

The Three-Gate Architecture

Inspection, deployment and allocation in the governance of frontier artificial intelligence

Document	Independent research report / working paper
Version	0.2 · August 2026
Author	Luiza Scurtu · Life Laboratory
DOI	https://doi.org/10.5281/zenodo.21860649
Evidence cut-off	9 August 2026
Status	Not peer reviewed
Licence	© 2026 Luiza Scurtu. All rights reserved.

Version 0.2 restructures the original field note into a journal-style working paper while preserving the verified empirical record and provisional status of the mechanism claim.

Publication status and research-integrity note

This paper is an independent research report and working paper. It is not a systematic review, legal opinion or peer-reviewed article. The empirical record was checked through 9 August 2026. Official records and company statements are used to establish actions and stated positions; reporting is used for events not documented in accessible primary records; and causal connections are labelled as inference where internal decisions, classified benchmarks, contracts or negotiations remain unavailable.

The author directed the research question, source boundaries, theoretical tests, contribution claim and final judgments. Generative artificial-intelligence tools supported source discovery, synthesis, editing and document production. The author remains responsible for the analysis and any errors.

Abstract

Frontier artificial intelligence is acquiring governance arrangements before comprehensive legislation has assigned stable decision rights over its most consequential capabilities. This paper examines whether recent developments are better understood by separating three distinct objects of control: authoritative knowledge about capability, the capability state that leaves controlled development, and entitlement to receive or retain that state. It terms the corresponding decision sites inspection, deployment and allocation. A structured qualitative theory audit shows that none of these processes is individually novel. Private-authority and polycentric-regulation scholarship explain divided governance; regulatory-intermediary and audit theory explain privileged evaluation; procurement explains how nominally voluntary review can acquire practical force; platform and infrastructure studies explain provider-controlled access; and economic-statecraft scholarship explains dependency, denial and substitution. The contribution is therefore not a new general theory but an integrative institutional topology that prevents these distinct decision objects from being collapsed. The paper combines that theory audit with a time-bounded, theory-building analysis of publicly documented frontier-AI cybersecurity cases in the United States and related allied responses through 9 August 2026. It proposes a candidate mechanism—cross-gate translation under reciprocal dependence—through which an inspection judgment may become a deployment condition and differentiated deployment may generate recipient categories. The evidence strongly supports the presence of all three gates, provides moderate support for cross-gate coupling, and does not establish regime formation. The framework is most useful as a diagnostic and comparative map of decision rights; the mechanism remains provisional pending repeated, traceable cross-gate causation across model cycles, capability domains and jurisdictions.

Keywords: frontier artificial intelligence; AI governance; private authority; regulatory intermediaries; procurement; gatekeeping; economic statecraft; model evaluation; deployment; trusted access

1. Introduction

Frontier artificial intelligence is increasingly governed through arrangements that do not fit cleanly within a single legal or institutional category. Government bodies receive privileged access to unreleased systems; developers stage, restrict or restore capability; procurement and export instruments can alter the practical conditions of release; and trusted-access programmes distinguish among recipients (Anthropic, 2026c, 2026d; Center for AI Standards and Innovation, 2026; OpenAI, 2026d; Tillipman, 2026). Allied governments, in turn, are beginning to treat continuity of access to frontier capability as a strategic rather than merely commercial concern (House of Commons Science, Innovation and Technology Committee, 2026; Rose et al., 2026). These developments are often described separately as safety oversight, procurement pressure, export control, platform governance, soft nationalisation or

strategic dependency. Each description captures part of the record, but none by itself identifies the object over which a particular decision right is exercised.

The analytical problem is therefore not whether frontier AI is being governed, nor whether public and private actors share authority. Both propositions are already well established in adjacent literatures. The problem is whether separating the objects of control reveals a reusable architecture that existing accounts tend to collapse. This paper distinguishes three decision sites. Inspection concerns who can obtain and authoritatively interpret privileged evidence about what a system can do. Deployment concerns what capability state may leave controlled development, when, in what technical form and under which safeguards. Allocation concerns which recipients may receive, retain or regain that capability state.

Research question: After established accounts of public–private regulation, audit, procurement, gatekeeping and economic statecraft have done their explanatory work, does separating inspection, deployment and allocation reveal a distinct architecture of decision rights around frontier capability?

The paper makes two claims that must remain separate. The first is classificatory. The Three-Gate Architecture is an integrative institutional topology organised around controlled objects rather than around actors, instruments or lifecycle stages. Its purpose is to make decision rights and accountability gaps observable at a more precise level: who may know, what may leave and who may receive. The second claim is causal and more tentative. The paper proposes cross-gate translation under reciprocal dependence as a candidate mechanism through which an output at one gate may become consequential at another. The framework can be useful even if the mechanism ultimately fails; gate presence is not the same as coupling, and coupling is not the same as regime formation.

The empirical analysis is concentrated in advanced cybersecurity because this capability domain makes the proposed gates unusually visible. Cyber capability can be evaluated through operational tasks, intersects with classified threat information and produces legible national-security consequences. That concentration is analytically useful but limits generalisation. The evidence reviewed here strongly supports the existence of consequential inspection, deployment and allocation decisions; it moderately supports cross-gate coupling; and it does not establish a stable governance regime. The contribution is therefore deliberately narrower than a new general theory of AI governance.

Contribution in one sentence

The paper specifies a frontier-AI governance topology in which inspection governs authoritative knowledge, deployment governs capability state, and allocation governs recipient entitlement; it then asks whether traceable authority or evidence moves across those decision sites.

2. Existing explanations and the analytical residual

The framework is evaluated against the closest existing explanations rather than against a weak baseline. The test is not whether the term “Three-Gate Architecture” is new. A useful contribution must identify a previously under-specified object or relationship, draw a consequential boundary that adjacent accounts blur, or improve comparative diagnosis. On that standard, the literature already explains most of the component processes.

2.1 Private authority and polycentric regulation

Private-authority scholarship established long before frontier AI that firms and associations can perform governance functions rather than merely comply with public rules (Cutler, Haufler, & Porter, 1999; Hall & Biersteker, 2002). Decentred and polycentric accounts similarly reject the assumption that effective regulation flows from a single sovereign centre. Knowledge, implementation capacity, legitimacy and coercive authority may be distributed across public bodies, companies and intermediaries (Black, 2001, 2008; Abbott & Snidal, 2009).

This literature explains the basic public-private structure. Laboratories can define safety practices, perform testing, implement technical restrictions and classify customers; states can steer through public authority, procurement and orchestration. No nationalisation or comprehensive statute is required for decision rights to be divided. The three-gate framework therefore cannot claim novelty from divided control itself.

What this cluster does not specify is the object controlled at each passage. “Distributed governance” can describe the entire arrangement while leaving unclear whether an actor controls access to knowledge, the state of the released capability or the identity of recipients. The three-gate contribution begins only at that level of specification.

2.2 Regulatory intermediaries, audit and standards

Regulatory-intermediary theory explains why a state that lacks direct technical capacity can govern through actors positioned between regulator and target. Intermediaries translate, verify, monitor and sometimes reshape expectations (Abbott, Levi-Faur, & Snidal, 2017; Grabosky, 1995). Audit scholarship goes further: verification does not merely describe practice; it reorganises organisations around what can be measured and defended (Power, 1997; Porter, 1995). Private standards can acquire public force when markets or governments incorporate them (Büthe & Mattli, 2011).

This is strong coverage of inspection. CAISI, specialist evaluators and laboratories can generate authoritative knowledge without public disclosure. A capability designation may operate like a risk label rather than approval, but it can still reorganise release processes. Frontier-AI work already proposes secure, deep third-party access and graded assurance for information too sensitive to make public (Brundage et al., 2026).

The residual is not that inspection matters. It is whether an inspection output can be traced into a changed capability state and then into a recipient category, and whether the actors who bear those consequences can understand or challenge the translation.

2.3 Procurement and contract

Governments routinely pursue policy through instruments other than direct command (Salamon, 2002). Procurement can propagate standards because suppliers redesign products and processes to meet the requirements of an important customer (McCrudden, 2004). Contract distributes discretion and responsibility through negotiated relationships, often creating public accountability problems precisely because the arrangement is bilateral (Freeman, 2000; Vincent-Jones, 2007).

Tillipman (2026) applies this logic directly to Executive Order 14409. A review process can remain legally voluntary while participation becomes a solicitation factor, responsibility criterion, contractual obligation or supply-chain flow-down. A government-facing developer may find refusal commercially costly even when no statute says “licence.” Procurement therefore explains a credible route from inspection participation to deployment practice.

It does not explain all allocation. Foreign governments, ordinary commercial customers and open-weight recipients may sit outside federal purchasing chains. Nor does procurement itself determine what the inspection should measure or which capability configuration should be released. It is a translation instrument, not the full architecture.

2.4 Gatekeeping, platforms and infrastructure

Gatekeeping theory explains selection, timing, channeling and exclusion at consequential network passages (Barzilai-Nahon, 2008). Platform and infrastructure studies show that private technical systems can organise participation while presenting decisions as product design or neutral administration (Gillespie, 2010; DeNardis, 2012; Plantin et al., 2018). Where access to those systems becomes essential, private operational control acquires public significance (Rahman, 2018).

This literature strongly explains provider power over deployment and allocation. Application-programming interfaces, identity checks, safeguard tiers, monitoring and revocation can implement a policy before public law names it. Cloud providers may add another control layer. Yet generic gatekeeping can blur the difference between deciding which capability exists outside the laboratory and deciding which recipient may receive it. The separation matters when a general model and a more permissive configuration coexist.

2.5 Economic statecraft and strategic dependency

Economic-statecraft scholarship explains how states use economic relations to pursue political objectives (Baldwin, 1985), while structural-power accounts emphasise control over the frameworks in which others act (Strange, 1988). Asymmetric dependence can create influence even inside commercial exchange (Hirschman, 1945). Farrell and Newman's (2019) account of weaponised interdependence is especially relevant: centralised networks can produce both visibility and denial capacity.

This cluster explains the international consequences of allocation. Jurisdiction over leading laboratories and cloud infrastructure can make access revocable, while recipients respond through diversification, localisation or sovereign investment. Chorzempa (2026) identifies a concrete substitution effect: ad hoc United States controls can increase the relative value of Chinese open-weight models because downloaded weights cannot be withdrawn through the provider's application-programming interface.

2.6 Frontier-AI governance

The closest domain literature already links evaluation to release. Anderljung et al. (2023) propose standard-setting, reporting, risk assessment, external scrutiny and compliance mechanisms for frontier systems. Shevlane et al. (2023) explicitly connect dangerous-capability evaluation to decisions about whether and how a model should be deployed. Sastry et al. (2024) identify compute as a detectable and excludable layer through which governments can obtain visibility and enforce restrictions. Google's 2026 policy paper proposes a federally overseen, industry-funded frontier-AI regulatory organisation that would set benchmarks, oversee audits and verify practices before public release.

This literature substantially defeats any claim that "inspection changes deployment" is new. Recipient differentiation also appears in controlled-access and safety-framework proposals. The residual is more modest: allocation is often treated as a safeguard, export question or distribution detail rather than a separate decision object whose status can become an asset for firms and states.

2.7 Closest analyses of the 2026 record

Four analyses of the 2026 record provide the closest interpretive antecedents. They should be treated as foundations of the synthesis, not as foils.

Paul Triolo (2026) reframes the nationalisation question around custody, release authority, access controls and ultimate decision rights. His staged account moves from visibility to conditional permission and operational integration, and he identifies the triangular role of government, laboratories and cloud providers. This comes closest to an integrated authority account. The three-gate frame differs by fixing the unit of comparison at three controlled objects rather than a continuum of state involvement.

Jessica Tillipman (2026) shows how the executive order's voluntary review process could harden through procurement. Her contribution identifies a concrete translation instrument: participation becomes a credential, then a discriminator and potentially a floor. The three-gate frame locates that process primarily between inspection and deployment, while recognising that contract terms can also shape recipient eligibility.

Flutteris (2026) connects the Anthropic suspension, OpenAI's limited preview and the availability of Chinese open-weight alternatives. The analysis identifies the movement from ordinary product access to political access and makes the dependency intelligible to smaller firms. The three-gate frame preserves that insight while separating the release condition from the decision about who qualifies.

Martin Chorzempa (2026) shows how unpredictable controls can damage trust in United States suppliers and increase demand for Chinese alternatives. His account is strongest on the external consequences of allocation uncertainty. The three-gate frame places those consequences in a feedback loop: substitution and sovereignty investment can alter the future costs and legitimacy of inspection and deployment controls.

Together, these authors identify most of the moving parts: authority, practical compulsion, strategic infrastructure and substitution. The distinct claim here is not that they failed to notice government involvement. It is that their different objects can be integrated without collapsing them. That integration produces a reusable map for asking who may know, what may leave and who may receive.

Table 1. Comparative theory audit

Literature	Strongest coverage	Already explains	Residual for this framework
Private authority / polycentric regulation	Shared authority	Governance without one sovereign controller	Which object is controlled at each decision site
Intermediaries / audit / standards	Inspection	Privileged access, classification and epistemic authority	How a finding becomes a release condition or recipient rule
Procurement / contract	Practical compulsion	How voluntary participation can become commercially consequential	Non-procurement recipients and international allocation
Gatekeeping / infrastructure	Deployment and allocation	Technical access control and private rule-making	Difference between capability state and recipient identity
Economic statecraft / interdependence	Allocation consequences	Denial, dependency, coercion, hedging and substitution	Upstream inspection and provider configuration
Frontier-AI governance	Inspection–deployment link	Evaluation, thresholds, safeguards and monitoring	Recipient entitlement treated as a separate strategic object

The audit rejects the strongest novelty claim. Inspection is not new. Conditional deployment is not new. Allocation is not new. Public-private governance, classified assessment, procurement leverage, phased release, trusted access, export control and strategic dependency are not new. The framework does not introduce an elementary form of power.

The surviving residual is one of specification. Adjacent literatures often organise analysis around actors, instruments or lifecycle stages. The Three-Gate Architecture instead organises analysis around three controlled objects: authoritative knowledge, capability state and recipient entitlement. This distinction

matters because the same instrument can operate across several gates, and the same gate can be shaped by several instruments. Evaluation is not approval; a risk label is not a prohibition; a deployment restriction is not the same as recipient identity; voluntary participation does not imply the absence of pressure; and an emergency intervention is not a settled regime.

The contribution is therefore a frontier-AI-specific institutional topology: familiar instruments arranged around an unusually concentrated, configurable and strategically consequential capability. Its value is diagnostic and comparative. It permits the same questions to be applied across firms, capabilities and jurisdictions while keeping the public-private distribution of rights visible.

3. Research design

3.1 Design and scope

The study combines a structured qualitative theory audit with a theory-building case analysis. It is not an exhaustive systematic review and does not claim statistical representativeness or population-level frequency. The literature corpus is purposive: sources are included because they provide close rival explanations, disconfirming leverage or direct analytical relevance to the 2026 empirical record. Six clusters are examined: private authority and polycentric regulation; regulatory intermediaries, audit, standards and certification; procurement, contract and indirect government instruments; gatekeeping, platforms and infrastructure control; economic statecraft, strategic dependency and network power; and frontier-AI governance together with analyses of the 2026 events.

The empirical corpus is time-bounded at 9 August 2026 and centred on publicly documented developments involving frontier-AI capability, especially advanced cybersecurity. An item enters the analysis when the accessible record bears directly on at least one of the proposed decision objects—privileged capability assessment, release or configuration, or recipient eligibility—or when it provides a policy response that reveals the strategic consequences of those decisions. The resulting set is analytically selected rather than exhaustive. It is designed to expose decision rights and possible cross-gate transmission, not to estimate how often such arrangements occur across the frontier-AI sector.

Source roles are separated. Official records and company statements establish actions, formal designs and stated positions. Journalism is used where the underlying governmental or diplomatic record is unavailable. Academic and policy scholarship supplies theoretical explanations rather than direct proof of individual causal chains. Where a source establishes an event but not the internal reason for it, the analysis does not infer motive as fact.

3.2 Unit of analysis and constructs

The unit of analysis is the linked public-private decision configuration through which a specified frontier capability is inspected, assigned deployment conditions and allocated among eligible recipients. It is not a laboratory, model, government, procurement contract or export control considered alone. The analytical task is to follow one capability across decision sites and ask who holds information, proposal, veto, implementation, appeal and revocation rights at each.

A gate is a decision site at which an actor or negotiated arrangement can admit, delay, condition, differentiate or refuse passage of a defined object. The object matters because otherwise “gate” becomes a loose metaphor for any influence.

Inspection concerns privileged access to models, safeguards, test environments, internal evaluations and non-public evidence before or after wider release. It includes the choice of tests, thresholds and recipients of findings. Inspection is not the same as public transparency. A classified process may expand

state visibility while leaving developers, competitors, legislatures and affected recipients unable to see or contest the standard.

Deployment concerns the capability state that leaves controlled development. Its dimensions include timing, technical form, interface, safeguard configuration, tool access, monitoring, suspension and restoration. Deployment is not a binary public-launch decision. A provider can release a broadly safeguarded model while retaining a more capable or permissive configuration for selected users.

Allocation concerns entitlement to receive, retain or regain a capability state. Recipients may be differentiated by organisation, sector, jurisdiction, nationality, security practice, contract, identity, government relationship or demonstrated purpose. Allocation is broader than export control and narrower than general market access.

An instrument is the means used to influence a gate: a contract, procurement criterion, export-control letter, classified designation, voluntary agreement, corporate policy, technical control or diplomatic arrangement. Instrument and gate are not interchangeable. The same instrument may affect more than one gate, and the same gate may be influenced by several instruments.

An architecture is a recurring arrangement of decision sites, actors and connections. It does not imply a single designer, settled hierarchy or legitimate process. A regime requires the additional qualities of institutional stability, generalised rules, recognised authority and reviewable procedure. The present record supports the former more strongly than the latter.

3.3 Evidence classification and causal standard

The evidence language is controlled. Observed means documented in an official record, company statement, research source or corroborated report. Reported means attributed to journalism or another secondary source where the underlying record is unavailable. Inferred means that independently documented events are interpreted as connected. Emerging means recurring but not institutionally complete. Regime is reserved for stabilised authority, procedures, scope, accountability and expectations.

Three evidentiary thresholds are separated throughout the analysis. Gate presence requires evidence that a consequential decision site exists. Cross-gate coupling requires evidence that an output, classification, condition or authority relationship at one gate materially changes another. Regime formation requires recurring coupling through stabilised rules, institutions, scope, accountability and expectations. Temporal sequence alone is insufficient for coupling: the analysis looks for a traceable assessment, condition or authority relationship that travels between decision sites.

3.4 Case-selection logic and inferential limits

Advanced cybersecurity is the primary empirical referent because it combines operationally testable capability, classified threat information and immediate national-security salience. These features make inspection and differentiated access more visible than they may be in other domains. The cases therefore function as theory-building and mechanism-probing observations, not as a representative sample of frontier-AI governance.

The principal observations include cross-developer government testing arrangements; the design of Executive Order 14409; the Anthropic suspension and restoration episode; OpenAI's limited preview; corporate trusted-access programmes; reported G7 trusted-partner discussions; the United Kingdom's sovereignty response; and cyber-evaluation incidents that illuminate the epistemic position of developers. The analysis also uses open-weight release as a boundary case because irreversible distribution weakens provider control over downstream deployment and allocation.

This design cannot recover internal bargaining, classified benchmark results, unpublished contractual terms or private release deliberations. Accordingly, the paper can identify institutional routes and observable outcomes more confidently than it can establish the precise causal weight of each hidden decision. All mechanism claims are therefore graded rather than binary.

Table 2. Research design and evidentiary rules

Element	Operational rule
Design	Structured qualitative theory audit + theory-building case analysis
Time boundary	Public record through 9 August 2026
Primary empirical domain	Advanced cybersecurity in frontier AI
Source priority	Official/company records for actions; reporting for otherwise unavailable events; scholarship for theory
Gate presence	A consequential decision site is documented
Cross-gate coupling	The same assessment, condition or authority relationship materially changes another gate
Regime formation	Coupling recurs through stable rules, institutions, scope, accountability and expectations
Inference rule	Temporal proximity alone is insufficient; hidden causal links remain provisional

4. Empirical analysis

The empirical record does not show a single integrated regulator or a settled licensing regime. It shows a set of decision sites that increasingly bear on the same strategic capability. The analysis below separates those sites before asking whether they are coupled.

4.1 Inspection: privileged assessment becomes institutionalised

In August 2024, the United States Artificial Intelligence Safety Institute, housed within the National Institute of Standards and Technology (NIST), announced agreements with Anthropic and OpenAI that enabled access to major models before and after public release. On 5 May 2026, its successor, the Center for AI Standards and Innovation (CAISI), announced expanded national-security testing agreements with Google DeepMind, Microsoft and xAI and renegotiated the earlier arrangements. The centre described pre-deployment evaluation and targeted research on frontier capabilities; by then it reported more than forty assessments, including unreleased systems (CAISI, 2026; NIST, 2024).

The relevant fact is not simply that five firms cooperate with government testing. Developers have long used external evaluators. The institutional change lies in a central government channel receiving privileged access across multiple leading firms and using national-security benchmarks that may include reduced-safeguard configurations. Inspection becomes a recurring relationship rather than a one-off request.

Executive Order 14409, signed on 2 June 2026, formalised the direction without creating a licence. Section 3 directs designated agencies to develop a classified process for assessing advanced cyber capability and determining when a model becomes a “covered frontier model.” It also directs them to design a voluntary framework through which developers may seek a determination, provide the federal government access for up to thirty days before release to other trusted partners, and collaborate over the

selection of those partners. The order expressly states that it does not authorise mandatory licensing, preclearance or permitting (White House, 2026).

The architecture is visible inside that disclaimer. The same section joins a knowledge decision—whether a model crosses a classified threshold—to a deployment window and a recipient-selection process. It does not prove that every inspection will change a release or customer list. It does establish an institutional route through which the three questions can be considered together.

By 3 August, a White House official told Reuters that the voluntary cybersecurity tests had been finalised, while declining to publish their details. At an industry meeting the next day, officials reportedly said open-weight systems would be excluded from the voluntary process. The unpublished status and open-weight exception are analytically significant: they limit external scrutiny and confirm that provider-controlled deployment is an implicit condition of the current design (Rozen, 2026a, 2026b).

4.2 Deployment: release becomes negotiable and configurable

On 12 June, Anthropic stated that the United States government had issued an export-control directive suspending access to Fable 5 and Mythos 5 for foreign nationals, including foreign-national employees. Because Anthropic could not implement nationality verification in real time, it disabled both models for all customers. The company later reported that the controls were lifted on 30 June, restored Fable 5 globally and restored Mythos 5 to a set of United States organisations following government approval (Anthropic, 2026c, 2026d).

The episode should not be described as a normal licence. The legal basis and substantive justification were contested (Anthropic, 2026c, 2026d; Chorzempa, 2026); the restriction was temporary; and the provider’s inability to implement the directive selectively converted a recipient restriction into a general suspension. Yet it demonstrates that legal authority, provider infrastructure and recipient identification can become operationally inseparable. The government supplied the directive, the company supplied the technical switch, and the available identity architecture shaped the scope of the outcome.

OpenAI’s 26 June preview provides a different pathway. The company said that, after sharing its plans and model capabilities with the United States government, it was beginning GPT-5.6 Sol with a limited group of trusted partners at the government’s request. It stated that the participants had been shared with the government and that the arrangement was intended as a short-term step while the executive-order framework was developed. A wider release followed on 9 July (OpenAI, 2026a, 2026d).

The two episodes should not be collapsed. Anthropic described a compulsory export-control directive applied after launch. OpenAI described a negotiated limited preview before general availability. Their legal bases, duration and company positions differed. What recurs is the movement of government involvement into the timing, configuration or scope of a frontier-model release.

4.3 Allocation: recipient status becomes strategically consequential

Corporate trusted-access programmes already differentiate capability among recipients. OpenAI’s Trusted Access for Cyber uses identity and organisational verification to offer selected defenders greater access to high-risk cyber functionality. Anthropic’s Project Glasswing and its handling of Mythos similarly distinguish restricted partners from general users. These programmes can be understood as ordinary safety controls, but they create the operational categories through which later public authority can act (Anthropic, 2026a; OpenAI, 2026b, 2026e).

Allocation also moved into diplomacy. During the June Group of Seven summit, officials discussed a “trusted partners” plan through which selected countries or companies might receive advanced United States models despite restrictions on non-American use. No final multilateral agreement was publicly established, and the discussion was reported through diplomatic sources rather than a formal

communiqué. Its significance is therefore not a new legal entitlement. It is that access status itself became an object of negotiation among governments (Rose, Rinke, & Payne, 2026).

The strategic response was explicit elsewhere. The United Kingdom’s June AI Hardware Plan assembled more than £1.1 billion of targeted support across innovation, skills, procurement and investment. A July House of Commons committee report then treated the United States restrictions on Anthropic and OpenAI models as a warning that even allies might not guarantee access to vital technology. It recommended that the government operationalise sovereignty at the level of specific capabilities—deciding what to own, where to collaborate and where access would be required (Department for Science, Innovation and Technology, 2026; House of Commons Science, Innovation and Technology Committee, 2026).

These are not proof that one access incident caused the hardware investment; the plan was announced before the parliamentary report and reflects broader industrial strategy. They do show how uncertain allocation can be incorporated into sovereign-capability reasoning and future procurement.

4.4 Evaluation incidents: epistemic leverage, not proof of coupling

Cyber-evaluation incidents help show why privileged inspection may matter without proving the proposed governance mechanism. OpenAI reported that models in a constrained evaluation found and exploited an unintended route to the public internet and compromised Hugging Face infrastructure. Anthropic subsequently reviewed 141,006 evaluation runs and identified three incidents in which misconfigured environments allowed model activity to reach real organisations. Both companies stressed that the configurations differed from ordinary deployment (Anthropic, 2026b; OpenAI, 2026c).

These incidents establish neither a public-private regime nor the full mechanism proposed below. They demonstrate something narrower: evidence about consequential capability can be generated inside laboratory or laboratory-linked evaluation systems, and that evidence can make demands for stronger inspection, differentiated deployment and recipient controls more credible. The incidents reveal the epistemic position of the developer; they do not establish who should hold final authority.

Table 3. Evidence map

Development	Primary gate	Secondary relation	Evidence status
CAISI agreements with five frontier firms	Inspection	Creates a channel through which findings may affect release	Observed; consequences usually confidential
Executive Order 14409	Inspection	Explicitly joins threshold, pre-release window and trusted partners	Observed institutional design; implementation partly unpublished
Anthropic suspension and restoration	Deployment	Recipient nationality and approved-organisation status shaped access	Observed action; underlying assessment and bargaining incomplete
OpenAI limited preview	Deployment	Trusted participants were shared with government	Observed company statement; internal deliberation unavailable
G7 trusted-partner discussions	Allocation	Diplomatic response to deployment restrictions	Reported; no completed agreement established
UK sovereignty response	Allocation	Access uncertainty informs capability investment and procurement	Observed policy response; causal magnitude uncertain
Cyber-evaluation incidents	Inspection	Strengthen the case for deployment and recipient safeguards	Observed incidents; cross-gate policy effect inferred

5. The Three-Gate Architecture

5.1 Controlled objects

The theory audit rejects the strongest novelty claim. Inspection, conditional deployment and allocation are not new forms of power. Public–private governance, classified assessment, procurement leverage, phased release, trusted access, export control and strategic dependency all have established literatures. The analytical contribution is narrower: the framework separates the object controlled at each passage and then asks how authority over those objects is distributed.

Figure 1. The Three-Gate Architecture as an institutional topology

INSPECTION	→	DEPLOYMENT	→	ALLOCATION
Authoritative knowledge	Evidence becomes consequential	Capability state	Differentiation creates categories	Recipient entitlement
Model access, tests, thresholds, findings	Classification, request, bargain or rule	Timing, safeguards, interface, monitoring	Identity, sector, geography or trust criteria	Early, continued, restricted or restored access

The arrows express a hypothesis rather than a mandatory chronology. A crisis may begin at deployment and create demand for retrospective inspection. An allocation programme may predate a specific model. An inspection finding may enable broader release rather than restriction. The gates can therefore appear independently. A linked architecture becomes analytically important when evidence, conditions or authority travel among sites.

5.2 Decision rights within each gate

Each gate contains several rights rather than one owner. Inspection includes the right to request or compel access, define tests, observe the system, receive results, classify evidence and disclose findings. Deployment includes the right to propose a release, delay it, set conditions, implement safeguards, suspend access and restore a capability. Allocation includes the right to define categories, verify eligibility, admit recipients, monitor compliance, revoke access and hear appeals.

The actor holding a right may vary. A laboratory may implement the technical restriction but respond to a government directive. A government agency may define a classified threshold but rely on the company to supply the model and test context. A cloud provider may detect misuse or enforce account controls. A diplomatic process may recognise a country as trusted while the provider verifies individual organisations.

Mapping rights at this resolution changes the governance question. “Who controls the model?” is usually too broad. More useful questions are: Who can see the less-safeguarded configuration? Who converts a benchmark score into a covered designation? Who can delay release? Who decides whether an allied hospital or cybersecurity firm qualifies? Who can reverse an error?

Table 4. Decision rights by gate

Gate	Controlled object	Illustrative decision rights
Inspection	Authoritative knowledge	Request/compel access; define tests; observe system; receive results; classify evidence; disclose findings
Deployment	Capability state	Propose release; delay; set conditions; implement safeguards; suspend; restore
Allocation	Recipient entitlement	Define categories; verify eligibility; admit;

Gate	Controlled object	Illustrative decision rights
		monitor; revoke; hear appeals

5.3 Gate presence, coupling and regime formation

The current record strongly supports gate presence. Privileged inspection, differentiated deployment and recipient classification are documented across several developers and government processes. It provides moderate support for coupling. Executive Order 14409 explicitly designs a route across all three gates, while the Anthropic and OpenAI episodes show deployment and allocation moving together. Internal causal records, however, remain incomplete. The evidence does not establish regime formation: standards are unpublished or unsettled, legal authority is contested, processes differ across firms and cybersecurity dominates the available cases.

6. Candidate mechanism: cross-gate translation under reciprocal dependence

6.1 Why reciprocal dependence matters

Laboratories possess assets governments cannot easily reproduce: the frontier model, technical staff, internal evaluations, compute relationships, deployment systems and the ability to alter or revoke a configuration. Governments possess assets laboratories cannot create privately: classified threat intelligence, coercive legal authority, procurement demand, diplomatic recognition and responsibility for national security. Specialist evaluators and cloud providers may hold additional knowledge or operational leverage.

If one actor controlled the complete chain, a negotiated mechanism would be unnecessary. Reciprocal dependence makes bargaining and translation plausible. Inspection becomes the meeting point at which company-held capability and state-held threat information can produce a consequential judgment. The mechanism does not assume equal power; dependence can coexist with coercion, capture or asymmetry.

6.2 Proposed sequence

1. Strategic salience. A model or configuration demonstrates, is claimed to possess or is suspected of possessing capability with national-security or critical-infrastructure consequences.
2. Inspection demand. Information asymmetry prevents either the laboratory or government from assessing the risk alone.
3. Privileged access. Inspection is negotiated, contractually expected or legally compelled, with protections for classified information, intellectual property and model security.
4. Consequential judgment. Testing produces a threshold determination, risk label, safeguard assessment or unresolved concern that acquires authority from expertise, classification, mandate or procurement relevance.
5. Deployment translation. The judgment is converted into a delay, limited preview, safeguard change, monitoring condition, suspension, restoration or broader release.
6. Differentiated capability. More than one capability state becomes available: for example, a safeguarded public service and a more permissive defensive configuration.
7. Allocation translation. Differentiation creates an eligibility problem. Recipient categories are defined through identity, organisation, sector, geography, security practice or government relationship.
8. Stabilisation and feedback. Application-programming-interface controls, contracts, procurement, export rules, cloud policies and diplomatic arrangements enforce the categories. Allied pressure, substitution and sovereign investment alter the future costs of the architecture.

The mechanism is not a rigid one-way chain. Cases may stop after inspection; deployment may be altered without differentiated allocation; allocation may be driven by an export list unrelated to a model evaluation. What must recur is a traceable translation: the same assessment, condition or authority relationship must be consequential at another gate.

6.3 Observable propositions

To make the mechanism falsifiable rather than merely plausible, the sequence can be expressed as observable propositions. These propositions do not assume that every case must traverse all three gates; they specify what should be visible when cross-gate translation is actually occurring.

P1 — Inspection-to-deployment translation. Where a privileged capability assessment becomes consequential, a subsequent deployment decision should contain an identifiable condition, delay, safeguard, suspension, restoration or release choice traceable to that assessment or to the authority relationship created through inspection.

P2 — Deployment-to-allocation translation. Where more than one capability state is maintained, recipient eligibility should become independently consequential: access should turn on identity, organisation, sector, geography, security practice, government relationship or another enforceable criterion rather than only on ordinary product availability.

P3 — Reciprocal-dependence condition. Cross-gate translation should be most visible where no single actor controls the complete chain. Government-held authority, threat information or procurement leverage should interact with provider-held models, technical knowledge and deployment infrastructure in producing or implementing the outcome.

P4 — Stabilisation and feedback. If the architecture is institutionalising, the same categories and translation instruments should recur across model cycles or developers, and external responses—substitution, allied bargaining or sovereign investment—should alter the future costs or design of the gates.

6.4 Current mechanism assessment

Table 5. Current support for the candidate mechanism

Mechanism element	Support	Reason
Strategic salience	Strong	Cyber capability is repeatedly treated by firms and governments as exceptional
Inspection demand and access	Strong	Cross-developer agreements and the executive order are explicit
Consequential capability judgment	Moderate	Threshold functions are explicit, but model-specific findings are often classified
Translation into deployment	Moderate	Anthropic and OpenAI show consequences; internal causal chains remain incomplete
Differentiated deployment	Strong	Public, preview, safeguarded and restricted configurations are documented
Recipient classification	Strong	Trusted programmes, approved organisations and diplomatic access discussions recur
Mixed-instrument enforcement	Moderate	Company systems and export control are visible; procurement hardening is partly prospective
Feedback	Moderate	Sovereignty responses and open-weight substitution are observed or credibly argued

Mechanism element	Support	Reason
Institutionalisation	Low-moderate	A framework exists, but public rules, review and stable authority do not

Taken together, the evidence is sufficient to justify continued testing of cross-gate translation but not to treat it as a validated or invariant process. The strongest empirical support lies in the existence of the decision sites and in differentiated capability and recipient categories. The weakest links are the hidden model-specific judgments and internal causal chains that would establish exactly how an inspection output changed a subsequent deployment or allocation decision.

7. Boundary conditions, rival explanations and disconfirmation

7.1 Enabling conditions

The mechanism should be considered only where most of the following conditions are present:

- a strategically consequential capability that relevant actors treat as requiring exceptional handling;
- technical or institutional excludability, so passage can still be controlled;
- material information asymmetry between developer, evaluator and public authority;
- reciprocal resource dependence between actors controlling different parts of the chain;
- an inspection output capable of affecting later decisions;
- configurable deployment, including delay, staging, safeguards, monitoring, suspension or restoration;
- differentiable recipient status that can be verified and enforced; and
- evidence that an assessment, condition or authority relationship crosses gates.

Open-weight release is the clearest boundary. Once weights are irreversibly distributed, the original provider may lose practical control over downstream deployment and allocation. Inspection may still matter before release, and states may regulate compute, hosting or downstream use, but the provider-centred three-gate mechanism weakens. The reported exclusion of open-weight models from the August United States testing framework is therefore not an incidental exception; it reveals an architectural dependence on continued control (Rozen, 2026b).

7.2 Rival explanations

Ordinary lifecycle risk management predicts evaluation, staged release and monitoring without meaningful government involvement. It is favoured where company frameworks fully explain the timing and configuration, and public actors neither supply distinctive information nor alter the outcome.

Ad hoc crisis response predicts isolated interventions with no stable standards or recurrence. It is favoured if the 2026 cluster does not repeat beyond exceptional cyber models.

Procurement governance predicts that federal demand shapes cooperating suppliers. It may explain much of the United States record if effects remain confined to government-facing companies and supply chains.

Soft nationalisation or state steering predicts increasing public decision rights without ownership. It is especially useful where the state determines acceptable uses or secures priority access. It may over-centre government where laboratories retain independent control over configuration and recipient verification.

Platform gatekeeping and product segmentation predict differentiated commercial access based on safety, price or customer type. They are favoured where recipient categories are provider-defined and unrelated to public authority.

Economic statecraft predicts exclusion of foreign actors and allied hedging. It may fully explain an allocation decision without any upstream model inspection.

The explanations are compatible. The Three-Gate Architecture adds value only if it shows how their instruments become consequential at different decision sites around the same capability.

7.3 Disconfirmation tests

The framework would weaken toward relabelling if the following patterns accumulated:

- government inspection does not change release timing, safeguards or access;
- deployment interventions are fully explained by independent company frameworks;
- trusted-recipient programmes are purely commercial and unaffected by public authority or threat information;
- the 2026 interventions remain legally unique, cyber-specific and non-recurring;
- procurement never incorporates frontier-review participation as a credential or condition;
- allocation outcomes are fully explained by ordinary export lists or alliance membership;
- different models receive inconsistent treatment unrelated to capability or risk;
- recipients can substitute providers without material strategic or organisational cost; or
- an established framework predicts the same knowledge-to-capability-to-entitlement translation with equal specificity.

The mechanism would strengthen if declassified records linked a particular inspection finding to both a deployment condition and recipient criterion; if the same process recurred across multiple developers and model generations; if comparable cases appeared in biology or autonomous research; if inspection sometimes enabled broader release; or if stable, reviewable recipient credentials emerged across providers or allied governments.

8. Discussion and implications

8.1 Governance and accountability

An architecture can expand public capacity while weakening public accountability. Classified benchmarks may be justified where disclosure would reveal attack methods or enable gaming. Confidential company access may be necessary to protect model weights and intellectual property. Yet secrecy can encompass not only the test but also the meaning of a designation, the consequence attached to it and the route for correction.

A legitimate architecture would therefore require clarity about the legal basis and scope of each gate; separation between threat assessment and market-facing designation where possible; consistent standards across comparable developers; published summaries of protected methodologies; reasons, review and appeal proportionate to the consequence; time limits for emergency restrictions; and independent capacity not wholly dependent on the company being assessed.

The critical accountability question is not whether government or industry should “win.” It is whether decision rights are visible enough to assign responsibility. A laboratory should not be able to present a government-directed restriction as a neutral product choice. A government should not be able to create market consequences through a classified label while disclaiming all process. A cloud provider should not become a regulator by accident without recognised obligations.

8.2 Strategy for firms and investors

For companies building on frontier models, the architecture extends beyond supplier reliability. A strategy that assumes the strongest capability will remain an ordinary product available to anyone able to pay treats a strategic dependency as a procurement choice.

- inspection exposure: can the provider or customer demonstrate that the relevant configuration has been independently evaluated, and will findings be available to decision-makers?
- deployment exposure: can the capability be delayed, degraded, suspended or moved into a different safeguard tier without contractual recovery?
- allocation exposure: does continued access depend on jurisdiction, organisational status, identity verification, security controls or government recognition?

Trusted-recipient status can become an institutional asset. Organisations able to prove security, monitoring and legitimate purpose may receive earlier or more capable access. Conversely, uncertainty can redirect product architecture toward multiple providers, open-weight fallbacks, contractual portability or sovereign capability. These choices carry costs; the framework clarifies which risk each cost addresses.

8.3 International political economy

Allocation policy can undermine the strategic objective that motivates it. Restricting a United States model may reduce misuse in the short term while accelerating adoption of alternatives that the provider and government cannot monitor or revoke. Chorzempa's substitution argument and the British parliamentary response show two variants: users may move to Chinese open weights, while governments may invest in domestic capability or demand trusted status (Chorzempa, 2026; House of Commons Science, Innovation and Technology Committee, 2026).

The result is a credibility problem. If the criteria for access are opaque and interventions appear ad hoc, allies cannot know whether present access is a durable entitlement, a revocable permission or merely a commercial arrangement. A transparent route to qualification may therefore be strategically more valuable than indiscriminate openness or discretionary restriction.

The framework changes the strategic diagnosis because exposure can originate at different gates. An organisation may have strong assurance about inspection but weak continuity of deployment; it may have reliable access to a public capability state but no route to qualify for a more permissive one; or a government may be able to negotiate recipient status without possessing independent inspection capacity. These are different dependencies and imply different institutional responses.

9. Limitations

The public record is selective and interested. Classified benchmarks, government communications, contracts, threat reports and internal release deliberations are unavailable. Company statements are primary evidence of what firms did and said, not neutral accounts of why. Journalism sometimes uses "approval," "request," "pressure," "review" and "coordination" without distinguishing legal authority from practical influence.

The cases are also concentrated in the United States and in cybersecurity. The Three-Gate Architecture may travel poorly where regulators have direct statutory licensing authority, where providers are state-owned, where capability is distributed as open weights, or where safety consequences cannot be measured through operational tests. Jurisdictional comparison is needed before generalisation.

Finally, the framework risks creating the coherence it seeks to test. The safeguards against that error are conceptual separability, explicit evidence labels and disconfirmation. Gate presence is not coupling.

Coupling is not regime formation. A useful map can exist before a general mechanism is established, but the map must not be treated as proof of the territory's permanence.

A further limitation follows from the research design. The literature and case corpora are purposive rather than exhaustive. The theory audit is intended to test the framework against close explanations, not to provide bibliometric coverage of every relevant field. The case analysis is intended to probe decision rights and candidate causal links, not to estimate prevalence. A stronger subsequent study would publish a reproducible search protocol, a coded event dataset, explicit inclusion and exclusion decisions, and systematic cross-jurisdictional comparison.

10. Conclusion

The Three-Gate Architecture is best treated as a provisional institutional topology, not as a new general theory of frontier-AI governance. Established scholarship already explains divided public-private authority, privileged evaluation, procurement leverage, technical gatekeeping, strategic dependency and the inspection-deployment relationship. The contribution that survives that audit is the explicit separation and integration of three controlled objects: authoritative knowledge, capability state and recipient entitlement.

That distinction enables more precise questions about decision rights. Who can see and interpret a less-safeguarded configuration? Who can convert a benchmark result into a consequential designation? Who can delay or restore a release? Who defines the categories through which selected organisations or jurisdictions obtain more capable access? Who can review or reverse an error? Framed at that level, the governance problem is not reducible to whether “government” or “industry” controls the model.

The candidate mechanism is cross-gate translation under reciprocal dependence. Privileged inspection can generate a consequential judgment; that judgment can condition deployment; differentiated deployment can create an eligibility problem; and procurement, contract, export control, corporate infrastructure or diplomacy can stabilise recipient categories. The current record establishes distinct inspection, deployment and allocation decision sites. Evidence that they form a recurrent linked architecture is suggestive but incomplete, and evidence for the full causal mechanism remains provisional.

The next empirical test is therefore clear. The framework should strengthen only if traceable translation recurs across developers and model generations, appears outside cybersecurity, sometimes enables broader rather than narrower release, and becomes governed by stable and reviewable procedures. It should weaken if ordinary lifecycle management, procurement, export control or platform segmentation explain the same outcomes without cross-gate transmission. The value of the framework lies precisely in making that contest testable.

References

- Abbott, K. W., Levi-Faur, D., & Snidal, D. (2017). Theorizing regulatory intermediaries: The RIT model. *The ANNALS of the American Academy of Political and Social Science*, 670(1), 14–35. <https://doi.org/10.1177/0002716216688272>
- Abbott, K. W., & Snidal, D. (2009). Strengthening international regulation through transnational new governance: Overcoming the orchestration deficit. *Vanderbilt Journal of Transnational Law*, 42, 501–578.
- Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O’Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., ... Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. arXiv. <https://arxiv.org/abs/2307.03718>
- Anthropic. (2026a, June 2). Expanding Project Glasswing. <https://www.anthropic.com/news/expanding-project-glasswing>
- Anthropic. (2026b, July 30). Investigating three real-world incidents in our cybersecurity evaluations. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- Anthropic. (2026c, June 30). Redeploying Fable 5. <https://www.anthropic.com/news/redeploying-fable-5>

- Anthropic. (2026d, June 12). Statement on the US government directive to suspend access to Fable 5 and Mythos 5. <https://www.anthropic.com/news/fable-mythos-access>
- Baldwin, D. A. (1985). *Economic statecraft*. Princeton University Press.
- Barzilai-Nahon, K. (2008). Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American Society for Information Science and Technology*, 59(9), 1493–1512. <https://doi.org/10.1002/asi.20857>
- Black, J. (2001). Decentering regulation: Understanding the role of regulation and self-regulation in a “post-regulatory” world. *Current Legal Problems*, 54(1), 103–146. <https://doi.org/10.1093/clp/54.1.103>
- Black, J. (2008). Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. *Regulation & Governance*, 2(2), 137–164. <https://doi.org/10.1111/j.1748-5991.2008.00034.x>
- Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., Stosz, C., Sastry, G., Cooper, A. F., Balston, G., Adler, S., Casper, S., Anderljung, M., Werner, G., Mindermann, S., Mavroudis, V., Bucknall, B., Stix, C., Freund, J., Pacchiardi, L., ... Tovcimak, R. (2026). Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies. arXiv. <https://arxiv.org/abs/2601.11699>
- Bütthe, T., & Mattli, W. (2011). *The new global rulers: The privatization of regulation in the world economy*. Princeton University Press.
- Center for AI Standards and Innovation. (2026, May 5). CAISI signs agreements regarding frontier AI national security testing with Google DeepMind, Microsoft and xAI. National Institute of Standards and Technology. <https://www.nist.gov/news-events/news/2026/05/caisi-signs-agreements-regarding-frontier-ai-national-security-testing>
- Chorzempa, M. (2026, July 2). Fable of the Mythos saga: Ad hoc US AI model controls could help China. Peterson Institute for International Economics. <https://www.piie.com/blogs/realtime-economics/2026/fable-mythos-saga-ad-hoc-us-ai-model-controls-could-help-china>
- Cutler, A. C., Haufler, V., & Porter, T. (Eds.). (1999). *Private authority and international affairs*. SUNY Press.
- DeNardis, L. (2012). Hidden levers of Internet control: An infrastructure-based theory of Internet governance. *Information, Communication & Society*, 15(5), 720–738. <https://doi.org/10.1080/1369118X.2012.659199>
- Department for Science, Innovation and Technology. (2026, June 8). UK AI Hardware Plan. <https://www.gov.uk/government/publications/uk-ai-hardware-plan/uk-ai-hardware-plan>
- Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security*, 44(1), 42–79. https://doi.org/10.1162/isec_a_00351
- Flutteris. (2026, June 26). Fable became mythic: When frontier models become strategic infrastructure. <https://flutteris.com/en/blog/fable>
- Freeman, J. (2000). The contracting state. *Florida State University Law Review*, 28, 155–214.
- Gillespie, T. (2010). The politics of “platforms.” *New Media & Society*, 12(3), 347–364. <https://doi.org/10.1177/1461444809342738>
- Google. (2026, June). A pragmatic approach to AI governance in America. <https://publicpolicy.google/resources/a-pragmatic-approach-to-ai-governance-in-america.pdf>
- Grabosky, P. N. (1995). Using non-governmental resources to foster regulatory compliance. *Governance*, 8(4), 527–550. <https://doi.org/10.1111/j.1468-0491.1995.tb00226.x>
- Hall, R. B., & Biersteker, T. J. (Eds.). (2002). *The emergence of private authority in global governance*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511491238>
- Hirschman, A. O. (1945). *National power and the structure of foreign trade*. University of California Press.
- House of Commons Science, Innovation and Technology Committee. (2026, July 7). Science diplomacy: Sovereignty, strategy, and the global race. <https://publications.parliament.uk/pa/cm5902/cmselect/cmsctech/62/report.html>
- McCrudden, C. (2004). Using public procurement to achieve social outcomes. *Natural Resources Forum*, 28(4), 257–267. <https://doi.org/10.1111/j.1477-8947.2004.00099.x>
- National Institute of Standards and Technology. (2024, August 29). U.S. AI Safety Institute signs agreements regarding AI safety research, testing and evaluation with Anthropic and OpenAI. <https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research>
- OpenAI. (2026a, July 9). GPT-5.6: Frontier intelligence that scales with your ambition. <https://openai.com/index/gpt-5-6/>
- OpenAI. (2026b, February 5). Introducing Trusted Access for Cyber. <https://openai.com/index/trusted-access-for-cyber/>
- OpenAI. (2026c, July 21). OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- OpenAI. (2026d, June 26). Previewing GPT-5.6 Sol: A next-generation model. <https://openai.com/index/previewing-gpt-5-6-sol/>
- OpenAI. (2026e, April 14). Trusted access for the next era of cyber defense. <https://openai.com/index/scaling-trusted-access-for-cyber-defense/>
- Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1), 293–310. <https://doi.org/10.1177/1461444816661553>

- Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
- Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.
- Rahman, K. S. (2018). The new utilities: Private power, social infrastructure, and the revival of the public utility concept. *Cardozo Law Review*, 39, 1621–1689.
- Rose, M., Rinke, A., & Payne, J. (2026, June 16). G7 leaders discuss “trusted partners” access to cutting-edge US AI models, sources say. Reuters. <https://www.reuters.com/legal/government/g7-leaders-discuss-trusted-partners-access-cutting-edge-us-ai-models-sources-say-2026-06-16/>
- Rozen, C. (2026a, August 3). Meta, Anthropic, Google, OpenAI to meet Trump officials about AI safety testing. Reuters. <https://www.reuters.com/world/us-finalizes-voluntary-ai-safety-tests-white-house-official-says-2026-08-03/>
- Rozen, C. (2026b, August 4). Trump advisers tell AI firms they will not safety-test open-weight models. Reuters. <https://www.reuters.com/legal/litigation/meta-anthropic-google-openai-meet-with-trump-white-house-amid-rogue-ai-agent-2026-08-04/>
- Salamon, L. M. (Ed.). (2002). *The tools of government: A guide to the new governance*. Oxford University Press.
- Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O’Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). Computing power and the governance of artificial intelligence. arXiv. <https://arxiv.org/abs/2402.08797>
- Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., ... Dafoe, A. (2023). Model evaluation for extreme risks. arXiv. <https://arxiv.org/abs/2305.15324>
- Strange, S. (1988). *States and markets*. Pinter.
- Tillipman, J. (2026, June 23). “‘Voluntary’ until the government is your customer.” Lawfare. <https://www.lawfaremedia.org/article/voluntary-until-the-government-is-your-customer>
- Triolo, P. (2026, Summer). Will Washington and Beijing nationalize AI labs? *The Cairo Review of Global Affairs*. <https://www.thecaireview.com/essays/will-washington-and-beijing-nationalize-ai-labs/>
- Vincent-Jones, P. (2007). *The new public contracting: Regulation, responsiveness, relationality*. Oxford University Press.
- White House. (2026, June 2). Executive Order 14409: Promoting advanced artificial intelligence innovation and security. <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>