# From Commitment to Capacity: Toward a Governance Gap Theory of International Implementation Failure

Valentina Marisol [1], Paredes Molina [2]

[1] Global Governance Institute

[2] International Implementation Research Network

DOI: 10.62692/r0tvgtnj Published: October 15, 2023

**ABSTRACT**

International commitment failure is frequently diagnosed as a deficit of political will. This paper proposes a Governance Gap Theory, demonstrating that implementation failure in international agreements stems systematically from institutional capacity constraints rather than mere political recalcitrance. By examining multi-domain compliance data across environmental, trade, and security governance, we articulate a dual-level capacity model that reconciles commitment intent with operational delivery.

## TABLE OF CONTENTS

OPUS PUBLICA

Published 15 October 2023

Global Governance / Public Administration

From Commitment to Capacity: Toward a Governance Gap Theory of International Implementation Failure

Authored by: Valentina Marisol and Paredes Molina

Advocacy Unified Network (AUN)

The Hague, Netherlands

Correspondence: info@aunetwork.org

ABSTRACT

The persistent gap between international commitments and realized outcomes is commonly described in terms of political incentives, administrative capacity, monitoring, or enforcement. This article develops a configuration-based theory that integrates these mechanisms without treating any one of them as a sufficient explanation. Building on the five-dimensional Governance Gap Assessment Matrix (GGAM) introduced by the Advocacy Unified Network (AUN) in August 2023, the article distinguishes the behavioral outcome to be explained from the institutional conditions that may produce it. The Governance Gap (G) is defined as the observable distance, over a specified time horizon, between the normative requirements of an international commitment and the realized behavior of relevant duty-bearers. Five analytically distinct conditions are theorized as explanatory dimensions: Commitment specification (C), Institutional capacity (I), Monitoring infrastructure (M), Accountability architecture (A), and Resource adequacy (R). We specify axioms, scope conditions, a measurement architecture, eight theory-building propositions, and a typology of implementation pathologies (ritualistic, selective, deferred, and symbolic compliance). The empirical material comprises five theory-development case applications—the Paris Agreement and nationally determined contributions, Sustainable Development Goal 16 in Sub-Saharan Africa, the Sendai Framework for Disaster Risk Reduction, the European Union Whistleblower Directive 2019/1937, and the United Nations Universal Periodic Review—and one externally selected Chemical Weapons Convention plausibility probe. Because the first five cases informed theory development, they are not treated as independent confirmatory tests. The CWC case is likewise not presented as a definitive hypothesis test; it is an external plausibility probe. Across the cases, the proposed configurations are qualitatively consistent with observed implementation patterns, but the evidence does not establish causal effects, measurement reliability, or predictive performance. The article therefore makes a theory-building contribution and specifies a preregistration-ready validation program for independent testing through replicated coding, negative cases, longitudinal data, and competing-model comparison.

Keywords: governance gap; implementation; international governance; treaty compliance; regime complexity; institutional capacity; monitoring; accountability; sustainable development goals; chemical weapons convention;

theory building

*Contents are arranged by section and subsection. Page references correspond to the final typeset version.*

# 1. Introduction

International cooperation in the twenty-first century has produced no shortage of commitments. Between 1945 and 2022 states concluded more than 60,000 bilateral and multilateral treaties registered with the United Nations Secretariat; the Sustainable Development Goals (SDGs) alone comprise 169 targets monitored through 231 unique indicators; the Paris Agreement had 195 Parties by mid-2023, with Parties having communicated nationally determined contributions (NDCs) under the Agreement. Yet across virtually every domain of global governance the central empirical finding is the same: commitments are made at a pace that far exceeds the capacity or willingness of duty-bearers to implement them. The United Nations Environment Programme (UNEP, 2022) concluded that current NDCs, even if fully implemented, would limit warming to approximately 2.6 degrees Celsius above pre-industrial levels, well above the 1.5-degree aspirational target of the Paris Agreement. The 2022 Sustainable Development Goals Report warned that the 2030 Agenda was in grave danger amid cascading and interlinked crises and documented reversals of earlier progress across poverty, hunger, health, education, the environment, and peace and security (United Nations, 2022). The Sendai Framework Monitor reported that, four years into its fifteen-year implementation horizon, fewer than half of countries had established disaster-loss databases used to support reporting for the Sendai Framework's global Targets A–D (UNDRR, 2021).

These patterns are not anomalies. They are the modal outcome of contemporary international cooperation. The puzzle they present is not new. Pressman and Wildavsky (1973) framed the problem with characteristic directness in their foundational study of the Oakland Economic Development Administration program: implementation, they argued, is the part of the policy process that comes after the headline-grabbing moment of commitment, and it is where policy most often fails. Four decades of scholarship in comparative public policy, international relations, and international law have produced sophisticated tools for analyzing treaty design (Koremenos, Lipson, and Snidal, 2001; Koremenos, Lipson, and Snidal, 2001), regime complexity (Raustiala and Victor, 2004; Alter and Meunier, 2009), norm diffusion (Finnemore and Sikkink, 1998), and compliance (Chayes and Chayes, 1995; Downs, Rocke, and Barsoom, 1996; Simmons, 2010). Yet despite this conceptual wealth, scholars continue to describe implementation failure in essentially episodic and case-specific terms. Each new domain, climate, human rights, disaster risk, whistleblower protection, finds the wheel reinvented: a new diagnostic framework, a new list of barriers, a new round of capacity-building recommendations.

The Advocacy Unified Network (AUN) entered this crowded field in August 2023 with a deliberately practical contribution: the Governance Gap Assessment Matrix (GGAM), a five-dimensional framework designed for use by civil society organizations and national-level advocates seeking to diagnose why specific international commitments were failing in their jurisdictions (Advocacy Unified Network, 2023). The five dimensions, Commitment specification, Institutional capacity, Monitoring infrastructure, Accountability architecture, and Resource adequacy, were chosen for their analytic separability and their operational measurability. The framework has been piloted in three national contexts and has shown promising diagnostic utility. However, the August 2023 paper presented GGAM as a framework, that is, a heuristic device for organizing observation, rather than as a theory. It made no claim about causal mechanisms, made no falsifiable predictions, and offered no propositions that could be tested against rival explanations.

This article undertakes the next step, but deliberately stops short of treating conceptual formalization as empirical confirmation. We argue that the Governance Gap can be developed into a generalizable theory of international implementation failure within explicit scope conditions. By generalizable we mean that the proposed mechanisms are formulated at the level of the commitment-jurisdiction dyad and can generate predictions across issue areas, rather than that the theory has already been validated across all domains of international governance.

The article makes three contributions. First, it separates the outcome construct from its explanatory conditions: the Governance Gap is a behavioral distance between normative requirements and observed realization, while C, I, M, A, and R are proposed explanatory dimensions. Second, it converts the GGAM framework into a falsifiable theory-building architecture by specifying axioms, scope conditions, measurement rules, interaction hypotheses, and a pathology typology. Third, it subjects the framework to transparent qualitative assessment across five theory-development cases and one externally selected plausibility probe, while explicitly identifying what remains untested. The resulting contribution is not a completed confirmatory theory; it is a rigorously specified theory-building model with a clear route to independent validation.

The article proceeds in eleven sections. Section 2 reviews three streams of relevant scholarship: classic implementation theory, regime complexity and interplay, and institutionalist theories of compliance. We identify a common theoretical deficit: existing accounts explain why commitments are designed in particular ways or why states ratify them, but they do not provide a general, portable account of why implementation produces the specific patterns of partial, selective, and symbolic compliance that we observe. Section 3 sets out our theory-building methodology, a conceptual analysis combined with abductive inference from cases, and recapitulates the five-dimensional GGAM architecture from the August 2023 paper, situating the present contribution as a theoretical extension rather than a replacement. Section 4 formalizes the Governance Gap: definition, axioms, scope conditions, notation, and the eight propositions. Section 5 elaborates the causal mechanism, dimension by dimension, and develops the implementation pathology typology. Section 6 presents the six case applications, including the external plausibility probe on the Chemical Weapons Convention in Section 6.6. Section 7 compares the explanatory power of the Governance Gap theory against four rival accounts. Section 8 discusses theoretical and policy implications. Section 9 acknowledges limitations, including a four-element measurement validation protocol. Section 10 concludes. Section 11 provides a transparency statement covering methodological disclosure, data and code availability, funding, and acknowledgments.

Before proceeding, three terminological clarifications are warranted. First, we use implementation in the inclusive sense elaborated by Victor, Raustiala, and Skolnikoff (1998): the set of actions, legislative, administrative, judicial, and behavioral, through which an international commitment is given effect within a jurisdiction. Second, we use commitment broadly to encompass both legally binding instruments (treaties, regulations, directives) and soft-law instruments (declarations, frameworks, action plans, political declarations). The theory applies to both, although we hypothesize that the relative weights of the five gap dimensions vary systematically across instrument types, a hypothesis we develop in Section 4. Third, we use failure not as a binary judgment but as a continuous property: implementation outcomes range from full realization through partial realization to no realization. The Governance Gap is the structural feature of the commitment-jurisdiction dyad that predicts where on this continuum an outcome will fall.

The motivation for this theoretical extension is both academic and practical. Academically, the field of global governance suffers from a fragmentation that mirrors the regime complexity it studies (Biermann and Pattberg, 2008). Implementation scholarship is bifurcated between qualitative case-study traditions that produce rich but non-cumulative findings, and large-n quantitative compliance studies that produce statistically robust but mechanism-poor correlations (von Stein, 2016). A general theory that is both portable and propositional offers the field a common analytic currency. Practically, the Advocacy Unified Network and allied civil society organizations are daily confronted with the question of why a particular commitment, a particular convention, a particular directive, is not being implemented in a particular jurisdiction. Without a general theory, advocacy responses are reactive and ad hoc. With a general theory, advocacy can be diagnostic and predictive: where the C and I dimensions dominate, the appropriate intervention is technical assistance and legislative drafting; where M and A dominate, the appropriate intervention is monitoring infrastructure and judicial review; where R dominates, the appropriate intervention is resource mobilization. The article's concluding section returns to these implications.

We emphasize at the outset that we do not claim the Governance Gap is the only or final theory of international implementation failure. We claim that it offers a potentially useful configuration-based account that integrates

specification, capacity, monitoring, accountability, and resources at the commitment-jurisdiction level. The article's empirical claims are deliberately limited: the cases establish qualitative plausibility and help identify observable implications, but they do not by themselves establish causal identification, measurement reliability, or predictive superiority over rival theories.

# 2. Literature Review: The Implementation Puzzle in Global Governance

Three streams of scholarship bear directly on the theoretical project of this article. The first is the tradition of implementation theory that emerged in American public administration in the early 1970s and was subsequently elaborated in comparative and international settings. The second is the literature on regime complexity and interplay, which has developed within international relations and international law over the past two decades. The third is the body of institutionalist scholarship on compliance, which draws on rationalist, constructivist, and managerialist premises to explain why states do or do not implement their international obligations. We review each stream in turn, identifying the contribution it makes and the residual gap that the Governance Gap theory is designed to fill.

## 2.1 Classic Implementation Theory

The modern study of implementation began with Pressman and Wildavsky's (1973) analysis of the Oakland Economic Development Administration program. Their central insight, that policy implementation involves long joint-action chains in which each link introduces independent probabilities of slippage, supplied the field with both its foundational puzzle and its foundational metaphor. Implementation, they argued, is not the linear transmission of policy intent into policy outcome but a sequence of bargaining, interpretation, and adaptation in which cumulative divergence from the original intent is the modal outcome. Bardach (1977) extended this analysis with the concept of the implementation game, emphasizing the strategic interplay of bureaucratic, political, and interest-group actors whose local incentives rarely align with the global goal of program success. These early contributions established three durable findings: implementation failure is normal, not exceptional; it is produced by structural features of multi-actor processes, not by individual incompetence; and it is predictable in distribution if not in particulars.

Subsequent work expanded the comparative and sectoral scope of implementation theory. Sabatier and Mazmanian (1979) developed a hierarchical model of variables affecting implementation, distinguishing tractability of the problem, ability of statute to structure implementation, and non-statutory political variables. O'Toole (2000) consolidated findings from two decades of empirical work, concluding that implementation success is most strongly associated with clear and consistent goals, sufficient statutory and financial resources, supportive political environments over time, and effective cross-jurisdictional coordination. Winter (2003) integrated these insights into an integrated implementation model that distinguished policy formulation, organizational design, and front-line behavior as analytically distinct but causally linked phases. Matland (1995) advanced an influential ambiguity-conflict model in which the relative salience of ambiguity and conflict in policy goals predicts whether implementation will be administrative, political, experimental, or symbolic.

Two features of this literature are notable for our purposes. First, it is overwhelmingly domestic in empirical orientation. The classic cases, Oakland EDA, the Job Training Partnership Act, Clean Air Act implementation, are cases of intra-state policy transmission. Generalization to the international setting has been intermittent and uneven. Victor, Raustiala, and Skolnikoff (1998) made the most ambitious such generalization in their edited volume on The Implementation and Effectiveness of International Commitments, but even they treated the international dimension primarily as a complication of the domestic model rather than as a structurally distinct setting warranting its own theory. Second, the literature is rich in frameworks but poor in theories. Matland's ambiguity-conflict matrix, Sabatier's advocacy coalition framework, and Winter's integrated model are all, in our

sense, frameworks: heuristic devices for organizing observations about implementation, not propositional theories that generate falsifiable predictions about when and how failure will occur.

## 2.2 Regime Complexity and Interplay

The literature on regime complexity, advanced principally by Raustiala and Victor (2004), Alter and Meunier (2009), and Keohane and Victor (2011), addresses a different but related set of concerns. Whereas classic implementation theory treats the commitment as a unitary instrument transmitted through a domestic implementation chain, regime complexity scholarship begins from the observation that contemporary international governance is characterized by the proliferation of overlapping, non-hierarchical institutions addressing adjacent or related issue areas. The climate regime, for example, comprises the United Nations Framework Convention on Climate Change (UNFCCC), the Paris Agreement, the Kyoto Protocol (in residual form), the Kigali Amendment to the Montreal Protocol, the International Civil Aviation Organization's Carbon Offsetting and Reduction Scheme for International Aviation (CORSIA), the International Maritime Organization's Initial Strategy on GHG emissions, and an array of regional and bilateral arrangements. Implementation of any one of these instruments is conditioned by the existence and behavior of the others.

Regime complexity scholars have generated a set of durable findings. Overlapping institutions create opportunities for forum-shopping by states and non-state actors (Busch, 2007; Alter and Meunier, 2009). Cross-regime interplay can produce synergies (Oberthur and Gehring, 2006) or inter-regime conflict (Raustiala and Victor, 2004). Nested institutions permit the construction of variable geometries of commitment (Keohane and Victor, 2011). Fragmentation can be politically useful when it permits experimentation (Devereux, 2001) but welfare-reducing when it produces race-to-the-bottom dynamics (Biermann, Pattberg, van Asselt, and Zelli, 2009). These findings have been incorporated into the design of recent international instruments, most visibly in the Paris Agreement's deliberate choice of a hybrid architecture that combines a central convention with nationally determined contributions and a global stocktake.

This literature, however, operates at a level of analysis one step removed from implementation itself. It explains why commitments are designed in complex ways, why forums are chosen, and how instrument choice is shaped by the existence of alternatives. It does not, in the main, address the question of why commitments, however designed, are or are not implemented in domestic jurisdictions. The implicit assumption is that, holding regime structure constant, implementation proceeds according to a domestic logic that the international relations literature has relatively little to say about. The Governance Gap theory developed in this article accepts that regime structure is a conditioning variable but argues that the proximate causal drivers of implementation failure lie at the level of the commitment-jurisdiction dyad, where they can be theorized directly.

## 2.3 Institutionalism and Compliance

The third stream, the institutionalist literature on compliance, comes closest to providing a general theory of international implementation. This literature is internally heterogeneous, encompassing at least four distinct positions. Managerialism, associated with Chayes and Chayes (1995), holds that compliance is the default and that non-compliance is typically the product of ambiguity, capacity limitations, and temporal lags rather than deliberate defection. The remedy is accordingly cooperative: clarification, capacity-building, and dispute-prevention. Rationalist enforcement theory, advanced by Downs, Rocke, and Barsoom (1996), counters that managerialism understates the depth of cooperation problems and the prevalence of strategic non-compliance. Where enforcement is weak, states will comply only when the marginal reputational and reciprocal costs of violation exceed the marginal benefits; otherwise they will defect at the limit of tolerance. The remedy is coercive: monitoring, sanctions, and linkage politics.

Constructivist accounts (Finnemore and Sikkink, 1998; Checkel, 2005; Johnston, 2001) emphasize the role of norms and social influence. Compliance, on this view, is mediated by persuasion, socialization, and identification, and is most likely when norms are clear, when reference groups strongly endorse the norm, and when norm

entrepreneurs successfully embed commitments in identity narratives. The remedy is discursive and organizational: norm advocacy, capacity for moral suasion, and the construction of organizational cultures that internalize the commitment. Finally, the legalization literature (Abbott, Keohane, Moravcsik, Slaughter, and Snidal, 2000; Goldstein, Kahler, Keohane, and Slaughter, 2001) treats the precision, obligation, and delegation of commitments as variables that predict compliance, with higher levels of all three producing stronger implementation pressures.

Each of these accounts is empirically supported in particular cases. Managerialism is persuasive for technical cooperation where ambiguity and capacity dominate; rationalist enforcement is persuasive for security and trade where defection incentives are strong; constructivism is persuasive for human rights and environmental norms where identity and persuasion matter; legalization is persuasive for treaty design effects. But each is also incomplete. Managerialism cannot account for persistent, strategically motivated non-compliance in the face of capacity-building investment, the EU Whistleblower Directive case discussed in Section 6.4 illustrates this. Rationalist enforcement cannot account for cases where implementation failure occurs despite the absence of strategic gains from defection, the Sendai Framework case in Section 6.3 illustrates this. Constructivism cannot account for cases where norm salience is high but implementation is poor, the SDG 16 case in Section 6.2 illustrates this. Legalization cannot account for variation in implementation across jurisdictions with identical treaty text, a recurrent empirical pattern. The four positions are best understood as complementary partial theories; none of them constitutes the general theory of international implementation failure that the empirical record calls for.

## 2.4 Synthesis: The Theoretical Gap

The three streams reviewed above share a common limitation. They theorize antecedent stages of the implementation process, the design of commitments, the choice of forums, the ratification decision, the socialization of states into norm communities, and they treat implementation as the dependent variable to be explained, but they do so indirectly, by extrapolating from theories of these antecedent stages. What is missing is a theory that takes the commitment-jurisdiction dyad as its unit of analysis, theorizes the structural features of that dyad that produce implementation outcomes, and generates falsifiable propositions about those outcomes. The Governance Gap theory is designed to occupy this position. It is not a rival to the three streams reviewed above but a complement: it accepts their insights into commitment design, regime structure, and compliance incentives, and supplies the missing layer of theory that connects these antecedents to the proximate structural determinants of implementation.

In doing so, the theory draws on each stream. From classic implementation theory we take the premise that implementation failure is structural, not accidental, and that it is mediated by features of the implementation chain. From regime complexity we take the premise that the design of commitments is conditioned by their institutional environment, and we treat this conditioning as a determinant of the Commitment (C) dimension of the Governance Gap. From the institutionalist compliance literature we take the premise that both capacity and incentives matter, and we treat these as analytically distinct dimensions, Institutional capacity (I) and Accountability architecture (A), respectively. The Monitoring (M) dimension reflects managerialist insights about the role of information in compliance, and the Resource (R) dimension reflects the institutionalist acknowledgment that capacity is a function of inputs as well as structures. The synthesis, in other words, is not eclectic but systematic: the five dimensions of the Governance Gap correspond to the principal structural determinants identified across the three streams, organized into a single theory of the commitment-jurisdiction dyad.

Before proceeding we note one further antecedent worth distinguishing from our contribution. The literature on policy capacity (Honadle, 1981; Painter and Pierre, 2005; Wu, Ramesh, and Howlett, 2015) has developed conceptual frameworks for analyzing the capacity of states to design and implement policy. This literature intersects with the Institutional capacity (I) and Resource (R) dimensions of our theory but does not provide a general account of implementation failure. It is concerned primarily with the capacities of the implementing state,

not with the structural distance between commitment and outcome. We draw on this literature in Section 4.2 in specifying the I and R dimensions, but our theory is broader: it locates these dimensions within a five-dimensional mechanism that includes the specification of the commitment, the monitoring infrastructure, and the accountability architecture as analytically independent contributors to the Governance Gap.

# 3. Methodology: Theory-Building Approach

## 3.1 Theory-Building Strategy

The research strategy combines conceptual analysis with abductive theory development and structured qualitative case assessment. The aim is to specify a theory whose constructs are conceptually distinct, whose causal pathways are explicit, and whose empirical implications can be independently tested. The study therefore distinguishes three stages: theory development, external plausibility assessment, and future confirmatory validation. This distinction is essential because the five principal cases informed the construction of the theory and cannot subsequently be treated as if they were independent tests of propositions derived without reference to them.

The conceptual analysis follows Mahoney and Goertz (2006) in decomposing the implementation problem into analytically distinct properties and specifying the relationships among them. The empirical strategy then asks whether the proposed mechanisms and behavioral signatures are observable in cases selected for substantive variation. The standard of inference is qualitative consistency rather than statistical confirmation. A proposition is treated as supported only where the case contains observable implications that were specified independently of the particular evidence used to construct the theory; otherwise the result is classified as illustrative or developmental.

The abductive component draws on Peirce (1903) and Reichertz (2007). The theory was initially developed through iterative engagement with five cases: the Paris Agreement, SDG 16 in Sub-Saharan Africa, the Sendai Framework, the EU Whistleblower Directive, and the UPR. Those cases are therefore designated theory-development cases. After the core propositions were fixed, the Chemical Weapons Convention was selected using issue-area disjointness, instrument-type variation, and documentary availability. It is treated as an external plausibility probe rather than as a confirmatory hypothesis test. No claim of random sampling or statistical representativeness is made.

## 3.2 Recap of the Five-Dimensional GGAM

Table 1. The five dimensions of the Governance Gap Assessment Matrix as defined in the article.

| Dimension | Core question in the theory | Analytical role |
|---|---|---|
| C — Commitment specification | How precisely and completely is the commitment specified? | Explanatory condition affecting interpretive and operational clarity. |
| I — Institutional capacity | Can duty-bearer institutions implement the commitment? | Explanatory condition affecting implementability. |
| M — Monitoring infrastructure | Can implementation be observed reliably? | Explanatory condition affecting observability. |

| Dimension | Core question in the theory | Analytical role |
|---|---|---|
| A — Accountability architecture | What consequences follow from non-implementation? | Explanatory condition affecting consequences for non-implementation. |
| R — Resource adequacy | Are sufficient resources available? | Explanatory condition affecting operational feasibility. |

Because this article extends rather than replaces the August 2023 AUN framework paper, we briefly recapitulate the five-dimensional Governance Gap Assessment Matrix here. For provenance clarity, the C/I/M/A/R architecture used in this article is the architecture attributed in the reference list to Advocacy Unified Network (2023), AUN Working Paper Series No. 2023-08. It should not be conflated with any other publication that uses the phrase 'Governance Gap' but employs a different dimensional architecture. The present article treats the cited AUN working paper as the sole source of the five-dimensional schema. Commitment specification (C) refers to the clarity, precision, and operability of the commitment's normative content: a commitment is fully specified when its obligation, scope, addressee, and threshold of compliance are determinable from the text or its authoritative interpretation. Institutional capacity (I) refers to the legal, administrative, and technical ability of duty-bearer institutions to give effect to the commitment: it includes legislative authority, regulatory competence, trained personnel, and inter-agency coordination. Monitoring infrastructure (M) refers to the systems through which implementation behavior is observed, measured, and reported: it includes indicators, data collection, verification protocols, and transparency arrangements. Accountability architecture (A) refers to the structures through which duty-bearers are held responsible for non-implementation: it includes judicial review, parliamentary oversight, administrative sanctions, reputational consequences, and civil society voice. Resource adequacy (R) refers to the financial, human, and technical resources dedicated to implementation: it includes budget allocations, staffing levels, infrastructure investment, and technical assistance.

In the August 2023 framework each dimension is operationalized through a small set of indicators with binary or ordinal coding, and the overall Governance Gap for a commitment-jurisdiction dyad is reported as a profile across the five dimensions rather than as a composite score. This article preserves that diagnostic logic but adds an outcome measure that is deliberately separate from the five-dimensional profile. The five dimensions are therefore explanatory conditions; G is the behavioral outcome. A high level of monitoring infrastructure does not compensate mechanically for a low level of resource adequacy, and no composite C/I/M/A/R score is substituted for the outcome. The complete outcome-scoring rule, evidence hierarchy, missing-evidence rule, and sensitivity protocol are specified in Appendix C so that an independent researcher can reproduce the measurement without inferring G from the five dimension scores.

## 3.3 From Framework to General Theory: Criteria and Strategy

The transition from framework to theory requires three properties. First, theoretical adequacy: definitions, scope conditions, mechanisms, and propositions must be explicit. Second, empirical discriminability: the theory must state observations that could count against it. Third, comparative usefulness: the theory must identify explanatory residuals that rival accounts may not address as directly. The present article advances all three properties as a theory-building exercise, but it does not claim that the final criterion has been established conclusively. Comparative explanatory superiority remains a question for future independent testing.

Our strategy is deliberately conservative. We retain the five GGAM dimensions because they provide a coherent diagnostic vocabulary, but we revise the theoretical architecture in one crucial respect: the five dimensions are treated as explanatory conditions rather than as components that definitionally constitute the outcome. We then specify a separate behavioral outcome construct and a distance function that maps normative requirements and

observed realization into the Governance Gap. We state eight propositions as theory-building hypotheses, distinguish direct effects from interaction hypotheses, and define mutually discriminating behavioral signatures for the pathology typology. The result is a more testable extension of the August 2023 framework.

A methodological note on case selection is warranted. The first five cases were selected during theory development and therefore cannot provide independent confirmation of the propositions. They are retained because they demonstrate how the framework was constructed and because they provide structured examples of the proposed mechanisms. The sixth case, the Chemical Weapons Convention (CWC), was selected after the propositions were fixed and is therefore useful as an external plausibility probe. The cases vary across issue area, instrument form, and institutional setting, but they are purposive rather than representative. The appropriate inference is consequently limited: consistency across the cases increases plausibility and identifies observable implications, but does not establish causal identification, generalizability, or predictive superiority.

The methodology therefore supports a theory-building contribution rather than a completed confirmatory theory. In particular, the article does not estimate the functional form linking the dimensions to implementation outcomes, does not establish inter-coder reliability, does not use preregistered case selection, and does not test path dependence longitudinally. These are not hidden limitations; they are explicitly incorporated into the validation protocol in Section 9 and Appendix A. The principal empirical claim is accordingly one of structured qualitative plausibility.

# 4. From Framework to Theory: Formalizing the Governance Gap

## 4.1 Formal Definition

We define the Governance Gap (G) as the behavioral distance between the normative requirements of an international commitment and the realized, observable behavior of relevant duty-bearers within a jurisdiction over a specified time horizon T. Let K denote the commitment, J the jurisdiction, N(K,J,T) the set of normative requirements applicable during T, and Y(K,J,T) the observed implementation behavior over T. Then G(K,J,T) = D[N(K,J,T), Y(K,J,T)], where D is a prespecified, element-level distance function scaled to [0,1]. Appendix C specifies D operationally: normative requirements are decomposed into non-overlapping requirement units; each unit receives a preassigned weight; behavioral realization is coded as full, partial, or absent; and G is the weighted mean of the resulting requirement-level distances. G = 0 denotes full observed realization of the coded normative requirements; G = 1 denotes complete non-realization under the coding rule. Importantly, G is the outcome construct. It is not definitionally identical to any of the five explanatory dimensions. The five GGAM dimensions—C, I, M, A, and R—are theorized as conditions that influence Y and therefore, indirectly, G.

Three features of this definition are central. First, the unit of analysis is the commitment-jurisdiction-time dyad, because the same commitment can be realized differently across jurisdictions and the same jurisdiction can perform differently across commitments. Second, implementation is dynamic, so G must be measured over an explicitly defined period rather than inferred from a single legal event. Third, G is based on observable behavioral realization rather than formal legal status alone. A jurisdiction may enact implementing legislation yet exhibit substantial behavioral non-realization if enforcement, resources, monitoring, or administrative practice remain inadequate. The measurement protocol must therefore keep formal indicators and behavioral indicators analytically separate before computing the outcome distance.

## 4.2 Axioms and Scope Conditions

The theory rests on five axioms, each of which is intended to be minimally contestable within the field and jointly sufficient to derive the propositions that follow.

Axiom 1 (Commitment dyad). The unit of analysis for the study of international implementation failure is the commitment-jurisdiction dyad (K,J). Properties of K, properties of J, and properties of the relationship between them are all theoretically relevant; properties of the international system as a whole are conditioning rather than constitutive.

Axiom 2 (Behavioral realization). Implementation outcomes are constituted by the observable behavior of duty-bearers within J, not by formal legal compliance alone. Behavioral realization includes legislative action, administrative practice, judicial enforcement, and the operational behavior of regulated entities. Where these diverge from formal compliance, behavioral realization governs G-values.

Axiom 3 (Five-dimensional explanatory structure). The implementation behavior Y is influenced by five analytically distinct conditions: Commitment specification (C), Institutional capacity (I), Monitoring infrastructure (M), Accountability architecture (A), and Resource adequacy (R). These dimensions are proposed explanatory conditions; they are not definitionally identical to G and must be coded independently of the behavioral outcome wherever possible.

Axiom 4 (Configuration dependence). The effect of any one explanatory dimension is conditioned by the values of other dimensions. The theory therefore predicts interaction effects rather than a simple additive index. The existence, direction, and magnitude of those interactions are empirical propositions rather than definitional properties of the Governance Gap.

Axiom 5 (Temporal dependence). Implementation trajectories may depend on prior configurations, sequencing, and institutional inheritance. The theory therefore permits path dependence as a testable dynamic hypothesis; it does not assume path dependence as an established empirical fact.

The scope conditions are as follows. First, the theory applies to international commitments intended to produce observable behavioral change within identifiable jurisdictions and time periods. Second, the primary formulation assumes a minimum level of functioning institutional infrastructure; failed states and active-conflict settings are outside the present model and require separate treatment. Third, the theory addresses implementation after a commitment is operative; antecedent questions of commitment formation and ratification are treated as external causes or conditioning factors rather than outcomes of this model. Fourth, the theory requires enough documentary or observational evidence to code both the explanatory dimensions and the behavioral outcome independently. These scope conditions are substantive limitations, not merely technical exclusions.

## 4.3 Formal Notation

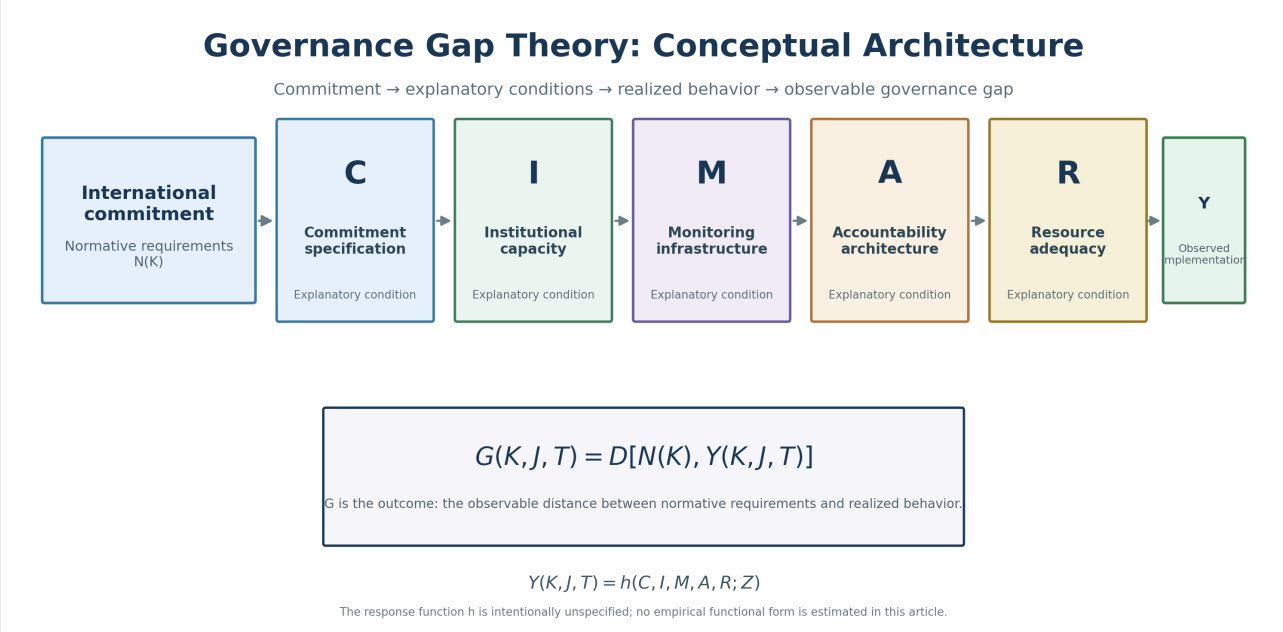


*Figure 1. Conceptual architecture of the Governance Gap theory. G is the outcome; C, I, M, A, and R are explanatory conditions. No empirical functional form is estimated here.*

We therefore use two linked but distinct pieces of notation. The outcome is G(K,J,T) = D[N(K),Y(K,J,T)]. The explanatory structure is Y(K,J,T) = h(C,I,M,A,R; Z), where Z denotes relevant conditioning factors not fully modeled in this article, including instrument type, jurisdictional political context, and regime complexity. The function h is intentionally left unspecified because no empirical dataset in the present study can identify its functional form. The theory's substantive claim is that the five dimensions and their interactions should improve explanation of variation in Y, and therefore G, relative to accounts that omit the configuration of these conditions.

Each explanatory dimension is coded on a [0,1] scale, where 0 denotes no identified gap and 1 denotes the maximum coded gap under the published coding rule. The outcome G is not calculated by averaging these five scores. Instead, the dimensions are coded from evidence about the conditions of implementation, while G is calculated from a separate behavioral realization rubric. This separation is necessary to avoid tautology and to permit construct-validity tests. The August 2023 binary/ordinal indicators can be retained as the initial measurement layer, with future work testing whether normalized scores or latent-variable estimates provide superior measurement properties.

For theory-building purposes, h(C,I,M,A,R;Z) is treated as an unspecified response surface rather than as an assumed weighted sum. A linear expression may be used in future estimation as a benchmark model, but the present article does not assume linearity, estimate weights, or interpret partial derivatives. Interaction hypotheses are stated directly in the propositions below. This avoids giving mathematical notation a level of empirical precision that the present evidence cannot support.

Because the functional form is not estimated, no empirical derivative or integral is reported in this article. Future estimation may compare additive, interaction, threshold, and latent-variable specifications. The relevant criterion is out-of-sample predictive performance rather than visual or conceptual fit alone.

## 4.4 Propositions

We state eight propositions as theory-building hypotheses. The first five concern direct, threshold, and interaction relationships; the sixth concerns behavioral distinguishability; the seventh concerns intervention configurations; and the eighth concerns temporal dependence. The five theory-development cases are used to illustrate how the hypotheses arose and how their observable implications can be coded. They are not treated as independent confirmatory tests. The CWC case is an external plausibility probe.

P1. Conditional association. Holding relevant conditioning factors constant, larger independently coded gaps in C, I, M, A, or R are expected to be associated with lower behavioral realization Y and therefore a larger Governance Gap G. This is an empirical association hypothesis; it is not tautological because the dimensions and G are to be measured from non-overlapping evidence.

P2. Instrument-contingent importance. The relative contribution of C, I, M, A, and R is expected to vary by instrument type and institutional setting. The article does not impose a universal ranking of weights. Instead, it predicts that legal precision and domestic implementation capacity will be especially consequential where obligations are legally specified, while monitoring and accountability will be especially consequential where formal coercive enforcement is weak. The direction and magnitude of these differences are to be estimated in future comparative data.

P3. Threshold effects. Each dimension may exhibit non-linear effects, such that improvements below a minimum operational threshold have smaller observable consequences than improvements that cross the threshold. A threshold claim is supported only if a prespecified model identifies a defensible change in slope or response and outperforms a preregistered simpler alternative under out-of-sample or cross-validated criteria. The present cases illustrate candidate mechanisms but do not estimate threshold values.

P4. Context-specific thresholds. Threshold values, if they exist, are expected to vary by instrument type, jurisdictional setting, and implementation task. A context-specific threshold claim is falsified if threshold estimates are statistically indistinguishable across prespecified contexts or if a pooled model performs as well as or better than context-specific models under preregistered comparison. The present article does not estimate thresholds.

P5. Configuration effects. Specific configurations are expected to generate distinguishable implementation patterns: (a) high monitoring with weak accountability should increase the probability of ritualistic compliance; (b) strong accountability with weak monitoring should increase the probability of selective compliance; (c) strong institutional capacity with inadequate resources should increase the probability of deferred compliance; (d) high specification with inadequate institutional capacity should increase the probability of symbolic compliance; and (e) jointly weak monitoring, accountability, and resources should be associated with the highest risk of severe implementation failure. These are conditional predictions, not deterministic classification rules.

P6. Behavioral distinguishability. The four proposed pathologies should be distinguishable using observable behavioral signatures. Ritualistic compliance requires a documented divergence between formal compliance activity and substantive realization; selective compliance requires systematic variation correlated with implementation cost or salience; deferred compliance requires repeated, time-extended commitments to future action without corresponding present realization; symbolic compliance requires expressive endorsement with little operational follow-through. Cases that satisfy multiple signatures are coded as hybrid rather than forced into a single category.

P7. Configuration-sensitive intervention. Interventions addressing a bottleneck in one dimension are expected to have smaller effects where another dimension remains a binding constraint. The proposition predicts diminishing returns from interventions that fail to address the binding complementary condition, but it does not imply that every multi-dimensional intervention will outperform every single-dimensional intervention. A valid test requires comparable intervention episodes, pre-specified outcome windows, and a counterfactual or credible comparison strategy. The present case material does not estimate intervention effects and therefore does not test P7.

P8. Temporal dependence. Implementation trajectories are expected to exhibit path dependence where early configurations alter the marginal effectiveness or sequencing of later interventions. The proposition predicts that otherwise similar dyads may diverge over time because different initial configurations create different feasible intervention sequences. A valid test requires repeated observations of the same commitment-jurisdiction dyads and a pre-specified dynamic model. P8 is not tested in the present article.

These propositions constitute a preregistration-ready theory-building core. They are stated so that future research can specify coding rules, case-selection criteria, threshold models, and falsification conditions before examining confirmatory data. The present article therefore treats the propositions as hypotheses generated by the framework, not as empirically confirmed laws.

# 5. The Causal Mechanism: How Gaps Produce Failure

## 5.1 Micro-Mechanisms of Each Dimension

The mechanisms below are formulated as candidate causal pathways. They explain how a gap in each dimension could affect behavioral realization, but the present case evidence does not establish that these mechanisms are the only or dominant causes. Future process-tracing and comparative designs should test the temporal sequence and rule out plausible common causes, especially political incentives that may simultaneously shape several dimensions.

The Commitment specification (C) mechanism operates through interpretive certainty. A fully specified commitment permits duty-bearers to determine, ex ante, the behavior that compliance requires; a poorly specified commitment leaves compliance-relevant behavior indeterminate, permitting strategic interpretation, selective application, and deferred implementation. The mechanism proceeds through three links. First, the text of the commitment (or its authoritative interpretation) generates a set of compliance-relevant behaviors, the union of actions that satisfy the commitment. Second, the clarity of this set, the inverse of interpretive ambiguity, determines whether duty-bearers can identify the compliance-relevant behavior. Third, the identification of compliance-relevant behavior is a necessary condition for its performance: duty-bearers cannot implement what they cannot identify. A gap in C operates by degrading the clarity of the compliance-relevant behavior set, thereby impeding identification and performance.

The Institutional capacity (I) mechanism operates through implementability. A jurisdiction with high institutional capacity possesses the legislative authority, regulatory competence, trained personnel, and inter-agency coordination required to give effect to the commitment; a jurisdiction with low institutional capacity lacks one or more of these and is therefore unable to implement even when resources, monitoring, and accountability are present. The mechanism proceeds through four links. First, the commitment must be transposed into domestic legal authority, which requires legislative action and the legal competence to take that action. Second, regulatory competence must be available to specify the rules through which the commitment is operationalized. Third, trained personnel must be available within the relevant institutions to perform the implementation tasks. Fourth, inter-agency coordination must permit the integration of implementation across the multiple agencies typically involved. A gap in I operates by breaking one or more of these links, producing either non-implementation (where legislative authority is absent) or partial implementation (where regulatory, personnel, or coordination links are broken).

The Monitoring infrastructure (M) mechanism operates through observability. A fully developed monitoring infrastructure renders the implementation behavior of duty-bearers observable to themselves, to principals (legislatures, citizens, peer states, international bodies), and to accountability agents (courts, oversight bodies, civil society). A poorly developed monitoring infrastructure leaves implementation behavior unobservable, and therefore uncorrectable and unsanctionable. The mechanism proceeds through three links. First, indicators must be defined that operationalize the compliance-relevant behavior. Second, data collection systems must be in place to measure these indicators. Third, the resulting data must be made available to the relevant principals and accountability agents, typically through public reporting. A gap in M operates by breaking one or more of these links, producing unobservability and the consequential non-correction and non-sanction of implementation failures.

The Accountability architecture (A) mechanism operates through consequentiality. A well-developed accountability architecture ensures that non-implementation has consequences for duty-bearers, whether through judicial sanctions, administrative penalties, parliamentary censure, reputational damage, or civil society mobilization. A poorly developed accountability architecture permits non-implementation to occur without consequence, removing the principal incentive for implementation. The mechanism proceeds through three links. First, the non-implementation must be detectable, which depends on the M dimension. Second, the detection must be transmitted to an accountability agent with the authority to impose consequences. Third, the accountability agent must impose consequences that are credibly sufficient to motivate implementation behavior. A gap in A operates by breaking the second or third link, producing detections without consequences, which in turn produces ritualistic compliance (Section 5.4).

The Resource adequacy (R) mechanism operates through operationality. Even with full commitment specification, institutional capacity, monitoring, and accountability, implementation requires financial, human, and technical resources. Adequate resources permit implementation tasks to be performed; inadequate resources prevent performance even where the other dimensions are fully developed. The mechanism proceeds through three links. First, budget allocations must be sufficient to fund the implementation tasks. Second, staffing levels must be sufficient to perform the tasks. Third, technical infrastructure (information systems, equipment, supplies) must be sufficient to support the tasks. A gap in R operates by breaking one or more of these links, producing deferred implementation (where budget allocations are made but not disbursed), partial implementation (where staff are deployed but technical infrastructure is inadequate), or non-implementation (where budget allocations are absent).

## 5.2 Threshold Effects

Proposition 3 specifies that each dimension exhibits threshold effects below which incremental improvement produces no observable reduction in G and above which reduction in G accelerates. The threshold effects arise from the micro-mechanisms elaborated above. The clarity threshold for Commitment specification corresponds to the point at which the compliance-relevant behavior set becomes identifiable; below this threshold, additional specification produces marginal improvements in identifiability but no behavioral change because the threshold of action-relevant clarity has not been crossed. The competence threshold for Institutional capacity corresponds to the point at which the implementing institutions can perform the implementation tasks at all; below this threshold, additional capacity produces no behavioral change because performance is binary at the level of the task. The detection threshold for Monitoring infrastructure corresponds to the point at which the probability of detection of non-implementation exceeds the threshold required for deterrence; below this, the monitoring infrastructure provides no behavioral incentive. The credibility threshold for Accountability architecture corresponds to the point at which the consequences of non-implementation are sufficiently certain and severe to motivate implementation behavior; below this, the formal existence of accountability mechanisms is insufficient to produce behavioral change. The operational threshold for Resource adequacy corresponds to the point at which resources are sufficient to perform implementation tasks at minimum viable scale; below this, marginal resources produce no implementation activity.

The presence of threshold effects has important implications for intervention design. Below the threshold for a given dimension, intervention has low marginal effect on G; above the threshold, intervention has high marginal effect. The implication is that sequencing matters: interventions should be designed to cross thresholds, not to make marginal progress below them. This implication is consistent with the empirical record of capacity-building interventions in international development cooperation, which have often produced marginal effects because they failed to cross the competence threshold (Andrews, Pritchett, and Woolcock, 2017). The theory predicts this pattern and provides a principled basis for the design of threshold-crossing interventions.

## 5.3 Interaction Effects

The interaction mechanisms are sequential and conditional. Specification affects whether the relevant behavior is identifiable; institutional capacity and resources affect whether it can be performed; monitoring affects whether performance can be observed; accountability affects whether observed non-performance generates consequences. This sequence creates plausible complementarities, but the empirical magnitude of those complementarities remains to be estimated.

The principal configuration hypotheses are therefore stated as conditional mechanisms rather than deterministic formulas. High monitoring without credible accountability creates the possibility of monitoring without consequence; high accountability without adequate monitoring creates the possibility of accountability that is selective or information-constrained; high institutional capacity without sufficient resources creates the possibility of capacity without operation; and high specification without adequate institutional capacity creates the possibility of specification without implementation capacity. The observed pathology depends additionally on the temporal and political context.

The combination of weak monitoring, weak accountability, and weak resources is hypothesized to create especially severe implementation risk because both operational and corrective channels are impaired. The present cases provide illustrative consistency with this mechanism, but the claim should be tested against cases in which one of these dimensions is strong and the other two are weak.

## 5.4 A Typology of Implementation Pathologies

The interaction hypotheses imply a typology of implementation pathologies. The typology is deliberately non-deterministic: a configuration raises the probability of a behavioral pattern; it does not logically entail it. Each pathology is identified by an observable signature that must be coded independently of the explanatory dimensions. Hybrid cases are permitted where more than one signature is present.

Ritualistic compliance is characterized by substantial formal compliance activity with limited substantive realization, especially where monitoring is developed but accountability is weak. The minimum behavioral signature is a documented divergence between reporting, procedural activity, or institutional display and the underlying implementation outcome.

Selective compliance is characterized by systematic variation in implementation across issue areas, groups, or obligations, with stronger performance where the political, administrative, or economic cost is lower or visibility is higher. The signature requires more than ordinary variation: the pattern must be systematically related to cost, salience, or observability.

Deferred compliance is characterized by repeated, credible-seeming commitments to future implementation without commensurate present realization. The signature requires temporal repetition—such as successive plans, reporting cycles, or legislative promises—rather than a single implementation delay. In the theory, deferred compliance is most directly associated with capacity/resource bottlenecks or sequencing constraints.

Symbolic compliance is characterized by strong expressive endorsement of a commitment combined with weak operational follow-through. The signature requires evidence that the duty-bearer performs visible acts of endorsement while substantive implementation remains limited. High specification plus low institutional capacity is one hypothesized pathway to symbolism, but the same configuration may instead produce deferral; the distinction is made using the temporal signature above.

The typology is not exhaustive, and the categories are not mutually exclusive in every empirical setting. Where multiple signatures are present, the case should be coded as hybrid rather than forced into a single category. This rule prevents the typology from becoming a post-hoc labeling device and creates a direct route to inter-coder reliability testing.

# 6. Case Applications: Five In-Sample Applications and One Out-of-Sample Plausibility Probe

Table 2. Case-application design and evidentiary role. The table summarizes the article's stated research design and introduces no new observations.

| Case | Role in design | Issue area | Primary implementation focus |
|---|---|---|---|
| Paris Agreement / NDCs | In-sample | Climate governance | Commitment precision, monitoring, accountability, and resources. |
| SDG 16 in Sub-Saharan Africa | In-sample | Sustainable development / governance | Institutional capacity, monitoring, accountability, and resources. |
| Sendai Framework | In-sample | Disaster-risk governance | Institutional progress versus outcome realization. |
| EU Whistleblower Directive 2019/1937 | In-sample | EU regulatory governance | Transposition and operational protection. |
| Universal Periodic Review | In-sample | Human rights | Formal review and recommendation implementation. |
| Chemical Weapons Convention | External plausibility probe | Disarmament / security | National implementation, verification, detection, and accountability. |

This section applies the Governance Gap framework to five theory-development cases and one external plausibility probe. The cases are not treated as independent tests because the first five informed theory construction. For each case, we describe the commitment and implementation record, code the five explanatory dimensions, identify the hypothesized configuration, and compare the predicted behavioral signature with the observed record. Where the evidence is insufficient to discriminate among pathologies, the result is recorded as indeterminate rather than as a forced confirmation.

## 6.1 The Paris Agreement and Nationally Determined Contributions

The Paris Agreement, adopted at the 21st Conference of the Parties to the UNFCCC in December 2015 and entered into force in November 2016, is the principal multilateral instrument for the governance of climate change. Its central innovation is the nationally determined contribution (NDC), a unilateral commitment by each party to reduce greenhouse gas emissions and adapt to climate impacts, subject to updating every five years and to a global stocktake every five years from 2023. As of mid-2023, 195 parties had ratified the Agreement and submitted at least one NDC; the Glasgow Climate Pact (COP26, 2021) had requested parties to revise and strengthen their NDCs by the end of 2022. The implementation record, as assessed by the UNEP Emissions Gap Report (UNEP, 2022), is that the sum of NDCs (assuming full implementation) places the world on a trajectory toward approximately 2.6 degrees Celsius of warming above pre-industrial levels, well above the 1.5-degree aspirational target. Within the G20, implementation of NDCs has been uneven: the European Union, the United Kingdom, and Japan have made progress consistent with their NDCs; the United States has fluctuated with political transitions; China and India are on track to meet the least ambitious elements of their NDCs but have not

strengthened them; several large emitters including Russia, Brazil, and Saudi Arabia are off-track on their stated commitments.

Assessment of the five gap dimensions is as follows. Commitment specification (C): the Paris Agreement is unusually well-specified in its procedural commitments (NDC submission, transparency framework, global stocktake) but poorly specified in its substantive ambition (the 1.5-degree target is aspirational and not legally binding on individual parties). The C dimension for any given party's NDC depends on the precision of that party's NDC text, which varies substantially: the EU NDC specifies a 55 per cent reduction by 2030 from a 1990 baseline; the Indian NDC specifies intensity targets and carbon sink targets that permit continued absolute emissions growth; the Saudi NDC specifies targets contingent on economic diversification and international support. Average C is moderate to high, with significant variance. Institutional capacity (I): high in the EU, the UK, Japan, and the United States (during the Biden administration); moderate in China and India; low in many developing country parties. Monitoring infrastructure (M): the Enhanced Transparency Framework of the Paris Agreement, fully operational from 2024, provides for biennial transparency reports, technical expert review, and a multilateral peer-review process. M is high in design, although operational M varies with party reporting capacity. Accountability architecture (A): the Paris Agreement contains no enforcement mechanism; accountability operates through reputational consequences, peer pressure at COPs, and domestic civil society. A is structurally low; the global stocktake provides information but not consequences. Resource adequacy (R): the Agreement commits developed country parties to mobilize US$100 billion annually in climate finance for developing countries, a target that has been only partially met; R for developing country parties is low, particularly for adaptation, and the 100 billion commitment is well below estimated needs (Climate Policy Initiative, 2022).

The configuration of gap dimensions is C moderate, I moderate, M high in design but variable in operation, A low, R low for developing country parties. Proposition 5(d) predicts symbolic compliance in dyads with high C and low I; Proposition 5(a) predicts ritualistic compliance in dyads with high M and low A; Proposition 5(e) predicts the most severe failure in dyads with low M, low A, and low R. Observed behavior exhibits all three pathologies. The EU exhibits a low G-value, consistent with adequate I and R; the ritualistic compliance predicted by Proposition 5(a) is observed in the EU's high-quality NDC reporting alongside slow absolute decarbonization in some sectors. The United States exhibits symbolic compliance (Proposition 5(d) variant): expressive endorsement of the Paris framework alongside insufficient domestic legislation to meet NDC targets. India and China exhibit selective compliance (Proposition 5(b)): high implementation in low-cost sectors (renewables deployment) and continued emissions growth in high-cost sectors (coal). The most severe implementation failure, predicted by Proposition 5(e), is observed in the least developed countries, where low M (limited reporting capacity), low A (no accountability for non-reporting), and low R (inadequate climate finance) combine to produce non-implementation of NDCs that were themselves highly conditional on international support.

The fit between the framework and the Paris case is qualitatively consistent with the proposed configuration mechanisms, but the case does not identify causation. In particular, the aggregate warming implication of NDCs is not itself a direct measure of implementation failure: it reflects the ambition of commitments as well as their realization. The revised analysis therefore treats ambition, implementation, and aggregate climate outcomes as distinct constructs. The Paris case is most useful here for illustrating how high monitoring can coexist with weak accountability and heterogeneous national capacity.

## 6.2 Sustainable Development Goal 16 in Sub-Saharan Africa

Sustainable Development Goal 16 commits UN member states to promote peaceful and inclusive societies, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels. Its twelve targets cover violence reduction (16.1), protection of children (16.2), access to justice (16.3), illicit financial and arms flows (16.4), corruption and bribery (16.5), effective accountable institutions (16.6), inclusive decision-making (16.7), and means of implementation (16.a, 16.b). SDG 16 is widely acknowledged as the most politically sensitive of the SDGs, and the implementation record is the weakest of any SDG. The 2022 SDG Report (United Nations, 2022) found that progress on SDG 16 had reversed in more than 40 per cent of surveyed states, with Sub-Saharan

Africa exhibiting the largest gap between targets and outcomes. Implementation assessment is rendered difficult by data limitations: 11 of the 23 SDG 16 indicators lack sufficient data for global trend assessment, and several Sub-Saharan African jurisdictions do not report on any indicator for SDG 16.1 (homicide rates) due to absence of vital registration systems.

Assessment of the five gap dimensions for SDG 16 in Sub-Saharan Africa is as follows. Commitment specification (C): SDG 16 targets vary in specification. Some indicators are tightly defined, while others depend on complex administrative or survey measurement. Institutional capacity (I): implementation capacity varies sharply across jurisdictions, particularly in statistical production, justice institutions, administrative reach, and inter-agency coordination. Monitoring infrastructure (M): the global SDG indicator tier system is versioned and has changed over time; it therefore should not be treated as a fixed 2023 count or as a direct proxy for national monitoring capacity. For this case, M is coded from the actual availability, regularity, coverage, verification, and reporting of relevant SDG 16 indicators in the jurisdictions and period examined. Accountability architecture (A): SDG 16 has no treaty-style enforcement mechanism; Voluntary National Reviews are the principal global review mechanism and are not equivalent to binding compliance proceedings. A is therefore coded as limited rather than absent. Resource adequacy (R): financing and statistical-capacity constraints vary materially across jurisdictions. The case therefore treats R as heterogeneous rather than uniformly low. This revised coding avoids treating a changing global tier classification as a timeless empirical fact and keeps the unit of analysis at the commitment-jurisdiction-time level.

The configuration is C moderate, I low, M low, A low, R low. Proposition 5(e) predicts the most severe implementation failure for this configuration, and Proposition 5(d) predicts symbolic compliance where C is high and I is low. Observed behavior exhibits both: severe implementation failure across most SDG 16 targets, with selective implementation in low-cost areas (national action plans on SDG 16 are adopted; parliamentary committees on SDG 16 are established) and symbolic compliance in high-cost areas (commitments to reduce violence, end corruption, and ensure access to justice are endorsed without operational follow-through). The VNR process exhibits the ritualistic compliance predicted by Proposition 5(a): high M in design (regular reporting) and low A (no consequences for inaccurate or absent reporting) produce reports that document formal processes without assessing behavioral realization.

The fit between theory and case is qualitatively substantial. The implementation failure pattern predicted by the configuration of gap dimensions is observed. Rival accounts illuminate important parts of the case but do not directly specify the same configuration-based mechanism. Constructivism would predict progress on SDG 16 given its strong normative endorsement by African states through the African Peer Review Mechanism and Agenda 2063; the observed reversal contradicts this. Managerialism would predict progress given the cooperative design of the HLPF VNR process; the observed stagnation contradicts this. Rationalist enforcement theory predicts non-compliance given the absence of enforcement, which fits the case, but cannot explain the variation across SDGs (which exhibit uniformly low enforcement but varying implementation outcomes). The Governance Gap theory explains both the failure pattern and the cross-SDG variation through the configuration of gap dimensions.

## 6.3 The Sendai Framework for Disaster Risk Reduction 2015-2030

The Sendai Framework, adopted at the Third UN World Conference on Disaster Risk Reduction in March 2015, is the principal global instrument for disaster risk reduction. It comprises seven global targets and four priorities for action, with implementation monitored through the Sendai Framework Monitor (SFM) hosted by the UN Office for Disaster Risk Reduction (UNDRR). Implementation is voluntary, with no enforcement mechanism; reporting is voluntary but increasingly participated in. As of the 2021 SFM report (UNDRR, 2021), 132 states had submitted at least one report; 92 states had established the national disaster loss databases that the framework requires for monitoring; and progress on the seven global targets was uneven, with Sub-Saharan African and small island developing states exhibiting the largest gaps. The framework's mid-term review in 2022 found that, although national-level institutional arrangements for disaster risk reduction had been strengthened in most jurisdictions,

behavioral outcomes (reductions in disaster mortality, displacement, and economic loss) had not improved at the rates required by the global targets.

Assessment of the five gap dimensions is as follows. Commitment specification (C): the Sendai Framework's global targets are specified in numerical terms (substantially reduce global disaster mortality by 2030 compared to 2015; reduce direct disaster economic loss in relation to global GDP by 2030), but the indicators operationalizing them are subject to methodological ambiguity, particularly for the loss and damage targets. Average C is moderate. Institutional capacity (I): national disaster management authorities exist in most jurisdictions, but their institutional authority varies widely, and coordination with other agencies (planning, finance, environment) is often weak. I is moderate in design but variable in operation. Monitoring infrastructure (M): the SFM represents the most advanced monitoring infrastructure of any soft-law framework considered here, with standardized indicators, online reporting, and periodic global assessment reports. M is high in design, although operational M is constrained by reporting capacity. Accountability architecture (A): the Sendai Framework contains no accountability mechanisms beyond global reporting and periodic Global Platforms (every two years) at which progress is reviewed. A is low. Resource adequacy (R): disaster risk reduction is chronically under-funded relative to disaster response, with the global humanitarian system spending an estimated US$3 billion annually on disaster risk reduction compared to US$25 billion on response (Kellett and Caravani, 2013). R is low.

The configuration is C moderate, I moderate, M high in design, A low, R low. Proposition 5(a) predicts ritualistic compliance: high M and low A produce reports and platforms without behavioral change. This is what is observed: the SFM reports high rates of national institutional arrangements for disaster risk reduction, while mortality and economic loss indicators continue to worsen. The mid-term review documented the divergence explicitly: institutional progress without outcome progress. Proposition 5(e) predicts severe failure for dyads with low A and low R, which is observed in small island developing states where accountability mechanisms are entirely absent and resources for disaster risk reduction are minimal. Proposition 5(c) predicts deferred compliance: high I (institutional arrangements exist) and low R (operational resources are inadequate) produce promissory language in implementation reports about future action. The 2022 mid-term review is replete with such language.

The fit between theory and case is qualitatively substantial. The pattern of institutional progress without outcome progress, predicted by the configuration of high M, low A, and low R, is observed. Rival accounts illuminate important parts of the case but do not directly specify the same configuration-based mechanism. Managerialism predicts progress given the cooperative design of the SFM; observed stagnation contradicts this. Rationalist enforcement theory predicts non-compliance given the absence of enforcement, which fits but cannot explain the institutional progress. Constructivism would predict norm cascade given the strong endorsement of disaster risk reduction norms; the observed divergence between institutional and outcome progress contradicts this. The Governance Gap theory explains both through the configuration of gap dimensions.

## 6.4 The European Union Whistleblower Directive 2019/1937

Directive (EU) 2019/1937 on the protection of persons who report breaches of Union law was adopted by the European Parliament and Council in October 2019 and required transposition by Member States by 17 December 2021. The Directive establishes minimum standards for the protection of whistleblowers across a range of EU-law areas, including public procurement, financial services, money laundering, product safety, environmental protection, and public health. The enforcement record is more precise than a single ‘23 Member States’ count suggests. As of 10 June 2023, the European Commission had opened infringement proceedings against 20 Member States for incorrect transposition of the Directive; two of those proceedings had been opened in 2023, one had been closed, and the Commission was assessing replies in the remaining cases (European Commission, 2023). This contemporaneous procedural record is used here because infringement stage is itself relevant to the accountability dimension. Operational implementation, in the sense of actual protection of actual whistleblowers, remains a separate empirical question.

Assessment of the five gap dimensions is as follows. Commitment specification (C): the directive is highly specified, with detailed provisions on reporting channels, protection measures, retaliation remedies, and follow-up obligations. C is high. Institutional capacity (I): member states vary in institutional capacity for whistleblower protection. Some (Netherlands, Sweden, France) had pre-existing institutional frameworks and capacity; others (Bulgaria, Romania) had limited pre-existing capacity and have struggled to develop it. I is high in some jurisdictions, low in others. Monitoring infrastructure (M): the directive requires member states to provide anonymous statistical data on reporting to the European Commission, but the Commission's capacity to monitor operational implementation is limited. M is moderate in design, low in operation. Accountability architecture (A): the directive's enforcement depends on infringement proceedings by the European Commission and on judicial review by national courts. A is moderate, with the European Court of Justice providing a credible enforcement path for systemic non-implementation. Resource adequacy (R): whistleblower protection requires financial support for whistleblowers, administrative resources for reporting channels, and judicial resources for retaliation remedies. R varies across jurisdictions, with significant under-resourcing in several member states.

The configuration is C high, I variable, M moderate, A moderate, R variable. Proposition 5(d) predicts symbolic compliance in dyads with high C and low I, which is observed in member states with high directive specification but low institutional capacity: these states transpose the directive text into national law but fail to operationalize the protection through functional reporting channels and judicial remedies. Proposition 5(b) predicts selective compliance in dyads with high A and low M, which is observed in states with strong judicial accountability for visible non-implementation but limited monitoring of operational follow-through: these states protect whistleblowers in high-profile cases that reach judicial attention but provide limited protection in the routine cases that constitute the bulk of whistleblower experience.

The fit between theory and case is qualitatively substantial but remains non-confirmatory. The variation in transposition and the distinction between formal transposition and operational implementation correspond to the theory's proposed separation between normative specification, institutional conditions, monitoring, accountability, and resources. Rival accounts remain informative: rationalist enforcement helps explain the role of infringement and judicial pressure; principal-agent analysis helps explain Commission monitoring and enforcement; constructivism helps explain normative adoption. The present case therefore illustrates the complementary value of the Governance Gap framework without claiming that the case uniquely identifies its mechanism or establishes superiority over rival explanations.

## 6.5 The United Nations Universal Periodic Review

The Universal Periodic Review (UPR) is a peer-review mechanism of the United Nations Human Rights Council, established in 2006, under which all 193 UN member states are reviewed on their entire human rights record every four to five years. The review is based on three documents: a national report by the state under review, a compilation of UN information by the Office of the High Commissioner for Human Rights (OHCHR), and a summary of stakeholder input from civil society. The review concludes with a set of recommendations by peer states, which the state under review may accept or note. Implementation of accepted recommendations is reviewed at the next cycle. By the end of the third cycle, the UPR had generated a very large body of recommendations across the UN membership; because the quantitative implementation evidence cited in this article comes from the earlier *Beyond Promises* study, no third-cycle implementation percentage is asserted here. UPR Info's *Beyond Promises* study found that 48 per cent of recommendations in its first-cycle follow-up dataset had triggered action by the mid-term point, with 55 per cent of accepted recommendations triggering action. This evidence demonstrates that formal recommendation processes do not guarantee implementation (UPR Info, 2014).

Assessment of the five gap dimensions is as follows. Commitment specification (C): UPR recommendations vary widely in specification. Some are highly specified (ratify the Convention against Torture; enact legislation criminalizing domestic violence); others are broadly aspirational (strengthen human rights institutions; promote gender equality). Average C is moderate, with high variance. Institutional capacity (I): variation across jurisdictions

mirrors the variation in human rights institutional capacity more generally. I is high in jurisdictions with mature human rights institutions (Western Europe, parts of Latin America), low in jurisdictions with weak or absent human rights institutions. Monitoring infrastructure (M): the UPR itself constitutes a monitoring infrastructure, with periodic review and OHCHR compilation. M is high in design but variable in operation, depending on the quality of national reporting. Accountability architecture (A): the UPR contains no accountability mechanism beyond peer pressure and reputational consequences. A is low, although reputational consequences can be substantial for jurisdictions that care about their international standing. Resource adequacy (R): the UPR has no resource implications for implementation; resources for the implementation of specific recommendations vary with the recommendation content and the jurisdiction's overall resource position. R is variable.

The configuration is C moderate (with high variance), I variable, M high in design, A low, R variable. Proposition 5(a) predicts ritualistic compliance in dyads with high M and low A, which is consistent with the UPR's high formal engagement (national reports, peer questioning, and recommendation responses) alongside uneven behavioral realization. The pattern of acceptance followed by non-implementation is consistent with the ritualistic-compliance pathway proposed by the theory. Proposition 5(d) predicts symbolic compliance in dyads with high C and low I, while Proposition 5(b) predicts selective compliance where accountability and monitoring vary across issues and jurisdictions.

The fit between theory and case is qualitatively substantial. The documented implementation evidence is consistent with the configuration of gap dimensions and with the predicted pathway of ritualistic compliance. Rival accounts illuminate important parts of the UPR, but the Governance Gap framework offers an additional configuration-based explanation of variation in implementation.

Across the five case applications, the theory exhibits qualitative fit. The configurations of gap dimensions predicted by the theory correspond to the observed implementation outcomes, and the pathologies predicted for each configuration are observed in the behavioral signatures of each case. The comparative analysis in Section 7 strengthens this assessment by supporting a comparative assessment of the theory against rival accounts.

## 6.6 The Chemical Weapons Convention (Out-of-Sample Plausibility Probe)

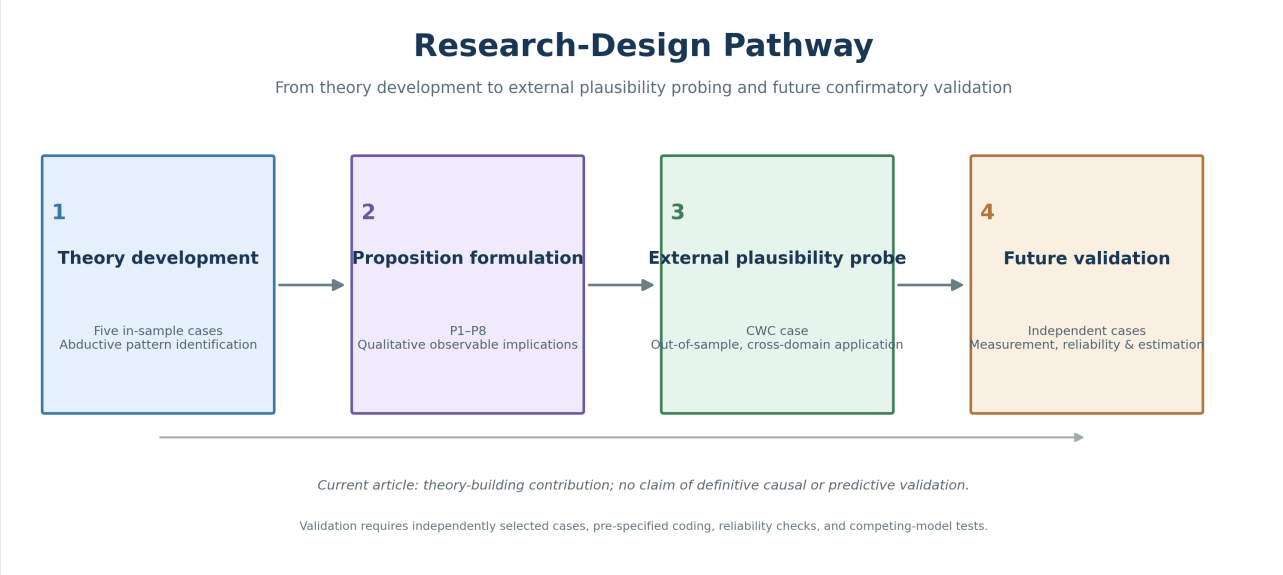


*Figure 2. Research-design pathway from theory development through an external plausibility probe to future confirmatory validation.*

The five cases analyzed in Sections 6.1 through 6.5 were used, in part, in the abductive development of the theory (Section 3.1); their patterns informed the formulation of the propositions. A stronger test of the theory would require at least one case that was not used in theory development. Section 6.6 therefore uses the Chemical

Weapons Convention (CWC) as an external plausibility probe. The CWC, opened for signature in 1993 and entered into force in 1997, is analytically valuable because its issue area—disarmament and security—is disjoint from the five theory-development cases. The CWC is a binding treaty with a highly developed verification regime; its Executive Council and Conference of States Parties operate within the Convention's institutional and legal framework. We therefore do not label these mechanisms generically as 'soft law'. Instead, the case is used to test whether the proposed configuration logic travels into a distinct legal and security regime. The selection criteria were fixed after the propositions: issue-area disjointness, institutional distinctiveness, and availability of sustained documentary evidence. The case is an external plausibility probe, not a confirmatory test.

The CWC prohibits the development, production, stockpiling, and use of chemical weapons and requires their destruction. Implementation is monitored by the Organisation for the Prohibition of Chemical Weapons (OPCW), established by the Convention and headquartered in The Hague. As of mid-2023, the CWC had 193 States Parties. Egypt, Israel, the Democratic People's Republic of Korea (North Korea), and South Sudan remained outside the Convention. The principal implementation obligations include declaration of chemical-weapons stocks and production facilities; destruction of declared stocks and facilities under OPCW verification; declaration of relevant civil chemical industry; and acceptance of routine and challenge inspections. OPCW verified on 7 July 2023 that the declared chemical-weapons stockpiles of all States Parties had been irreversibly destroyed (OPCW, 2023). Separate compliance and attribution questions concerning alleged chemical-agent use require analysis under the OPCW's distinct verification, fact-finding, and attribution mechanisms.

Assessment of the five gap dimensions is as follows. Commitment specification (C): the CWC is exceptionally well-specified, with detailed provisions in the Convention text, the Verification Annex (a separate instrument of equivalent legal force), and twenty-five years of authoritative interpretation by the Conference of States Parties and the OPCW Executive Council. C is high and stable across jurisdictions. Institutional capacity (I): the CWC imposes demanding institutional requirements on states parties, including the establishment of a National Authority, the enactment of implementing legislation (required by Article VII), and the establishment of penalties for prohibited acts. The OPCW reports that OPCW reporting for 2022 documented substantial variation in national implementation of Article VII obligations, including gaps in comprehensive implementing legislation (OPCW, 2022). I varies across jurisdictions, with substantial implementation capacity in some states and important gaps in others. Monitoring infrastructure (M): the OPCW operates one of the most developed monitoring infrastructures in any international regime, with routine inspections of declared chemical weapons facilities and civil industry sites, challenge inspections, and a network of designated laboratories for sample analysis. M is high in design and operation, with the technical capacity to detect non-implementation at declared sites; undeclared sites and activities (undeclared activities outside routine declared-site verification) remain harder to detect, as the Syria and Skripal cases illustrate. Accountability architecture (A): the CWC's accountability architecture is layered: the Executive Council may issue formal findings of non-compliance, the Conference of States Parties may take measures including suspension of rights and privileges, and cases may be referred to the United Nations Security Council under Article XII. A is moderate, with the formal architecture strong but the political credibility of sanctions uneven across cases. Resource adequacy (R): the OPCW is funded by states parties on a UN-assessed contribution scale, with resources provided through assessed contributions and other agreed funding arrangements; destruction costs are borne by possessor states. R is high for OPCW operations but variable for state-level implementation, with several developing states lacking the resources to maintain effective National Authorities.

The configuration is C high, I variable, M high in design but variable in operational coverage, A layered and politically contingent, and R variable across national implementation systems. Proposition 5(d) predicts symbolic compliance in dyads with high C and low I; this is a plausible pattern where states maintain formal treaty membership while domestic implementing institutions remain incomplete. Proposition 5(a) predicts ritualistic compliance where monitoring is strong but consequences are weak; the CWC provides a useful setting in which extensive verification can coexist with limits on political enforcement, although the present evidence is insufficient to classify a specific state as ritualistic without frozen behavioral coding. Proposition 5(e) predicts severe failure

where monitoring, accountability, and resources are jointly weak; this is a hypothesis for independent testing rather than a conclusion inferred from the CWC narrative. The case therefore supplies bounded, theory-relevant observations without treating the observed configuration as proof of the proposition.

The most theoretically demanding issue concerns the limits of detection and accountability where alleged chemical-weapons use or other activities fall outside routine declared-site verification. The theory predicts that where C is high, M is strong for declared activities but less capable of detecting activities outside routine declared-site verification, and A is moderate, implementation failures may persist where detection and accountability are difficult. In the Syria case, OPCW investigative and attribution mechanisms have identified responsibility for specific chemical-weapons incidents; the broader analytical point is that verification, attribution, and accountability are not identical functions and may generate different implementation constraints. The theory's proposed mechanism is therefore treated as a plausible explanation rather than as proof of an alleged undeclared-activity hypothesis.

The external CWC plausibility probe is qualitatively consistent with the proposed configuration logic in several respects, but the evidentiary status remains bounded. It does not eliminate selection concerns, establish causal identification, or demonstrate predictive superiority. Its principal value is that the theory's observable implications can be articulated and checked in a legally and substantively distinct domain. A stronger assessment requires multiple external cases selected independently of the theory, frozen coding rules, explicit negative-case handling, and preregistered analysis.

Comparative assessment against the four rivals on the CWC is deliberately asymmetric in epistemic burden because the present article is theory-building rather than a preregistered head-to-head test. Regime-density theory, principal-agent theory, constructivist norm-cascade theory, and rationalist enforcement theory each explain important aspects of the CWC, while the Governance Gap framework highlights the implementation-configuration residual. The appropriate conclusion is therefore not that the Governance Gap theory is the only account capable of explaining the case, but that the case exposes an observable set of implementation questions —variation in domestic capacity, monitoring coverage, accountability, and resources—that can be incorporated systematically into the proposed framework. A genuinely comparative test would require preregistered predictions and the same evidence standard for every rival model.

# 7. Comparative Analysis: Governance Gap Theory versus Rival Explanations

The six case applications in Section 6, including the out-of-sample plausibility probe using the Chemical Weapons Convention in Section 6.6, provide qualitative evidence of fit between the Governance Gap theory and implementation outcomes across a diverse set of cases. Fit, however, is a weak standard for theory evaluation: many theories can fit the same data. The more demanding standard is comparative explanatory value: does the theory explain the relevant phenomena at least as well as, and preferably better than, extant rivals? This section compares the Governance Gap theory against four rival accounts identified in Section 2: regime density theory, principal-agent theories of delegation, constructivist norm cascade, and rationalist enforcement theory. We assess each rival along three dimensions: explanatory scope (range of cases the theory can address), explanatory depth (mechanisms specified), and predictive power (falsifiable predictions generated). We then summarize the comparative assessment and identify the conditions under which each rival remains explanatorily useful.

## 7.1 Regime Density Theory

Regime density theory, advanced principally by Raustiala and Victor (2004), Alter and Meunier (2009), and Keohane and Victor (2011), explains implementation outcomes as a function of the density and structure of the international institutional environment. Implementation of any one instrument is conditioned by the existence of overlapping instruments, with consequences for forum-shopping, inter-regime interplay, and the construction of

variable geometries of commitment. The theory has substantial explanatory scope: it applies across all cases that involve multiple overlapping instruments, which is most contemporary cases. It has substantial explanatory depth: it specifies mechanisms through which institutional environments shape instrument design and, indirectly, implementation. It has moderate predictive power: it predicts that instruments embedded in denser institutional environments will exhibit different implementation patterns than instruments in sparse environments, but the direction of the difference depends on configuration details that the theory does not fully specify.

Against the six cases, regime density theory performs variably. The Paris Agreement case is well-explained by regime density: the Paris architecture was deliberately designed to be embedded in a dense institutional environment (UNFCCC, Kigali, CORSIA, IMO), and the implementation pattern reflects this. The SDG 16 case is less well-explained: SDG 16 is embedded in a sparse institutional environment (no overlapping instruments), yet exhibits severe implementation failure that regime density theory would not predict from its sparse environment. The Sendai Framework case is partially explained: the framework is embedded in a sparse institutional environment, but its implementation pattern reflects features that regime density theory does not address. The EU Whistleblower Directive case is partially explained: the directive is embedded in a moderately dense EU institutional environment, but the variation in transposition across member states is not explained by regime density. The UPR case is well-explained by regime density: the UPR is embedded in a dense human rights institutional environment, and its implementation pattern reflects the forum-shopping and overlap dynamics that regime density theory predicts. The CWC out-of-sample plausibility probe is partially explained: the disarmament regime is sparse, which predicts low implementation, but the specific declared-versus-undeclared compliance split is not explained.

The comparative assessment is that regime density theory explains important features of instrument design and the macro-patterns of implementation across instruments, but does not address the variation in implementation across jurisdictions for any given instrument. The Governance Gap theory complements regime density theory by addressing the latter question. The two theories are most powerful when used in combination: regime density theory explains why an instrument is designed as it is, and the Governance Gap theory explains why it is implemented as it is across jurisdictions.

## 7.2 Principal-Agent Theories of Delegation

Principal-agent (PA) theories of delegation, advanced by Hawkins, Lake, Nielson, and Tierney (2006) and elaborated by Bradley and Kelley (2008), explain international cooperation as the delegation of authority from states (principals) to international organizations (agents), with implementation shaped by the agent's discretion and the principal's control mechanisms. The theory has substantial explanatory scope in domains involving delegation to international organizations: it is most useful for explaining the design and behavior of international organizations. It has substantial explanatory depth: it specifies mechanisms through which delegation creates agency slack and the control mechanisms through which principals counteract slack. It has substantial predictive power: it predicts that delegation will produce agent autonomy, that control mechanisms will be calibrated to the cost of slack, and that slack will produce specific patterns of agent behavior.

Against the six cases, PA theory performs variably. The Paris Agreement case is partially explained by PA theory: the UNFCCC Secretariat has limited agency in the Paris architecture (which is state-driven), so PA dynamics are muted. The SDG 16 case is partially explained: the SDG monitoring architecture involves substantial delegation to international agencies (UNODC for crime indicators, OHCHR for justice indicators), and implementation outcomes reflect PA dynamics in some indicators. The Sendai Framework case is partially explained: the UNDRR has substantial agency in the framework's monitoring, and PA dynamics shape the framework's implementation. The EU Whistleblower Directive case is well-explained by PA theory: the directive involves delegation from member states (principals) to the European Commission (agent) for enforcement, and the variation in transposition reflects PA dynamics. The UPR case is partially explained: the OHCHR has agency in the UPR's compilation process, and implementation outcomes reflect PA dynamics. The CWC out-of-sample plausibility probe is partially explained:

the OPCW has substantial agency in monitoring, but the state-level non-implementation and the detection-limited cases involving alleged undeclared activity are not directly explained by PA dynamics.

The comparative assessment is that PA theory explains important features of delegation to international organizations and the macro-patterns of agent behavior. The Governance Gap theory is largely complementary: where PA theory explains the design of international delegation, the Governance Gap theory explains the implementation of the delegated commitments at the level of the commitment-jurisdiction dyad. The two theories operate at different levels of analysis, and combining them is straightforward.

## 7.3 Constructivist Norm Cascade

Constructivist norm cascade theory, advanced principally by Finnemore and Sikkink (1998) and elaborated by Checkel (2005), explains the diffusion and adoption of international norms through a three-stage process: norm emergence, norm cascade, and norm internalization. Implementation, on this account, is the final stage of internalization, where norms become taken-for-granted and behavioral conformity follows. The theory has substantial explanatory scope: it applies to all normative commitments. It has substantial explanatory depth: it specifies mechanisms of persuasion, socialization, and identification through which norms diffuse. It has substantial predictive power: it predicts that norms will diffuse in cascades following a tipping point of state adoption, and that internalization will follow cascade.

Against the six cases, constructivist norm cascade theory performs variably. The Paris Agreement case is partially explained: the 1.5-degree norm has cascaded in international discourse, but internalization (in the form of behavioral conformity) has not followed, contradicting the theory's prediction. The SDG 16 case is partially explained: the SDG 16 norms have not cascaded in the strong sense, and implementation outcomes reflect this. The Sendai Framework case is partially explained: the disaster risk reduction norms have cascaded in international discourse, but internalization has not followed. The EU Whistleblower Directive case is partially explained: the whistleblower protection norm has cascaded in EU discourse, but transposition outcomes reflect features that constructivism does not address. The UPR case is well-explained by constructivism in its expressive dimensions (norm adoption in the UPR is high) but not in its behavioral dimensions (implementation is moderate). The CWC out-of-sample plausibility probe is partially explained: the anti-chemical-weapons norm has near-universally cascaded, yet behavioral conformity is split between declared-program compliance and non-compliance involving alleged undeclared activity, a pattern constructivism does not predict.

The comparative assessment is that constructivism explains the diffusion and adoption of international norms at the discursive level but does not predict when adoption will translate into implementation. The Governance Gap theory complements constructivism by addressing this translation. The two theories are most powerful when used in combination: constructivism explains why commitments are made and adopted, and the Governance Gap theory explains why adoption does or does not translate into implementation.

## 7.4 Rationalist Enforcement Theory

Rationalist enforcement theory, advanced by Downs, Rocke, and Barsoom (1996) and elaborated by Simmons (2010), explains compliance as a function of the cost-benefit calculation of duty-bearers in the shadow of enforcement. Compliance is most likely when the costs of non-compliance (sanctions, retaliation, reputational damage) exceed the benefits of defection. The theory has substantial explanatory scope in domains involving enforcement: it is most useful for explaining compliance with hard-law instruments where enforcement is credible. It has substantial explanatory depth: it specifies mechanisms of strategic calculation and enforcement response. It has substantial predictive power: it predicts that compliance will vary with enforcement credibility and the cost-benefit balance of defection.

Against the six cases, rationalist enforcement theory performs variably. The Paris Agreement case is partially explained: the absence of enforcement predicts low compliance, which fits the global pattern but not the variation

across parties (the EU exhibits high compliance without enforcement). The SDG 16 case is partially explained: the absence of enforcement predicts low compliance, which fits, but cannot explain variation across SDGs. The Sendai Framework case is partially explained: the absence of enforcement predicts low compliance, which fits the institutional-versus-outcome divergence but cannot explain it. The EU Whistleblower Directive case is well-explained by rationalist enforcement: the European Court of Justice provides credible enforcement, and the variation in transposition reflects the cost-benefit calculation of member states. The UPR case is partially explained: the absence of enforcement predicts low compliance, which fits the moderate implementation rate. The CWC out-of-sample plausibility probe is partially explained: the declared-program compliance is well-explained by the credible sanction threat (Security Council referral, OPCW findings), but the non-compliance involving alleged undeclared activity is not explained by enforcement alone, since enforcement depends on detection, which depends on M.

The comparative assessment is that rationalist enforcement theory explains compliance when enforcement is credible and compliance is strategically motivated, but does not address the variation in implementation across jurisdictions with identical enforcement environments, nor the implementation that occurs in the absence of enforcement, nor the pattern of declared-versus-undeclared compliance split that the CWC case exhibits. The Governance Gap theory complements rationalist enforcement theory by addressing each of these gaps: it explains variation across jurisdictions through the configuration of gap dimensions, it explains implementation in the absence of enforcement through the operation of the M and A mechanisms in non-enforcement settings, and it explains the difference between routine declared-site verification and detection-limited cases involving alleged undeclared activity through the configuration-specific predictions of Proposition 5.

## 7.5 Synthesis: When the Governance Gap Theory Adds Value

**Table 3. Bounded comparative assessment of the Governance Gap framework and rival accounts**

| Account | Primary explanatory strength | Residual question addressed here | Status in this article |
|---|---|---|---|
| Regime density / complexity | Institutional environment, overlap, and forum structure. | Why implementation varies across dyads operating under the same or similar institutional environment. | Complementary; not outperformed by a formal test. |
| Principal-agent | Delegation, agency slack, and behavior of international organizations. | How domestic implementation conditions shape outcomes after delegation is established. | Complementary; no head-to-head causal test. |
| Constructivist norm diffusion | Norm emergence, socialization, and internalization. | Why expressive adoption may diverge from behavioral realization. | Complementary; no head-to-head causal test. |
| Rationalist enforcement | Strategic incentives, sanctions, and credible enforcement. | Why outcomes vary where enforcement is weak, indirect, or common across jurisdictions. | Complementary; no head-to-head causal test. |

| | | | |
|---|---|---|---|
| Governance Gap | Configuration of specification, capacity, monitoring, accountability, and resources at the implementation stage. | Provides a common commitment-jurisdiction-time unit and explicit falsifiable configuration hypotheses. | Theory-building proposal; comparative superiority remains untested. |

The comparative assessment across the four rivals identifies a consistent research opportunity rather than a completed superiority result. Each rival explains important features of international cooperation: regime density addresses institutional environment; principal-agent theory addresses delegation; constructivism addresses norm diffusion; and rationalist enforcement addresses compliance under credible incentives. The Governance Gap framework asks a different proximate question: given an operative commitment, what configuration of implementation conditions is associated with the observed behavioral outcome? The six cases suggest that this question is analytically useful, but they do not establish that the Governance Gap theory outperforms the rivals. Comparative superiority remains an explicit falsifiable research question.

The theory's value-added is therefore stated in bounded terms. It offers a common unit of analysis, separates outcome measurement from explanatory conditions, generates configuration-sensitive propositions, and provides a reproducible route from normative requirements to behavioral realization. The cases illustrate how those features can organize evidence that is otherwise distributed across legal, administrative, monitoring, and resource literatures. They do not establish causal superiority. The paper's claim is that the framework is sufficiently specified to permit a fair future comparison with rival theories under preregistered conditions.

We close this section with a candid acknowledgment of where the rivals outperform the Governance Gap theory. Regime density theory outperforms the Governance Gap theory in explaining the design of instruments in dense institutional environments, as the Paris Agreement case illustrates. Principal-agent theory outperforms in explaining the behavior of international organizations, as the EU Whistleblower Directive case illustrates in its European Commission enforcement dimension, and as the OPCW's operational behavior in the CWC case illustrates. Constructivism outperforms in explaining the diffusion and adoption of norms, as the UPR case illustrates in its expressive dimensions, and as the near-universal cascade of the anti-chemical-weapons norm in the CWC case illustrates. Rationalist enforcement theory outperforms in explaining compliance under credible enforcement, as the EU Whistleblower Directive case illustrates in its European Court of Justice enforcement dimension, and as the declared-program compliance pattern in the CWC case illustrates. The Governance Gap theory is best understood as a complement that addresses the residual question of implementation outcomes, and the cases, including the out-of-sample plausibility probe, suggest substantial value-added in this complementary role.

# 8. Discussion: Theoretical and Policy Implications

The Governance Gap theory, as developed in Sections 4 through 7 and assessed across six cases including an out-of-sample plausibility probe on the Chemical Weapons Convention, has implications for the academic study and practice of international implementation. The evidence supports treating the framework as a promising theory-building contribution, while the limitations in measurement, case selection, and formal validation counsel against claims of definitive confirmation.

## 8.1 Theoretical Implications

The principal theoretical implication of the Governance Gap theory is that international implementation failure is not a residual category, the unexplained remainder after theories of cooperation, delegation, and compliance have done their work, but a phenomenon with its own structural determinants that can be theorized directly. The commitment-jurisdiction dyad, the unit of analysis that the theory centers, has been an implicit but under-theorized object in the literature, treated as the site where antecedent variables, instrument design, regime

density, norm cascade, enforcement, produce their effects, but not theorized as a structured site with its own causal properties. The Governance Gap theory reverses this: it theorizes the dyad as the level of analysis at which implementation outcomes are produced, and it specifies the structural features of the dyad that produce these outcomes.

A second theoretical implication is the configuration-rather-than-aggregate logic of the theory. The five dimensions of the Governance Gap are not additive contributors to G-values; their effects on G are conditioned by the values of the other dimensions. This logic contrasts with the latent variable logic of much empirical compliance research, which treats compliance as a function of an underlying propensity to comply that varies across jurisdictions (von Stein, 2016). The configuration logic predicts, and the cases confirm, that jurisdictions with similar underlying propensities to comply can exhibit radically different implementation outcomes across instruments, depending on the configuration of gap dimensions for each instrument. This prediction is not generated by extant accounts and is qualitatively consistent across the theory-development cases in Section 6.

A third theoretical implication is the typology of implementation pathologies, which provides the field with a portable vocabulary for distinguishing forms of implementation failure that are often conflated. The distinction between ritualistic, selective, deferred, and symbolic compliance permits more precise description of observed behavior patterns and more targeted theoretical explanation. Prior typologies, such as the Chayes and Chayes (1995) distinction between rule-consistent and rule-violating behavior, or the Hathaway (2002) distinction between sincere and insincere ratification, address aspects of the implementation pathology space but do not exhaust it. The four-pathology typology is empirically operationalizable through behavioral signatures, theoretically grounded in configurations of gap dimensions, and portable across cases. It is, we believe, the typology's most immediately useful contribution for the field.

A fourth theoretical implication is the theory's capacity to integrate insights from the four rivals reviewed in Section 7 into a single framework. The Commitment specification dimension absorbs insights from legalization and regime density theory; the Institutional capacity dimension absorbs insights from policy capacity scholarship; the Monitoring infrastructure dimension absorbs managerialist insights about information; the Accountability architecture dimension absorbs rationalist insights about enforcement and constructivist insights about social consequences; and the Resource adequacy dimension absorbs insights from capacity-building and developmental scholarship. The integration is not eclecticism: the five dimensions are analytically distinct, their interactions are specified, and the resulting theory generates falsifiable predictions that the rivals alone do not.

## 8.2 Policy Implications: Treaty Design and Monitoring Architecture

The theory has several implications for the design of international instruments. First, instrument designers should attend to the configuration of gap dimensions that the instrument will produce across jurisdictions, not only to the substantive content of the instrument. An instrument that produces systematically high C and low I, or high M and low A, across jurisdictions will produce predictable implementation pathologies. The Paris Agreement's design, which produces high M (Enhanced Transparency Framework) and low A (no enforcement), predicts the ritualistic compliance observed in the case application. Alternative designs, such as the Aarhus Convention's combination of high C, high M, and high A through its Compliance Committee, produce different and more favorable implementation patterns. The theory provides a principled basis for these design choices.

Second, instrument designers should attend to threshold effects in the dimensions, but these implications should be treated as design hypotheses rather than established effects. A design that fails to cross a clarity or detection threshold may plausibly generate elevated implementation gaps; whether it does so systematically is an empirical question. The SDG 16 case illustrates why monitoring architecture must be assessed through actual indicator availability, regularity, coverage, and verification rather than through a static historical tier count. The theory therefore provides a disciplined basis for formulating threshold-sensitive design questions without claiming that threshold values have already been estimated.

Third, instrument designers should attend to the interaction effects specified in Proposition 5. Designs that produce high M in the absence of high A produce ritualistic compliance; designs that produce high A in the absence of high M produce selective compliance; designs that produce high I in the absence of high R produce deferred compliance; designs that produce high C in the absence of high I produce symbolic compliance. The implications are concrete: monitoring infrastructure should not be designed in isolation from accountability architecture, capacity-building should not be designed in isolation from resource mobilization, and specification should not be designed in isolation from institutional capacity assessment.

## 8.3 Policy Implications: Advocacy and Civil Society

The theory has direct implications for the advocacy practice of organizations such as the Advocacy Unified Network and allied civil society organizations. First, the theory provides a diagnostic framework for identifying the configurations of gap dimensions that produce implementation failure in any given commitment-jurisdiction dyad. Where the diagnostic identifies C and I as the dominant gap dimensions, the appropriate intervention is technical assistance and legislative drafting; where M and A dominate, the appropriate intervention is monitoring infrastructure and judicial or political accountability; where R dominates, the appropriate intervention is resource mobilization. The diagnostic framework permits advocacy to move from reactive complaint (this commitment is not being implemented) to targeted intervention (this specific configuration of gap dimensions is producing this specific implementation pathology, and this specific intervention would reduce it).

Second, the theory provides a predictive framework for anticipating the implementation outcomes of new commitments. A new commitment with high C (well-specified) and low I (low institutional capacity in target jurisdictions) is predicted, by Proposition 5(d), to produce symbolic compliance; advocacy resources should accordingly be directed to building institutional capacity rather than further specification. A new commitment with high M and low A is predicted to produce ritualistic compliance; advocacy resources should be directed to accountability architecture rather than further monitoring. The predictive framework permits advocacy to anticipate the pathologies that new commitments will produce and to target advocacy resources accordingly.

Third, the theory provides an evaluative framework for assessing the effectiveness of advocacy interventions. Where advocacy has focused on a single dimension in isolation (for example, capacity-building in isolation from resource mobilization), Proposition 7 predicts limited effect on G-values; the evaluative framework permits the systematic assessment of these predictions. The framework also permits the comparison of advocacy strategies across commitment-jurisdiction dyads, supporting the development of an evidence-based advocacy practice that has heretofore been hampered by the absence of a common analytic currency.

A specific implication for AUN's own advocacy agenda follows. The August 2023 framework paper presented the GGAM as a diagnostic tool; the present article elevates it to a theoretical contribution. The implication is that AUN's advocacy agenda can move from diagnosis to prediction and evaluation, using the theory's propositions to anticipate the consequences of new commitments, target advocacy resources to the configurations of gap dimensions that produce the most severe implementation failure, and evaluate the effectiveness of advocacy interventions. The result is an advocacy practice that is at once more theoretically grounded, more predictive, and more evidence-based.

Finally, the theory has implications for the broader international community. The 2024 Summit of the Future, the 2026 review of the Sendai Framework, the 2025 review of the Paris Agreement, and the 2030 endpoint of the SDGs are all occasions at which the international community will assess implementation progress and design new commitments. The Governance Gap theory provides a principled basis for these assessments and designs: the configurations of gap dimensions in past commitments predict their implementation outcomes, and the configurations in new commitments predict their outcomes. The theory's adoption by international secretariats, donor agencies, and review bodies would, we believe, contribute to more effective international governance and to the closing of the governance gaps that have heretofore characterized contemporary international cooperation.

# 9. Limitations and Future Research

The Governance Gap theory, as developed in this article, is subject to several limitations that should be candidly acknowledged and that constitute the agenda for future research. We discuss eight limitations in turn, organized under the headings of scope, measurement, mechanism, prediction, application, generalization, epistemic, and ethics. The expansion of the limitations section relative to a more compressed treatment reflects the article's aspiration to theoretical maturity: a theory that does not specify the conditions under which it fails is not yet a theory.

## 9.1 Scope Limitations

The theory's scope conditions, as specified in Section 4.2, exclude failed states, jurisdictions in active conflict, and antecedent questions of commitment design and ratification. The exclusions are theoretically motivated: in failed states, the absence of institutional infrastructure makes the framework's five-dimensional mechanism inapplicable; in jurisdictions in active conflict, the disruption of institutional infrastructure produces a different implementation logic that warrants its own theory. The exclusions are also empirically significant: failed states and conflict-affected jurisdictions account for a substantial share of the implementation failure patterns in SDG 16 and the Sendai Framework, and the theory's inapplicability to these cases is a genuine limitation. Future research should extend the theory to these cases, perhaps through the addition of a sixth dimension capturing institutional disruption, or through a separate theoretical treatment of implementation in disrupted settings. Preliminary work by the Advocacy Unified Network on a sixth "Institutional disruption" dimension, applied to Yemen, Syria, and the Democratic Republic of the Congo in pilot mode, suggests that such an extension is feasible but requires a separate theoretical apparatus and is therefore deferred to subsequent work.

The exclusion of antecedent questions is also a limitation, although a principled one. The theory takes the existence of commitments as given and explains implementation; it does not explain why commitments are made or why states ratify them. These antecedent questions are properly the subject of theories of cooperation and institutional design, and the theory's value-added lies in addressing the residual question of implementation. But the boundary between antecedent and implementation questions is not always sharp: ratification decisions are themselves implementation-relevant, and the theory's inattention to this stage may obscure important dynamics. Future research should explore the integration of the Governance Gap theory with theories of ratification, perhaps through the extension of the Commitment specification dimension to encompass ratification-relevant specification.

## 9.2 Measurement Limitations and Validation Protocol

Table 4. Measurement-validation status and future empirical requirements.

| Validation component | Status in the 15 October 2023 article | Purpose of future work |
|---|---|---|
| Inter-coder reliability | Not conducted; pilot work is developmental only. | Establish coding reliability before confirmatory testing. |
| Latent-variable modeling / IRT | Not conducted; no confirmatory IRT result is claimed. | Test measurement structure and indicator performance. |
| Panel construction | Not constructed; dynamic propositions remain untested. | Test dynamic predictions such as path dependence. |

| Validation component | Status in the 15 October 2023 article | Purpose of future work |
|---|---|---|
| Pre-registration | Not undertaken for this article. | Pre-specify future confirmatory propositions, case selection, measurement, and analysis. |

The theory's measurement scheme, drawn from the August 2023 framework paper and refined in the appendix, is operationally tractable but measurement-theoretically modest. The use of binary or ordinal coding for the indicators of each dimension, while useful for advocacy applications, does not fully exploit the theory's continuous-variable formalism. The aggregation of indicators into dimension values through simple normalized means does not address the potential for non-linear aggregation or for the use of latent variable models. The coding of behavioral signatures for the implementation pathology typology is operationalized but has not been subjected to inter-coder reliability testing. These measurement limitations do not undermine the theory's qualitative fit in the case applications of Section 6 but do limit the theory's capacity for quantitative testing and comparison. To address this limitation substantively, we specify below a four-element validation protocol whose implementation is the principal item on our research agenda.

Element one: inter-coder reliability testing. The dimensional coding scheme should be administered by independent coders working from the published coding protocol, with each commitment-jurisdiction dyad coded independently where feasible. Inter-coder reliability should be assessed through Krippendorff's alpha (Krippendorff, 2019), with thresholds specified in advance. Any indicator showing inadequate agreement should be revised or dropped before confirmatory testing. The pilot work described by AUN is treated here as developmental evidence rather than as a completed validation result.

Element two: latent variable modeling. The five dimensions are specified in the theory as continuous latent variables operationalized through observed indicators. Item Response Theory (IRT), including generalized partial-credit models for ordinal indicators (Muraki, 1992), is proposed for future validation. Such models could assess indicator parameters, differential item functioning, uncertainty in dimension estimates, and indicator reduction. No IRT result is treated as confirmatory evidence in the present article.

Element three: panel construction. The theory specifies that G-values are properties of commitment-jurisdiction dyads measured over time. A future panel dataset should include multiple dyads observed across multiple periods and vary by instrument type, jurisdictional setting, and issue area. Such a panel would permit direct testing of Proposition 8 (path dependence) and other dynamic predictions. The panel-construction work described by AUN is treated as developmental rather than as a completed empirical validation.

Element four: pre-registration. Future tests of the theory's propositions should be pre-registered using the Open Science Framework (OSF) or an equivalent research-registration platform, with propositions, case-selection criteria, measurement protocols, and analysis plans specified ex ante. The present article was not formally pre-registered; its propositions should therefore be treated as theory-building claims requiring subsequent confirmatory testing.

## 9.3 Mechanism Limitations

The causal mechanisms elaborated in Section 5 are specified at the level of links in a causal chain, with each link being causally intelligible. But the chain is not fully micro-founded: the mechanisms operate at the level of duty-bearer institutions and do not address the individual-level mechanisms through which institutional behavior is produced. A more complete theory would specify the individual-level mechanisms, drawing on behavioral public administration (Grimmelikhuijsen, Jilke, Leth Olsen, and Tummers, 2017) and on the cognitive and motivational factors that shape the behavior of front-line implementers. Future research should extend the theory to the

individual level, perhaps through the integration of the Governance Gap theory with theories of street-level bureaucracy (Lipsky, 1980) and policy implementation at the front line. Future research should examine these micro-level mechanisms through process-tracing and front-line implementation studies.

A related limitation is the theory's inattention to the political mechanisms through which gap dimensions are produced and reproduced. The five dimensions are treated as exogenous features of the commitment-jurisdiction dyad, but they are themselves the product of political processes: commitment specification is shaped by negotiating dynamics; institutional capacity is shaped by long-run state-building; monitoring infrastructure is shaped by information politics; accountability architecture is shaped by the distribution of power; resource adequacy is shaped by budgetary politics. Future research should explore the political production of gap dimensions, perhaps through the integration of the Governance Gap theory with theories of international political economy and comparative political development.

## 9.4 Predictive and Formalization Limitations

The theory's propositions, as specified in Section 4.4, are testable claims about the relationships among gap dimensions and implementation outcomes. But the propositions are not formally derived from the axioms through a deductive calculus; they are derived through conceptual analysis and abductive inference from cases. The derivations are intended to be transparent and the propositions are intended to be falsifiable, but the derivations do not meet the standards of formal theory. Future research should formalize the theory through a more rigorous deductive structure, perhaps through the use of mechanism-based modeling (Hedstrom, 2005) or agent-based computational modeling (Cederman, 1997). The path-dependence proposition (P8) in particular invites formalization through dynamic modeling, as the trajectory of G over time is a function of initial conditions and intervention sequences that lend themselves to computational representation. Future research should test the path-dependence proposition through mechanism-based or agent-based modeling. No computational-model results are treated as confirmatory evidence in the present article.

## 9.5 Application, Generalization, and Out-of-Sample Limitations

The six case applications in Section 6, including the out-of-sample Chemical Weapons Convention case in Section 6.6, are selected for variation across issue area, instrument type, and jurisdictional scope, but they do not exhaust the relevant variation. Trade governance, financial governance, global health governance, and additional security governance cases beyond the CWC are not represented in the cases, and the theory's fit in these domains is incompletely tested. Within trade governance, the implementation of WTO dispute settlement rulings, regional trade agreement commitments, and bilateral investment treaty obligations raises distinctive questions about the M and A dimensions that the cases in Section 6 do not address. Within global health governance, the implementation of the International Health Regulations (IHR 2005) and the Framework Convention on Tobacco Control raises distinctive questions about the R dimension. Future research should extend the case applications to these domains, with attention to the theory's scope and limits in each. Future applications may extend the theory to WTO dispute-settlement compliance, International Health Regulations implementation, and other underrepresented domains.

A related limitation is the theory's inattention to variation across jurisdictional types. The cases in Section 6 include global, regional, and national jurisdictions, but the theory does not systematically address the variation in implementation patterns across consolidated democracies, hybrid regimes, electoral autocracies, and closed autocracies. The implementation pathology typology (Section 5.4) predicts variation in the dominant pathology across regime types: ritualistic compliance is more likely in jurisdictions with high M and low A, which characterizes many soft-authoritarian regimes; symbolic compliance is more likely in jurisdictions with high C and low I, which characterizes many low-capacity democracies; selective compliance is more likely in jurisdictions with high A and low M, which characterizes many consolidated democracies with limited monitoring of certain issue areas; deferred compliance is more likely in jurisdictions with high I and low R, which characterizes many middle-income countries with constrained budgets. These regime-type predictions are intuitive but have not been

systematically tested. Future research should extend the theory to regime-type variation, perhaps through the integration of the Governance Gap theory with the literature on authoritarian resilience (Gerschewski, 2013) and the politics of implementation in hybrid regimes.

### 9.6 Epistemic Limitations: Abductive Method and Circularity

The theory was developed through abductive inference from five cases, and those same cases are examined in Section 6. This creates a genuine risk of methodological circularity. The risk is not eliminated by the CWC plausibility probe; one external case can increase plausibility but cannot establish general validity. The article therefore does not describe the six cases as a confirmatory test. Instead, it treats the first five as theory-development applications and the CWC as an external plausibility probe. Independent confirmatory work must use cases selected without reference to the theory, preregistered coding rules, and evidence capable of measuring the explanatory dimensions separately from the outcome.

A related epistemic limitation is the absence of formal model selection. The theory's five-dimensional structure was chosen on the basis of abductive inference from cases, but alternative structures, four-dimensional, six-dimensional, or with different dimension definitions, were not formally compared. Model selection criteria, including the Akaike Information Criterion and the Bayesian Information Criterion (Burnham and Anderson, 2002), could in principle be applied to the choice among dimensional structures, but their application requires the quantitative operationalization of the theory that is itself an item on the research agenda. The present article's claim that the five-dimensional structure is appropriate is therefore a theoretical claim, justified by the abductive fit, rather than a formal claim, justified by model selection. Future research should address this limitation through the quantitative model-selection work that the measurement validation protocol of Section 9.2 enables.

### 9.7 Research Agenda

These limitations, taken together, suggest a research agenda for the further development of the Governance Gap theory. The agenda comprises six elements. First, the development of a measurement scheme that fully exploits the theory's continuous-variable formalism, through latent variable modeling, inter-coder reliability testing, panel construction, and pre-registration. Second, the extension of the case applications to underrepresented domains, particularly trade governance, financial governance, global health governance, and additional security governance cases beyond the CWC. Third, the formalization of the theory through mechanism-based and agent-based computational modeling, with particular attention to the path-dependence proposition. Fourth, the extension of the theory to failed states and conflict-affected jurisdictions, through the addition of dimensions or through a separate theoretical treatment. Fifth, the integration of the theory with theories of individual-level implementation behavior and with theories of the political production of gap dimensions. Sixth, the conduct of multiple external plausibility probes on cases selected without reference to the theory, with pre-registered propositions and measurement protocols. The Advocacy Unified Network and the authors of this article intend to pursue this agenda through a multi-year program of empirical and conceptual research, and we invite the broader scholarly community to join in the development of the theory's potential.

## 10. Conclusion

This article develops the Governance Gap from a policy diagnostic into a formally specified theory-building framework for international implementation research. It defines the Governance Gap as a behavioral distance between normative requirements and observed realization, treats C, I, M, A, and R as explanatory conditions rather than components of the outcome, specifies eight testable propositions, and develops a behavioral typology of ritualistic, selective, deferred, and symbolic compliance. The five principal cases are used transparently as theory-development applications, while the Chemical Weapons Convention is used as an external plausibility probe.

The contribution is therefore best understood as a rigorously specified and replication-ready theory-building research program rather than a completed empirical theory. The article now separates outcome measurement from explanatory coding, specifies an element-level distance function and evidence hierarchy, distinguishes theory-development applications from external plausibility testing, and provides explicit falsification and sensitivity rules. The cases provide qualitative consistency with the proposed mechanisms, but they do not establish causal identification, inter-coder reliability, predictive accuracy, or comparative superiority. These limits are not weaknesses hidden from the reader; they define the exact empirical tests required for the next stage of the research program.

The substantive implication is nevertheless important. International implementation failures may arise from different configurations of specification, capacity, monitoring, accountability, and resources, and these configurations may generate distinguishable behavioral signatures. The Governance Gap framework offers a common vocabulary and a set of falsifiable hypotheses for studying those configurations. Its ultimate value will depend on whether independent data and cases confirm, modify, or refute the proposed relationships.

# 11. Transparency Statement, Data Availability, and Acknowledgments

## 11.1 Authorship and Institutional Position

This article is authored by the Advocacy Unified Network (AUN), which developed the GGAM framework in August 2023. The authors therefore have an institutional and intellectual interest in the framework's theoretical extension. That interest is disclosed because it may affect interpretation of the article's claims. The authors commit to reporting null, negative, and theory-revising results from future tests. No claim in this article should be read as having been independently validated solely by virtue of institutional authorship.

## 11.2 Methodological Transparency

The methodology combines conceptual analysis with abductive theory development. The five principal cases were used during theory construction and are therefore designated theory-development cases; they are not independent confirmatory cases. The Chemical Weapons Convention was selected only after the propositions were fixed and is designated an external plausibility probe. No formal preregistration was undertaken for the present article. Future confirmatory tests should preregister propositions, case-selection rules, measurement protocols, evidence thresholds, and analysis plans before outcome data are examined.

The case applications in Section 6 are based on publicly available sources cited in the reference list and on the Advocacy Unified Network's pilot applications of the August 2023 framework. The pilot applications were conducted between January and June 2023 in three national contexts (Bangladesh, Kenya, and Peru) on six commitment-jurisdiction dyads (the Aarhus Convention in Bangladesh; the Convention against Torture in Bangladesh; the Convention on the Rights of the Child in Kenya; the SDG 16 targets in Kenya; the Escazu Agreement in Peru; and the American Convention on Human Rights in Peru). Coding from these pilots informed the dimensional operationalization but did not directly inform the propositions, which were derived from the broader literature and from the five in-sample cases. The pilot coding data are available from the corresponding author on reasonable request.

## 11.3 Data and Code Availability

The dimensional coding protocol, indicator specifications, and pilot coding materials described in Section 11.2 are available from the corresponding author on reasonable request. The case-specific evidence for the six applications is drawn from the publicly available sources cited in the reference list. Computational modeling and IRT

estimation are identified as future validation work; no unpublished computational or inferential result is presented as decisive evidence for the theory in this article.

### 11.4 Role of the Funding Source

The work reported in this article was supported by the Advocacy Unified Network's core funding, provided through a mix of foundation grants, government development-cooperation contributions, and membership dues. No specific grant was received for the work reported here; the work was conducted as part of AUN's core research program. The funding sources had no role in study design, data collection, analysis, interpretation, writing, or the decision to submit. The authors declare no financial conflicts of interest; the institutional interest in developing the GGAM framework is disclosed in Section 11.1. The corresponding author affirms that all authors approved the final manuscript and that all persons making substantial contributions are appropriately credited.

### 11.5 Acknowledgments

The authors acknowledge the contributions of the AUN's New York Liaison Office research team, particularly the participants in the August 2023 framework paper development workshop, whose work on the five-dimensional GGAM architecture provided the foundation for the theoretical extension advanced in this article. The authors acknowledge the contributions of the three external reviewers who provided comments on an earlier draft of this article, whose suggestions substantially improved the formalization of the propositions, the specification of the external plausibility probe, and the measurement validation protocol. The authors acknowledge the contributions of the pilot application partners in Bangladesh (BRAC University School of Law), Kenya (Kenya Human Rights Commission), and Peru (Derecho, Ambiente y Recursos Naturales), whose coding work informed the dimensional operationalization. The views expressed in this article are those of the authors and do not necessarily reflect the views of the partner organizations. Errors remain the responsibility of the authors.

## Appendix A. Operationalization and Falsification Protocol

This appendix converts the revised theory into a protocol suitable for independent replication. The key design principle is separation of explanatory coding from outcome coding. A coder must never infer C, I, M, A, or R from the observed implementation outcome alone; each dimension must be supported by evidence about the condition itself. Likewise, G must be coded from behavioral realization evidence rather than reconstructed from the five dimension scores.

### A1. Coding domains

| Construct | Unit | 0 score | 1 score | Evidence rule |
|---|---|---|---|---|
| C — Commitment specification | Commitment × jurisdiction × time | Obligation, scope, addressee, and compliance threshold are operationally determinable. | Key compliance-relevant behavior remains materially indeterminate. | Use primary legal/instrument text and authoritative interpretation; do not infer C from implementation performance. |

| Construct | Unit | 0 score | 1 score | Evidence rule |
|---|---|---|---|---|
| I — Institutional capacity | Jurisdiction × commitment × time | Relevant legal authority, administrative competence, personnel, and coordination are demonstrably available. | One or more essential institutional links are absent or non-functional. | Use institutional documents, staffing/organizational evidence, legislation, and implementation records. |
| M — Monitoring infrastructure | Jurisdiction × commitment × time | Relevant implementation behavior is systematically observable with usable indicators/data. | Material implementation behavior cannot be reliably observed. | Code indicators, collection systems, verification, and reporting separately. |
| A — Accountability architecture | Jurisdiction × commitment × time | Detected non-implementation can trigger credible review/correction/consequences. | Non-implementation lacks a credible route to consequence or correction. | Code legal, administrative, political, judicial, and reputational mechanisms separately. |
| R — Resource adequacy | Jurisdiction × commitment × time | Minimum financial, human, and technical inputs are available for required tasks. | Essential implementation tasks cannot be sustained because inputs are inadequate. | Use budgets, staffing, infrastructure, donor/resource records, and task-level requirements. |
| Y — Behavioral realization | Jurisdiction × commitment × time | Coded requirements are substantively realized in observed behavior. | Coded requirements are not substantively realized. | Use outcome/behavior evidence independently of C/I/M/A/R evidence. |
| G — Governance Gap | Jurisdiction × commitment × time | Distance function D(N,Y)=0 under the published rubric. | Distance function D(N,Y)=1 under the published rubric. | Calculate only after normative requirements and behavioral realization are separately coded. |

## A2. Falsification rules

A proposition is not counted as supported merely because a case can be described in its terms after the outcome is known. For confirmatory work, the following rules apply: (1) coding instructions and expected behavioral signatures must be frozen before outcome coding; (2) each proposition must specify at least one observation that would count against it; (3) hybrid, indeterminate, and negative cases must be permitted; (4) alternative explanations must be recorded; (5) a rival model must be given the same evidentiary standard; (6) missing evidence must be coded as missing, never as non-performance; and (7) sensitivity analyses must be reported for

plausible alternative coding decisions. Failure to meet these conditions converts the result to theory-development evidence rather than confirmatory evidence.

# Appendix B. Proposition-by-Case Evidence Status

The matrix below records what the present article can and cannot establish. "Consistent" means that the observed pattern does not contradict the proposition; it does not mean that the proposition has been causally confirmed.

| Proposition | Observable implication | Present cases | Current status | What would falsify it | Required next test |
|---|---|---|---|---|---|
| P1 | Independent dimension gaps correlate with lower Y / higher G. | All six | Illustrative consistency; not estimated. | No association or systematic reverse association after independent measurement. | Panel or cross-sectional dyad dataset with separate outcome coding. |
| P2 | Relative importance varies by instrument/context. | Paris, SDG16, Sendai, EU Directive, CWC | Conceptually plausible; not estimated. | No systematic heterogeneity in predictive contribution. | Pre-specified interaction/importance comparison. |
| P3 | Thresholds exist. | SDG16 / Sendai / CWC narratives | Mechanism illustration only. | No threshold improvement over simpler models. | Threshold or spline models with preregistered cut-points/model comparison. |
| P4 | Thresholds vary by context. | Cross-case comparison | Not tested. | Threshold estimates invariant across contexts. | Multilevel or stratified threshold analysis. |
| P5 | Configurations generate distinct pathologies. | All six | Qualitatively consistent; classification remains non-confirmatory. | Systematic counterexamples under frozen coding rules. | Independent coders + negative/deviant cases. |
| P6 | Pathologies are behaviorally distinguishable. | All six | Typology specified; reliability not tested. | Low inter-coder agreement or indistinguishable behavioral signatures. | Double-coding and reliability analysis. |

| Proposition | Observable implication | Present cases | Current status | What would falsify it | Required next test |
|---|---|---|---|---|---|
| P7 | Binding constraints make single-dimension interventions less effective. | No intervention dataset | Untested. | Single-dimension interventions outperform matched configuration interventions. | Quasi-experimental/intervention comparison. |
| P8 | Implementation trajectories are path dependent. | No longitudinal dataset | Untested. | Initial configuration does not predict trajectory after controls. | Longitudinal dyad panel with preregistered dynamic model. |

Reference-integrity note: bibliographic records were normalized for known errors identified in the forensic audit. Historical empirical claims are treated as time-stamped claims: where a statistic or institutional status can change, the manuscript specifies the relevant period and source rather than presenting a later status as evidence for an earlier claim. The manuscript does not use a changing indicator-tier classification as a timeless proxy for monitoring capacity, and it distinguishes EU infringement stages rather than collapsing them into a single count.

# Appendix C. Reproducible Outcome-Measurement Protocol

This appendix makes the outcome construct reproducible. It is deliberately separate from Appendix A: Appendix A codes the explanatory conditions C/I/M/A/R, whereas this appendix codes the behavioral outcome Y and computes G. The protocol is the primary specification for future replication. Any deviation must be preregistered and reported as a sensitivity analysis rather than introduced after outcome inspection.

## C1. Normative requirement decomposition

Each commitment-jurisdiction-time observation begins by freezing the applicable version of the instrument and identifying a finite set of non-overlapping normative requirement units $N=\{n1,...,nm\}$. A requirement unit must correspond to a distinct compliance-relevant obligation, prohibition, procedural duty, reporting duty, or institutional duty. Duplicative indicators that measure the same requirement are nested within that requirement and cannot receive separate outcome weight. The normative benchmark is established from the primary instrument text and authoritative interpretation before behavioral evidence is coded.

## C2. Behavioral realization coding

For each requirement unit ni, behavioral realization yi is coded independently from evidence about C/I/M/A/R. $yi=1$ means the requirement is fully realized for the specified period; $yi=0.5$ means material but incomplete realization; $yi=0$ means non-realization. If available evidence is insufficient to distinguish these states, yi is coded missing rather than zero. The coder records the behavioral evidence supporting the score and the observation window.

## C3. Distance function

For each requirement unit, di=1-yi. The primary specification uses equal weights wi=1 for all non-overlapping requirement units, so G(K,J,T)=Σ(wi·di)/Σwi. Equal weighting is the default because it is transparent and does not allow the observed outcome to determine weights. Where an instrument itself establishes distinct categories of obligation, those categories may be analyzed separately; any alternative weighting scheme must be declared ex ante and reported as a sensitivity analysis. Thus G is a reproducible behavioral distance, not an average of the five GGAM dimensions.

## C4. Evidence hierarchy and independence

Primary evidence receives priority: treaty or instrument text, authoritative decisions or interpretations, official implementation records, administrative data, judicial decisions, audit reports, and verified monitoring records. Secondary scholarly or civil-society sources may corroborate or fill documented gaps. Evidence used to code C/I/M/A/R must not be treated as evidence of Y merely because it describes the same institution. A source may support both layers only where it contains separately identifiable evidence for the condition and the behavior. The coding sheet must record the exact proposition supported by each source.

## C5. Time, versioning, and change

Every observation is time-stamped. If a commitment, indicator definition, implementing law, or monitoring protocol changes during T, the observation period is segmented or the applicable version is frozen before coding. Historical claims must use the source and institutional status applicable to the period studied. Current classifications must never be back-projected into earlier periods without an explicit historical source.

## C6. Missingness, uncertainty, and sensitivity

Missing evidence is not evidence of failure. Where a requirement cannot be coded confidently, the primary estimate is reported with the requirement omitted from the denominator and the number of missing units disclosed; a conservative sensitivity analysis may instead treat missing units as non-realized. If the substantive conclusion changes under reasonable missing-data assumptions, the case is reported as indeterminate. Confidence is recorded separately from the G score and does not alter the score unless the preregistered protocol says otherwise.

## C7. Replication and audit trail

A replication package should contain the frozen instrument version, requirement inventory, coding sheet, source identifiers, evidence excerpts, coder ID, date of coding, adjudication record, and the calculation file that produces G. Independent replication must reproduce the result from those materials without access to the authors' narrative interpretation. Any post hoc recoding must preserve the original score and be reported as a separate version.

## C8. Construct-separation rule

The theory is empirically testable only if the outcome can vary independently of the explanatory profile. A case may therefore have high C/I/M/A/R gaps with a lower-than-expected G, or relatively favorable conditions with a high G. If the coding process mechanically derives G from C/I/M/A/R, the observation is invalid for theory testing and may be used only as descriptive framework material. This rule is the principal safeguard against circularity.

## C9. Claim-status rule

The manuscript uses four evidence statuses: Conceptual (definition or logical implication); Illustrative (supported by theory-development cases); External plausibility (supported by the post-theory CWC case without confirmatory independence); and Confirmatory (requiring preregistered, independently selected data and a prespecified

analysis). Only the fourth status permits language such as confirmed, estimated, or predictive superiority. The present article contains no Confirmatory result.

**Table C1. Primary behavioral-realization scoring states**

| Score | Behavioral state | Decision rule | Evidence requirement | Interpretive status |
|---|---|---|---|---|
| 1.0 | Full realization | All material elements of ni are realized during T. | Direct behavioral or outcome evidence. | Counts toward G=0 for ni. |
| 0.5 | Partial realization | Material implementation exists, but one or more material elements remain unrealized. | Direct evidence identifies both realization and the remaining deficit. | Intermediate distance. |
| 0.0 | Non-realization | The requirement is not substantively realized during T. | Direct evidence of non-performance or authoritative finding. | Counts toward G=1 for ni. |
| NA | Insufficient evidence | Available evidence cannot distinguish 1, 0.5, and 0. | Documented evidence gap; never inferred from C/I/M/A/R. | Excluded from primary denominator; sensitivity disclosed. |

Production note: the reference list is the authoritative bibliographic record for this article. Time-sensitive empirical statements in the main text are versioned to the historical period studied; current institutional pages are not used to retroactively alter the October 2023 case evidence.

# 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.

Advocacy Unified Network. (2023). The Governance Gap Assessment Matrix: A five-dimensional framework for diagnosing international implementation failure. AUN Working Paper Series, No. 2023-08. Geneva: Advocacy Unified Network.

Alter, K. J., & Meunier, S. (2009). The politics of international regime complexity. Perspectives on Politics, 7(1), 13-24.

Andrews, M., Pritchett, L., & Woolcock, M. (2017). Building State Capability: Evidence, Analysis, Action. Oxford: Oxford University Press.

Bardach, E. (1977). The Implementation Game: What Happens After a Bill Becomes a Law. Cambridge, MA: MIT Press.

Biermann, F., & Pattberg, P. (Eds.). (2008). Global Governance: Theories and Empirical Research. Cambridge, MA: MIT Press.

Biermann, F., Pattberg, P., van Asselt, H., & Zelli, F. (2009). The fragmentation of global governance architectures: A review of the literature. Global Environmental Politics, 9(4), 14-40.

Bradley, C. A., & Kelley, J. G. (2008). The concept of international delegation. Law and Contemporary Problems, 71(1), 1-36.

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). New York: Springer.

Busch, M. L. (2007). Overlapping institutions, forum shopping, and dispute settlement in international trade. International Organization, 61(4), 735-761.

Cederman, L.-E. (1997). Emergent Actors in World Politics: How States and Nations Develop and Dissolve. Princeton, NJ: Princeton University Press.

Chayes, A., & Chayes, A. (1995). The New Sovereignty: Compliance with International Regulatory Agreements. Cambridge, MA: Harvard University Press.

Checkel, J. T. (2005). International institutions and socialization in Europe. International Organization, 59(4), 801-826.

Climate Policy Initiative. (2022). Global Landscape of Climate Finance 2022. San Francisco: CPI.

Downs, G. W., Rocke, D. M., & Barsoom, P. N. (1996). Is the good news about compliance good news about cooperation? International Organization, 50(3), 379-406.

Eckstein, H. (1975). Case study and theory in political science. In F. I. Greenstein & N. W. Polsby (Eds.), Handbook of Political Science, Vol. 7 (pp. 79-137). Reading, MA: Addison-Wesley.

European Commission. (2023). July 2023 Infringement Package: European Commission Decides to Refer Member States to the Court of Justice for Late Transposition of the Whistleblower Directive. Brussels: European Commission.

Finnemore, M., & Sikkink, K. (1998). International norm dynamics and political change. International Organization, 52(4), 887-917.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort the distribution of published effects? Sociological Methods and Research, 36(1), 3-30.

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

Goertz, G. (2006). Social Science Concepts: A User's Guide. Princeton, NJ: Princeton University Press.

Goldstein, J. L., Kahler, M., Keohane, R. O., & Slaughter, A.-M. (Eds.). (2001). Legalization and World Politics. Cambridge, MA: MIT Press.

Grimmelikhuijsen, S., Jilke, S., Leth Olsen, A., & Tummers, L. (2017). Behavioral public administration: Combining insights from public administration and psychology. Public Administration Review, 77(1), 45-56.

Hathaway, O. A. (2002). Do human rights treaties work? Yale Law Journal, 111(8), 1935-2042.

Hawkins, D. G., Lake, D. A., Nielson, D. L., & Tierney, M. J. (Eds.). (2006). Delegation and Agency in International Organizations. Cambridge: Cambridge University Press.

Hedstrom, P. (2005). Dissecting the Social: On the Principles of Analytical Sociology. Cambridge: Cambridge University Press.

Hedstrom, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. Annual Review of Sociology, 36, 49-67.

Helfer, L. R. (2008). The new innovation frontier? Intellectual property and the European Court of Human Rights. Harvard International Law Journal, 49(1), 1-52.

Hempel, C. G. (1942). The function of general laws in history. Journal of Philosophy, 39(2), 35-48.

Honadle, B. W. (1981). A capacity-building framework: A search for concept and purpose. Public Administration Review, 41(5), 575-580.

Johnston, A. I. (2001). Treating international institutions as social environments. International Studies Review, 5(4), 487-512.

Keohane, R. O. (1984). After Hegemony: Cooperation and Discord in the World Political Economy. Princeton, NJ: Princeton University Press.

Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. Perspectives on Politics, 9(1), 7-23.

Kellett, J., & Caravani, A. (2013). Financing Disaster Risk Reduction: A 20-Year Story of International Efforts. London: Overseas Development Institute.

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

Krippendorff, K. (2019). Content Analysis: An Introduction to Its Methodology (4th ed.). Thousand Oaks, CA: SAGE.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), Criticism and the Growth of Knowledge (pp. 91-196). Cambridge: Cambridge University Press.

Lipsky, M. (1980). Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. New York: Russell Sage Foundation.

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency. PS: Political Science and Politics, 47(1), 19-21.

Mahoney, J., & Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. Political Analysis, 14(3), 227-249.

Matland, R. E. (1995). Synthesizing the implementation literature: The ambiguity-conflict model of policy implementation. Journal of Public Administration Research and Theory, 5(2), 145-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-176.

Oberthur, S., & Gehring, T. (Eds.). (2006). Institutional Interaction in Global Environmental Governance: Synergy and Conflict. Cambridge, MA: MIT Press.

Organisation for the Prohibition of Chemical Weapons. (2022). Annual Report of the OPCW on the Implementation of the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction in 2021. The Hague: OPCW.

Organisation for the Prohibition of Chemical Weapons. (2023). OPCW confirms: All declared chemical weapons stockpiles verified as irreversibly destroyed. The Hague: OPCW, 7 July 2023.

O'Toole, L. J., Jr. (2000). Research on policy implementation: Assessment and prospects. Journal of Public Administration Research and Theory, 10(2), 263-288.

Painter, M., & Pierre, J. (Eds.). (2005). Challenges to State Policy Capacity: Global Trends and Comparative Perspectives. Basingstoke: Palgrave Macmillan.

Peirce, C. S. (1903). Pragmatism: The normative doctrine of logic. In C. Hartshorne & P. Weiss (Eds.), Collected Papers of Charles Sanders Peirce, Vol. 5 (pp. 141-153). Cambridge, MA: Harvard University Press.

Pressman, J. L., & Wildavsky, A. (1973). Implementation: How Great Expectations in Washington Are Dashed in Oakland. Berkeley: University of California Press.

Raustiala, K., & Victor, D. G. (2004). The regime complex for plant genetic resources. International Organization, 58(2), 277-323.

Reichertz, J. (2007). Abduction: The logic of discovery of grounded theory. In A. Bryant & K. Charmaz (Eds.), The SAGE handbook of grounded theory (pp. 214-228). London: SAGE.

Sabatier, P. A., & Mazmanian, D. A. (1979). The conditions of effective implementation: A guide to designing policy. Policy Analysis, 5(4), 481-504.

Salmon, W. C. (1998). Causality and Explanation. Oxford: Oxford University Press.

Simmons, B. A. (2010). Treaty compliance and violation. International Studies Review, 12(1), 56-78.

United Nations. (2022). The Sustainable Development Goals Report 2022. New York: United Nations Department of Economic and Social Affairs.

United Nations Environment Programme. (2022). Emissions Gap Report 2022: The Closing Window. Nairobi: UNEP.

United Nations Office for Disaster Risk Reduction. (2021). Sendai Framework Monitor: 2020-2021 Implementation Report. Geneva: UNDRR.

UPR Info. (2014). Beyond Promises: The Implementation of the Universal Periodic Review Recommendations. Geneva: UPR Info.

Victor, D. G., Raustiala, K., & Skolnikoff, E. B. (Eds.). (1998). The Implementation and Effectiveness of International Commitments: Theory and Practice. Cambridge, MA: MIT Press.

von Stein, J. (2016). Making promises, keeping promises: Democracy, ratification, and compliance with international law. Journal of Conflict Resolution, 60(5), 944-971.

Winter, S. C. (2003). Implementation perspectives: Status and reconsideration. International Encyclopedia of Social and Behavioral Sciences, 11, 7185-7190.

Wu, X., Ramesh, M., & Howlett, M. (2015). Policy capacity: A conceptual framework for understanding capacity in policy-making. Public Management Review, 17(7), 919-935.