

The Governability Threshold: When Algorithmic Systems Exceed the Capacity of Existing Regulatory Institutions

Yong DING

Weifang University of Science and Technology, Weifang, China · ddy2004@126.com

Abstract

This article introduces the concept of the “**governability threshold**”—the point beyond which intrinsic properties of algorithmic systems exceed institutional capacity for meaningful oversight, rendering conventional regulatory approaches structurally insufficient regardless of design quality. Drawing on governability theory, state capacity literature, and regulatory failure analysis, we construct a diagnostic framework organized around four system properties: determinism, explainability, agency, and scope. A Governability Threshold Matrix maps each property across governability configurations, while an Instrument Preconditions Matrix identifies the minimum system-property requirements for each conventional regulatory tool—revealing the conditions under which the entire instrument toolkit collapses. Analysis of three paradigmatic cases—high-frequency trading, large language models, and lethal autonomous weapon systems—demonstrates that certain system configurations produce institutional failure patterns persisting regardless of regulatory sophistication. Comparative analysis reveals that regulatory practice has independently converged on threshold logic, exemplified by the EU AI Act’s 10^{25} FLOPS compute threshold. Four governance adaptation strategies—prohibition, regulatory sandboxes, outcome-based liability, and adaptive regulation—are evaluated across jurisdictional responses.

Keywords governability threshold, institutional capacity, regulatory limits, algorithmic complexity, state capacity, ungovernability, policy instruments, autonomous systems

1 Introduction

1.1 The Assumption of Governability

Modern regulatory theory rests on a foundational assumption so deeply embedded that it is rarely examined: that the objects of governance are, in principle, governable. The regulatory state, as [Majone \(1997\)](#) influentially argued, emerged precisely because governments developed the institutional capacity to oversee complex economic and social activities through rule-making, standard-setting, monitoring, and enforcement. The history of regulation is, in significant part, a history of expanding governance capacity—from factory inspections in the nineteenth century to financial market oversight in the twentieth to data protection in the twenty-first.

Algorithmic governance scholarship inherits this assumption. The burgeoning literature on AI regulation—spanning risk-based classification ([European Parliament, 2024](#)), principles-based frameworks ([Jobin et al., 2019](#); [Hagendorff, 2020](#)), accountability architectures ([Mittelstadt, 2019](#)), and instrument design—uniformly presupposes that effective governance is achievable with appropriate institutional design. The question animating the field is consistently *how* to govern algorithmic systems, not *whether*

governance is possible. Even the most sophisticated treatments of regulatory failure (Coglianese, 2017; Black, 2001) frame failure as a correctable design flaw rather than a structural impossibility.

This article interrogates that assumption. It proposes that certain algorithmic systems have crossed a “governability threshold” beyond which existing regulatory institutions cannot adequately function—not because those institutions are poorly designed, but because the intrinsic properties of the systems themselves exceed what governance institutions can meaningfully oversee. The governability threshold is not a claim about regulatory incompetence or political will. It is a structural claim about the relationship between system characteristics and institutional capacity.

1.2 The Governability Question

Consider three concrete illustrations. High-frequency trading algorithms execute thousands of trades per millisecond—far faster than any human regulator can observe, comprehend, or intervene. The 2010 Flash Crash, in which the Dow Jones Industrial Average plunged nearly 1,000 points in minutes before partially recovering, demonstrated that algorithmic trading can generate systemic market consequences faster than any institutional response mechanism can activate (Kirilenko et al., 2017). Large language models with hundreds of billions of parameters generate outputs through processes that their own developers cannot fully explain; emergent capabilities—abilities not present in smaller models and not predictable from training objectives—appear at scale thresholds without being anticipated (Wei et al., 2022; Bommasani et al., 2021). Lethal autonomous weapon systems raise the prospect of targeting decisions made faster than human cognition permits, in environments too complex for reliable algorithmic classification, with lethal consequences that are irreversible (Scharre, 2018; Horowitz, 2019).

In each case, the governance challenge is not primarily one of instrument selection or institutional design. It is a prior question: does the system possess properties that place it beyond the reach of any conventional governance mechanism? If an algorithm operates faster than human cognition, no oversight regime premised on human review can be effective. If a system’s outputs cannot be explained, no transparency requirement produces meaningful comprehension. If capabilities emerge unpredictably, no pre-deployment assessment can provide assurance. These are not failures of regulatory ambition; they are structural limits on what governance institutions—as presently constituted—can accomplish.

While prior work on governability (Kooiman, 2003) and regulatory capacity (Lodge & Wegrich, 2012) identifies structural mismatches between governing systems and governed objects, no framework has yet operationalized the specific system properties that produce ungovernability in algorithmic contexts or provided a replicable diagnostic tool for threshold determination. The governability threshold concept fills this gap. It provides a diagnostic framework for determining when algorithmic systems approach or exceed the boundaries of institutional capacity, thereby reframing the central question of algorithmic governance: not merely “how should we regulate?” but “can we regulate?”—and when the answer is negative, what alternative architectures might be required.

1.3 Scope and Limitations

Unlike sociological approaches that examine the cultural reception of algorithmic systems, this article restricts its analysis to the institutional architecture of governance and regulatory design. It does not address questions of public perception, democratic legitimacy in the deliberative sense, or the normative desirability of specific technologies. The analysis is diagnostic: its purpose is to equip policymakers with tools for recognizing when conventional regulatory approaches have reached their limits, not to resolve the substantive question of which systems should or should not exist.

The article also does not claim that ungovernable systems are necessarily undesirable, nor that governability problems are permanent. Institutional innovation may expand governance capacity over time.

The governability threshold is a dynamic concept: what is ungovernable today may become governable through new institutional forms. But the threshold exists at any given moment, and failing to recognize it leads to governance arrangements that are nominal rather than effective—regulatory theatre that provides the appearance of control without its substance.

The remainder of the article proceeds as follows. Section 2 reviews the literatures on governance, governability, and regulatory failure, identifying the theoretical gap that the governability threshold concept addresses. Section 3 develops the theoretical framework, defining the governability threshold through four system properties—determinism, explainability, agency, and scope—and constructing the diagnostic instruments. Section 4 examines policy instruments and their failure modes through three paradigmatic cases. Section 5 provides comparative analysis of jurisdictional responses to ungovernable systems. Section 6 discusses institutional implications and advances testable propositions. Section 7 concludes.

2 Literature Review: The Limits of the Regulatory State

2.1 Governance, Governability, and State Capacity

The concept of governability—the susceptibility of a society, sector, or phenomenon to governance—has a distinguished lineage in political science, though it has been surprisingly underutilized in algorithmic governance scholarship. [Mayntz \(1993\)](#) identified “governing failures” as a structural feature of modern societies, distinguishing between failures attributable to inadequate governing capacity and those attributable to the inherent resistance of governance objects to external direction. This distinction is crucial: it implies that some governance failures cannot be remedied through institutional improvement because the problem lies in the governed object rather than the governing institution.

[Kooiman \(2003\)](#) systematized this insight through the concept of “governing interactions,” arguing that governance outcomes depend on the properties of the system being governed as much as on the properties of the governing apparatus. Governability, in Kooiman’s framework, is a relational concept: it describes the fit between the governing system and the system-to-be-governed. When the complexity, diversity, and dynamism of the governed system exceeds the governance system’s capacity to comprehend and direct, governing failures become structural rather than contingent.

[Fukuyama’s \(2013\)](#) influential reconceptualization of state capacity provides complementary analytical resources. Fukuyama distinguishes between the scope of state activities (what the state attempts to do) and the strength of state capacity (how effectively it can accomplish those activities). High-scope, low-capacity states—those that attempt to govern domains beyond their institutional competence—produce worse outcomes than more modest governance ambitions calibrated to actual capacity. This framework has direct implications for algorithmic governance: regulatory regimes that claim jurisdiction over systems they cannot effectively oversee may produce outcomes worse than no regulation at all, by creating false assurance without genuine control.

[Jessop’s \(1998\)](#) analysis of “the risks of failure” in governance arrangements adds a further dimension. Jessop argues that governance failure is not merely probable but systematically likely when governance mechanisms are applied to problems whose structural properties resist the logic of those mechanisms. Market-based governance fails for public goods; hierarchical governance fails for distributed systems; network governance fails for zero-sum conflicts. Each governance mode has inherent limitations that cannot be overcome through improved application of the same mode. The question, then, is whether algorithmic systems present properties that resist all conventional governance logics.

2.2 Regulatory Failure Theory

Regulatory theory offers a well-developed taxonomy of failure modes. [Baldwin, Cave, and Lodge \(2012\)](#) identify five principal failure types: regulatory capture (regulators internalize industry perspectives), information asymmetry (regulated entities possess superior knowledge), regulatory lag (rules become obsolete before enforcement takes effect), enforcement gaps (detection and sanction probabilities are too low for deterrence), and regulatory overreach (rules impose costs disproportionate to benefits). [Stigler's \(1971\)](#) foundational work on regulatory capture demonstrated how regulatory institutions can be systematically redirected toward private rather than public interests. Each of these failure modes has been extensively analyzed and, critically, each admits corrective interventions: independence mechanisms address capture, disclosure requirements address information asymmetry, sunset provisions and adaptive frameworks address lag, and enhanced monitoring addresses enforcement gaps.

The governability threshold identifies a different category of failure—one that persists despite optimal institutional design. When the governed system operates at velocities exceeding human cognition, information asymmetry becomes absolute rather than relative: no disclosure requirement can bridge a comprehension gap that is inherent to the system's architecture rather than contingent on information distribution. When system behavior is non-deterministic, regulatory standards cannot specify acceptable performance because the system's outputs are irreducibly probabilistic. These are not correctable failures; they are structural impossibilities. [Lodge and Wegrich \(2012\)](#) approach this insight in their analysis of "regulatory capacity," arguing that regulatory effectiveness depends not merely on institutional design but on the cognitive and organizational resources available to regulatory authorities. But they stop short of the governability threshold concept: the possibility that certain regulated objects may be inherently resistant to governance within any feasible institutional arrangement.

2.3 The Collingridge Dilemma and Technological Control

The technology governance literature introduces a further complication. [Collingridge \(1980\)](#) famously identified a dilemma of control: at the early stages of a technology's development, effective governance is possible because the technology is still malleable, but it is difficult because the consequences are unknown; at later stages, consequences become apparent, but governance is difficult because the technology is deeply embedded and resistant to change. The Collingridge dilemma implies a temporal dimension to governability—a point at which technological momentum may exceed institutional capacity for redirection.

[Perrow's \(1984\)](#) "normal accidents" framework provides a structural complement. Perrow demonstrated that in tightly coupled, complex systems, catastrophic failures are inevitable regardless of safety measures because the interaction effects among components exceed any actor's capacity to anticipate or control them. The relevance to algorithmic systems is striking: modern AI systems are both complex (their behavior emerges from the interaction of billions of parameters) and tightly coupled (their deployment in critical infrastructure creates cascading dependencies). If Perrow is correct that certain system architectures produce inevitable failures, the governance implication is profound: some algorithmic system configurations may be inherently resistant to reliable oversight.

[Stilgoe, Owen, and Macnaghten \(2013\)](#) responded to these challenges with the concept of "responsible innovation," proposing anticipatory, reflexive, inclusive, and responsive governance mechanisms. Their framework assumes, however, that innovation can be steered through institutional engagement at the design stage—an assumption that may not hold when capabilities emerge unpredictably or when system properties are determined by mathematical structure rather than design choice.

2.4 Gaps in Existing Approaches

The existing literature thus approaches but does not quite reach the governability threshold concept. Governability theory (Mayntz, Kooiman) establishes that governance failure can be structural rather than contingent but does not systematically identify the system properties that produce structural ungovernability. Regulatory failure theory (Baldwin et al.) provides sophisticated failure typologies but treats all failures as correctable. The Collingridge dilemma identifies temporal constraints on governance but does not extend to the system-property analysis required for algorithmic systems. Normal accidents theory (Perrow) demonstrates that certain system architectures produce inherent failures but was developed for industrial systems rather than computational ones.

The present framework advances beyond these prior contributions in three specific respects. First, whereas Mayntz and Kooiman identify governability as a relational concept without specifying the system properties that determine it, this article isolates four measurable dimensions—determinism, explainability, agency, and scope—that enable structured diagnostic assessment. Second, whereas Perrow’s normal accidents theory and Collingridge’s dilemma were developed for industrial and early-stage technologies respectively, the governability threshold framework addresses computational systems whose properties (stochasticity, opacity, autonomous agency, cross-jurisdictional deployment) differ categorically from those of industrial infrastructure. Third, whereas governance principles articulated in defense and security contexts—such as the responsibility, reliability, and governability criteria employed in Department of Defense and NATO autonomous weapons frameworks—address governability as an institution-side design objective (“how should we build governable systems?”), the present framework treats governability as a system-side diagnostic (“does this system possess properties that defeat governance regardless of institutional design?”). This diagnostic reorientation is the article’s central contribution.

Several scholars bridge the gap between traditional governance theory and algorithmic governance, approaching but not formalizing the threshold concept. Yeung (2018) proposed “algorithmic regulation” as a distinct governance modality, identifying how algorithmic decision-making transforms the objects of governance in ways that challenge conventional regulatory logic—but framed the challenge as one of regulatory adaptation rather than structural limit. Danaher (2016) analyzed the “threat of algocracy,” arguing that algorithmic decision systems fundamentally undermine democratic accountability by rendering governance processes opaque—a diagnosis that implies threshold-crossing on the explainability dimension but does not generalize to a multi-dimensional framework. Binns (2018) examined the philosophical foundations of algorithmic fairness and concluded that competing fairness criteria create irreconcilable normative tensions when applied to opaque systems—an insight that suggests structural limits to standards-based governance but does not operationalize the boundary condition. Each of these contributions identifies symptoms of threshold-crossing without naming the underlying structural mechanism or providing a diagnostic tool for its identification.

What is missing is a framework that identifies the specific properties of algorithmic systems that determine whether they can be governed effectively—a diagnostic tool that enables regulators to assess governability before investing in governance mechanisms that may prove structurally inadequate. The governability threshold concept fills this gap.

3 Theoretical Framework: The Governability Threshold

3.1 Defining the Governability Threshold

The governability threshold is the point at which the intrinsic properties of an algorithmic system exceed institutional capacity for meaningful oversight, rendering conventional governance approaches structurally insufficient regardless of design quality.

This definition requires unpacking. “Intrinsic properties” refers to characteristics of the system itself—its decision-making speed, interpretability, autonomy, and deployment scope—rather than contingent features of the regulatory environment. “Institutional capacity” encompasses the cognitive, temporal, monitoring, and enforcement resources available to governance institutions. “Meaningful oversight” means governance that achieves more than nominal compliance—oversight that enables institutions to understand, monitor, intervene in, and hold accountable the systems under their jurisdiction. “Structurally insufficient” distinguishes threshold failures from correctable design failures: beyond the threshold, no feasible improvement in regulatory design produces effective governance.

The critical distinction is between **correctable** and **structural** governance failures. Correctable failures respond to improved institutional design: more expert staff, better information systems, enhanced enforcement powers, or refined regulatory standards. Structural failures persist regardless of design improvement because they reflect a mismatch between the inherent properties of the governed system and the inherent capabilities of governance institutions. The governability threshold marks the boundary between these two categories.

A terminological clarification is essential. Throughout this article, “ungovernability” denotes the *consequence state* that obtains when a system exceeds the governability threshold—it is the diagnostic output, not a synonym for the threshold itself. The threshold is a boundary condition defined by system properties; ungovernability is the governance condition that obtains beyond that boundary.

A further clarification guards against misinterpretation. Ungovernability as used in this framework is a diagnostic and context-dependent condition, not a normative judgment about technological legitimacy. It is contextual because it reflects the relationship between *current* system properties and *current* institutional capacities—a relationship that can change as institutions innovate or as system properties are constrained. It is time-bound because advances in interpretability research, monitoring infrastructure, or institutional design may shift a system from “Beyond Threshold” to “Threshold-Adjacent” in subsequent periods. And it does not imply that ungovernable systems should be tolerated: to the contrary, the appropriate response to a threshold-crossing diagnosis is either to constrain the system’s properties until governability is restored, or to prohibit the ungovernable configuration entirely. The framework thus counsels neither regulatory fatalism nor technological pessimism, but diagnostic precision as the prerequisite for effective governance response.

The governability threshold must also be distinguished from adjacent frameworks that analyze governance deficits from the *institution side*. Concepts such as state capacity (Fukuyama, 2013), regulatory capacity (Lodge & Wegrich, 2012), and institutional absorptive capacity address how institutions can be redesigned, resourced, or restructured to handle governance challenges more effectively. These are institution-side frameworks: they ask “how can governance institutions upgrade their capacity?” The governability threshold, by contrast, is a system-side diagnostic: it asks “does this object possess properties that defeat governance regardless of institutional capacity?” The distinction is analytically consequential because it determines the appropriate policy response. When governance failure is institution-side, the remedy is institutional reform—more resources, better design, enhanced coordination. When governance failure is system-side (threshold-crossing), institutional reform alone is insufficient; the remedy must include constraining the system’s own properties—reducing speed, limiting scope, mandating interpretability safeguards, requiring human-in-the-loop—to pull the system back within the governable zone.

This distinction has direct implications for regulatory strategy: a finding that a system has crossed the governability threshold does not authorize regulatory abdication. To the contrary, threshold-crossing should trigger **mandatory system-property constraints**—forced re-entry mechanisms that require the system to be modified until its properties fall back within the governable zone. Prohibition of the ungovernable configuration, mandatory speed limitations, scope restrictions, or interpretability requirements

Table 1: Object-Limits vs. Institution-Limits: A Diagnostic Distinction

Dimension	Object-Side Limit (Governability Threshold)	Institution-Side Limit (Capacity Deficit)
Diagnostic question	Can this system be governed by <i>any</i> feasible institutional arrangement?	Can this institution be improved to govern the system effectively?
Root cause	Intrinsic system properties (speed, opacity, autonomy, scope)	Institutional design, resources, expertise, or coordination
Failure persistence	Persists regardless of institutional improvement	Correctable through institutional reform
Policy response	Constrain system properties (speed limits, scope restrictions, mandatory interpretability) or prohibit	Reform institutions (staff, resources, coordination, mandate)
Example	LLM emergent capabilities defeat any pre-deployment verification regime	Financial regulator lacks AI expertise to evaluate algorithmic trading systems

are all legitimate responses to threshold-crossing.

Table 2: Definition Standard: Core Concepts

Concept	Definition	Positive Test	Negative Test
Governability	The capacity of institutions to meaningfully oversee, direct, and hold accountable an algorithmic system	Institutions can understand system behavior, monitor operations, intervene when necessary, and assign responsibility for outcomes	Institutional oversight is nominal, performative, or systematically ineffective despite adequate resources
Governability Threshold	The point at which system properties exceed institutional capacity for meaningful oversight	Conventional regulatory instruments fail systematically across multiple institutional contexts regardless of design quality	Governance challenges addressable through improved instrument design, increased resources, or institutional reform
Ungovernability	A state in which no feasible combination of conventional regulatory instruments achieves meaningful oversight	Fundamental capacity gap exists; no realistic allocation of regulatory resources produces effective oversight; failure pattern is structural and persistent	Failure results solely from insufficient resources, poor instrument selection, or correctable institutional design

3.2 Four Dimensions of System Governability

Governability is determined by the relationship between system properties and institutional capacity. Four system properties are decisive: determinism, explainability, agency, and scope. Each creates distinct governance challenges, and each can independently push a system toward or beyond the governability threshold.

These four dimensions do not exhaust the factors that influence governance outcomes. Economic concentration, environmental externalities, sociotechnical embedding, and political salience all affect the feasibility and effectiveness of regulatory interventions. However, these factors operate as secondary or derivative influences that shape the governance *context* rather than defining the structural *boundary* of governability itself. The four dimensions identified here are threshold-defining because they determine whether conventional regulatory instruments can *function at all*, not merely whether they function *well*.

Determinism determines whether governance can rely on testing, certification, and performance standards. If a system is deterministic, regulators can verify behavior through input-output testing: given input X, the system should produce output Y, and any deviation constitutes a verifiable violation. Non-deterministic systems—including most contemporary neural networks—defeat this approach because their outputs are irreducibly variable. Critically, determinism encompasses not only input-output consistency but also **temporal verifiability**: a system whose behavior is deterministic in principle but whose operations occur faster than any feasible verification process presents the same threshold challenge as a non-deterministic system. High-frequency trading algorithms may produce deterministic outputs for given inputs, but when those operations execute in microseconds, no human or institutional verification process

Table 3: The Governability Threshold Matrix

System Property	Low Governability (Threshold Risk)	Medium Governability (Manageable with Adaptation)	High Governability (Standard Regulation Sufficient)
Determinism	Non-deterministic: identical inputs produce variable outputs; behavior is irreducibly stochastic	Semi-deterministic: outputs vary within bounded ranges; probabilistic behavior is statistically predictable	Fully deterministic: identical inputs always produce identical outputs; behavior is completely reproducible
Explainability	Opaque: no available method can explain why specific outputs were produced; causal mechanisms are inaccessible	Partially interpretable: statistical correlations or approximate explanations are available but mechanistic understanding is incomplete	Fully interpretable: complete causal chain from input to output is accessible, comprehensible, and verifiable
Agency	Fully autonomous: system selects goals, modifies strategies, and executes actions without human intervention	Semi-autonomous: system operates independently within human-defined parameters; human oversight exists but is not decision-by-decision	Human-controlled: every consequential decision requires explicit human authorization; system functions as decision-support tool
Scope	General-purpose: deployed across multiple domains, jurisdictions, and use cases with no predetermined application boundary	Domain-general: operates across related fields within identifiable sectoral or jurisdictional boundaries	Narrow-domain: operates within a single, well-defined application context with clear jurisdictional authority

can confirm compliance in real time.

Explainability determines whether governance can rely on transparency, disclosure, and accountability mechanisms. If a system’s decision process is interpretable, regulators can evaluate whether decisions were made on legitimate grounds, affected parties can contest adverse decisions, and auditors can verify compliance with substantive standards. Opaque systems—including deep neural networks with billions of parameters operating through complex pattern recognition—defeat transparency-based governance because no available method can explain why specific outputs were produced. Disclosure requirements produce data without comprehension; accountability mechanisms lack the causal attribution they require (Pasquale, 2015).

Agency determines whether governance can rely on human oversight, authorization, and control mechanisms. If a system operates under human direction—as a decision-support tool providing recommendations that humans evaluate and act upon—governance can focus on the human decision-maker: their training, accountability, and institutional constraints. Autonomous systems that select goals, modify strategies, and execute actions without human intervention in the decision loop render human-oversight governance approaches structurally inadequate.

Scope determines whether governance can rely on jurisdictional authority, sectoral expertise, and domain-specific regulation. Scope operates along two analytically distinct sub-dimensions. **Domain scope** refers to the range of sectors, fields, and use cases in which the system is deployed. **Jurisdictional scope** refers to the territorial reach of deployment. Either sub-dimension alone can produce jurisdiction failure. General-purpose systems deployed at global scale—such as foundation models—present challenges on both sub-dimensions simultaneously, creating compound jurisdiction failures that are structurally resistant to governance by any existing institutional arrangement (Gorwa, 2019; Gillespie, 2018).

3.3 Institutional Failure Diagnosis

When a system approaches or exceeds the governability threshold, specific patterns of institutional failure emerge. These patterns are diagnostic: they indicate which governance dimension has been exceeded and therefore where conventional approaches are structurally inadequate.

These failure types frequently co-occur, creating compound governability challenges. A large language model, for example, may simultaneously exhibit verification failure (non-deterministic outputs), attribution failure (opaque reasoning), and jurisdiction failure (global deployment). When multiple failure types coincide, the governability challenge compounds geometrically rather than additively.

3.4 Threshold Determination: An Operationalizable Scoring Protocol

Determining whether a specific system exceeds the governability threshold requires systematic assessment across all four dimensions. Each dimension is scored on a scale of 0–2, where the score reflects the degree to

Table 4: Institutional Failure Diagnosis

Failure Type	Cause	Illustrative Example	Threshold Dimension
Verification Failure	System outputs cannot be predicted, replicated, or tested against performance standards because behavior is non-deterministic	Stochastic language model generating variable responses to identical queries	Determinism
Attribution Failure	Decision processes cannot be traced, explained, or evaluated because internal mechanisms are opaque	Deep learning classifier denying loan applications based on inscrutable pattern recognition	Explainability
Control Failure	Human oversight cannot be exercised because system operates faster than, or independently of, human decision-making	Autonomous trading algorithm executing strategies without human authorization	Agency
Jurisdiction Failure	No single regulatory authority possesses the expertise, mandate, or territorial reach to govern the system's full operational scope	General-purpose AI model deployed simultaneously across multiple sectors and jurisdictions	Scope

which the system's properties are amenable to governance through conventional regulatory instruments:

- **Score 2 (High Governability):** System property is fully compatible with standard regulation.
- **Score 1 (Medium Governability):** System property presents governance challenges that require significant adaptation but remain manageable.
- **Score 0 (Low Governability—Structural):** System property defeats conventional governance instruments regardless of design quality.

The four dimension scores are summed to produce a **composite governability score** (range: 0–8). The threshold determination depends on both the composite score and the number of structural (zero-score) dimensions:

Table 5: Governability Scoring Protocol: Zone Definitions

Zone	Composite Score	Structural Dimensions (Score = 0)	Classification	Policy Implication
Governable	≥ 7	None	System is within institutional capacity	Standard regulatory instruments apply; routine oversight sufficient
Threshold-Adjacent	5–6	At most 1	System approaches the governability threshold	Governance adaptation required; enhanced monitoring and adapted instruments; regulatory sandboxes appropriate
Beyond Threshold	≤ 4 , or any score with ≥ 2 structural dimensions	≥ 2	System exceeds the governability threshold	Conventional instrument toolkit collapses; system-property constraints or prohibition required

The threshold is thus a configurational concept: it depends on the combination of system properties rather than any single property in isolation.

Applying this protocol to the three paradigmatic cases analyzed in Section 4 yields the following threshold determinations:

Table 6: Case Application: Governability Threshold Scoring

System	Determ.	Explain.	Agency	Scope	Composite	Struct. Dims	Zone
HFT	0	0	0	1	1	3	Beyond Threshold
LLM (Foundation Model)	0	0	1	0	1	3	Beyond Threshold
LAWS	1	0	0	1	2	2	Beyond Threshold

All three paradigmatic cases score deep in the “Beyond Threshold” zone, confirming the framework's diagnostic validity. The scoring protocol also reveals diagnostic nuances: HFT's threshold-crossing is driven by temporal verifiability failure and autonomous speed; LLM threshold-crossing is driven by the

compound of non-determinism, opacity, and unbounded scope; LAWS threshold-crossing combines lethal autonomy with environmental opacity.

A non-deterministic system that operates in a narrow domain under human control (e.g., a medical diagnostic AI) might score $0-1-2-2 = 5$ with one structural dimension—Threshold-Adjacent, requiring adapted governance but not fundamentally ungovernable. The protocol thus provides graduated diagnostic output rather than a binary governable/ungovernable determination.

The scoring protocol is designed as a structured heuristic rather than a mechanical metric, and its application inevitably involves interpretive judgment. To enhance reliability, scoring should be conducted by panels combining regulatory expertise, technical knowledge, and domain-specific experience, with explicit calibration against the anchor cases provided in Table 6. Future empirical validation—including inter-rater reliability testing across expert panels, sensitivity analysis of alternative threshold definitions, and systematic application to emerging system categories such as agentic AI and multi-agent architectures—will be essential to refine the protocol’s diagnostic precision.

4 Policy Instrument Analysis: When Conventional Tools Fail

4.1 Why Standard Instruments Are Insufficient Beyond the Threshold

Conventional governance instruments—prohibition, licensing, disclosure, audit, and liability—each presuppose specific conditions for effectiveness. Prohibition requires the capacity to detect and prevent prohibited activities. Licensing requires the capacity to evaluate applicants against defined standards. Disclosure requires that disclosed information is comprehensible and actionable. Audit requires observable, verifiable criteria against which performance can be measured. Liability requires attributable causation linking specific decisions to specific harms.

Beyond the governability threshold, one or more of these preconditions fails. If system behavior is non-deterministic, audit against performance standards becomes unreliable because the same system may perform differently on the same test at different times. If decision processes are opaque, disclosure produces data without comprehension. If the system operates autonomously, human-oversight licensing conditions become nominal rather than operative. If deployment scope exceeds jurisdictional reach, enforcement becomes impossible in significant portions of the system’s operating territory.

This does not mean governance is impossible beyond the threshold—it means that governance requires fundamentally different approaches than the conventional instrument toolkit provides.

Table 7: Governance Adaptation Strategies

Strategy	Description	Applicable When	Trade-offs	Threshold Dimension Addressed
Prohibition	Categorical ban on system configurations that exceed governability thresholds	Unacceptable risk; no governance adaptation can restore effective oversight	Innovation foreclosed; regulatory rigidity; enforcement challenges	All dimensions (last resort)
Regulatory Sandbox	Controlled experimentation environments with bounded deployment and enhanced monitoring	High uncertainty about system behavior; governability status unclear	Limited scope restricts benefits; graduated deployment delays innovation diffusion	Determinism, Explainability
Outcome-Based Liability	Strict or presumptive liability for harms, shifting burden from ex ante oversight to ex post accountability	Attribution gaps preclude traditional fault-based liability; system complexity defeats direct oversight	Innovation deterrence; actuarial gaming; compensation without prevention	Explainability, Agency
Adaptive Regulation	Iterative, learning-based regulatory frameworks with built-in review cycles and real-time monitoring	Rapid technological change renders static rules obsolete; system capabilities evolve faster than revision cycles	Regulatory uncertainty; compliance costs for evolving standards	Determinism, Scope

The grammatical structure of this analysis can now be made explicit. Each conventional regulatory instrument presupposes minimum levels of system governability on specific dimensions. When a system’s properties fall below these preconditions, the instrument becomes structurally ineffective. The following matrix maps these preconditions:

Table 8: Instrument Preconditions $\times$ System Properties: The Grammar of Instrument Collapse

Regulatory Instrument	Determinism Req.	Explainability Req.	Agency Req.	Scope Req.	Collapse Condition
Performance Standards / Testing	$\geq$ Medium	—	—	—	Non-deterministic systems produce variable outputs; testing cannot establish compliance
Disclosure / Transparency	—	$\geq$ Medium	—	—	Opaque systems produce data without comprehension; disclosure becomes performative
Audit / Certification	$\geq$ Medium	$\geq$ Medium	—	—	Non-deterministic or opaque systems defeat audit methodology
Human Oversight / Authorization	—	—	$\leq$ Medium	—	Fully autonomous systems operate outside human decision loops
Sectoral Regulation	—	—	—	$\leq$ Medium	General-purpose systems span multiple regulatory mandates
Fault-Based Liability	—	$\geq$ Medium	$\leq$ Medium	—	Opaque, autonomous systems defeat causal attribution and agent identification
Jurisdictional Enforcement	—	—	—	$\leq$ Medium	Global deployment defeats territorial enforcement

The critical insight emerges from reading the matrix vertically by system type. A system scoring 0 on determinism defeats performance standards, testing, audit, and certification—four instruments collapse simultaneously. A system scoring 0 on explainability defeats disclosure, audit, and fault-based liability. When a system scores 0 on two or more dimensions, the instrument collapses compound: essentially the entire conventional regulatory toolkit becomes structurally ineffective. This is the grammatical logic of threshold-crossing: it is not that regulation becomes *harder*—it is that the preconditions for conventional instruments *cease to exist*.

4.2 Case Application: High-Frequency Trading

High-frequency trading (HFT) represents the paradigmatic case of threshold-crossing along the agency and determinism dimensions. HFT systems execute trading strategies in microseconds—approximately one thousand times faster than human perceptual capacity and one million times faster than institutional decision-making cycles. The 2010 Flash Crash demonstrated that algorithmic trading can generate massive market effects within a time horizon that precludes any feasible human intervention (Kirilenko et al., 2017).

The empirical record is instructive. On May 6, 2010, the Dow Jones Industrial Average plunged approximately 998.5 points—roughly 9%—in thirty-six minutes, with the most severe decline occurring in a five-minute window between 2:42 and 2:47 p.m. Eastern Time. During this interval, approximately 2 billion shares changed hands. Individual securities experienced price dislocations of extraordinary magnitude: Accenture shares traded at \$0.01 while Apple shares briefly traded at \$100,000. The subsequent joint CFTC-SEC investigation required five months to produce preliminary findings—a temporal asymmetry that epitomizes the governability challenge. Kirilenko et al. (2017) demonstrated that the crash was triggered by a single algorithmic sell order of 75,000 E-Mini S&P 500 futures contracts, but the cascading effects resulted from interactions among thousands of algorithms responding to each other's behavior.

Budish, Cramton, and Shim (2015) identified a deeper structural problem: the continuous limit order book mechanism that governs modern electronic markets creates an arms race in which trading firms invest billions in speed advantages measured in millionths of a second. The resulting market microstructure operates at timescales where the concept of “human oversight” becomes physically impossible.

Governability assessment: HFT systems score low on the governability matrix along multiple dimensions. Their trading strategies are proprietary and complex (low explainability). Many employ adaptive algorithms that modify behavior in response to market conditions (reduced determinism). They

operate autonomously, executing thousands of decisions per second without human authorization (high agency). And they interact with global electronic markets across multiple jurisdictions (broad scope).

Institutional failure pattern: The primary failure type is control failure: human oversight cannot operate at the speeds required for meaningful intervention. A human trader requires approximately 200 milliseconds to perceive a visual stimulus; during that interval, an HFT algorithm may execute over 1,000 trades. Secondary failure types include attribution failure and verification failure.

Governance response: The regulatory response to HFT illustrates governance adaptation to threshold challenges. Circuit breakers—automated trading halts triggered by extreme price movements—represent structural constraints that restore governability by imposing pauses in system operation. The Limit Up–Limit Down (LULD) mechanism, implemented in 2012, automatically pauses trading when prices move beyond specified bands. Minimum resting times for orders reduce effective speed. Order-to-trade ratio limits constrain the volume of algorithmic activity (Budish et al., 2015; MacKenzie, 2021).

The HFT case demonstrates a crucial principle: governance adaptation to threshold challenges requires constraining system properties rather than enhancing oversight capacity. Several exchanges have introduced asymmetric speed bumps (deliberate latency for certain order types) and the transition to T+1 settlement in 2024 reduced some systemic coupling, but these adaptations operate through the same logic of property constraint rather than enhanced oversight.

4.3 Case Application: Large Language Models

Large language models (LLMs) represent threshold-crossing along the determinism and explainability dimensions. Contemporary foundation models—with parameter counts in the hundreds of billions—operate through pattern recognition processes that are fundamentally opaque (Bommasani et al., 2021). Moreover, LLMs exhibit emergent capabilities that appear at certain scale thresholds, creating verification challenges that no pre-deployment testing regime can fully address (Wei et al., 2022).

Weidinger et al. (2022) identified six primary risk categories for language models: discrimination and exclusion, information hazards, misinformation harms, malicious uses, human-computer interaction harms, and environmental and socioeconomic harms. This taxonomy is itself diagnostic of the governability challenge: the range of potential harms spans categories governed by entirely different regulatory regimes, none of which was designed to govern a single computational system capable of producing harms across all categories simultaneously.

Governability assessment: LLMs present low governability along the determinism dimension (identical prompts may yield different outputs due to sampling mechanisms) and the explainability dimension (output generation processes resist mechanistic explanation). Agency remains moderate in most current deployments (humans initiate interactions, though agentic LLM deployments are rapidly emerging). Scope is broad: foundation models serve as general-purpose infrastructure deployed across healthcare, education, finance, law, and other domains.

Institutional failure pattern: The primary failure is attribution failure: when an LLM generates harmful content, no mechanistic explanation can identify why this specific output was produced rather than a harmless alternative. The responsibility chain—developer, deployer, user—is distributed across multiple actors, none of whom exercises the kind of directed decision-making that traditional accountability frameworks presuppose. Secondary failure is verification failure: because outputs are non-deterministic and capabilities emerge unpredictably, pre-deployment testing cannot provide comprehensive assurance (Wei et al., 2022).

Governance response: Current governance responses include red-teaming, responsible scaling policies, output filtering, and usage policies. These responses implicitly acknowledge the explainability threshold: absent mechanistic understanding, governance relies on empirical testing and behavioral constraints

rather than direct regulatory oversight of decision processes. Since 2024, major AI laboratories have adopted responsible scaling commitments ([Anthropic, 2024](#)), which represent a significant development precisely because they internalize the threshold logic.

The EU AI Act's transparency and risk-classification requirements for general-purpose AI models represent the most ambitious regulatory attempt to impose governance structure despite threshold challenges ([European Parliament, 2024](#); [Smuha, 2021](#)). The Act mandates governance of inputs, processes, and outputs precisely because governance of internal operations is structurally infeasible. Model evaluation frameworks proposed by [Shevlane et al. \(2023\)](#) represent emerging approaches to standardizing pre-deployment threshold assessment.

4.4 Case Application: Lethal Autonomous Weapon Systems

Lethal autonomous weapon systems (LAWS)—weapon systems capable of selecting and engaging targets without human intervention—represent the most consequential governability threshold case because ungovernable operation produces irreversible lethal outcomes. LAWS combine threshold challenges across the agency dimension, the determinism dimension, and the explainability dimension ([Scharre, 2018](#); [Horowitz, 2019](#)).

Governability assessment: LAWS present extreme governance challenges on the agency dimension: by definition, they select and engage targets without human intervention. Explainability is low: targeting decisions based on complex sensor fusion and pattern classification resist transparent reconstruction. Determinism is reduced by the complexity of real-world environments. Scope varies but can be broad.

Institutional failure pattern: The primary failure type is control failure: if targeting decisions occur faster than human cognition permits, no “human-in-the-loop” requirement can be substantively meaningful. This is compounded by attribution failure: when an autonomous system causes civilian casualties, tracing the causal chain from system design through deployment context to specific targeting decision is prohibitively complex ([Crootoof, 2015](#); [Heyns, 2017](#)).

Governance response: International responses to LAWS illustrate two threshold-responsive strategies. First, prohibition—several states and civil society organizations advocate a preemptive ban on autonomous weapons systems that cannot ensure meaningful human control. Second, redefinition of the governance object—rather than attempting to govern individual targeting decisions, some proposals redefine the governance target as the decision to deploy an autonomous system in a particular context. This latter approach shifts governance from the ungovernable level (individual algorithmic targeting) to a governable level (human deployment authorization).

The LAWS case also reveals the limitations of analogical governance reasoning. Autonomous weapons present both humanitarian and epistemic challenges simultaneously: their effects may be harmful not only because of the damage they cause but because no actor may be able to predict, explain, or attribute specific outcomes. The governability threshold concept clarifies what analogical reasoning obscures: LAWS present a qualitatively different governance challenge from previous weapon categories because their system properties simultaneously exceed governance capacity across multiple dimensions.

5 Comparative Analysis: Jurisdictional Responses to Ungovernable Systems

5.1 European Union: Risk-Tier Classification as Threshold Proxy

The EU AI Act (Regulation 2024/1689) represents the most ambitious attempt to institutionalize governability assessment in regulatory design. The Act's four-tier risk classification—unacceptable risk, high risk, limited risk, and minimal risk—functions as an implicit governability threshold mechanism. Systems classified as “unacceptable risk” (Article 5) are effectively deemed to exceed the governability threshold: they

are prohibited not because they cannot be designed responsibly but because no governance mechanism can provide adequate assurance (European Parliament, 2024).

The Act's treatment of general-purpose AI models (GPAI) is particularly revealing. The Act introduces an explicit computational threshold for enhanced governance: GPAI models whose training involves cumulative compute exceeding 10^{25} floating point operations (FLOPS) are presumed to present "systemic risk" (Article 51) and face additional obligations. This FLOPS-based threshold is, in effect, a legislative operationalization of the governability threshold concept. The significance of this legislative development cannot be overstated: it demonstrates that regulatory practice has independently converged on the structural insight that certain system properties create categorical governance discontinuities rather than merely quantitative governance challenges.

More broadly, the Act's GPAI provisions recognize the explainability and determinism threshold challenges that foundation models present: unable to govern the models' internal operations, the Act mandates governance of inputs, processes, and outputs (Yeung, 2018; Smuha, 2021). However, the compute-based threshold is not without limitations: algorithmic efficiency improvements may enable models to achieve equivalent capabilities at compute levels below the 10^{25} FLOPS threshold.

5.2 United States: Sectoral Responses to Threshold Challenges

The United States has addressed governability thresholds primarily through sector-specific regulation rather than comprehensive framework legislation. Financial regulation illustrates sophisticated threshold adaptation. The SEC and the CFTC confronted HFT's velocity threshold directly following the 2010 Flash Crash. Their response—circuit breakers, consolidated audit trails, and registration requirements—represents structural constraints on system properties rather than enhanced oversight capacity. The Dodd-Frank Act's Volcker Rule further illustrates scope constraints.

The NHTSA's approach to autonomous vehicles provides a contrasting example. Rather than comprehensive pre-market regulation, NHTSA has pursued a framework that combines voluntary manufacturer commitments, incident reporting, and investigative authority.

The absence of comprehensive federal AI legislation means that threshold challenges are addressed ad hoc, with significant governance gaps in domains where no sector-specific regulator possesses relevant authority. Late 2025 executive orders signaled federal intent to consolidate AI oversight, but the resulting framework remains sectoral in orientation.

The contrast with Japan's approach is instructive. Japan's AI Promotion Act (enacted May 2025) adopted a deliberately "light-touch" regulatory posture, encouraging voluntary cooperation rather than imposing binding obligations (National Diet of Japan, 2025). The governability threshold framework would characterize this as accepting threshold-crossing in exchange for innovation benefits, with deferred governance costs.

5.3 International Governance: The Convention on Certain Conventional Weapons

The most explicit engagement with governability thresholds in international governance occurs in the Convention on Certain Conventional Weapons (CCW) process on LAWS. The Group of Governmental Experts (GGE) has debated since 2014 whether autonomous weapons systems require new international legal instruments, with the central question being precisely a governability question: can existing international humanitarian law effectively govern systems that select targets without human intervention?

The debate crystallizes a governability threshold dispute. States advocating prohibition argue that certain levels of autonomy in targeting decisions cross a threshold beyond which IHL compliance cannot be ensured. States opposing prohibition argue that meaningful human control can be maintained through system design, deployment contexts, and command structures.

The CCW process has produced a significant conceptual contribution: the articulation of “meaningful human control” as the operative governance standard. This concept represents an attempt to define the governability threshold for lethal autonomous systems in operational terms.

5.4 Comparative Synthesis

Three patterns emerge from comparative analysis. First, jurisdictions converge on the recognition that certain algorithmic system configurations require governance approaches qualitatively different from conventional regulation. Each jurisdiction has, independently, concluded that some algorithmic systems cannot be governed through conventional oversight mechanisms.

Second, approaches diverge on the strategic response to threshold-exceeding systems: prohibition (EU unacceptable risk; proposed LAWS ban), structural constraint (US HFT regulation), or governance-level shifting (CCW meaningful human control at the deployment rather than operational level).

Third, all jurisdictions struggle with scope challenges: algorithmic systems that operate transnationally expose the limits of territorially-bounded governance institutions. The EU AI Act attempts to address this through extraterritorial application, but enforcement remains practically challenging.

A fourth pattern deserves mention: the convergence on **governance-level shifting** as the default adaptation to threshold-crossing. When individual algorithmic decisions exceed governance capacity, all three jurisdictional approaches shift governance to higher levels of abstraction. The theoretical significance is substantial: it suggests that governance beyond the threshold requires a reconceptualization of what governance governs—not individual system behaviors but the structural conditions under which systems operate.

6 Institutional Implications

6.1 New Regulatory Competencies

Governance beyond the threshold requires institutional capabilities that current regulatory agencies largely lack. Four competency gaps are critical.

Technical comprehension capacity. Regulators must develop the capacity to assess system properties—determinism, explainability, agency, scope—with sufficient depth to conduct threshold determinations. Existing regulatory agencies were designed to oversee activities performed by identifiable human actors within bounded jurisdictions; algorithmic governance requires comprehension of computational systems whose behavior emerges from mathematical operations on massive datasets.

Temporal matching capacity. Regulatory institutions must develop the capacity to operate on timescales matched to the systems they govern. When algorithmic systems operate in milliseconds and iterate in weeks, regulatory review cycles measured in months or years produce governance that is systematically retrospective (Dafoe, 2018; Calo, 2017).

Jurisdictional coordination capacity. General-purpose algorithmic systems deployed across multiple sectors and jurisdictions require governance coordination mechanisms that transcend traditional regulatory silos.

Anticipatory assessment capacity. Governance of systems exhibiting emergence ungovernability requires institutions capable of assessing capabilities that do not yet exist. Responsible innovation frameworks (Stilgoe et al., 2013) provide conceptual foundations, but operationalizing anticipation in regulatory institutions requires mechanisms for horizon scanning, scenario planning, and graduated deployment.

6.2 The Anticipatory Governance Challenge

The governability threshold concept reveals a structural tension in contemporary governance: the systems most urgently requiring governance are precisely those whose properties make them most resistant to governance. Algorithmic systems become more powerful—and therefore more consequential—as they become faster, more complex, more autonomous, and more broadly deployed. But each of these property changes pushes the system further toward and potentially beyond the governability threshold.

This tension cannot be resolved within conventional regulatory frameworks. Prohibition resolves it by sacrificing capability for governability—an approach appropriate for unacceptable-risk applications but impractical as a general solution. Adaptive regulation attempts to maintain pace with capability development but faces inherent temporal limits. Outcome-based liability indirectly governs through incentive structures but provides no mechanism for preventing harms, only compensating for them.

6.3 The Governance Innovation Deficit

The structural tension between capability development and governability produces what we term the **governance innovation deficit**: the persistent gap between the rate of algorithmic capability development and the rate of governance innovation. This deficit is not merely quantitative (governance is slower than technology) but qualitative (governance innovation and technological innovation operate according to fundamentally different logics).

Technological innovation in algorithmic systems is driven by competitive dynamics, massive capital investment, and the mathematical properties of scaling laws. Governance innovation, by contrast, is driven by democratic deliberation, institutional precedent, and stakeholder consensus. These structural differences ensure that the governance innovation deficit is not a transient condition but a persistent feature of the algorithmic governance landscape.

The deficit manifests in three observable ways. First, **regulatory latency**: the time between the deployment of a threshold-crossing system and the development of adequate governance mechanisms consistently exceeds the time during which the system can cause significant harm. Second, **conceptual latency**: governance frameworks consistently lag behind the conceptual innovations required to understand threshold-crossing systems. Third, **institutional rigidity**: existing regulatory institutions resist the organizational transformation required to govern threshold-crossing systems.

Addressing the governance innovation deficit requires institutional experimentation at a scale and pace that democratic governance systems have rarely attempted. The governability threshold framework suggests that incremental institutional reform is insufficient: governance of threshold-crossing systems requires governance institutions that are themselves innovative.

A fourth dimension of the deficit deserves acknowledgment: the **political economy of threshold-crossing**. Even when governability thresholds are identified and diagnostic tools are available, powerful economic interests may resist the system-property constraints that threshold-crossing demands. This political economy dimension does not invalidate the diagnostic framework, but it does imply that threshold identification alone is insufficient: effective governance requires both diagnostic capacity and the political will to act on diagnostic findings.

6.4 Testable Propositions

The governability threshold framework generates empirically testable propositions.

Proposition 1 (Compound Threshold): Systems exhibiting low governability on two or more dimensions simultaneously will produce more severe governance failures than systems exhibiting low governability on any single dimension, controlling for other factors.

Proposition 2 (Governance-Level Shifting): When systems exceed governability thresholds at the operational level, governance tends to shift to higher levels of abstraction—from overseeing individual decisions to governing deployment contexts, system design standards, or organizational processes.

Proposition 3 (Threshold Dynamics): Governability thresholds are dynamic rather than fixed: institutional innovation can expand governance capacity, moving previously ungovernable systems within the threshold boundary.

Proposition 4 (Threshold Externalization): Systems that exceed the governability threshold on internal dimensions (determinism, explainability) but remain governable on external dimensions (agency, scope) will generate governance strategies oriented toward output monitoring rather than process oversight. The EU AI Act’s FLOPS-based threshold provides initial evidence: the Act mandates governance of inputs, processes, and outputs precisely because governance of internal operations is structurally infeasible.

7 Conclusion

This article has introduced the concept of the **governability threshold**—the point at which algorithmic system properties exceed institutional capacity for meaningful oversight—and developed a diagnostic framework for identifying when systems approach or exceed this threshold.

The framework is organized around four system properties that determine governability: determinism (whether system behavior can be predicted, tested, and verified), explainability (whether system reasoning can be understood, attributed, and evaluated), agency (whether human oversight can be meaningfully exercised), and scope (whether regulatory authority can encompass the system’s full operational range). Threshold-crossing occurs when challenges in two or more dimensions simultaneously exceed what institutional design can address.

Analysis of three paradigmatic cases illustrates the framework’s diagnostic value. High-frequency trading demonstrates velocity-driven threshold-crossing. Large language models demonstrate complexity-driven threshold-crossing, where non-determinism and opacity defeat verification and attribution mechanisms. Lethal autonomous weapon systems present the most consequential threshold case, combining agency, determinism, and explainability challenges with irreversible lethal consequences.

Comparative analysis reveals convergent recognition across jurisdictions that certain system configurations require qualitatively different governance approaches, combined with divergent strategic responses: prohibition, structural constraint, and governance-level shifting.

The governability threshold concept does not counsel regulatory despair. Many algorithmic systems remain well within governance capacity, and institutional innovation can expand the boundaries of what institutions can effectively govern. By 2026, institutional innovations—mandatory model registries, third-party red-teaming infrastructure, and liability allocation mechanisms—have begun to close the governance gap. These innovations constitute structural governance adaptations that shift systems from “Beyond Threshold” toward “Threshold-Adjacent.” The trajectory illustrates that threshold-crossing, while structurally determined, is not necessarily permanent—but that recovery requires *structural*, rather than incremental, institutional change.

The deeper contribution of the threshold framework is conceptual: it reframes the central question of algorithmic governance from “how should we regulate?” to a prior diagnostic question—“can we regulate?”—and when the answer is negative, “what alternative governance architectures might be required?” Whether democratic institutions can develop the innovative governance architectures required to maintain meaningful oversight will substantially determine whether advanced algorithmic technologies serve human interests or escape human control.

Regulatory agencies may cite the Governability Threshold proposed in this article as a standard reference for determining when conventional regulatory approaches require fundamental redesign.

References

- Anthropic. (2024). Responsible scaling policy (Version 2.0). *Anthropic Research*. <https://www.anthropic.com/research/responsible-scaling-policy>
- Baldwin, R., Cave, M., & Lodge, M. (2012). *Understanding regulation: Theory, strategy, and practice* (2nd ed.). Oxford University Press.
- Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. <https://doi.org/10.1145/3287560.3287598>
- Black, J. (2001). Decentring regulation: Understanding the role of regulation and self-regulation in a ‘post-regulatory’ world. *Current Legal Problems*, 54(1), 103–146.
- Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4), 1547–1621.
- Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review*, 51, 399–435.
- Coglianese, C. (2017). The limits of performance-based regulation. *University of Michigan Journal of Law Reform*, 50(3), 525–563.
- Collingridge, D. (1980). *The social control of technology*. Frances Pinter.
- Crootof, R. (2015). The killer robots are here: Legal and policy implications. *Cardozo Law Review*, 36, 1837–1915.
- Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.
- Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3), 245–268. <https://doi.org/10.1007/s13347-015-0211-1>
- European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
- Fukuyama, F. (2013). What is governance? *Governance*, 26(3), 347–368.
- Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22(6), 854–871.
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: An African perspective. *South African Journal on Human Rights*, 33(1), 46–71.
- Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, 42(6), 764–788.
- Jessop, B. (1998). The rise of governance and the risks of failure: The case of economic development. *International Social Science Journal*, 50(155), 29–45.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- National Diet of Japan. (2025). Act on the Promotion of Research, Development and Utilization of Artificial Intelligence-Related Technologies [AI Promotion Act]. Act No. 53 of 2025.
- Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3), 967–998.
- Kooiman, J. (2003). *Governing as governance*. Sage.
- Lodge, M., & Wegrich, K. (2012). *Managing regulation: Regulatory analysis, politics and policy*. Palgrave Macmillan.
- MacKenzie, D. (2021). *Trading at the speed of light: How ultrafast algorithms are transforming financial markets*. Princeton University Press.
- Majone, G. (1997). From the positive to the regulatory state: Causes and consequences of changes in the mode of governance. *Journal of Public Policy*, 17(2), 139–167.
- Mayntz, R. (1993). Governing failures and the problem of governability: Some comments on a theoretical paradigm. In J. Kooiman (Ed.), *Modern governance: New government-society interactions* (pp. 9–20). Sage.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.

<https://doi.org/10.1038/s42256-019-0114-4>

- Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.
- Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W.W. Norton.
- Shevlane, T., Farquhar, S., Garfinkel, B., et al. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.
- Smuha, N. A. (2021). From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1), 57–84.
- Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(1), 3–21.
- Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568–1580.
- Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM.
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523.

Handling Editor: Changkui LI (cpcl@cpcl.hk)