable once it is politically owned. Sovereignty-based obligations do not draw on the same long-standing normative capital accumulated through the liberal international order (Bae & Salvati, 2026). Unlike human rights, whose authority has been reinforced through decades of institutionalization and socialization within liberal international organizations, sovereignty claims (especially in the context of non-interference) do not benefit from this inherited legitimacy (Lake et al., 2021).

Second, sovereignty-based claims in this context align with a status-quo policy position, returning refugees rather than extending protection. Because the recommendation involves limiting redistribution and maintaining national discretion, respondents may perceive the policy as relatively low-cost. When legal claims align with intuitive or protectionist preferences, audiences may rely less on messenger credibility and more on the substantive direction of the policy itself. Human rights have long been embedded in liberal advocacy networks and associated with identifiable state coalitions, making messenger identity highly diagnostic. Sovereignty, by contrast, has only recently been systematically invoked to justify restrictive migration policies, particularly by non-democracies.

Indeed, mixed reaction to two different legal content shows precisely why message variation is needed to even understand the messenger effect. A key feature of the design was that both the content of the legal argument (human rights vs. sovereignty) and the identity of the messenger were experimentally varied. Without this factorial structure, messenger effects would be difficult to interpret. For example, if only a sovereignty message were used, null differences across messengers could reflect the inherent persuasiveness of the sovereignty frame rather than the irrelevance of messenger credibility. Conversely, variation in the human-rights message reveals that messenger identity strongly conditions persuasion. By manipulating both dimensions, the experiment separates rejection of the message from indifference to the messenger and allows message × messenger interactions to be interpreted cleanly.

These findings are consistent with expectation-based accounts of source credibility, which argue that audiences evaluate messages in light of prior associations between actors and issues. When a legal claim is strongly associated with a particular political identity, attribution may trigger skepticism about strategic motivation rather than signal neutral legal obligation (Chapman, 2009; Druckman, 2001, 2004). Similarly, Chaudoin and Lim (2026) shows that citizens are skeptical of ICC allegation when it is from the United States, but not from ”swing states” that are not geopolitically aligned. Human-rights principles are closely linked to liberal democratic advocacy, making their political attribution more diagnostic and potentially more contestable. Sovereignty claims, by contrast, do not activate the same set of expectations in this context, reducing the salience of messenger identity. In this sense, messenger effects appear most pronounced when legal norms are normatively loaded and politically associated with identifiable actors.

4.4 Heterogeneous Treatment Effects

I next examine whether the persuasive force of international law varies across individual characteristics. For each moderator, I estimate marginal-contrast effects from an interaction model and compute the difference in predicted approval between each treatment condition and the pure control group. Across most attitudinal and ideological traits (including cosmopolitanism, nationalism, international trust, political awareness, and ideology), treatment effects remain substantively similar to the main results. Sovereignty messages reduce approval broadly across subgroups, while human-rights effects remain concentrated in specific messenger conditions. More meaningful heterogeneity emerges along three dimensions; prior moral evaluations of deportation, beliefs about government responsibility toward refugees, and democratic efficacy. I convert 5-point attitudinal scales into three-category ordinal bins (Low / Medium / High). [9]

[9] This binning procedure follows methodological considerations. Without binning, many treatment–moderator combinations would include too few respondents to support stable marginal-contrast estimates, especially once split by country, message type, and messenger type. Second, binning improves interpretability by mapping the original scales

For political engagement, I construct a measure of democratic efficacy by averaging two closely related items, political efficacy ("How much would you say the political system in your country allows people like you to have a say in what the government does?") and perceived electoral influence ("Some people think that having honest elections makes a lot of difference in their lives; other people think that it doesn't matter much. How important would you say is having honest elections for you?"). These items capture an individual's perceived agency and participation within democratic processes. After averaging the two variables, I round the resulting continuous value to the nearest integer and collapse it into a binary indicator (Low / High). This approach reduces measurement noise inherent in single-item scales and produces a more reliable measure of political engagement that aligns conceptually with the literature on political efficacy.

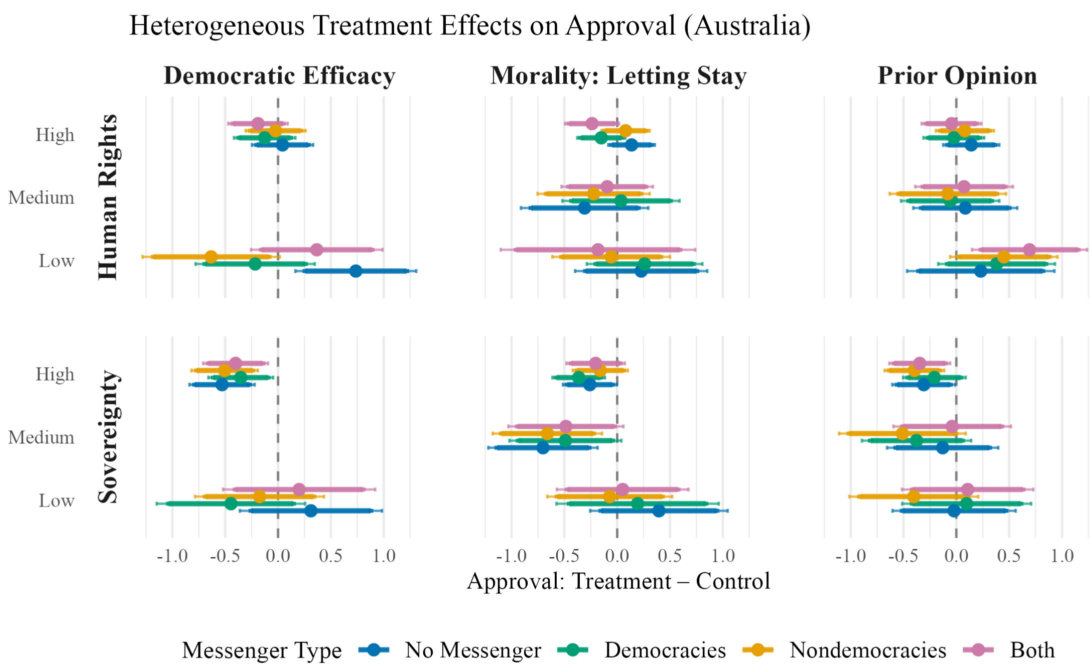


Figure 3: Heterogeneous Effects

In Australia, heterogeneous treatment effects follow a clear and internally consistent pattern across moral, normative, and political predispositions. First, moral evaluations strongly structure baseline attitudes toward refugee acceptance, yet sovereignty-based arguments continue to exert

onto substantively meaningful distinctions.

influence even among individuals with strong ethical commitments. As shown in Figure 3, the sovereignty message reliably reduces approval across all moral subgroups—including respondents who rated deportation as “somewhat immoral” or “extremely immoral.” While these individuals maintain the highest overall support for letting refugees stay, the downward shift in predicted approval remains sizable and consistent.

A similar pattern emerges among respondents who believe it is an important government responsibility to offer refuge to those fleeing violence. These respondents who are normatively predisposed toward acceptance experience some of the largest declines in approval when confronted with sovereignty frames, except for democracies’ endorsement. In contrast, those who view refugee protection as unimportant show relatively little movement under sovereignty cues but respond noticeably to human rights messages. In this latter group, human rights arguments strengthen support for acceptance, particularly when the message is endorsed by both regime types or by non-democracies. These between-messenger differences indicate that individuals with weaker priors are more sensitive to the source of the legal argument, whereas individuals with strong pro-refugee priors react most strongly to the substantive content of the sovereignty appeal.

Patterns in the Philippines mirror those observed in Australia but exhibit greater variability in magnitude. As in Australia, sovereignty cues consistently reduce approval across all moral subgroups, including respondents who view deportation as morally wrong before the treatment. The magnitude of these shifts is often larger than in Australia, especially among respondents with strong moral commitments to allowing refugees to stay. This indicates that sovereignty appeals are capable of overriding substantial moral predispositions, even among individuals who would otherwise strongly oppose returning refugees to danger.

Prior beliefs about government responsibility show the same directional pattern. Respondents who believe the government should provide refuge experience notable declines in approval when presented with sovereignty-based arguments. Those who view refugee protection as important also shift upward, when presented with human rights cue, but *only* when it was provided by both democracies and nondemocracies and without any messenger specification.

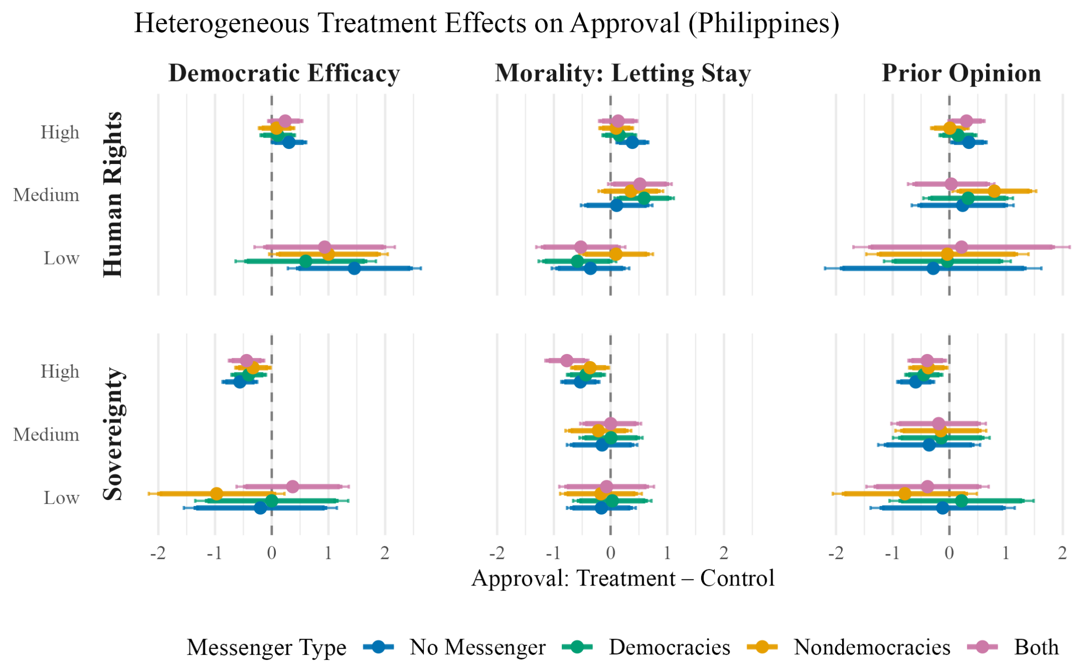


Figure 4: Heterogeneous Effects

Political interest and political efficacy similarly moderate treatment effects. Politically engaged respondents in the Philippines show clear and sometimes amplified negative reactions to sovereignty messages. As in Australia, these individuals begin with higher baseline support for acceptance but nonetheless shift meaningfully toward rejection under sovereignty framing. Given that politically attentive respondents typically consume more information about government action, this result suggests that sovereignty arguments may hold particular rhetorical power among those most invested in understanding political processes.

Turning to legality, heterogeneous treatment effects in Australia are weaker than for approval but follow a similar pattern. As shown in Figure 5, respondents who view deportation as morally wrong are already less likely to say that sending refugees back is legally permissible, yet sovereignty-based arguments further depress legality judgments within these groups. Among those who rated deportation as "somewhat" or "extremely" immoral, sovereignty messages, especially when endorsed by non-democracies or by both regime types, produce some of the largest declines in the perceived legality of rejection. By contrast, human-rights messages induce only small and statistically fragile changes in legality across moral subgroups, with no consistent evidence that they strengthen the belief that accepting refugees is legally required.

Respondents who believe refugees deserve protection start from higher baseline legality for acceptance, but sovereignty messages again reduce the share who view acceptance as legally appropriate, particularly under non-democratic or mixed-endorsers. Human-rights messages generate at most modest increases in legality among those with weak prior sense of responsibility and are generally indistinguishable from zero elsewhere.

Democratic efficacy exhibits the clearest moderation. Respondents with high democratic efficacy, those who believe they have a voice in political decision-making, show sharper declines in legality under sovereignty cues than those with low efficacy. In contrast, human-rights messages remain flat across efficacy levels, indicating that politically empowered individuals do not interpret human-rights information as changing the legal permissibility of refugee acceptance.

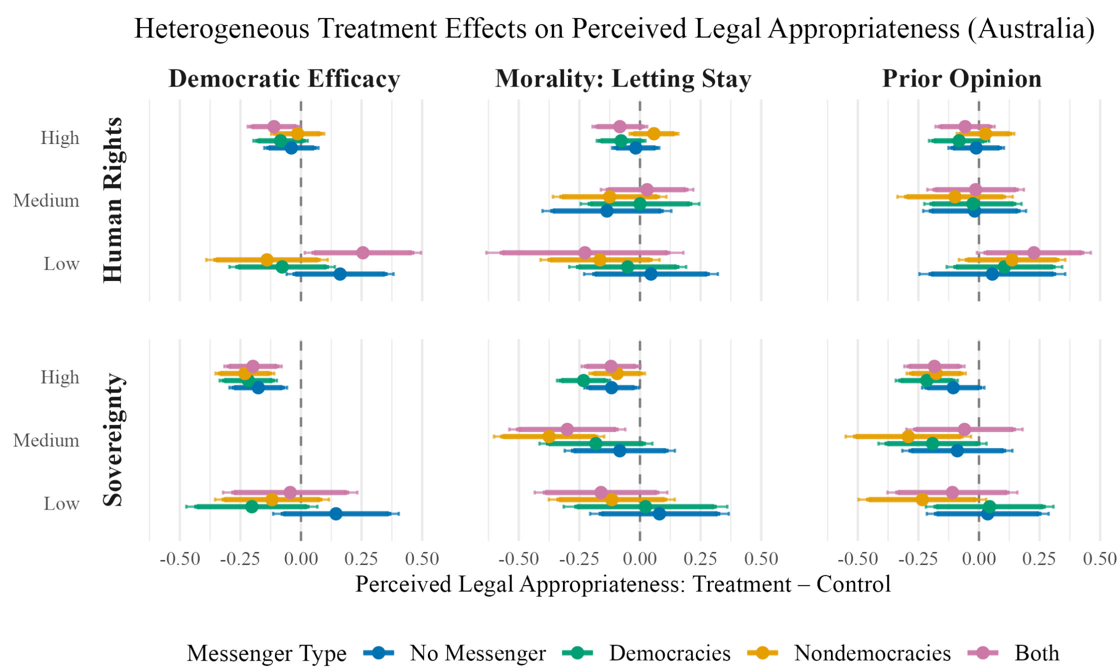


Figure 5: Heterogeneous Effects

In the Philippines, heterogeneous legality effects are more clearly aligned with the main message results. As shown in Figure 6, sovereignty messages consistently reduce the perceived legality of accepting refugees across moral subgroups, with especially large declines among respondents who judged deportation to be "somewhat immoral" or "extremely immoral." These effects are most pronounced when the sovereignty obligation is attributed to non-democracies or to a mixed group of regimes, indicating that sovereignty frames can override strong moral objections and push respondents toward the view that rejection is legally justified.

Human-rights messages, by contrast, substantially increase legality judgments among Filipinos who already hold strong pro-refugee priors. Respondents who view refugee protection as an important government responsibility, or who morally oppose deportation, become more likely to say that acceptance is legally appropriate when exposed to human-rights obligations, particularly under the no-messenger and mixed-endorser conditions. Among those with weaker priors, messenger identity plays a larger role. Human-rights messages attributed solely to non-democracies produce little legal updating, whereas the same content endorsed by both regime types yields a meaningful boost in perceived legality.

Democratic efficacy again moderates responses. High-efficacy respondents consistently register a positive shift in legality when given a human-rights frame, whereas low-efficacy respondents show smaller or null effects. Sovereignty messages reduce legality across all efficacy levels, but the drop is largest among high-efficacy respondents.

4.5 Indirect Mechanism Audit

To probe the pathways through which messenger identity might shape responses to international legal claims, the survey included four post-treatment measures capturing commonly theorized mechanisms: perceived impartiality, trust in the endorsing coalition, concern for international image, and concern about retaliation. I estimate causal mediation models following Tingley et al. (2014), decomposing each treatment effect into a direct component and an indirect component operating through each mediator (ACME). This allows us to assess whether messenger-specific persuasive effects operate through neutrality, trust, reputation, or retaliation concerns, or whether the substantive content of the legal message itself accounts for most of the shift in preferences

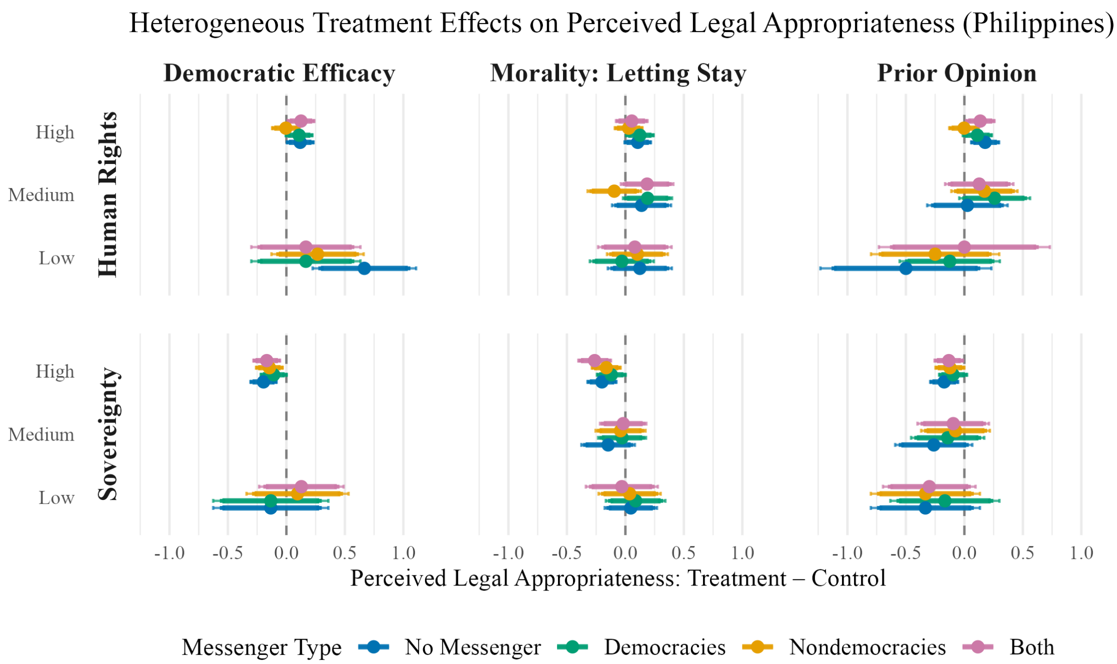


Figure 6: Heterogeneous Effects

(Brutger & Kertzer, 2018; Kim, 2019).

Appendix Figure 7 summarizes the estimated ACMEs by message and messenger type. Across both countries, indirect effects are generally small and statistically fragile relative to the total effects. In particular, the human-rights message produces little systematic mediation through any of the four measures, even in conditions where the message shifts legality judgments in the Philippines. The most visible pattern is a modest increase in retaliation concern in some sovereignty conditions (especially when the message is endorsed by non-democracies), but these indirect effects remain substantively small compared to the direct effect of the legal claim itself.

The mediation results remain largely null even in conditions where messenger identity substantively conditions outcomes, such as the human-rights treatment in the Philippines. This suggests that messenger effects are not primarily operating through generalized trust, perceived bias, reputational concern, or fear of retaliation. Rather, attribution may influence how respondents assess the legal claim itself, without substantially altering their broader evaluations of the endorsing coalition. Full results are reported in Appendix E.

Because treatment conditions varied in informational complexity, I examine whether differential comprehension explains the main results. Respondents completed factual manipulation checks (FMCs) regarding both the legal frame and the messenger identity. Passage rates varied across treatment conditions, indicating that some message–messenger combinations were more cognitively demanding.

To assess whether the substantive findings depend on respondent attention, I replicate all primary analyses using only the subsample of respondents who passed all manipulation checks. The core patterns remain substantively unchanged. Sovereignty messages continue to reduce approval and legality across messenger conditions, and human-rights effects in the Philippines remain concentrated in the no-messenger and mixed-endorser conditions. Estimated effect sizes are similar in direction and magnitude to those reported in the full sample. Full results for attentive-only analyses are reported in Appendix G, and further discussion is in Appendix F.

5 Conclusion

International law is often understood to persuade because it carries legitimacy, yet existing evidence largely assumes a liberal endorsement environment. This study examines whether international law retains persuasive force when endorsed by non-democratic states and whether messenger identity conditions its effects. Using a factorial survey experiment in Australia and the Philippines, I vary both legal content (human rights versus sovereignty/non-interference) and messenger type (democracies, non-democracies, mixed coalitions, or no attribution).

The findings demonstrate that international law can shape both policy preferences and legality judgments even when endorsed by non-democratic states. However, these effects are conditional. Sovereignty-based claims exert consistent influence across messenger types, whereas human-rights arguments are more sensitive to attribution. The persuasive authority of international law therefore reflects an interaction between legal content and messenger identity rather than deriving solely from either dimension.

These results contribute to the literature on the microfoundations of legal legitimacy by showing that compliance pull persists beyond a purely liberal endorsement context, while also clarifying the limits of messenger-based credibility. They further inform research on authoritarian engagement with international institutions. Although illiberal governments increasingly invoke international legal principles, the effectiveness of such rhetoric depends on how legal norms and endorsers jointly shape public interpretation. More broadly, the findings underscore that the legitimacy of international law is not fixed, but contingent on the interaction between substantive norms and political cues. Understanding when legal rhetoric persuades therefore requires attention not only to the content of international law, but also to the political identities of those who invoke it.

The introductory example of Myanmar's appeal to sovereignty illustrates the broader stakes of this inquiry. The results suggest that sovereignty-based legal arguments can influence public evaluations even when advanced by non-democratic actors. This finding also implies that authoritarian invocations of international law should not be dismissed as purely symbolic. Their persuasive impact depends on how legal principles and political identities interact in the minds of audiences.

References

Ambrosio, T. (2008). Catching the 'Shanghai Spirit': How the Shanghai Cooperation Organization Promotes Authoritarian Norms in Central Asia. *Europe-Asia Studies*, *60*(8), 1321–1344. https://doi.org/10.1080/09668130802292143

Arias, S. B., & Hulvey, R. A. (2025). China's leadership in the United Nations: Image management and institutional legitimacy. *The Review of International Organizations*. https://doi.org/10.1007/s11558-025-09604-4

Bae, J., & Moon, D. (2025). Regime-type composition of an intergovernmental organisation and dispute settlement mechanisms. *The British Journal of Politics and International Relations*, 13691481251329774. https://doi.org/10.1177/13691481251329774

Bae, J., & Salvati, F. (2026). United Nations Discourse on Liberalism across Regime Types. *Global Studies Quarterly*, *6*(1), ksag001. https://doi.org/10.1093/isagsq/ksag001

Bansak, K., Hainmueller, J., & Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science*, *354*(6309), 217–222. https://doi.org/10.1126/science.aag2147

Barsoom, P. N. (2022). Is the Good News about Compliance Good News about Cooperation?, 29.

Bayram, A. B. (2017). Due Deference: Cosmopolitan Social Identity and the Psychology of Legal Obligation in International Politics. *International Organization*, *71*(S1), S137–S163. https://doi.org/10.1017/S0020818316000485

Bicchieri, C. (2006). The grammar of society : The nature and dynamics of social norms [ISBN: 0521573726 Publication Title: The grammar of society : the nature and dynamics of social norms]. http://www.myilibrary.com?id=41708

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). Explaining norms [Edition: 1st ed. ISBN: 9780199654680 Publication Title: Explaining norms]. http://catdir.loc.gov/catdir/enhancements/fy1406/2013939268-t.html

Brutger, R., & Kertzer, J. D. (2018). A Dispositional Theory of Reputation Costs. *International Organization*, *72*(3), 693–724. https://doi.org/10.1017/S0020818318000188

Brutger, R., Kertzer, J. D., Renshon, J., Tingley, D., & Weiss, C. M. (2023). Abstraction and Detail in Experimental Design. *American Journal of Political Science*, *67*(4), 979–995. https://doi.org/10.1111/ajps.12710

Chapman, T. L. (2007). International Security Institutions, Domestic Politics, and Institutional Legitimacy. *Journal of Conflict Resolution*, *51*(1), 134–166. https://doi.org/10.1177/0022002706296177

Chapman, T. L. (2009). Audience Beliefs and International Organization Legitimacy. *International Organization*, *63*(4), 733–764. https://doi.org/10.1017/S0020818309990154

Chaudoin, S. (2014). Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions. *International Organization*, *68*(1), 235–256. https://doi.org/10.1017/S0020818313000386

Chaudoin, S., & Lim, T. (2026). Persuasion and Backlash from International Law in Global Swing States.

Chayes, A., & Chayes, A. H. (1993). On Compliance. *International Organization*, *47*(2), 175–205. http://www.jstor.org/stable/2706888

Chilton, A., & Linos, K. (2021). Preferences and Compliance with International Law. *Theoretical Inquiries in Law*, *22*(2), 247–298. https://doi.org/10.1515/til-2021-0023

Chilton, A., & Tingley, D. (2013). Why the Study of International Law Needs Experiments. *Columbia Journal of Transnational Law*, *52*, 64.

Chilton, A. S., & Versteeg, M. (2016). International law, constitutional law, and public support for torture. *Research & Politics*, *3*(1), 2053168016636413. https://doi.org/10.1177/2053168016636413

Chong, D., & Druckman, J. N. (2007). Framing Theory. *Annual Review of Political Science*, *10*(1), 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

Chong, D., & Druckman, J. N. (2013). Counterframing Effects. *The Journal of Politics*, *75*(1), 1–16. https://doi.org/10.1017/S0022381612000837

Chu, J. A. (2025, April). *Social Cues: How the Liberal Community Legitimizes Humanitarian War* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009557276

Claude, I. L. J. (1966). Collective legitimization as political function of the united nations. *20*(3), 367–379.

Cope, K. L., & Crabtree, C. (2022). Migrant-Family Separation and Higher-Order Laws' Diverging Normative Force. *The Journal of Legal Studies*, *51*(2), 403–426. https://doi.org/10.1086/721617

Cottiero, C., Hafner-Burton, E. M., Haggard, S., Prather, L., & Schneider, C. J. (2025). Illiberal regimes and international organizations. *The Review of International Organizations*, *20*(2), 231–259. https://doi.org/10.1007/s11558-024-09556-1

Cottiero, C., & Haggard, S. (2023). Stabilizing Authoritarian Rule: The Role of International Organizations. *International Studies Quarterly*, *67*(2), sqad031. https://doi.org/10.1093/isq/sqad031

Cottiero, C., & Schneider, C. J. (2026). International Financial Institutions and the Promotion of Autocratic Resilience. *International Organization*, 1–38. https://doi.org/10.1017/S0020818325101276

Dafoe, A., Zhang, B., & Caughey, D. (2018). Information Equivalence in Survey Experiments. *Political Analysis*, *26*(4), 399–416. https://doi.org/10.1017/pan.2018.9

Dai, X. (2005). Why Comply? The Domestic Constituency Mechanism. *International Organization*, *59*(2), 363–398. http://www.jstor.org/stable/3877908

Debre, M. J. (2022). Clubs of autocrats: Regional organizations and authoritarian survival. *The Review of International Organizations*, *17*(3), 485–511. https://doi.org/10.1007/s11558-021-09428-y

Dellmuth, L. M., & Tallberg, J. (2015). The social legitimacy of international organisations: Interest representation, institutional performance, and confidence extrapolation in the United Nations. *Review of International Studies*, *41*(3), 451–475. https://doi.org/10.1017/S0260210514000230

Druckman, J. N. (2001). On the Limits of Framing Effects: Who Can Frame? *The Journal of Politics*, *63*(4), 1041–1066. https://doi.org/10.1111/0022-3816.00100

Druckman, J. N. (2004). Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects. *American Political Science Review*, *98*(4), 671–686. https://doi.org/10.1017/S0003055404041413

Druckman, J. N. (2022a). A Framework for the Study of Persuasion. *Annual Review of Political Science*, *25*, 65–88.

Druckman, J. N. (2022b, May). *Experimental Thinking: A Primer on Social Science Experiments* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108991353

Ecker-Ehrhardt, M., Dellmuth, L., & Tallberg, J. (2024). Ideology and Legitimacy in Global Governance. *International Organization*, *78*(4), 731–765. https://doi.org/10.1017/S0020818324000304

Ecker-Ehrhardt, M., Dellmuth, L., & Tallberg, J. (2025). Peer opinion and the legitimacy of international organizations. *The Review of International Organizations*. https://doi.org/10.1007/s11558-025-09600-8

Fariss, C. J. (2014). Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review*, *108*(2), 297–318. https://doi.org/10.1017/S0003055414000070

Finnemore, M. (1993). International organizations as teachers of norms: The United Nations Educational, Scientific, and Cutural Organization and science policy. *International Organization*, *47*(4), 565–597. https://doi.org/10.1017/S0020818300028101

Franck, T. M. (1990). *The Power of Legitimacy among Nations*. Oxford University Press.

Fuller, L. L. (1969). Human Interaction and the Law. *The American Journal of Jurisprudence*, *14*(1), 1–36. https://doi.org/10.1093/ajj/14.1.1

Gangl, A. (2003). Procedural Justice Theory and Evaluations of the Lawmaking Process. *Political Behavior*, *25*(2).

Ghassim, F. (2024). Effects of Self-Legitimation and Delegitimation on Public Attitudes toward International Organizations: A Worldwide Survey Experiment. *International Studies Quarterly*, *68*(2), sqae012. https://doi.org/10.1093/isq/sqae012

Ghassim, F., Koenig-Archibugi, M., & Cabrera, L. (2022). Public Opinion on Institutional Designs for the United Nations: An International Survey Experiment. *International Studies Quarterly*, *66*(3), sqac027. https://doi.org/10.1093/isq/sqac027

Ginsburg, T. (2020). Authoritarian International Law? *American Journal of International Law*, *114*(2), 221–260. https://doi.org/10.1017/ajil.2020.3

Grieco, J. M., Gelpi, C., Reifler, J., & Feaver, P. D. (2011). Let's Get a Second Opinion: International Institutions and American Public Support for War1: Let's Get a Second Opinion. *International Studies Quarterly*, *55*(2), 563–583. https://doi.org/10.1111/j.1468-2478.2011.00660.x

Hollyer, J. R. (2011). Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance. *Quarterly Journal of Political Science*, *6*(3-4), 275–327. https://doi.org/10.1561/100.00010059

Hurd, I. (2005). The Strategic Use of Liberal Internationalism: Libya and the UN Sanctions, 1992–2003. *International Organization*, *59*(03). https://doi.org/10.1017/S0020818305050186

Kane, J. V., & Barabas, J. (2019). No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments. *American Journal of Political Science*, *63*(1), 234–249. https://doi.org/10.1111/ajps.12396

Kertzer, J. D., & Brutger, R. (2016). Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12201]. *American Journal of Political Science*, *60*(1), 234–249. https://doi.org/10.1111/ajps.12201

Kim, M. D. (2019). REPUTATION AND COMPLIANCE WITH INTERNATIONAL HUMAN RIGHTS LAW: EXPERIMENTAL EVIDENCE FROM THE US AND SOUTH KOREA. *Journal of East Asian Studies*, *19*(2), 215–238. https://doi.org/10.1017/jea.2019.20

Kreps, S. E., & Wallace, G. P. (2016). International law, military effectiveness, and public support for drone strikes. *Journal of Peace Research*, *53*(6), 830–844. https://doi.org/10.1177/0022343316657405

Lake, D. A., Martin, L. L., & Risse, T. (2021). Challenges to the Liberal Order: Reflections on *International Organization*. *International Organization*, *75*(2), 225–257. https://doi.org/10.1017/S0020818320000636

Levi, M., Sacks, A., & Tyler, T. (2009). Conceptualizing Legitimacy, Measuring Legitimating Beliefs. *American Behavioral Scientist*, *53*(3), 354–375. https://doi.org/10.1177/0002764209338797

Lupu, Y., & Wallace, G. P. R. (2019). Violence, Nonviolence, and the Effects of International Human Rights Law. *American Journal of Political Science*, *63*(2), 411–426. https://doi.org/10.1111/ajps.12416

Morse, J. C., & Pratt, T. (2022). Strategies of Contestation: International Law, Domestic Audiences, and Image Management. *The Journal of Politics*, *84*(4), 2080–2093. https://doi.org/10.1086/719418

Morse, J. C., & Pratt, T. (2025). Smoke and Mirrors: Strategic Messaging and the Politics of Noncompliance. *American Political Science Review*, 1–19. https://doi.org/10.1017/S0003055425100877

Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton University Press.

Mutz, D. C. (2021, April). Improving Experimental Treatments in Political Science. In J. Druckman & D. P. Green (Eds.), *Advances in Experimental Political Science* (1st ed., pp. 219–238). Cambridge University Press. https://doi.org/10.1017/9781108777919.016

Pauselli, G. (2023). Look Who Is Talking: Direct and Indirect Effects of Criticism on LGBT Rights. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4317082

Putnam, T. L., & Shapiro, J. N. (2017). International Law and Voter Preferences: The Case of Foreign Human Rights Violations. *Human Rights Review*, *18*(3), 243–262. https://doi.org/10.1007/s12142-017-0452-8

Simmons, B. A. (2009, October). *Mobilizing for Human Rights: International Law in Domestic Politics* [Google-Books-ID: qmDxDMruTMMC]. Cambridge University Press.

Simmons, B. A., & Danner, A. (2010). Credible Commitments and the International Criminal Court. *International Organization*, *64*(2), 225–256. https://doi.org/10.1017/S0020818310000044

Strezhnev, A., Simmons, B. A., & Kim, M. D. (2019). Rulers or Rules? International Law, Elite Cues and Public Opinion. *European Journal of International Law*, *30*(4), 1281–1302. https://doi.org/10.1093/ejil/chaa002

Tallberg, J., & Vikberg, C. (2025). Democracy, Autocracy, and the Design of International Organizations. *International Studies Quarterly*, *69*(2), sqaf034. https://doi.org/10.1093/isq/sqaf034

Tallberg, J., & Zürn, M. (2019). The legitimacy and legitimation of international organizations: Introduction and framework. *The Review of International Organizations*, *14*(4), 581–606. https://doi.org/10.1007/s11558-018-9330-7

Thibaut, J. W., & Walker, L. (1975). *Procedural justice : A psychological analysis*. L. Erlbaum Associates.

Thompson, A. (2006). Coercion Through IOs: The Security Council and the Logic of Information Transmission. *International Organization*, *60*(01). https://doi.org/10.1017/S0020818306060012

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). **mediation** : *R* Package for Causal Mediation Analysis. *Journal of Statistical Software*, *59*(5). https://doi.org/10.18637/jss.v059.i05

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

Tyler, T. R. (2001). A psychological perspective on the legitimacy of authorities and institutions. In *J. T. Jost & B. Major, The psychology of legitimacy: Emerging perspective on ideology, justice, and intergroup relations* (pp. 416–436). Cambridge University Press.

Tyler, T. R. (2006). *Why People Obey the Law*. Princeton University Press.

Voeten, E. (2008). The Impartiality of International Judges: Evidence from the European Court of Human Rights. *American Political Science Review*, *102*(4), 417–433. https://doi.org/10.1017/S0003055408080398

Wallace, G. P. (2013). International Law and Public Attitudes Toward Torture: An Experimental Study. *International Organization*, *67*(1), 105–140. http://www.jstor.org/stable/43282154

Weeks, J. L. (2008). Autocratic Audience Costs: Regime Type and Signaling Resolve. *International Organization*, *62*(01). https://doi.org/10.1017/S0020818308080028

Appendices

A Additional Plots and Figures

Table 2: Mean approval ratings (raw 1–5 scale) by country, messenger, and message type, including country-specific Control baselines

		Australia						Philippines					
Messenger	Message Type	Mean	95% CI	N	Control Mean	95% CI	N	Mean	95% CI	N	Control Mean	95% CI	N
Control	Control	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Both	Human Rights	3.69	3.49	[3.51, 3.88]	[3.26, 3.73]	124	114	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Democracies	Human Rights	3.61	3.35	[3.40, 3.83]	[3.10, 3.60]	119	112	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
No Messenger	Human Rights	3.94	3.58	[3.73, 4.15]	[3.35, 3.82]	119	113	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Non-democracies	Human Rights	3.71	3.35	[3.48, 3.93]	[3.12, 3.59]	119	108	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Both	Sovereignty	3.49	2.81	[3.26, 3.72]	[2.56, 3.06]	118	117	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Democracies	Sovereignty	3.45	2.84	[3.22, 3.67]	[2.60, 3.07]	119	116	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
No Messenger	Sovereignty	3.39	2.68	[3.17, 3.62]	[2.43, 2.93]	119	115	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229
Non-democracies	Sovereignty	3.32	2.84	[3.08, 3.56]	[2.59, 3.09]	119	110	3.74	3.21	[3.57, 3.91]	[3.03, 3.38]	200	229

B Percentage change table

Table 3: Percent change in approval and percentage-point change in legality, relative to country-specific Control baselines.

Country	Messenger	Message	N	Approval % change	Approval CI low	Approval CI high	Legality (pp change)	Legality CI low	Legality CI high
Australia	Both	Human Rights	124	-1.24	-6.16	3.68	-4.03	-12.05	3.99
Australia	Democracies	Human Rights	119	-3.38	-9.12	2.35	-7.77	-16.24	0.70
Australia	No Messenger	Human Rights	119	5.38	-0.16	10.91	0.63	-7.12	8.38
Australia	Non-democracies	Human Rights	119	-0.91	-6.92	5.10	-1.05	-8.97	6.87
Australia	Both	Sovereignty	118	-6.64	-12.75	-0.54	-15.68	-24.58	-6.78
Australia	Democracies	Sovereignty	119	-7.88	-13.90	-1.86	-19.54	-28.51	-10.57
Australia	No Messenger	Sovereignty	119	-9.23	-15.13	-3.32	-11.13	-19.80	-2.47
Australia	Non-democracies	Sovereignty	119	-11.25	-17.56	-4.93	-20.38	-29.36	-11.40
Philippines	Both	Human Rights	114	8.92	1.56	16.28	12.94	4.00	21.87
Philippines	Democracies	Human Rights	112	4.46	-3.25	12.17	11.37	2.28	20.45
Philippines	No Messenger	Human Rights	113	11.82	4.38	19.26	15.26	6.40	24.11
Philippines	Non-democracies	Human Rights	108	4.57	-2.72	11.87	1.58	-7.89	11.05
Philippines	Both	Sovereignty	117	-12.27	-20.16	-4.38	-14.30	-22.98	-5.62
Philippines	Democracies	Sovereignty	116	-11.51	-18.73	-4.29	-10.55	-19.46	-1.65
Philippines	No Messenger	Sovereignty	115	-16.44	-24.16	-8.72	-18.91	-27.36	-10.46
Philippines	Non-democracies	Sovereignty	110	-11.51	-19.31	-3.71	-12.98	-22.01	-3.95

C Descriptive Statistics for Australia

Inattentive respondents who were speeding through questions were dropped out from samples before given the experiment vignettes. Total of 1227 responses were collected. Overall passage rate is 46% for treatment groups, 55.8% for all respondents.

Table 4: Manipulation check pass rates by treatment group

Treatment	N	Passes	Pass Rate (%)
Control	218	218	100.0
T1	128	58	45.3
T2	129	54	41.9
T3	123	66	53.7
T4	126	67	53.2
T5	125	48	38.4
T6	127	37	29.1
T7	122	66	54.1
T8	129	70	54.3

Table 5: Descriptive statistics by treatment group (Mean approval with 95% CI)

Group	Treatment	Mean Approval	95\% CI	N
Control	Control	3.74	[3.59, 3.90]	218
T1	No Messenger - Sovereignty	3.41	[3.20, 3.63]	128
T2	No Messenger - Human Rights	3.97	[3.77, 4.17]	128
T3	Democracies - Sovereignty	3.42	[3.20, 3.64]	123
T4	Democracies - Human Rights	3.58	[3.37, 3.80]	125
T5	Non-democracies - Sovereignty	3.36	[3.13, 3.59]	125
T6	Non-democracies - Human Rights	3.74	[3.53, 3.95]	127
T7	Both - Sovereignty	3.45	[3.22, 3.68]	122
T8	Both - Human Rights	3.69	[3.51, 3.87]	129

D Descriptive Statistics for Philippines

Online survey was fielded from February 5th until March 8th in 2025. Respondents were enrolled adult panels at the TGM company. Sampling frame is adult Philippines nationals, and panel respondents are representative of the national demographic. Respondents were recruited through phone, and email. Overall passage rate is 42% for treatment groups, 53.7% for all respondents.

Table 6: Descriptive statistics by treatment group (95% CIs)

Group	Treatment	Mean Approval	95\% CI	N
Control	Control	3.21	[3.03, 3.38]	229
T1	No Messenger – Sovereignty	2.68	[2.43, 2.93]	115
T2	No Messenger – Human Rights	3.58	[3.35, 3.82]	113
T3	Democracies – Sovereignty	2.84	[2.6, 3.07]	116
T4	Democracies – Human Rights	3.35	[3.1, 3.6]	112
T5	Non-democracies – Sovereignty	2.84	[2.59, 3.09]	110
T6	Non-democracies – Human Rights	3.35	[3.12, 3.59]	108
T7	Both – Sovereignty	2.81	[2.56, 3.06]	117
T8	Both – Human Rights	3.49	[3.26, 3.73]	114

Table 7: Manipulation check pass rates by treatment group

Treatment	N	Passes	Pass Rate (%)
Control	229	229	100.0
T1	115	42	36.5
T2	113	28	24.8
T3	116	70	60.3
T4	112	54	48.2
T5	110	56	50.9
T6	108	23	21.3
T7	117	68	58.1
T8	114	39	34.2

E Mediation Analysis

To identify the pathways through which messenger identity may shape public preferences, the survey includes four follow-up questions asked immediately after treatment and outcome measurement. These items capture distinct theoretical mechanisms that may mediate the effect of messenger identity on approval and legality judgments.

Mediation analysis decomposes each messenger × message treatment effect into a direct effect and an indirect effect through each mediator (ACME), using the causal mediation framework of Tingley et al. (2014). For each outcome, we fit one model predicting the mediator from the treatment and another predicting the outcome from both mediator and treatment.

Respondents may find a legal claim more legitimate when endorsement comes from a group of countries that appear impartial. If messenger identity affects perceived neutrality, mixed coalitions should yield stronger persuasive effects. This mechanism is measured by impartiality ratings toward democratic, non-democratic, and mixed coalitions.

Respondents may be affected more strongly if the message comes from the endorsing coalition that seems similar to "countries like us." This mechanism is measured through a trust questionnaire that asks how trustworthy the endorsing coalition is. If in-group trust dominates, domestic perceptions of national identity should shape sensitivity to democratic versus non-democratic messenger.

I also include two explanations based on instrumental motivations. Respondents may worry that rejecting international law harms their country's image, regardless of the messenger. Respondents may comply out of fear of sanctions from the home country, which is independent of messenger credibility. This instrumental pathway should operate similarly across messenger types. It is measured by concern that rejecting the request would trigger economic or diplomatic retaliation.

Figure 7 reports ACME estimates for each mechanism under each messenger × message combination. Across all messenger conditions under human rights message, none of the four mechanisms explain variation in approval.

Sovereignty messages produce large and consistent reductions in support across messengers and countries. Unlike human-rights messages, sovereignty frames activate several measurable

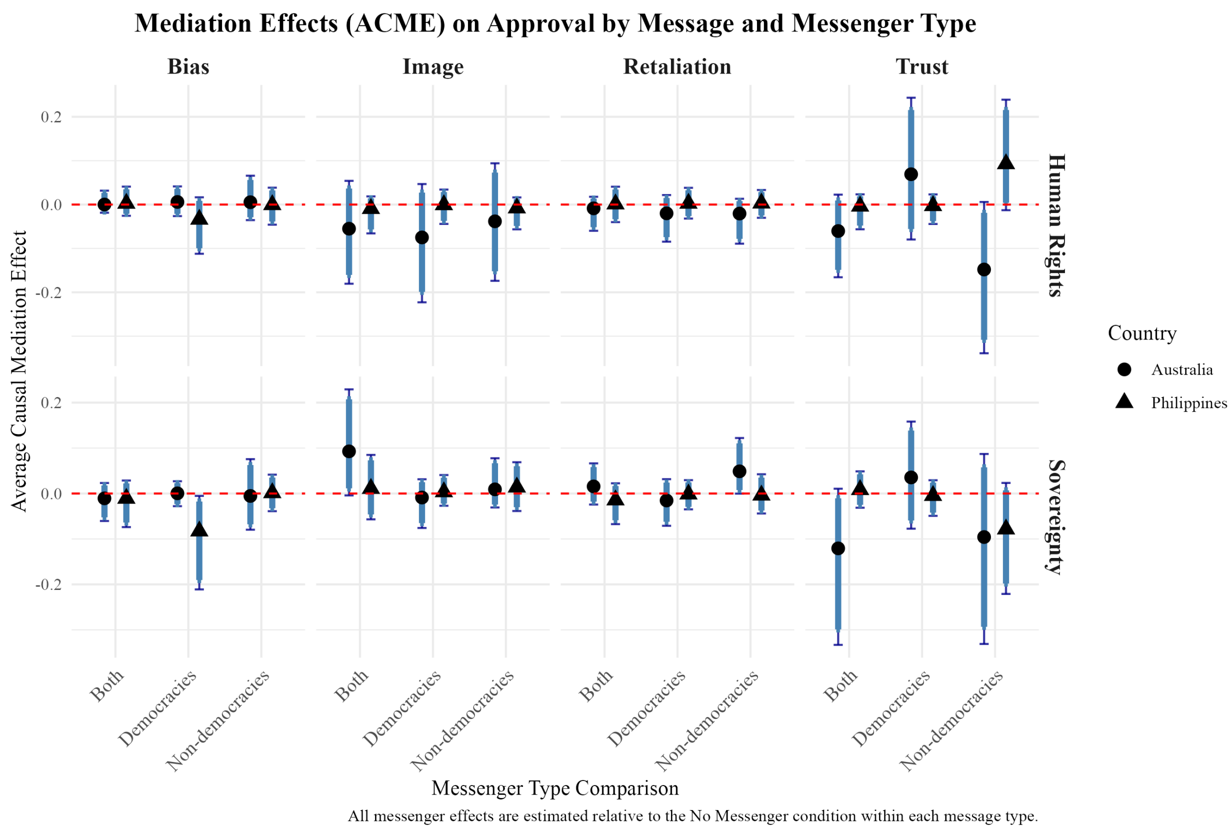


Figure 7: The thick bars show the 90% confidence intervals, and the thin bars show the 95% confidence intervals.

pathways, but these indirect effects rarely account for more than a small share of the total treatment effect.

The most consistent pathway is concern for retaliation for Australia. When non-democracies endorse the sovereignty message, respondents report higher expectations of punishment for rejecting the request, producing a clear indirect effect. Mixed coalitions also generate small ACMEs through increased concern about international image and decreased trust. However, none of these indirect effects are large enough to alter the overall negative effect of the sovereignty message on approval. Most of the treatment effect force operates through the direct pathway.

A similar pattern emerges in the Philippines. Messenger identity occasionally shifts perceived retaliation, trust, or image concerns, but these indirect effects do not meaningfully mediate approval. In nearly all combinations, ACMEs are close to zero, and the total effect is driven by the substantive content of the sovereignty message itself. Respondents reduce support for refugee acceptance largely regardless of why they believe the endorsing state is making the request.

The mediation patterns for legality closely parallel those observed for approval, small indirect effects. Across both countries, neither human-rights nor sovereignty messages produce meaningful ACMEs through trust, bias, image, or retaliation. While a handful of messenger–mediator combinations reach marginal significance at the 90% level, these movements are substantively trivial and do not account for the shifts in legality judgments.

In Australia, sovereignty messages from non-democracies slightly increase perceived retaliation risk, and mixed-regime messengers modestly reduce trust. In the Philippines, democratic messengers under sovereignty produce a minor decrease in perceived bias. However, none of these mediators translate into detectable changes in legality assessments. The estimated indirect effects remain near zero, and the total effects are virtually unchanged after conditioning on any mediator. The limited mediation effects suggest that legitimacy in this context operates less through strategic heuristics about trust, retaliation, or image and more through direct interpretation of the legal claim itself.

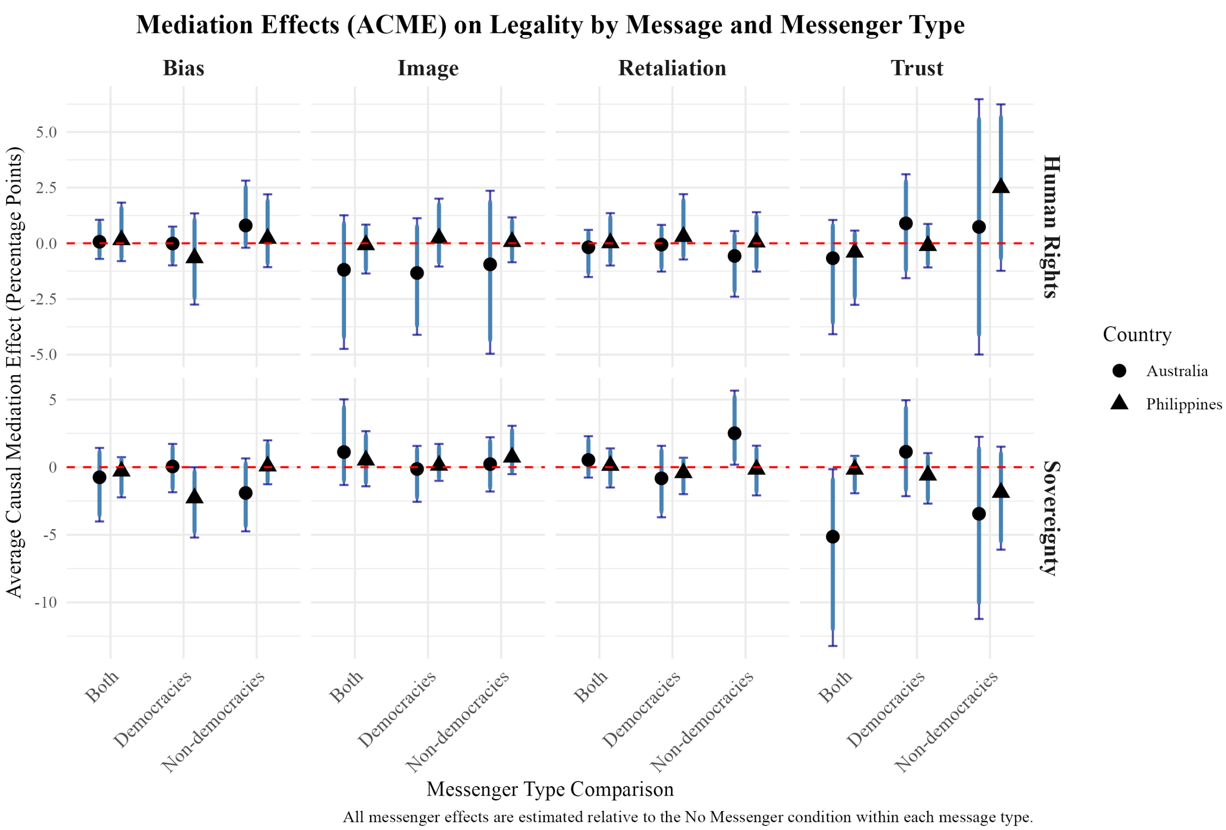
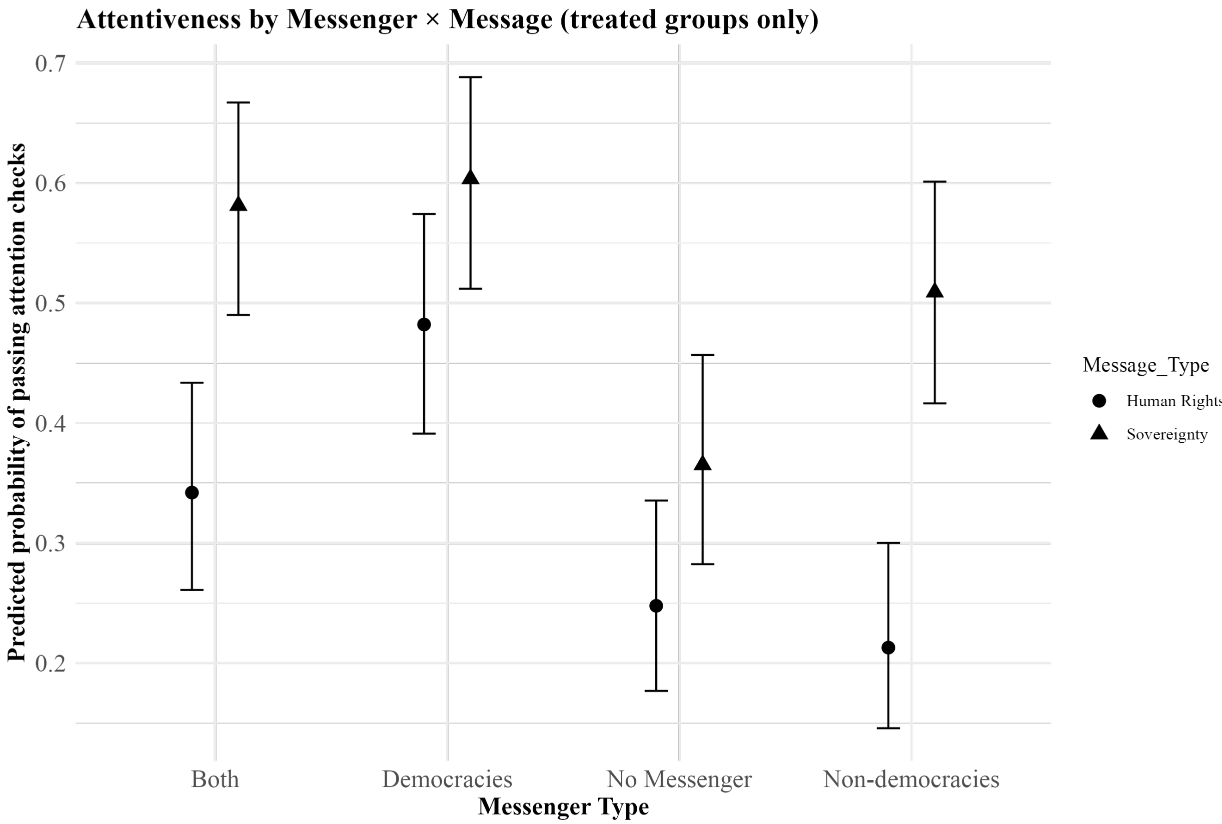


Figure 8: Mediation Analyses for Legality

F Attention and Manipulation Checks

To assess whether assignment to treatment groups systematically affected respondents' likelihood of passing the manipulation checks, I conducted a chi-square test of independence between treatment group and the binary indicator for passing. The results indicate a highly significant association ($\chi^2(8) = 316.82$, $p < 0.001$), suggesting that attention and comprehension were not evenly distributed across experimental conditions. In other words, the probability of passing the manipulation checks varied by treatment group, which underscores the importance of accounting for differential attrition and attention when interpreting the substantive results.



To examine whether respondents were more or less attentive depending on the messenger–message pairing, I estimated a logistic regression model predicting whether individuals passed the attention checks as a function of Messenger Type (democracies, non-democracies, both, or no messenger), Message Type (human rights vs. sovereignty), and their interaction. Because the control group was not exposed to a manipulation check, I excluded it from the analysis. I then generated predicted probabilities with confidence intervals across all messenger–message combinations. This approach allows us to see whether participants were systematically less likely to pass the attention checks in counterintuitive, "off-diagonal" conditions (for instance, when democracies endorsed a restrictive sovereignty message or when non-democracies endorsed a liberal human rights message) relative to more intuitive messenger–message pairings. For full transparency, I replicate the entire process using only a subsample of responses that have passed all three FMCs.

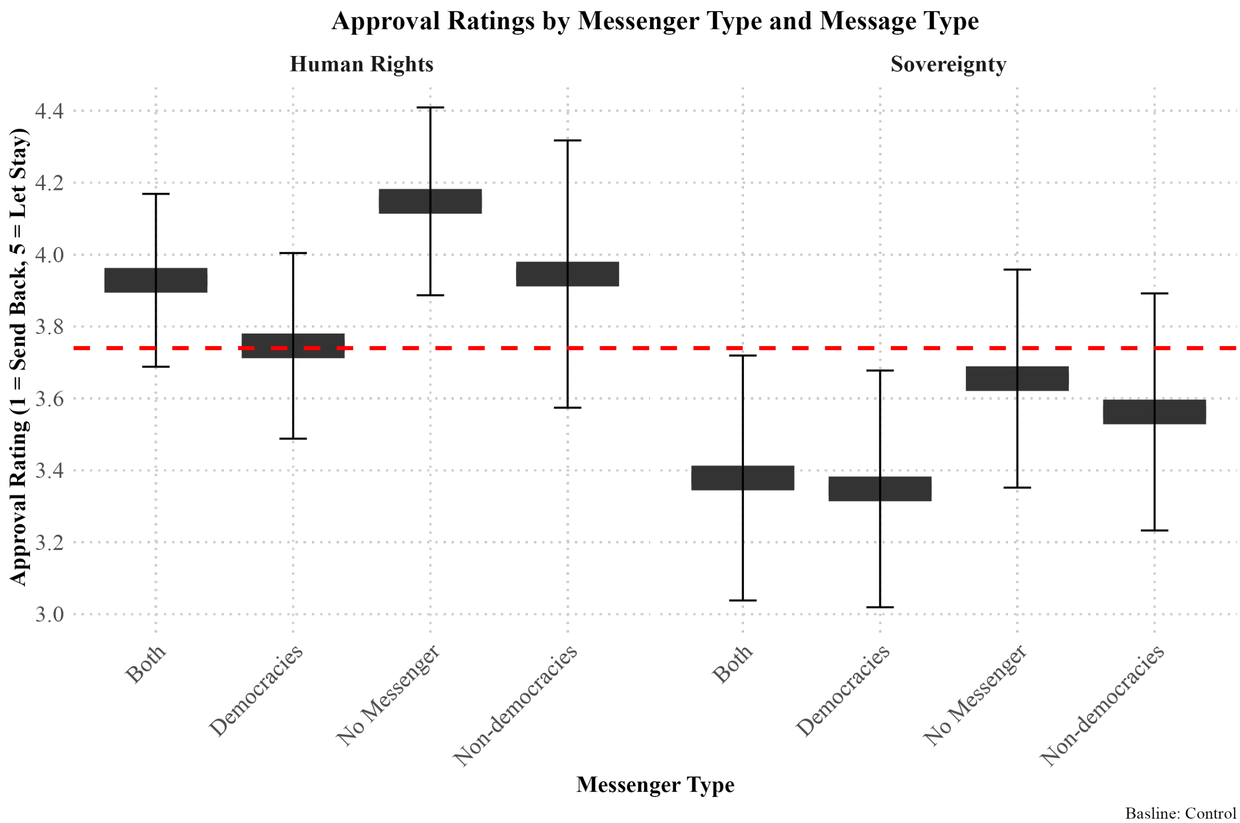
G Attentive Samples Only

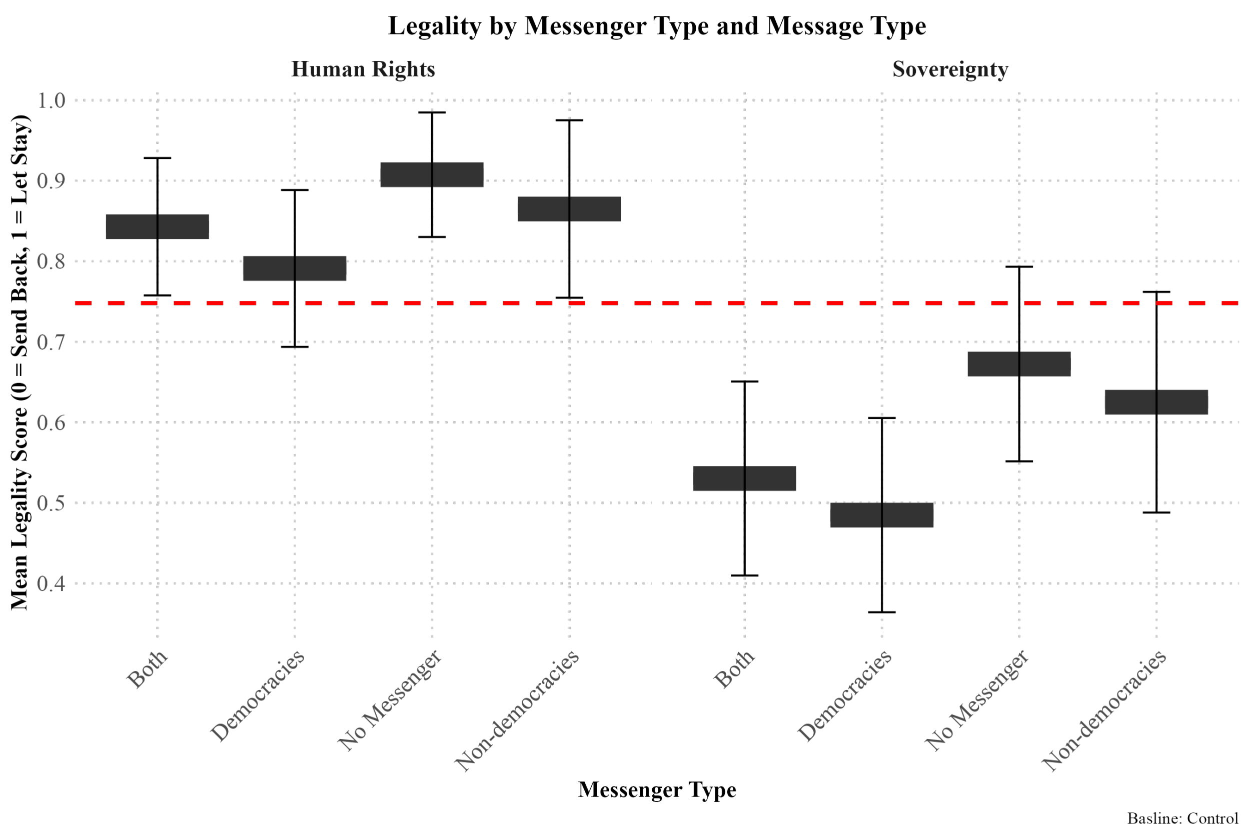
Across experimental conditions, messenger type and message framing jointly shaped both approval and perceived legality. Human rights messages increased higher approval and legality judgments than sovereignty messages, regardless of messenger. Among sovereignty frames, approval was notably lower when delivered by democracies (M = 3.35, 95% CI [3.02, 3.68]) and both democracies and non-democracies (M = 3.378, 95% CI [3.03, 3.71]).

G.1 Attentive sample for Australia

Table 8: Descriptive statistics by treatment group (Mean approval with 95% CI)

Group	Treatment	Mean Approval	95% CI	N
Control	Control	3.74	[3.59, 3.90]	218
T1	No Messenger - Sovereignty	3.66	[3.35, 3.96]	58
T2	No Messenger - Human Rights	4.15	[3.89, 4.41]	54
T3	Democracies - Sovereignty	3.35	[3.02, 3.68]	66
T4	Democracies - Human Rights	3.75	[3.49, 4.00]	67
T5	Non-democracies - Sovereignty	3.56	[3.23, 3.89]	48
T6	Non-democracies - Human Rights	3.95	[3.57, 4.32]	37
T7	Both - Sovereignty	3.38	[3.04, 3.72]	66
T8	Both - Human Rights	3.93	[3.69, 4.17]	70


Approval Ratings by Messenger Type and Message Type
Human Rights
Sovereignty
Approval Rating (1 = Send Back, 5 = Let Stay)
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Both
Democracies
No Messenger
Non-democracies
Both
Democracies
No Messenger
Non-democracies
Messenger Type
Basline: Control


Legality by Messenger Type and Message Type
Human Rights
Sovereignty
Mean Legality Score (0 = Send Back, 1 = Let Stay)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Both
Democracies
No Messenger
Non-democracies
Both
Democracies
No Messenger
Non-democracies
Messenger Type
Basline: Control

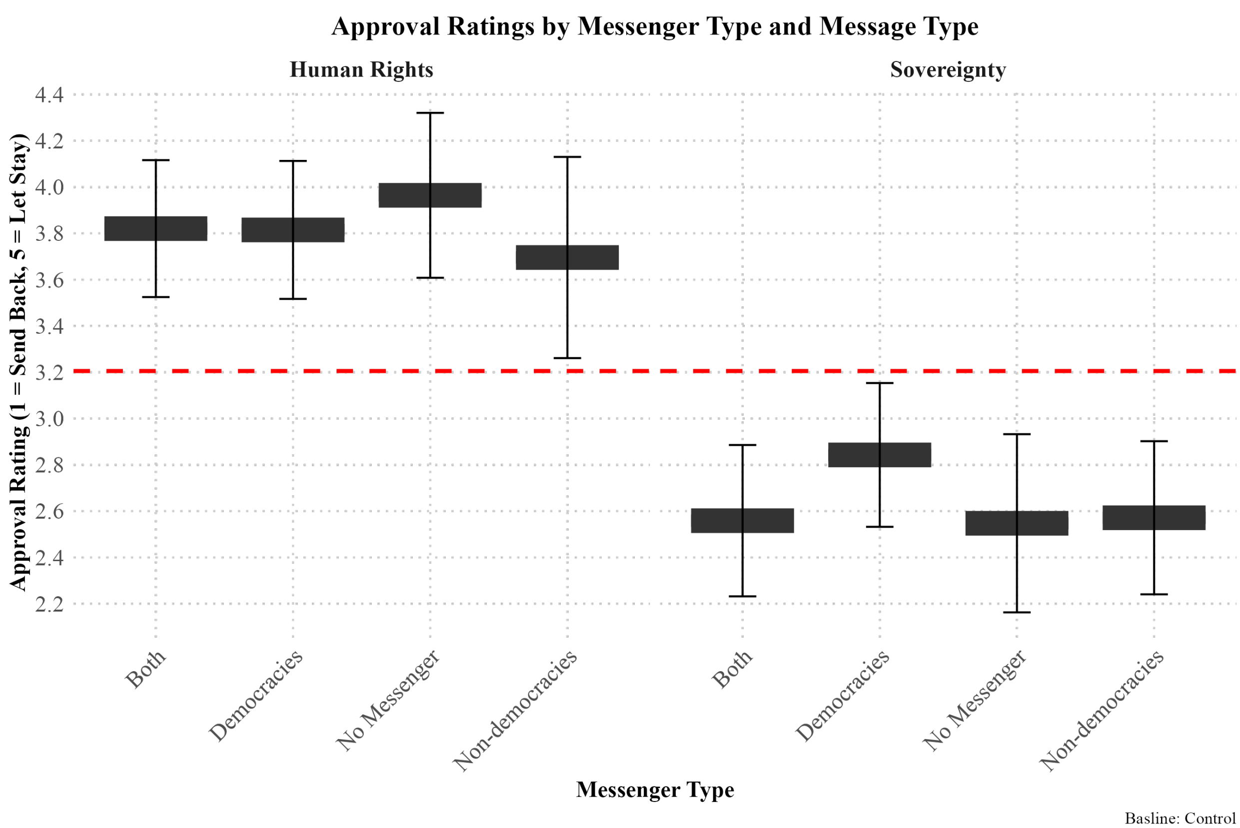
Legality followed a similar pattern where human rights messages led respondents to evaluate it is less aligned with legal appropriateness to send refugees back. Notably, no-messenger human rights messages produced the highest legality (0.91, 95% CI [0.83, 0.98]), underscoring that messenger choice can depress legitimacy perceptions, while human rights content broadly enhances them. Respondents who received the human rights message without any regime type specification reported approval that was 1.7% higher than the control group's mean legality rating.

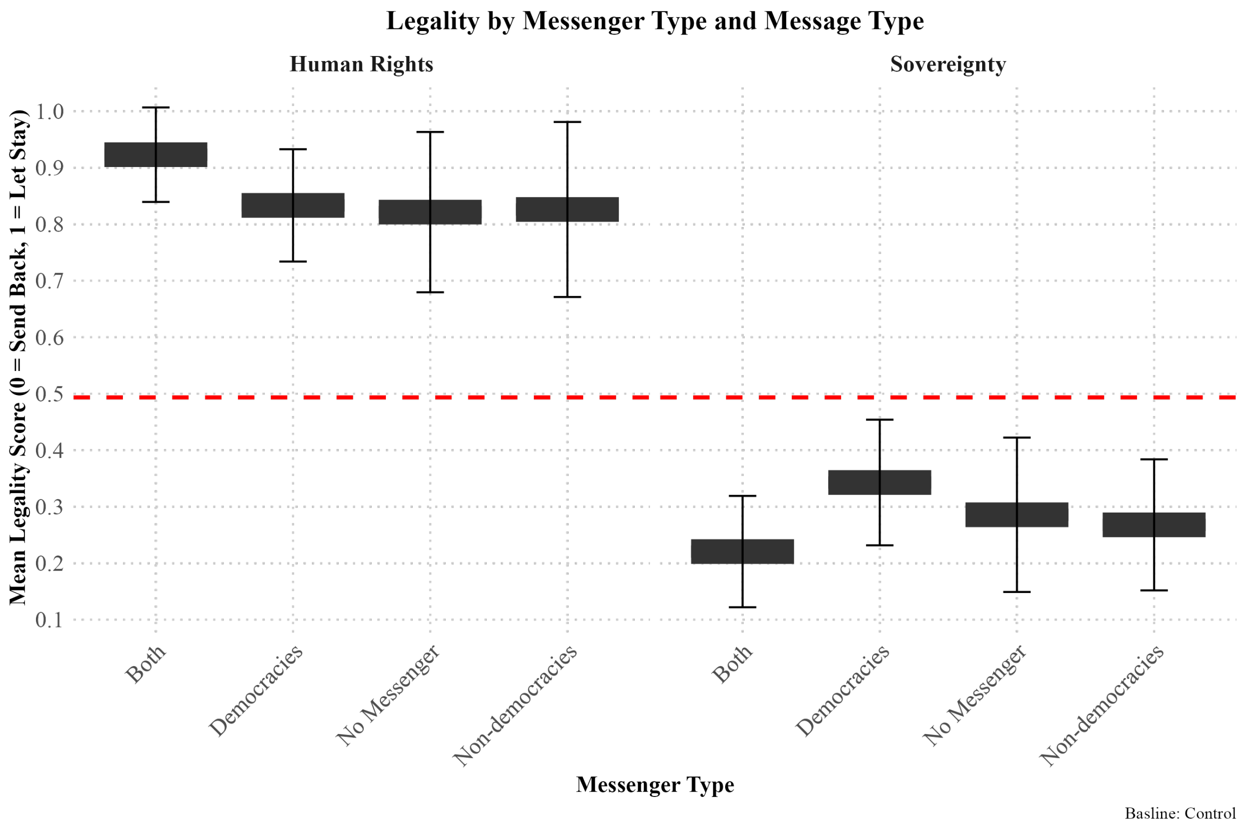
On the other hand, sovereignty messages showed sharper conditional differences depending on messenger type. When conveyed by democracies, the sovereignty message produced the lowest perceived legality (0.48, 95% CI [0.36, 0.61]) compared to the same message delivered by non-democracies (0.63, 95% CI [0.49, 0.76]) or both messenger types jointly (0.53, 95% CI [0.41, 0.65]). This pattern suggests that respondents update their judgments of legal appropriateness most strongly when sovereignty-based arguments are advanced by democracies, indicating that messenger amplifies the perceived legal appropriateness of rejecting refugees. By contrast, non-democratic messengers struggled to shift legality perceptions. When they delivered the sovereignty frame, respondents were relatively unconvinced, rating the message's legal appropriateness at 0.63 (95% CI [0.49, 0.76]), slightly stronger than the control group but falling short of having a statistical significance. However, when non-democracies endorsed the message with democracies, the impact closely mirrored the democracy-only condition (0.53, 95% CI [0.41, 0.65]).

Table 9: Descriptive statistics by treatment group (95% CIs)

Group	Treatment	Mean Approval	95% CI	N
Control	Control	3.21	[3.03, 3.38]	229
T1	No Messenger – Sovereignty	2.55	[2.16, 2.93]	42
T2	No Messenger – Human Rights	3.96	[3.61, 4.32]	28
T3	Democracies – Sovereignty	2.84	[2.53, 3.15]	70
T4	Democracies – Human Rights	3.81	[3.52, 4.11]	54
T5	Non-democracies – Sovereignty	2.57	[2.24, 2.9]	56
T6	Non-democracies – Human Rights	3.70	[3.26, 4.13]	23
T7	Both – Sovereignty	2.56	[2.23, 2.89]	68
T8	Both – Human Rights	3.82	[3.52, 4.12]	39

G.2 Attentive sample for the Philippines




Legality by Messenger Type and Message Type
Human Rights
Sovereignty
Mean Legality Score (0 = Send Back, 1 = Let Stay)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Both
Democracies
No Messenger
Non-democracies
Both
Democracies
No Messenger
Non-democracies
Messenger Type
Basline: Control

Conclusion

Through three papers, this dissertation addresses a central question: "What happens when authoritarian states actively shape international law and institutions?" International cooperation has often been theorized as a project of liberal democracies, grounded in shared commitments to rule-of-law governance and legal constraint. Yet authoritarian regimes take leadership in international organizations, sign treaties, and invoke international legal principles in diplomatic discourse. Rather than treating authoritarian participation in international institutions as anomalous, this dissertation argues that authoritarian regimes engage strategically with international law and institutions in ways that reflect their distinct political incentives and constraints.

Across three papers, the dissertation investigates how regime type shapes the design, interpretation, and political reception of international law. By examining institutional rules within intergovernmental organizations, rhetorical invocations of sovereignty in speech, and public reactions to international legal messages delivered by different regimes, the project provides a multi-level account of how authoritarian governments interact with the international order. Findings show that authoritarian states do not simply reject international law, but rather, they adapt and deploy legal institutions and principles in ways that advance regime interests.

The three chapters approach the question from complementary perspectives, each examining a different dimension of authoritarian engagement with international law. In my first paper, I tackle the foundational assumption that authoritarian states create and shape their own organizations, setting the scope of my inquiry apart from studies of authoritarian participation in existing institutions. Existing approaches to measuring IGO membership rely on mean-based regime scores, which obscure important variation in how member states are distributed.

To address this, I develop a Bayesian beta-distribution model that captures the full distributional shape of member-state regime types, distinguishing organizations that are skewed toward autocracies from those that are merely heterogeneous. Substantively, the analysis shows that authoritarian-skewed IGOs are not ad hoc alliances but formalized entities with codified rules, and that membership in such organizations is associated with longer incumbent survival and a dampening of democratic quality among member states. These findings establish that regime composition within IGOs has meaningful consequences for how international cooperation is structured and maintained.

Chapter Two shifts attention from organizations to legal discourse. Drawing on text analysis of United Nations General Debate speeches from 1945 until 2022, the chapter examines how governments invoke the principle of sovereignty in international diplomacy. Authoritarian regimes invoke sovereignty more frequently than democracies and associate it more strongly with concepts such as territorial integrity, non-interference, and external threat. These results suggest that authoritarian states emphasize a more restrictive, Westphalian interpretation of sovereignty that prioritizes regime security and insulation from external intervention.

Beyond illustrating these empirical patterns, the chapter contributes to long-standing debates in international relations regarding the meaning and evolution of sovereignty. The analysis situates contemporary diplomatic rhetoric within the broader historical development of the sovereignty principle (Bellamy, 2010; Glanville, 2013; Krasner, 2001). Beginning with its canonical association with the Peace of Westphalia, sovereignty has long served as a foundational organizing principle of the international system. Over time, however, its meaning has evolved as states and international organizations have sought to reconcile sovereignty with emerging norms of collective responsibility and humanitarian protection. The chapter highlights how states have historically navigated this tension, leaving an interpretive gap that authoritarian governments can readily exploit in advancing more restrictive understandings of sovereignty. Sovereignty simultaneously enables states to restrict the behavior of others in pursuit of collective goals while also limiting the extent to which external actors may intervene in domestic affairs. This dual and sometimes contradictory character makes sovereignty particularly susceptible to reinterpretation by states with different political incentives.

A second contribution of the chapter is to shift the focus of the emerging literature on authoritarian international law. Much of this scholarship has debated whether principles such as non-interference should be understood as inherently illiberal or, alternatively, as historically liberal norms that authoritarian states have appropriated for their own purposes (Mallat, 2020). Rather than adjudicating the origins of the principle, this chapter redirects attention toward how sovereignty is invoked and used in practice. By examining patterns of diplomatic speech, the analysis demonstrates how authoritarian governments strategically emphasize particular interpretations of sovereignty that align with regime interests.

Methodologically, the chapter also introduces a computational approach to studying legal discourse that allows researchers to track the evolution of legal concepts rather than simply the frequency of individual words. Using word embedding techniques, the analysis measures the semantic relationships between sovereignty and related concepts within a large corpus of diplomatic speech. This approach enables the study to capture how the meaning of sovereignty shifts across contexts and political actors. In doing so, the chapter traces not merely the use of a word, but the deployment of a broader legal *concept* within international discourse.

Finally, Chapter Three examines the downstream political effects of international legal rhetoric. Through a survey experiment conducted in two national contexts of Australia and the Philippines, the chapter investigates whether the credibility of international legal arguments depends on the regime type of the state delivering them. The results indicate that the persuasive power of international law varies depending on the identity of the messenger, suggesting that authoritarian and democratic endorsements of legal norms can shape public perceptions in different ways.

Beyond demonstrating the compliance pull of authoritarian international law, the chapter also contributes to broader theoretical discussion in international law. Classical theories of international law often treat legal authority as deriving from the content of legal norms or from institutionalized procedures through which rules are created (Tyler, 2006). I empirically test the ”given” authority of international law by asking whether the perceived legitimacy of international legal arguments may depend on the political identity of the actors invoking them. In this sense, the study offers an empirical approach to examining whether the source of international law lies at the message or the messenger. The credibility of legal claims does not arise solely from the substance of the rule being invoked, but also from the characteristics of the actors who mobilize those rules in political discourse, especially when the message like human rights has been politicized for a long time.

The chapter also speaks to scholarship in comparative politics on the relationship between regime type and public opinion in foreign policy. While existing research often focuses on how domestic institutions constrain leaders' foreign policy choices, this study highlights how the international identity of political actors shapes the reception of legal arguments among domestic audiences. Citizens interpret international legal claims through political heuristics that incorporate information about regime type, credibility, and normative alignment.

The broader dissertation demonstrates that authoritarian regimes actively shape international institutions and legal discourse. Chapter Three shows that these efforts are not merely symbolic. They can influence how international legal arguments are received by the public. In this sense, the political use of international law by authoritarian governments matters not only for institutional design or diplomatic rhetoric but also for the formation of public preferences about foreign policy and global governance.

The dissertation therefore concludes with a broader implication. Authoritarian engagement with international law is consequential precisely because it can shape how international legal norms are interpreted, contested, and ultimately supported by political audiences.

Contribution

This project makes both theoretical and methodological contributions to the literature on international institutions as well as authoritarian politics. First, taking authoritarian states' role in global governance seriously allows researchers to have a full picture of multilateral international order. While earlier work has laid out a stalwart foundation on institutions, autocrats' involvement in global governance has been treated as outlier or anomaly. This dissertation's overarching approach relaxes the assumption that benefits of international institutions are exclusively reserved for achieving goals of liberal democracies.

The first two chapters advance methodological approaches to address questions of international law, refining existing tools in ways suited to the study of authoritarian engagement with international institutions. Chapter 1 provides new empirical evidence that IGOs vary systematically in their authoritarian skew, and that this variation has meaningful consequences for incumbent survival and democratic quality among member states. Chapter 2 demonstrates empirically that authoritarian regimes invoke sovereignty more frequently and in semantically distinct ways compared to democracies. Chapter 3 also extends the empirical scope of this literature by incorporating an experimental component, exposing respondents to not just illiberal endorsers but also an 'illiberal' message that mandates states to deport political refugees. The project further brings attention to Southeast Asia, where the coexistence of authoritarian and democratic states highlights the tensions of international cooperation most vividly.

This dissertation carries broader implications for understanding the evolution of the international legal order. Rather than rejecting international institutions altogether, authoritarian regimes may seek to shape them from within. This can occur through the creation of new organizations, the reinterpretation of legal principles, or the strategic use of international law in diplomatic discourse. In this sense, the growth of authoritarian participation in international institutions may not necessarily lead to the erosion of international law itself but rather to the emergence of alternative interpretations of its principles.

This perspective suggests that the future of international cooperation may involve ongoing contestation over the meaning and function of legal norms. More broadly, this means that international law is not simply a static set of rules imposed on states but a contested arena in which different regime types advance competing interpretations of legal principles. Authoritarian governments may seek to reinterpret legal norms around sovereignty in ways that emphasize non-interference, territorial control, and regime autonomy. In doing so, they (wittingly or unwittingly) contribute to an evolving landscape of international law in which the meaning and function of legal principles are shaped by the political characteristics of the states that invoke them.

The expansion of international institutions has frequently been interpreted as part of a broader liberal international order. The evidence presented in the dissertation complicates this narrative. Authoritarian states do not simply stand outside international institutions or participate reluctantly in them. Instead, they actively shape institutions, and reinterpret legal principles in diplomatic discourse. These patterns suggest that international institutions should not be understood solely as instruments of liberal cooperation, but as arenas in which states with different regime types pursue distinct political objectives.

The political foundations of international cooperation are more heterogeneous than commonly assumed. International institutions may persist not because states share identical normative commitments, but because different types of governments are able to pursue their interests through the same institutional frameworks.

In addition to its theoretical contributions, this dissertation advances the methodological toolkit available to scholars of international law and global governance. While existing approaches have generated important insights, they often face limitations when attempting to systematically examine how legal principles evolve, or how institutions reflect underlying political distributions. This dissertation contributes to the emerging field of empirical international law by developing and refining methodological approaches that allow scholars to study these dynamics more systematically.

First, the project develops a new measurement strategy for capturing the political composition of intergovernmental organizations. Existing studies often rely on simple averages of member states' regime characteristics, such as the mean democracy score of an organization's membership. While informative, these measures obscure important distributional features within institutions. An organization composed of a few highly democratic states and many highly authoritarian states may have the same average score as an organization composed of politically moderate regimes, yet the political dynamics within these organizations may differ substantially.

To address this limitation, the first chapter introduces a measurement framework that captures the full distributional shape of regime characteristics within international organizations. By modeling the distribution of member states' regime types rather than relying solely on central tendency, this approach allows researchers to examine how institutional outcomes may depend on factors such as regime polarization, clustering, or dominance by particular regime types. This measurement strategy unlocks empirical testing of how political heterogeneity within institutions affects institutional design, and decision-making procedures.

Second, the dissertation demonstrates how computational text analysis can be applied to the study of international legal discourse. While scholars have long recognized the importance of diplomatic rhetoric in shaping international norms, empirical research has often relied on manual coding or simple word-frequency approaches. The analysis presented in Chapter Two moves beyond these methods and joins the burgeoning computational approach by employing word embedding techniques to capture the semantic relationships between legal concepts within large corpora of diplomatic speeches (Bae & Salvati, 2026; Kraft & Klemmensen, 2024; Rodriguez et al., 2023).

This approach allows researchers to examine how the meaning of legal principles evolves across political contexts rather than simply measuring how frequently particular words appear in diplomatic texts. In doing so, the study provides one of the first empirical tests of long-standing theoretical debates regarding the interpretation of sovereignty in international relations. By tracing the semantic proximity between sovereignty and related *concepts* (not words), such as territorial integrity, non-interference, the analysis demonstrates how different regime types emphasize distinct interpretations of the same legal principle.

Third, the dissertation builds on a growing body of work employing experimental methods into the study of international legal legitimacy. Much of the existing literature on international law treats the legitimacy of international law as an assumed property of legal rules or procedures. Chapter Three adopts a different approach by experimentally examining how individuals respond to international legal arguments delivered by different types of political actors.

By varying the regime type of the state invoking a legal principle, the experiment allows researchers to isolate the causal effect of messenger identity on public perceptions of legal legitimacy. This design bridges the literature on international law and comparative politics by demonstrating how citizens evaluate international legal claims through heuristic cues related to regime type and political credibility. In doing so, the study illustrates how experimental methods can complement traditional institutional analyses in the study of international law.

Policy Implications

The findings of this dissertation also carry several implications for policymakers and practitioners engaged in global governance. As authoritarian states increasingly participate in international institutions, the interpretation and application of international legal norms may become more contested. Policymakers should therefore recognize that legal principles such as sovereignty are not interpreted uniformly across political systems. Authoritarian governments may strategically emphasize interpretations that prioritize regime autonomy and non-interference, particularly in response to external criticism regarding domestic governance practices.

Non-state actors' efforts to promote international legal norms, including human rights protections and humanitarian intervention frameworks, often rely on appeals to widely shared legal principles (Hall et al., 2025). However, I suggest that these principles may face resistance from competing norm of state-centric sovereignty. Policymakers seeking to build multilateral consensus may therefore need to anticipate how authoritarian governments frame legal arguments and adjust diplomatic strategies accordingly.

Chapter Three further suggests that the political identity of states invoking international law can influence how legal claims are received by domestic audiences. This has implications for the communication strategies of governments, international organizations, and advocacy groups seeking to mobilize support for international legal norms. If public responses to international law are shaped by perceptions of the messenger as well as the message, the credibility of actors promoting legal principles becomes an important factor in the effectiveness of international legal advocacy.

A related implication concerns the audiences to whom international legal arguments are directed. The survey experiment presented in Chapter Three demonstrates that citizens in the Philippines, which is not a consolidated liberal democracy, can also respond to international legal arguments when forming political opinions. This finding suggests that international law may shape public discourse beyond the traditional settings in which it is most frequently studied. For policymakers, international organizations, and advocacy groups, this implies that appeals to international legal principles may resonate with broader global audiences than is often assumed.

Finally, as authoritarian regimes continue to create institutions, reinterpret legal principles, and engage in diplomatic discourse, their influence may shape the future development of international law. Policymakers and international organizations therefore face the challenge of navigating an increasingly pluralistic international legal order in which multiple political models seek to shape the meaning and application of global norms. The literature on forum shopping may provide a useful direction for further investigating specific policy implications. Overlapping institutions with similar agendas can shape states' preferences for choosing one institution over others (Busch, 2007; Reinsberg & Westerwinter, 2023).

Limitations

Despite these contributions, the studies presented in this dissertation also have several limitations. First, the empirical analyses focus primarily on observable institutional rules, diplomatic rhetoric, and survey responses. While these measures provide valuable insight into how states engage with international law, they may not fully capture the internal decision-making processes that shape government behavior at the time of negotiation. Future research drawing on archival materials, elite interviews, or diplomatic correspondence could provide a more detailed understanding of how governments formulate their legal strategies.

Second, the analysis of diplomatic discourse relies on textual data from United Nations General

Debate speeches. While these speeches provide a unique and comprehensive record of state rhetoric, they represent a particular form of diplomatic communication and may not fully reflect how governments interpret legal norms during a more inchoate stage. Future studies could explore how these interpretations emerge during treaty negotiations or institutional reform processes and how they affect the development of international legal norms.

Third, the survey experiment examines public reactions to international legal arguments in specific national contexts. While these cases provide useful insight into how messenger effects operate, the results may not generalize fully to other political environments or cultural contexts. Expanding the analysis to additional countries and political systems would provide a more comprehensive understanding of how audiences respond to international legal claims.

Avenues for Future Research

Building on addressing some of the limitations noted in the prior section, this dissertation further opens several avenues for future research. One promising direction concerns the role of authoritarian states in shaping institutional agendas within international organizations. While this dissertation focuses on the impacts on authoritarian survival, future work could examine how authoritarian governments influence agenda-setting, amendment flexibility, leadership selection, and policy priorities within international institutions. The empirical component of the dissertation is particularly helpful for this new avenue in that the new measurement offers not just central tendency but also distributional features. This allows researchers to directly test hypotheses based on how power disparity and bipolarity affect institutional design features (Koremenos & Betz, 2012; Koremenos et al., 2001).

Another area for further research involves the dynamics of legal interpretation of other prominent principles. The analysis presented in the dissertation demonstrates that authoritarian regimes emphasize particular interpretations of sovereignty in diplomatic rhetoric. But regime-contingent legal rhetoric is not limited to sovereignty. Indeed, there are other principles that are prominently used by autocrats to justify contentious policies. For example, the anti-terrorism norm is similarly ambiguous in scope, and autocrats may exploit that ambiguity to repress domestic dissidents.

Additional research could examine how international legal arguments are received by different audiences across the world. Public attitudes toward international law may vary depending on political institutions, media environments, and historical experiences with international organizations. Investigating these variations would provide a deeper understanding of how the legitimacy of international law is shaped in a global system characterized by diverse political regimes.

Future research could also further investigate the causal mechanisms linking authoritarian international cooperation to regime durability. While this dissertation highlights how authoritarian governments use international institutions and legal discourse strategically, authoritarian regimes typically face their most immediate political threats from domestic actors rather than external adversaries. Understanding how participation in authoritarian-skewed IGOs helps governments manage these internal challenges, whether through legitimacy building, elite coordination, or the diffusion of governance practice, would deepen our understanding of the relationship between international cooperation and authoritarian survival.

Closing Remarks

This dissertation carries several implications for the study of international relations and international law. First, findings suggest that theories linking legalization primarily to democratic governance may overlook important sources of variation in how international institutions operate. If authoritarian states shape institutional rules and legal discourse in systematic ways, then the global expansion of international organizations cannot be understood solely through the lens of liberal cooperation.

Second, the results highlight the importance of regime diversity within international institutions. Intergovernmental organizations increasingly include members with divergent political systems and normative preferences. As this heterogeneity grows, institutional rules and legal principles may reflect compromise or contestation among different regime types. Understanding how authoritarian and democratic states interact within these institutions therefore becomes central to explaining the evolution of international cooperation.

So what happens when authoritarian states take leadership in global governance? This dissertation offers answers from three angles sequentially. First, it directly tests the implicit assumption that authoritarian states create and shape their own organizations. The analysis shows that authoritarian governments indeed establish intergovernmental organizations whose membership is systematically skewed toward authoritarian regimes. Second, authoritarian governments strategically reinterpret foundational legal principles such as sovereignty, emphasizing meanings centered on territorial control and non-interference. Third, the political consequences of these legal strategies extend beyond diplomatic discourse. When authoritarian governments invoke international law to justify contested policies, such appeals can shape how citizens in other countries evaluate the legitimacy of those policies.

Lastly, why does the argument advanced here matter? Much of the literature on international institutions has implicitly treated legalization and institutionalization as products of liberal democratic governance. International law and organizations are often understood as mechanisms through which democracies promote rule-based cooperation, constrain state behavior, and diffuse liberal norms. The evidence presented in this dissertation complicates that view by showing that authoritarian states also actively engage with international law and institutions as tools of governance.

As authoritarian states assume a more proactive role in global governance, international institutions cannot be understood solely as vehicles of liberal cooperation. Instead, they increasingly function as arenas in which states with different regime types seek to shape the meaning and application of legal principles. The politics of international law therefore lies not only in whether states comply with legal rules, but also in how those rules are contested. Understanding this process is essential for explaining how international law evolves in a world where democratic and authoritarian governments coexist within the same institutional frameworks.

Finally, the dissertation speaks to ongoing discussion about the future of world order (Leeds et al., 2025). The growing influence of authoritarian states in global governance has raised concerns that international institutions may experience drift or gradual hollowing as illiberal actors gain influence. The evidence presented in thi s dissertation suggests a more complex dynamic, in that authoritarian governments are not merely eroding existing regimes but reshaping international institutions and legal discourse from within. As a result, many of these changes may be difficult to observe directly, unfolding gradually through shifts in interpretation and rhetoric before producing more visible institutional consequences.

References

Bae, J., & Salvati, F. (2026). United Nations Discourse on Liberalism across Regime Types. *Global Studies Quarterly*, *6*(1), ksag001. https://doi.org/10.1093/isagsq/ksag001

Bellamy, A. J. (2010). The Responsibility to Protect—Five Years On. *Ethics & International Affairs*, *24*(2), 143–169. https://doi.org/10.1111/j.1747-7093.2010.00254.x

Busch, M. L. (2007). Overlapping Institutions, Forum Shopping, and Dispute Settlement in International Trade. *International Organization*, *61*(04). https://doi.org/10.1017/S0020818307070257

Glanville, L. (2013). *Sovereignty and the responsibility to protect: A new history*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226077086.001.0001

Hall, N., Reiners, N., Basu, S., Chaudhry, S., Henry, L., Medie, P. A., Sundstrom, L. M., Vilán, A., & Zvobgo, K. (2025). Challenges and Opportunities for Transnational Advocacy. *International Studies Review*, *27*(3), viaf015. https://doi.org/10.1093/isr/viaf015

Koremenos, B., & Betz, T. (2012, December). The Design of Dispute Settlement Procedures in International Agreements. In J. L. Dunoff & M. A. Pollack (Eds.), *Interdisciplinary Perspectives on International Law and International Relations* (1st ed., pp. 371–393). Cambridge University Press. https://doi.org/10.1017/CBO9781139107310.019

Koremenos, B., Lipson, C., & Snidal, D. (2001). The Rational Design of International Institutions. *International Organization*, *55*(4,), 761–799. http://www.jstor.org/stable/3078615

Kraft, P. W., & Klemmensen, R. (2024). Lexical Ambiguity in Political Rhetoric: Why Morality Doesn't Fit in a Bag of Words. *British Journal of Political Science*, *54*(1), 201–219. https://doi.org/10.1017/S000712342300008X

Krasner, S. D. (2001, December). 1. Problematic Sovereignty. In S. D. Krasner (Ed.), *Problematic Sovereignty* (pp. 1–23). Columbia University Press. https://doi.org/10.7312/kras12178-003

Leeds, B. A., Mosley, L., Rosendorff, B. P., & Zarakol, A. (2025). The Future of Global Governance and World Order. *International Organization*, *79*(S1), S1–S11. https://doi.org/10.1017/S0020818325101185

Mallat, C. (2020). The Limits of Authoritarian International Law. *AJIL Unbound*, *114*, 247–251. https://doi.org/10.1017/aju.2020.48

Reinsberg, B., & Westerwinter, O. (2023). Institutional Overlap in Global Governance and the Design of Intergovernmental Organizations. *The Review of International Organizations*, *18*(4), 693–724. https://doi.org/10.1007/s11558-023-09488-2

Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding Regression: Models for Context-Specific Description and Inference. *American Political Science Review*, *117*(4), 1255–1274. https://doi.org/10.1017/S0003055422001228

Tyler, T. R. (2006). *Why People Obey the Law*. Princeton University Press.