

Who Watches the Watchers? Governance Agency in AI Verification and Monitoring Regimes

By
**Parveen Sheikh &
Meenakshi Neeharika Mangipudi**
With Apart Research
Regional track: **Asia**

Abstract

Artificial intelligence is advancing at a pace that could reshape the trajectory of our shared future, for better or for worse. Alongside its promise, it carries catastrophic risks through both misuse and misalignment, and it tends to concentrate power in the hands of a few. In response, the world has begun to debate an "IAEA for AI", a body to monitor and verify advanced AI. But this debate has focused almost entirely on what to monitor and how, and has skipped the prior political question of who holds verification authority, who can contest it, and who carries its burden. This paper argues that verification is not a neutral technical instrument but a structure that distributes power. We bring an interdisciplinary lens to the question, combining a quantitative analysis of the Alva Myrdal Centre database of 128 arms-control agreements (1817 to 2021) with a qualitative framework we call the Governance Agency. We find that direct verification is rare, that recourse is underdeveloped, and that monitoring authority and compliance burden rise together while the power to challenge does not. Reading these patterns through history and power, and through a comparison of the NPT and the CWC, we show that asymmetric verification is not only unjust but unstable. We argue that an inclusive verification regime, one that gives middle powers and the Global South real authority rather than tokenism, is necessary if humanity is to govern AI together.

1. Introduction

Artificial intelligence is advancing rapidly, and its growth has the capacity to change the trajectory of our future for better or for worse. Alongside its benefits, advanced AI carries catastrophic risks through both misuse and misalignment, and it concentrates power in ways that few technologies before it have. The danger is not only that something goes wrong, but that a small group could come to hold durable control over the lives of everyone else [1]. Because these risks fall on collective humanity, the institutions we build to govern them are a matter for collective humanity, and not only for the two states where frontier AI is being built.

In response to these risks, proposals for international AI monitoring have multiplied. Many reach for the same model: an agency in the mould of the IAEA that would monitor and verify compliance with rules on advanced AI [2]. Recent technical work has begun to map how such verification might actually work [3]. But almost all of this conversation asks what should be monitored and how. It rarely asks the question that comes first, which is a political one: who designs the monitor, who gets to inspect, who can contest a finding, and on whose terms. This matters because power in AI is already concentrated, with frontier capability and compute held by a very small number of actors [4]. A monitoring regime built under these conditions will inherit that imbalance unless it is deliberately designed not to.

This paper reframes the question. The issue is not whether to monitor AI, but what kind of monitoring we build, and who holds real authority within it. To answer this, we look to the one domain that has two centuries of evidence on how verification regimes are actually designed: international arms control. We pair a quantitative analysis of 128 arms-control agreements from the Alva Myrdal Centre database with a qualitative framework called Governance Agency.

Our empirical analysis reveals three core patterns. First, direct verification is rare: verified compliance mechanisms appear in only 26.6% of agreements, while consultation mechanisms appear in 50.8% [10][28]. Second, recourse mechanisms are consistently underdeveloped [10][19]. Third, verification architecture varies across institutional settings: bilateral agreements exhibit higher Access Scores (1.35 vs. 0.72) and Burden Scores (2.33 vs. 1.49) than multilateral ones, while multilateral agreements show higher Recourse Scores (1.61 vs. 0.94) [10][28]. Most significantly, monitoring authority and compliance burdens are strongly correlated ($r = 0.808$), but monitoring authority and recourse are weakly correlated ($r = 0.359$) [10]. Oversight capacity and compliance obligations evolve together, while opportunities for challenge do not.

Read through our Governance Agency framework—Access (the right to inspect), Recourse (the right to challenge), and Burden (the obligation to comply)—these patterns reveal a recurring logic. Access concentrates where authority already sits. Recourse remains underdeveloped. Burden falls unevenly. This is the architecture of systems built by the powerful for their own assurance [20]. As Foucault observed, power operates through the asymmetry of the gaze itself: the observer defines the standard against which the observed is judged [5].

To understand this asymmetry, we compare two treaties from the same era [11][12]. The NPT divides the world into nuclear-weapon states and non-nuclear-weapon states, placing inspection on the latter while exempting the former's arsenals [11]. The CWC chose a symmetric approach, subjecting all parties to the same inspection regime and allowing any member to request a challenge inspection [2][12]. The NPT is Burden without Authority; the CWC shows a symmetric design was always possible. Asymmetry is a design choice, not an inevitability.

Asymmetric verification is also unstable, because it makes defection rational. India declined the NPT on the ground that it institutionalised a discriminatory order, developed nuclear weapons, and is now treated as a de facto nuclear power [11]. The precedent is corrosive: capability bought India more standing than compliance would have. But the nuclear past and the AI present diverge. A nuclear weapon is a sovereign

possession; a state can defect, acquire it, and be more secure. AI's catastrophic risks do not respect borders [3]. No state can defect its way to safety from a collective risk. The cooperation imperative survives, indeed intensifies.

Our main contributions are, first, the Governance Agency framework, which evaluates participation in verification systems across three quantitative dimensions, Access, Recourse, and Burden, and two qualitative dimensions that the data cannot capture, legitimacy and exit. Second, an empirical analysis of how historical verification regimes have distributed authority and obligation. Third, an application of this framework to the AI monitoring proposals now being designed, which shows that without deliberate intervention they are likely to repeat a familiar and unstable pattern. The arms-control dataset is existing work; the framework, the coding and analysis, the comparison of the NPT and the CWC, the qualitative extension, and the entire mapping onto AI governance are our contributions during this hackathon.

2.Related Work

Arms-control verification. A substantial literature examines how international agreements verify compliance through inspection, monitoring, and reporting [2, 10]. Studies of the nuclear and chemical-weapons regimes analyse how these mechanisms were designed and how reliably they detect violations. This work treats verification primarily as a problem of effectiveness and security: can a regime catch a violator in time. It less often asks who holds verification authority, who is subject to it, and whether that distribution is equitable. Even close institutional analysis tends to focus on the difficulty of enforcement that the IAEA inspects but cannot itself punish, referring cases upward to states and the UN Security Council [2]. The question of representation within verification design remains largely unexamined.

AI verification and monitoring. A newer body of work asks whether comparable verification can be constructed for AI [2, 3]. It proposes compute accounting, hardware-level features, model auditing, and personnel-based measures, and it debates technical feasibility. This research is candid that verification carries risks: a leading technical overview cautions that such mechanisms "require guardrails to protect against abuse and power concentration" [3]. But the question of who would hold that power, and who would be excluded from it, remains secondary to whether the mechanisms can be built.

Global South, middle powers, and inequality. A third strand argues that AI governance excludes most of the world. Middle powers struggle to influence frontier development, and the global-governance route has, on one influential account, largely faltered [7]. Most states face deep dependence on US and Chinese compute, models, and infrastructure, leaving the open question of "on which nations, to what extent and on what terms" [4]. And the harms of technology tend to track existing inequalities, falling hardest on already-marginalised groups [8]. This literature names exclusion clearly but discusses it in broad terms. It does not connect exclusion to the specific architecture of verification regimes.

The gap. These three literatures rarely meet. Arms-control scholarship studies verification but not representation. AI-governance work studies feasibility but not authority. Global South scholarship studies exclusion but not verification design. No existing work uses historical verification data, read through a structural account of global power, to ask whether emerging AI-monitoring regimes are positioned to reproduce the representational asymmetries of past arms control. This paper addresses that gap.

3. Methods

3.1 Research Design

This study adopts an exploratory mixed-methods research design combining quantitative analysis of historical arms-control agreements with qualitative interpretation of their implications for emerging AI monitoring and verification regimes. The primary objective is not to establish causal relationships between treaty design and governance outcomes, but rather to identify recurring patterns in verification architectures and assess how these patterns may inform contemporary debates surrounding AI governance, international monitoring institutions, and representation within emerging verification systems [22][23].

The research is motivated by the observation that many current proposals for AI monitoring, compute governance, frontier-model oversight, and international AI verification institutions focus primarily on technical questions of monitoring and compliance while devoting comparatively less attention to questions of participation, recourse, and governance representation [13]. Historical arms-control agreements provide a useful empirical foundation for examining these issues because they contain a diverse range of verification, consultation, reporting, and compliance mechanisms developed across different geopolitical contexts over more than two centuries [30].

This study therefore proceeds in three stages. First, historical verification architectures are examined using the Arms Control Agreements Database developed by the Alva Myrdal Centre for Nuclear Disarmament (AMC). Second, a Governance Agency Framework is developed to operationalize key dimensions of participation within verification systems. Third, empirical findings are interpreted in relation to contemporary discussions concerning AI monitoring institutions, international verification proposals, and the governance challenges facing middle powers and Global South actors [29].

3.2 Data Source and Dataset Selection

The empirical component of this study utilizes the Arms Control Agreements Database developed by the Alva Myrdal Centre for Nuclear Disarmament (AMC), Uppsala University. The database contains systematically coded information on 128 arms-control agreements spanning the period from 1817 to 2021 and was designed to facilitate comparative analysis of compliance, verification, consultation, and monitoring arrangements embedded within international agreements. Four datasets from the AMC repository were selected based on their direct relevance to the research question [10]. The Agreement Information dataset provides treaty-level metadata including agreement year, participation characteristics, consultation provisions, and compliance-related indicators. Because this dataset contains one observation per agreement, it serves as the primary analytical foundation and facilitates comparison across treaties. The Verified Compliance dataset captures formal verification mechanisms embedded within agreements, including inspection authorities, access provisions, monitoring arrangements, and trigger conditions. These variables are particularly important because verification authority constitutes a central component of monitoring architectures and directly influences how oversight is operationalized within international governance systems. The Consultation dataset records consultation and referral mechanisms through which treaty parties may seek clarification, raise concerns, initiate reviews, or engage in dispute-resolution processes. Because meaningful participation within governance systems often depends upon the availability of channels through which actors can challenge or contest monitoring outcomes, consultation mechanisms provide an important proxy for institutional recourse. Finally, the Demonstrated Compliance dataset contains information regarding reporting requirements, notification procedures, compliance submissions, and related administrative obligations. These variables provide insight into how monitoring systems distribute compliance responsibilities and institutional burdens across participating actors. Other datasets contained within the AMC repository, including weapons-classification and treaty-association datasets, were excluded because they focus primarily on substantive treaty content rather than

governance and verification structures. Restricting the analysis to datasets directly relevant to monitoring, consultation, and compliance improves conceptual consistency and ensures alignment with the study's analytical objectives [28].

3.3 Governance Agency Framework

To examine verification architectures systematically, this study develops a Governance Agency Framework. Governance Agency is defined as the capacity of actors to participate meaningfully within verification systems through access to monitoring processes, institutional recourse mechanisms, and engagement with compliance structures.

The framework is informed by scholarship on international governance, institutional accountability, treaty compliance, and verification systems, which emphasizes that effective governance arrangements depend not only on monitoring capacity but also on opportunities for participation, consultation, review, and procedural accountability [20]. Verification systems are not merely technical mechanisms for detecting compliance; they are institutional arrangements that shape who can observe, who can challenge, and who bears responsibility for demonstrating compliance. This reading is sharpened by Foucault's account of surveillance, in which power is exercised through the asymmetry of the gaze itself, so that the observer defines the standard against which the observed is judged and an inspection clause is never merely technical but constitutive of authority [5].

Accordingly, the framework focuses on three dimensions that can be consistently observed across historical agreements and operationalized using available treaty-level data.

Access Agency refers to the degree to which monitoring institutions possess authority to obtain information and conduct verification activities. Verification systems depend upon access to facilities, locations, materials, and relevant information in order to assess compliance and monitor obligations. Consequently, access arrangements represent one of the most fundamental components of verification architecture [15]. Within this study, Access Agency is operationalized using four verification variables obtained from the Verified Compliance dataset: area access, facility access, item access, and item-section access. Collectively, these variables capture the breadth of inspection authority granted under a given agreement [10].

Recourse Agency refers to the availability of institutional mechanisms through which actors may challenge findings, request clarification, initiate consultations, seek mediation, or pursue independent review. Prior research on compliance and governance suggests that monitoring systems lacking meaningful avenues for consultation and dispute resolution may concentrate authority within oversight institutions while limiting opportunities for affected actors to contest decisions [16]. This dimension is operationalized using consultation mechanisms, referral provisions, and third-party referral pathways extracted from the Consultation dataset. These variables provide a proxy for the extent to which verification systems incorporate formal opportunities for dialogue, review, and challenge.

Burden Agency captures the extent to which participating actors are subject to compliance obligations, reporting requirements, notification procedures, and administrative responsibilities. Although such mechanisms are necessary for effective monitoring, they may also impose significant institutional and resource burdens upon participating entities [17]. Accordingly, Burden Agency is operationalized using demonstrated compliance mechanisms, reporting requirements, notification provisions, and compliance-submission structures recorded within the Demonstrated Compliance dataset.

Table 1. Governance Agency Framework

Framework Dimension	Operational Definition	Variables Used
Access Agency	Degree of monitoring and inspection access	Area Access, Facility Access, Item Access, Item-Section Access
Recourse Agency	Ability to challenge, consult, or seek review	Consultation Count, Referral Count
Burden Agency	Compliance and reporting obligations imposed upon participants	Compliance Mechanisms, Verified Compliance Requirements

3.4 Data Preparation and Variable Construction

The four datasets were linked using the agreement identifier variable, which serves as a common key across all AMC datasets [10]. Because the Agreement Information dataset contains one record per agreement while the remaining datasets contain multiple records per agreement, an aggregation procedure was required to facilitate agreement-level analysis. Mechanism-level variables were therefore aggregated by agreement identifier and subsequently merged into a unified analytical dataset. This approach enables direct comparison across agreements while preserving information concerning verification intensity, consultation structures, and compliance obligations.

Because the AMC datasets record individual verification, consultation, and compliance mechanisms rather than agreement-level governance characteristics, aggregation was necessary to facilitate systematic comparison across agreements. Composite indicators were therefore constructed to summarize the relative intensity of monitoring authority, recourse opportunities, and compliance obligations within each agreement. Similar aggregation approaches are widely employed in institutional and governance research when multiple observable mechanisms contribute to a broader conceptual construct [18].

To operationalize the Governance Agency Framework, three composite indices were constructed based on standard index-building techniques in treaty-design scholarship [28]:

The Access Score was calculated as the sum of area access, facility access, item access, and item-section access provisions recorded for each agreement:

Access Score = Area Access + Facility Access + Item Access + Item-Section Access

This measure serves as a proxy for the breadth of inspection authority embedded within a verification architecture, mirroring established metrics for monitoring intrusiveness [15].

The Recourse Score was calculated as the sum of consultation mechanisms and referral provisions:

Recourse Score = Consultation Count + Referral Count

This indicator reflects the extent to which actors possess formal opportunities to initiate consultations, challenge findings, or seek institutional review, capturing the legal flexibility and dispute-resolution capacity of the regime [19].

The Burden Score was calculated using the total number of demonstrated compliance mechanisms combined with verified compliance requirements:

Burden Score = Compliance Mechanisms + Verified Compliance Requirements

This measure provides a proxy for the administrative, reporting, and logistical obligations imposed upon participating actors [17].

These indicators are not intended to represent definitive measures of governance quality or institutional effectiveness. Rather, in line with exploratory mixed-methods practices, they function as analytical proxies that facilitate cross-case comparison and enable systematic examination of different dimensions of governance agency within historical verification architectures.

3.5 Analytical Strategy

The empirical analysis was conducted using Python within the Google Colab cloud computing environment [27]. Data processing, aggregation, and variable construction were performed using the Pandas library [24], while statistical analysis and exploratory data visualization were conducted using Matplotlib [25] and Seaborn [26].

In accordance with standard frameworks for exploratory and convergent mixed-methods research designs, the analysis proceeded in five distinct stages [22]:

First, descriptive statistics were used to evaluate the prevalence of consultation mechanisms, demonstrated compliance mechanisms, and verified compliance mechanisms across the agreement corpus. This stage provides an overview of the relative prominence of different governance instruments within historical verification systems.

Second, comparative analysis was conducted between bilateral and multilateral agreements. Bilateral agreements often operate under different institutional constraints than multilateral arrangements and may therefore exhibit distinct verification and consultation structures. Comparing these agreement types enables exploration of how governance architectures vary across different participation models [19].

Third, temporal analysis was conducted by grouping agreements according to decades of adoption. This approach allows examination of how verification architectures have evolved over time and whether governance mechanisms have become increasingly institutionalized, complex, or differentiated across historical periods [21].

Fourth, correlation analysis was conducted to examine relationships among Access Agency, Recourse Agency, and Burden Agency. Understanding how these dimensions interact provides insight into whether monitoring authority, opportunities for challenge, and compliance obligations tend to develop together or evolve independently within verification systems.

Finally, a qualitative case illustration was conducted using two prominent international verification regimes: the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) and the Chemical Weapons Convention (CWC) [11][12]. These agreements were selected because they represent widely recognized approaches to international verification and provide an opportunity to illustrate how monitoring authority, consultation mechanisms, and compliance obligations may be configured differently across verification architectures.

The purpose of this comparison is not to draw causal conclusions but to contextualize the quantitative findings and demonstrate how governance agency dimensions manifest within specific institutional arrangements. Collectively, these analyses provide an empirical foundation for assessing how historical verification architectures distribute monitoring authority, recourse mechanisms, and compliance obligations, and how these patterns may inform ongoing discussions concerning AI monitoring, international verification institutions, and governance participation.

4. Results

4.1 Prevalence of Governance Mechanisms Across Agreements

The first stage of analysis evaluated the structural distribution of institutional instruments across the 128 arms-control agreements contained within the AMC database, focusing specifically on consultation mechanisms, demonstrated compliance mechanisms, and verified compliance mechanisms. The empirical results reveal substantial variation in adoption rates across these categories: consultation mechanisms represent the most prevalent governance tool, appearing in 50.8% of agreements, followed closely by demonstrated compliance mechanisms at 45.3%. In contrast, verified compliance mechanisms are considerably rarer, appearing in only 26.6% of the treaty corpus as shown in Figure 1. This distribution demonstrates that historical arms-control architectures have fundamentally relied more heavily on diplomatic consultation and administrative reporting structures than on intrusive, direct verification regimes. Formal oversight systems involving onsite inspections or active technical monitoring represent a highly specialized tool of treaty design rather than a default or universal institutional standard. This relative scarcity of verified compliance mechanisms challenges the central, default role often attributed to objective verification within contemporary international monitoring debates. From an institutional governance perspective, these patterns indicate that state actors frequently design international frameworks to prioritize diplomatic coordination, information-sharing channels, and voluntary reporting over high-intensity monitoring architectures. By establishing that direct verification is historically the exception rather than the rule, this empirical baseline highlights a critical tension within verification design. It raises vital questions regarding how oversight authority is distributed, how technical monitoring capacity intersects with institutional resource burdens, and ultimately, how these structural patterns might inform equitable participation and recourse within emerging AI governance frameworks.

	Metric	Percentage
0	Consultation Mechanism Present	50.78125
1	Verified Compliance Present	26.56250
2	Demonstrated Compliance Present	45.31250

Figure 1. Percentage of agreements containing consultation mechanisms, demonstrated compliance mechanisms, and verified compliance mechanisms.

4.2 Comparative Analysis: Bilateral and Multilateral Agreements

To examine how governance architectures vary across institutional arrangements, agreements were partitioned into bilateral and multilateral categories and evaluated across the three core dimensions of the Governance Agency Framework: Access Agency, Recourse Agency, and Burden Agency. The analysis reveals stark differences in how these two agreement types balance institutional authority and state obligations. Bilateral agreements exhibit significantly higher average Access Scores (1.35) than their multilateral counterparts (0.72), and similarly display higher average Burden Scores (2.33) compared to

multilateral arrangements (1.49). Conversely, multilateral agreements reverse this trend entirely by demonstrating substantially higher average Recourse Scores (1.61) than bilateral treaties (0.94). These diverging metrics reveal that verification architectures are far from uniform across international settings; rather, the configuration of monitoring authority, administrative obligations, and consultative structures varies fundamentally according to the scale and diversity of participating actors as shown in Figure 2. This statistical divergence suggests that bilateral and multilateral regimes operate under entirely different institutional logics. While bilateral architectures prioritize direct physical oversight, intensive monitoring access, and heavy compliance burdens to manage localized rivalries, multilateral regimes place their primary structural emphasis on diplomatic coordination, broad consultation, and inclusive procedural engagement. This structural distinction proves that verification systems do not scale linearly; instead, they require distinct trade-offs between centralized oversight capacity and participatory governance. An expansion in monitoring authority does not naturally coincide with robust opportunities for institutional recourse, nor do strong consultation channels imply widespread monitoring access. Recognizing these historical trade-offs is highly instructive for contemporary policy, as it demonstrates that emerging verification regimes must intentionally choose whether to optimize for deep technical compliance or broad, legitimate participation.

	laterality	agreements	avg_access	avg_recourse	avg_burden
0	Bilateral	54	1.351852	0.944444	2.333333
1	Multilateral	74	0.716216	1.608108	1.486486

Figure 2. Average Access, Recourse, and Burden Scores across bilateral and multilateral agreements.

4.3 Evolution of Verification Architectures Over Time

The final stage of analysis evaluated how verification architectures have shifted chronologically by grouping agreements by their decade of adoption and tracking them across the Access, Recourse, and Burden dimensions of the Governance Agency Framework as shown in Figure 3. The results indicate that international verification architectures underwent a major structural transition following the mid-twentieth century. While earlier nineteenth- and early twentieth-century agreements generally relied on narrow compliance criteria with minimal monitoring access, a clear inflection point emerged in the 1960s. Moving through the 1990s, agreements increasingly incorporated dense, multi-layered governance instruments, including onsite inspection rights, formalized consultative bodies, and mandatory administrative reporting. The 1990s in particular marked the historical zenith of this trend, witnessing the emergence of highly comprehensive governance architectures characterized by simultaneous peaks in monitoring access, institutional recourse pathways, and explicit compliance burdens. This specific decade-level surge aligns with the broader post-Cold War expansion of global institutionalization and the rising demand for intrusive multilateral transparency. This long-term trajectory suggests that international verification regimes have steadily evolved from simple, ad-hoc state arrangements focused on narrow legal constraints into sophisticated, multi-dimensional governance systems. This historical expansion reflects a gradual but clear institutionalization of global oversight, where treaties are systematically engineered to integrate overlapping mechanisms for monitoring, state-level transparency, and diplomatic coordination. However, because several early historical periods contain a relatively low volume of treaties, these decade-level fluctuations must be

interpreted with caution. These metrics are most appropriately viewed as indicative of macro-historical shifts toward complex regulatory design rather than linear, lockstep temporal trends. Collectively, this temporal analysis demonstrates that as security regimes matured over time, they increasingly moved away from singular technical checks, choosing instead to build diverse governance architectures capable of balancing monitoring capacity with formal consultation and reporting.

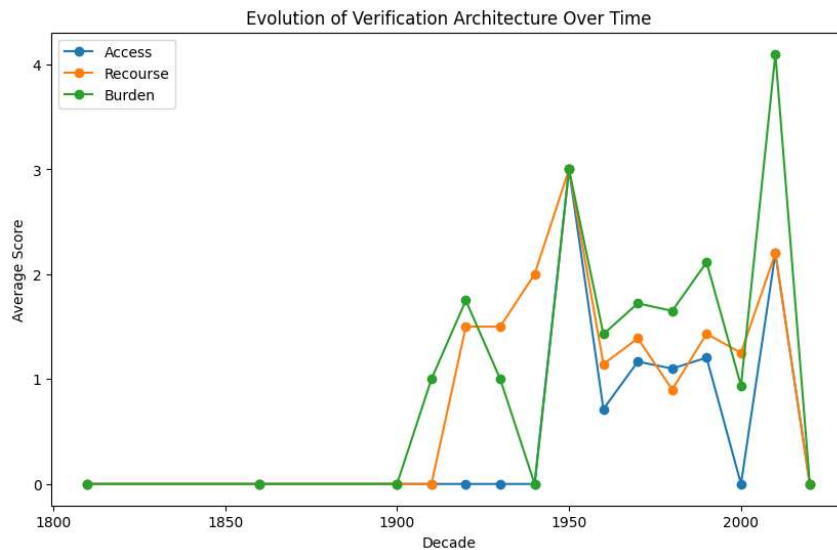


Figure 3. Average Access, Recourse, and Burden Scores across historical decades.

4.4 Relationships Between Governance Agency Dimensions

To examine how different structural components of governance agency interact within verification architectures, a Pearson correlation analysis was conducted using the constructed Access, Recourse, and Burden indicators. The statistical results reveal that the strongest relationship exists between Access Agency and Burden Agency ($r = 0.808$, $p < 0.001$). This robust positive association indicates that agreements granting greater monitoring and inspection authority also systematically impose more extensive administrative and reporting obligations, demonstrating that monitoring access and compliance burdens develop in tandem within international verification systems. Additionally, a moderate positive relationship was observed between Recourse Agency and Burden Agency ($r = 0.450$, $p < 0.001$), suggesting that agreements with heavier compliance architectures frequently incorporate supplemental consultation and review mechanisms to handle the resulting institutional complexity. By contrast, Access Agency and Recourse Agency exhibit a much weaker positive relationship ($r = 0.359$, $p < 0.001$). While agreements with stronger monitoring architectures do tend to include basic consultation mechanisms, this link is remarkably weaker than the bond observed between monitoring access and operational burdens as shown in Figure 4 and Figure 5.

This statistical divergence is highly significant because it proves that monitoring authority and formal opportunities for institutional challenge do not necessarily evolve together. Historical verification architectures have consistently strengthened oversight capacity and compliance obligations far more aggressively than they have expanded channels for consultation, review, or procedural contestation. From a global governance perspective, this pattern reveals a structural asymmetry: verification systems are frequently optimized to extract compliance data and monitor behavior without a proportionate expansion in opportunities for legitimate participation and recourse. These institutional imbalances heavily

concentrate authority within monitoring bodies while limiting the agency of monitored actors. For contemporary policy, this finding serves as a stark warning that unless explicitly designed otherwise, emerging regimes—including international AI verification frameworks—will naturally tend to build deep technical compliance structures while marginalizing mechanisms for equitable participation and institutional recourse.

```
access_score vs recourse_score
Correlation = 0.359
P-value = 0.0000
-----
access_score vs burden_score
Correlation = 0.808
P-value = 0.0000
-----
recourse_score vs burden_score
Correlation = 0.450
P-value = 0.0000
-----
```

Figure 4: Correlation Significance Between Governance Agency Dimensions

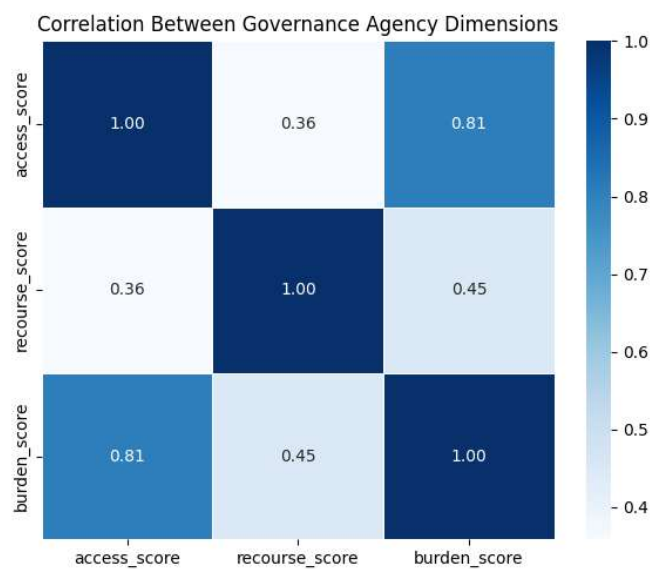


Figure 5. Correlation matrix illustrating relationships between Access Agency, Recourse Agency, and Burden Agency.

To complement the broader quantitative analysis, a comparative case illustration was conducted using two of the most prominent international verification regimes: the 1968 Treaty on the Non-Proliferation of Nuclear Weapons (NPT) and the 1993 Chemical Weapons Convention (CWC), as detailed in Table 2. While these agreements represent only two historical cases within the broader database, they provide highly instructive examples of how verification architectures can be configured around entirely distinct institutional configurations of monitoring authority, recourse mechanisms, and compliance obligations.

Agreement	Year	Access Score	Recourse Score	Burden Score
NPT	1968	0	3	0
CWC	1993	6	4	4

Table 2. Governance Agency Comparison: NPT and CWC

The comparison reveals substantial structural divergence between the two regimes. Within the AMC coding framework, the NPT exhibits relatively strong recourse mechanisms but limited treaty-level monitoring access and direct compliance obligations. In stark contrast, the CWC demonstrates a maximized structural footprint, featuring a significantly higher Access Score (6), stronger compliance burdens (4), and expanded consultation mechanisms (4). These differences illustrate, as shown in Figure 6, how verification architectures embody distinct governance logics. While the NPT reflects a mid-twentieth-century approach relying heavily on externalized agency (delegating technical safeguards to the IAEA), the CWC reflects the highly institutionalized, intrusive multilateral standard of the 1990s, built with its own dedicated verification organ (the OPCW).

Importantly, this case-level comparison reinforces the broader quantitative findings of this study by proving that verification systems vary widely in how they balance monitoring capacity against compliance overhead. Verification should therefore not be conceptualized as a static, singular institutional model, but rather as a diverse spectrum of governance arrangements that actively distribute authority, administrative friction, and procedural participation in fundamentally different ways.

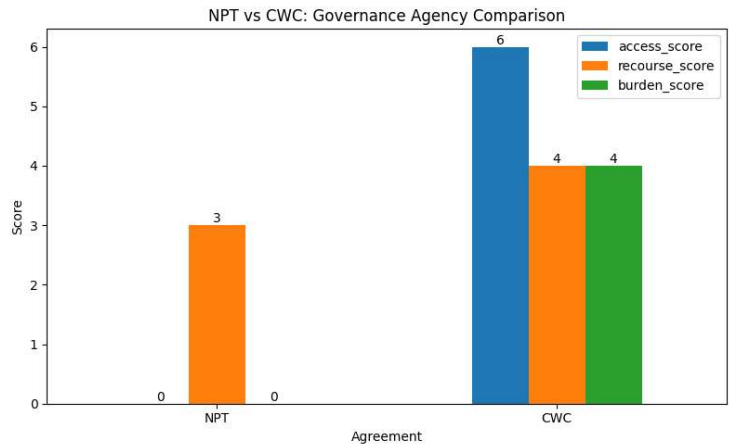


Figure 6. Comparison of Access Agency, Recourse Agency, and Burden Agency between the NPT and CWC.

Summary of Findings. Five major findings emerge from the analysis. First, consultation mechanisms are substantially more prevalent than formal verification mechanisms across historical arms-control agreements. While approximately half of all agreements contain consultation provisions, fewer than one-third contain verified compliance mechanisms, suggesting that direct verification remains relatively uncommon within historical governance architectures. Second, bilateral and multilateral agreements exhibit distinct governance structures, with bilateral agreements placing greater emphasis on monitoring access and compliance obligations, whereas multilateral agreements incorporate stronger consultation and recourse mechanisms. Third, verification architectures appear to have become increasingly complex and institutionalized over time, particularly during the latter half of the twentieth century, with agreements increasingly combining monitoring, consultation, and compliance mechanisms rather than relying on a single governance approach. Fourth, strong positive relationships exist between monitoring authority and compliance burdens, while the relationship between monitoring authority and opportunities for consultation remains comparatively weaker. This finding suggests that oversight capacity and compliance obligations tend to evolve together more consistently than participatory mechanisms. Finally, the illustrative comparison between the NPT and the CWC demonstrates that verification architectures can differ substantially in how they allocate monitoring authority, recourse opportunities, and compliance responsibilities. Collectively, these findings indicate that verification systems are not institutionally uniform and that historical agreements have adopted diverse approaches to balancing oversight authority, consultation mechanisms, and compliance obligations within governance arrangements.

5. Discussion

5.1 What the Patterns Suggest

Our analysis of the AMC database reveals three core patterns that together suggest how historical verification architecture functioned [10]. First, direct verification mechanisms are less common in treaties. Across 128 arms-control agreements (1817-2021), verified compliance mechanisms appear in only 26.6% of cases, while consultation mechanisms appear in 50.8% and reporting-style compliance in 45.3% [10][28]. This suggests that historically, the international community preferred consultation and reporting over direct verification. Formal verification systems involving inspections, monitoring access, or other verification procedures appear considerably less common than governance arrangements centered on communication and compliance reporting [15].

Second, recourse mechanisms, the channels through which a party can contest a finding such as formal referral and review mechanisms, are relatively underdeveloped across the corpus [10]. While consultation mechanisms appear in nearly half of all agreements, formal referral and review provisions are considerably rarer [19]. This is important to note because recourse is a significant part of any agreement. Without it, an agreement is just a one-way surveillance apparatus. Without recourse, weaker stakeholders are effectively placed at the mercy of the verifying authority [16]. Recourse gives stakeholders a voice, and that voice enriches the governance structure.

Third, the architecture of verification is not uniform across institutional settings. As per our analysis, bilateral agreements exhibit higher Access Scores (1.35 vs. 0.72) and Burden Scores (2.33 vs. 1.49) than multilateral ones, while multilateral agreements show higher Recourse Scores (1.61 vs. 0.94) [10][28].

This suggests that the number of participants fundamentally dictates how the verification architecture is distributed [18]. Bilateral agreements, often negotiated between roughly equal powers, tend to embed stronger monitoring access and compliance obligations. Multilateral agreements, by contrast, involve more diverse parties and consequently develop more elaborate consultation and review mechanisms, but at the cost of weaker direct verification [19].

Read through the Governance Agency lens, these patterns reveal a recurring logic. Access, the power to inspect, monitor, and obtain information, tends to concentrate where authority already sits. Recourse, the power to challenge and contest, remains the least-developed dimension across the entire corpus. Burden, the obligation to report, comply, and be monitored, falls unevenly [17]. This is precisely the architecture we would expect of systems built by the powerful for their own assurance, not systems designed to distribute governance authority equitably [20].

From a governance perspective, this finding is significant because verification institutions do more than assess compliance; they shape who can access information, who can challenge decisions, and who bears the costs of compliance [16][20]. The distinction observed between bilateral and multilateral agreements further reinforces this point. Bilateral agreements tended to emphasise monitoring access and compliance obligations, whereas multilateral agreements relied more heavily on consultation and recourse mechanisms [19]. Together, these findings indicate that verification architectures can embody different governance logics, balancing oversight authority and participation in distinct ways.

5.2 What the Bilateral-Multilateral Split Tells Us About AI

The structural finding is very consequential because it explains what kind of choice we have to make for AI governance [10][28]. Bilateral agreements concentrate Access and Burden: two parties, often rivals of comparable strength, inspect each other intrusively and accept heavy mutual obligations, but they keep the arrangement closed. Multilateral agreements invert this: as participation broadens, Recourse rises because many parties demand channels to raise concerns, yet Access shrinks because hard inspection is harder to agree among many unequal states [18][19]. In short, broadening the table tends to buy voice at the cost of inspection power.

This matters for AI because the emerging proposals are, almost without exception, multilateral in aspiration. An "IAEA for AI" is imagined as a near-universal body [2]. Our data suggests what such breadth typically produces: relatively strong consultation and relatively weak, reserved inspection, with real oversight power likely retained by the few capable of exercising it [10]. The implication is double-edged. Broad participation may indeed yield more recourse mechanisms than a closed great-power arrangement would, which is a point in favour of inclusive design. But broad participation does not, on its own, distribute Access. A Global South state could be granted a seat, a voice, and a complaint procedure, while the capacity to actually inspect, verify, and define compliance remains concentrated where compute and intelligence already sit [4]. Inclusion in participation is not the same as inclusion in authority, and our data shows these can be, and historically often are, pulled apart.

5.3 The Governance Agency Framework: Why These Patterns Are About Power

We interpret these findings through what we call Governance Agency: the capacity of a state to participate meaningfully in a verification system, not merely to be subject to it [20]. The framework separates that capacity into Access (the right to see and inspect), Recourse (the right to challenge), and Burden (the obligation to report and submit). The reason this separation matters is that the three can be distributed independently, and the gap between them is where power lives [5].

Let us try to understand power through the work of Michel Foucault. Michel Foucault's analysis of Jeremy Bentham's Panopticon prison design offers a precise vocabulary for what our data describes [5]. Foucault argues that the Panopticon's effectiveness rests on a single design principle: "power should be visible and unverifiable" [5]. In the case of the Panopticon, this meant that an inmate must always know that they are being watched all the time, whether it is true or not. Even if they are not being watched, they succumb to self-discipline in the observed. Power, on this account, operates through the asymmetry of the gaze itself, prior to any specific act of inspection, and the party who controls visibility holds the more fundamental position [5].

Applied to verification regimes, the Panopticon's lesson is that the question "who can inspect" is never merely technical. It is a question about who occupies the tower and who occupies the cell [5]. A treaty's inspection clause does not simply detect violations; it installs a relationship in which one party is structured as visible and accountable, and the other as the position from which visibility is exercised. Our Access and Recourse dimensions can be read as an attempt to measure exactly this: Access asks who occupies the tower, Recourse asks whether the cell has any means of answering back.

5.4 Study of NPT Versus CWC

To understand the patterns, let us look at two arms control treaties of the same era facing the same structurally similar problems but with fundamentally different approaches [11][12]. The Nuclear Non-Proliferation Treaty divides the world by design into nuclear-weapon states and non-nuclear-weapon states [11]. Inspection falls on the latter, whose facilities are subject to IAEA safeguards, while the arsenals of the former are not held to the same intrusive verification [2]. The disarmament obligation that was meant to balance this bargain by the nuclear states, Article VI, has been weakly fulfilled, a grievance the Global South has voiced for decades [30]. In our framework, the NPT is the archetype of Burden without Authority. The many accept the gaze and carry the reporting load, while the few retain both their weapons and their oversight position.

The Chemical Weapons Convention, facing the comparable task of verifying a banned weapons class, chose a different approach [12]. It prohibits the entire category for everyone, subjects all parties to the same inspection regime under the OPCW, distributes Recourse by allowing any member to request a challenge inspection that can launch within 12 hours of notification [2][12], and structures its executive council around rotating regional representation. It is, by design, far more symmetric in how it allocates Access, Recourse, and Burden.

The historical injustice of the NPT, that nuclear states exempted their own arsenals while non-nuclear states accepted intrusive inspection, is invisible to a coding scheme that counts mechanisms rather than asking on whom they fall [30]. A purely quantitative reading cannot see the asymmetry that matters most.

This is why we extend the Governance Agency with two qualitative dimensions that the textual coding cannot capture but the historical record can [28]. The first is legitimacy: does a regime's authority rest on consent or on coercion. The second is exit: can a state refuse or leave without losing everything. These are not new data; they are interpretive lenses applied to cases. Together they let us read what the numbers omit, not merely whether a mechanism exists, but whether those subject to it had any meaningful say in, or escape from, its terms.

The key takeaway from the comparative analysis of both treaties of the same era with similar problems is that asymmetry is a design choice [11][12][30]. The deeper problem with this pattern is not only that it is unjust; it is that it is unstable, because it makes defection rational for the very states it seeks to bind [16]. When the powerful exempt themselves from verification while imposing the full onus of inspection on the weak, and when the capability being monitored is itself a source of sovereignty and standing, the regime

teaches the excluded that security lies in acquiring the capability, not in complying with a regime that offers them no authority.

India is the clearest case. India declined the NPT on the explicit ground that it institutionalised a discriminatory order, and subsequently developed nuclear weapons; it is now treated as a de facto nuclear power [11]. Through our framework, India exercised both exit, it could refuse without collapse, and a legitimacy verdict, it named the regime as unjust. The precedent is corrosive: capability bought India more standing than compliance would have.

This diagnosis must not be mistaken for an endorsement of proliferation. Here the nuclear past and the AI present diverge in a way that resolves the tension. A nuclear weapon is a sovereign possession; a state can defect, acquire it, and be more secure. AI's catastrophic risks, especially a misaligned AI, do not respect the borders of the state that builds them [3]. No state can defect its way to safety from a collective risk. The very feature that made defection rational under the NPT is absent for AI. Asymmetric verification will still breed the incentive to defect, and so will still fail as a security strategy, but now defection buys no safety for anyone. The cooperation imperative therefore survives, indeed intensifies, even amid the cynicism the historical record justifies.

5.5 Middle Powers: Real Leverage, Not Yet Authority

If cooperation is necessary, the question becomes whether those outside the great-power core can secure genuine authority within it. In the era of globalisation, "the widening, deepening and speeding up of global interconnectedness" means middle powers are no longer confined to the binary of aligning or resisting [6]. They can specialise, share, and hedge across an interdependent supply chain [4]. India's roughly \$1.25 billion sovereign-AI strategy and Vietnam's positioning as a regional innovation hub illustrate this widened space [4][33][34].

But we resist easy optimism. This leverage is real, and yet realness is not the same as power. A state may host data centres and so enter the AI supply chain without thereby gaining any authority over the models trained within them, and it may carry the monitorable burden while the design authority sits elsewhere [4]. This is the NPT asymmetry reassembled for the AI era. It is also why we read Leicht's prescription as insufficient rather than wrong [7]. He argues that middle powers, unable to shape frontier development, should pivot from global governance toward national interest [7]. We take his scepticism seriously: the path pursued from weakness has indeed failed. But to accept the retreat is to accept the asymmetry as fixed. We argue instead that the structure itself must change and that middle powers should hold real authority in the design of verification, not as the tokens of an inclusive-sounding regime but as parties with genuine access and recourse. Leicht himself concedes that, brought to strength, they might become "a real third center of gravity" [7]; our claim is that verification design is precisely where that strength could be built.

Countries such as India, Vietnam, and members of ASEAN are increasingly active participants in global AI governance discussions [32][33][34]. India's "Seven Sutras" framework emphasises innovation-oriented governance, while ASEAN has promoted cooperative approaches through the ASEAN Guide on AI Governance and Ethics [32]. Vietnam has recently introduced binding AI legislation, reflecting growing engagement with AI governance within Southeast Asia [34]. These developments demonstrate that middle powers are not passive recipients of AI governance; they are actively shaping national frameworks. However, participation in domestic governance does not automatically translate into authority within international verification regimes. The challenge is to convert national capacity into institutional authority within the design of global monitoring systems.

5.6 Applying the Framework to AI Proposals, and the Enforcement Ceiling

Turning Governance Agency on the leading technical proposal is clarifying. Baker and colleagues organise verification into six layers, from on-chip features to compute accounting to intelligence-based measures [3]. Read through our dimensions, the default distribution is familiar: access concentrates among states with intelligence and compute capacity, with the authors themselves noting that intelligence-based verification is not equally feasible for all states and would leave weaker ones depending on stronger ones [3]; recourse is largely absent, with no evident way for a state to contest a chip-level or intelligence-derived finding; and burden falls on whoever hosts the physical layer. The most sophisticated AI-verification proposal available reproduces, almost line for line, the asymmetry our historical reading describes, a concern its own authors anticipate [3].

The historical evidence suggests that the design of monitoring institutions also shapes governance participation and procedural recourse [16][20]. The strong relationship observed between monitoring authority and compliance obligations suggests that future AI monitoring systems may naturally evolve toward increasingly sophisticated oversight structures [10][28]. Yet the comparatively weaker relationship between monitoring authority and recourse mechanisms raises an important governance question: will future AI verification systems provide meaningful opportunities for challenge, review, and participation, or will oversight capabilities expand more rapidly than procedural safeguards [10]?

This concern is particularly relevant for proposals involving centralised AI monitoring bodies, frontier-model reporting requirements, compute governance mechanisms, and international verification arrangements [13][14][31]. Historical verification architectures demonstrate that monitoring authority and participatory governance do not necessarily develop together [10][19]. Consequently, future AI governance frameworks should consider not only how AI systems are monitored but also how affected actors can engage with monitoring institutions, contest decisions, and participate in governance processes [9].

Even symmetric design, however, meets a harder ceiling, and honesty requires naming it. Verification distributes the power to see, not the power to enforce. Both the IAEA and the OPCW verify but cannot punish, referring non-compliance to the Security Council, where enforcement “relies on the preferences of individual nations, especially global superpowers” [2]. A veto-wielding power can shield an ally or withdraw from an agreement it once championed. Any credible AI regime must reckon with this constraint rather than wish it away. The key takeaway is that inclusive verification design is necessary but not sufficient, and must be pursued alongside, not in place of, reform of how enforcement authority is held.

5.7 Governance Agency Risks

The Governance Agency framework provides a useful lens for identifying potential risks associated with future AI monitoring systems [10][20]. Three risks appear particularly relevant.

First, monitoring authority may become concentrated within a small number of institutions possessing technical expertise, computational resources, or regulatory authority [3][4]. Such concentration could reduce transparency regarding how monitoring decisions are made and limit opportunities for external scrutiny [16]. This risk is heightened in AI governance because frontier capabilities are already concentrated among a small number of actors [4].

Second, compliance obligations may disproportionately affect actors with fewer institutional resources [17]. As demonstrated by the strong relationship between monitoring authority and compliance burdens, increasingly sophisticated monitoring systems often require corresponding reporting and administrative

commitments [10][28]. Smaller organisations, developing economies, and resource-constrained actors may therefore face higher relative compliance costs [8], reproducing the pattern of Burden without Authority that characterised the NPT [11].

Third, recourse mechanisms may develop more slowly than monitoring capabilities [10]. The comparatively weaker relationship between monitoring authority and recourse mechanisms observed in the dataset suggests that systems designed primarily for oversight may not automatically incorporate equivalent opportunities for consultation, challenge, or review [10][19]. Over time, this imbalance could reduce procedural legitimacy and weaken stakeholder trust in governance institutions [16]. Collectively, these risks highlight the importance of designing AI monitoring systems that balance oversight effectiveness with participation, accountability, and procedural fairness [20].

6. Limitations

Several limitations should be acknowledged when evaluating the findings of this study.

- First and foremost, arms control agreements are not the perfect analogy for AI governance. Nuclear and chemical capabilities are physical and countable; AI's inputs are comparatively intangible, which is itself a central critique of the IAEA model. Historical regimes should be read as structural analogies, not blueprints. AI governance differs significantly in its underlying technological characteristics, breakneck pace of development, and deeply commercialised institutional environment.
 - Second, our sample size is modest, limited to 128 agreements, and fewer after disaggregation. Our analysis is descriptive and correlational, not causal. The observed relationships between governance dimensions must therefore be interpreted as structural patterns within the dataset rather than evidence of direct causal mechanisms linking monitoring authority, recourse systems, and compliance obligations.
 - Third, and most importantly, our data measure the presence of verification mechanisms at the level of treaty text, not their distribution between powerful and weaker parties. The Governance Agency Framework employs composite indicators that function as analytical proxies rather than fully validated, psychometric measures of governance quality or institutional effectiveness. While these additive indicators successfully facilitate systematic comparison across agreements, alternative operationalisations or variable weights may yield different results.
 - Fourth, the bilateral-multilateral distinction is a structural proxy, not a direct measure of the power differentials between specific parties. Our central claim about asymmetry therefore rests on qualitative and historical interpretation, which the quantitative coding informs but cannot establish. We have tried to mark clearly where the evidence ends and the interpretation begins.
 - Fifth, our analysis focuses strictly on the formal provisions and text codified within the agreements, rather than their real-world implementation. The legal existence of a consultation, verification, or compliance mechanism does not automatically imply its effective, fair, or operational execution in historical practice.
 - Sixth, while the study discusses critical implications for Global South participation and contemporary AI governance, the database itself does not directly measure state representation, demographic inclusion, or voting power. As a result, our conclusions regarding the structural pathways toward governance exclusion remain interpretive and theoretical rather than quantitatively established.
-

7. Dual-Use Considerations

Our central diagnosis, that asymmetric verification makes defection rational, could be misread as an argument for proliferation, or as a justification for states to refuse AI oversight on sovereignty grounds. We reject that reading explicitly. The argument is that asymmetry is a legitimacy problem to be designed against, not a licence to defect, and as we argue in 5.4, the collective and un-defectable nature of AI's catastrophic risks means that no state secures its own safety by opting out. A second risk is that the framework's language of burden and exclusion could be used by powerful actors to resist any verification at all. The mitigation is the paper's own thesis: the goal is not less verification but more symmetric and genuinely representative verification.

8. Future Work

Several avenues for future empirical and theoretical research emerge from this study.

The most important extensions of this work are, in our view, interdisciplinary rather than purely quantitative. The qualitative dimensions we introduce, legitimacy and exit, deserve to be developed into a fuller analytical method and applied across a wider set of cases than the two we examine here, so that the question of consent and coercion in verification can be studied systematically rather than illustratively. We also see value in re-coding the agreements by the involvement of major powers, which would let us test the distribution of authority more directly than the bilateral-multilateral split allows.

Beyond the historical record, the framework can be applied to the AI proposals being designed right now, from frontier-model oversight to compute governance, to ask of each one who holds access, who holds recourse, and who carries the burden. Future work should apply the Governance Agency Framework directly to contemporary, non-binding, and binding AI governance proposals. This includes conducting structural analyses of frontier-model oversight frameworks, hardware-level compute governance initiatives, national AI Security Institutes, and emerging international monitoring arrangements. Such research would enable a direct evaluation of how these newly minted systems distribute Access, Burden, and Recourse agency across state and non-state actors.

Future research can also expand the dataset by developing more comprehensive, weighted measures of governance participation. This involves explicitly encoding variables for institutional voting structures, veto distributions, geographic and demographic representation, and formal decision-making authority within international secretariats. This would allow for a more precise, quantitative assessment of structural equity and institutional inclusion.

A comparative analysis across multiple specialised governance regimes, including arms control, cybersecurity protocols, global climate governance, and international financial oversight, could help identify recurring institutional patterns. Mapping these domains would reveal whether certain structural trade-offs between rigorous technical monitoring and flexible participatory mechanisms are unique to AI or represent systemic traits of global technocratic governance.

As international governance shifts from abstract policy principles to concrete enforcement, the growing availability of empirical datasets, tracking global compute distribution, regulatory sandboxes, hardware export compliance, and model training metrics, presents a critical opportunity. Future studies can leverage these emerging datasets to move beyond historical analogies and analyse active, real-time AI governance institutions directly.

Two further questions are worth opening, though we cannot settle them with our current evidence. The first is whether the shape of verification in each era reflects the wider distribution of power in the world order of that time, which is a question our data hints at but cannot prove. The second is whether middle powers and the Global South could build genuine collective leverage in AI governance, in the spirit of earlier non-aligned cooperation, rather than choosing between the two dominant powers. Both deserve careful, evidence-led study.

9. Conclusion

History offers us both a warning and a choice. Across two centuries of arms control, our analysis shows that verification has been rare, that recourse has been scarce, and that authority has been distributed by the structure of regimes which powerful states designed for their own assurance. Our Governance Agency framework, which extends the quantitative coding with the qualitative dimensions of legitimacy and exit, lets us see what the numbers on their own cannot. The injustice of asymmetric verification does not lie in whether mechanisms exist on paper, but in who designs them, who is bound by them, and who is allowed to refuse. Read through this interdisciplinary lens, the NPT stands as a pattern that is at once unjust and unstable, because it teaches the states it excludes that real security lies in capability rather than in compliance. The CWC reminds us that a more symmetric design was always possible.

For AI, the stakes are higher and the logic has changed. These risks fall on all of us, on collective humanity, on our planet, and on the species we share it with, and no state can defect its way to safety from a risk that is genuinely collective. This is why we read the historical record not as a reason for despair but as a call. An IAEA-style regime may be a reasonable place to begin, but only if it is built to be symmetric, and only if it brings middle powers and the Global South to the table as parties with real authority and not as tokens of an inclusive-sounding arrangement. The question facing AI governance is not whether to watch, but who watches the watchers, and whether those who carry the burden of being watched will have any genuine say in the watching. We have to answer this together, and we have to answer it now.

10. Code and Data

Data: Alva Myrdal Centre Arms Control Agreements Database, Uppsala University [10] (publicly available). **Code:**

https://colab.research.google.com/drive/1AsIyIYZ_McDrCkelohHZE5QOps1ZYYLO?usp=sharing

11. Author Contributions

M.M. led the quantitative workstream: operationalizing the framework dimensions into coded variables, constructing the composite indices, and running and interpreting the statistical analysis. P.S. conceived and led the project, developed the qualitative framework, conducted the historical and political analysis, the NPT–CWC comparison, the AI-governance mapping, and led the writing. Both authors reviewed and approved the final manuscript.

12. LLM Usage Statement

We used an AI assistant (Claude) for literature-summary support, editorial structuring, and drafting assistance with some sections. All arguments, the analytical framework, and the conclusions are the authors' own; all sources, quotations, and statistics were independently verified against primary sources by the authors.

13. References

- [1] Longview Philanthropy, "Request for Proposals on Extreme Power Concentration," 2025.
- [2] A. Wasil, "Do We Want an 'IAEA for AI'?", *Lawfare*, 20 Nov 2024.
- [3] M. Baker, G. Kulp, O. Marks, M. Brundage, L. Heim, "Verifying International Agreements on AI: Six Layers of Verification," arXiv:2507.15916, 2025.
- [4] F. J. Varela Sandoval, I. Wilkinson, A. Krasodomski, R. Wilkinson, "How Middle Powers Can Weather US and Chinese AI Dominance: The Case for 'Sovereign AI' Strategies," Chatham House, Feb 2026. DOI: 10.55317/9781784136710.
- [5] M. Foucault, *Discipline and Punish: The Birth of the Prison*, trans. A. Sheridan, New York: Vintage Books, 1977 (p. 201).
- [6] D. Held, A. McGrew, D. Goldblatt, J. Perraton, *Global Transformations: Politics, Economics and Culture*, Cambridge: Polity Press, 1999.
- [7] A. Leicht, "How AI Safety Is Getting Middle Powers Wrong," 2026.
- [8] Amnesty International, *Digitally Divided: Technology, Inequality, and Human Rights*, 2023.
- [9] I. Puscas, *Confidence-Building Measures for Artificial Intelligence: A Multilateral Perspective*, UNIDIR, 2024.
- [10] Alva Myrdal Centre for Nuclear Disarmament, *Arms Control Agreements Database*, Uppsala University.
- [11] *Treaty on the Non-Proliferation of Nuclear Weapons* (NPT), 1968.
- [12] *Chemical Weapons Convention* (CWC), 1993.
- [13] H. Belfield, "Domestic frontier AI regulation, an IAEA for AI, an NPT for AI, and a US-led Allied Public-Private Partnership for AI," arXiv:2507.06379, 2025.
- [14] L. Ho, J. Barnhart, R. Trager, et al., "International Institutions for Advanced AI," arXiv:2307.04699, 2023.
- [15] R. Avenhaus and M. J. Canty, *Compliance Quantified: An Introduction to Data Verification*, Cambridge: Cambridge University Press, 2009.
- [16] R. O. Keohane, "Global Governance and Legitimacy," *Review of International Political Economy*, 18(1), 99–109, 2011.
- [17] D. G. Victor, K. Raustiala, E. B. Skolnikoff, *The Implementation and Effectiveness of International Environmental Commitments*, MIT Press, 1998.
- [18] B. Koremenos, C. Lipson, D. Snidal, "The Rational Design of International Institutions," *International Organization*, 55(4), 2001.
- [19] B. Koremenos, *The Continent of International Law*, Cambridge University Press, 2016.
- [20] T. G. Weiss, R. Wilkinson, *Global Governance Futures*, Routledge, 2021.
- [21] R. B. Mitchell et al., "What We Know (and Could Know) About International Environmental Agreements," *Global Environmental Politics*, 20(1), 103–121, 2020.
- [22] J. W. Creswell, V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, Sage, 2018.
- [23] R. B. Johnson, A. J. Onwuegbuzie, L. A. Turner, "Toward a Definition of Mixed Methods Research," *Journal of Mixed Methods Research*, 2007.
- [24] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python in Science*

Conf., 2010.

[25] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, 9(3), 2007.

[26] M. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, 6, 2021.

[27] E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 2019.

[28] R. Bauman, M. Riley, Y. Wang, D. P.-L. Miars, "Trust but Quantify: A Composite Scoring Approach to Verification Intensity Across Global Arms Control Treaties," Alva Myrdal Centre, Uppsala University, 2025.

[29] S. Grape et al., "Non-Proliferation and Safeguards Activities Within the Alva Myrdal Centre for Nuclear Disarmament," 2025.

[30] S. Herzog, "Dual-track Nuclear Disarmament: Institutional Design of the Treaty on the Prohibition of Nuclear Weapons," *The Nonproliferation Review*, 32(1–3), 55–73, 2025.

[31] A. Scher, D. Abecassis, P. Barnett, B. Abeyta, "An International Agreement to Prevent the Premature Creation of Artificial Superintelligence," arXiv:2511.10783, 2025.

[32] ASEAN, *ASEAN Guide on AI Governance and Ethics*, 2024.

[33] S. Ghosh, "India's US\$1.25 billion push to power AI," *Nature India*, 17 Mar 2024.

[34] L. Le, "Vietnam's New AI Law Balances Innovation Push With Tight State Control," *Tech Policy Press*, 2026.