Standards Body · Foundation paper, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/foundation-paper/progressive-standards-and-requirements/

FOUNDATION_06_PROGRESSIVE_STANDARDS_AND_REQUIREME

Foundation 6: Progressive Standards and Requirements

Series: Foundations for Frontier AI Evaluation Infrastructure
Version: 1.0
Status: Canonical working white paper
Project: Standards Body
Primary domain: standardsbody.ai
Core line: Foundations for Frontier AI
Research basis reviewed through: July 16, 2026
Document owner: Standards Body
Review cycle: Annual review, with event-triggered revision after material changes in frontier capabilities, evaluation maturity, standards, market practice, or law

---

Document Purpose

This paper defines the Standards Body position on how frontier AI evaluation practices may responsibly develop from research and voluntary commitments into progressively more formal standards, assurance expectations, contractual duties, procurement requirements, and legal obligations.

It is intended to serve as:

- A first-principles explanation of why institutional requirements should mature in stages
- A framework for deciding when voluntary practice is sufficient and when stronger requirements are justified
- A guide to the transition from emerging evaluation methods to standards and conformity assessment
- A model for proportionality, phase-in, testing, review, enforcement, revision, and retirement
- A bridge between technical evidence, private ordering, standards development, and public law
- A reference for future standards, governance frameworks, working groups, partnerships, and policy analysis
- A durable source document from which shorter articles, implementation guides, and formal proposals can be developed

This paper is not legislation.

It does not propose that every voluntary practice should become mandatory.

It does not establish a universal regulatory threshold or enforcement authority.

It defines the conditions under which progressively stronger requirements may become justified, workable, and legitimate.

Executive Summary

Frontier AI institutions face two opposite risks.

The first is acting too late.

A consequential capability may emerge before evaluation methods, independent reviewers, standards, safeguards, procurement rules, or public institutions are ready. In that case, society may rely on improvised decisions after deployment or after harm becomes visible.

The second is formalizing too early.

Immature methods can become rigid requirements. Poor benchmarks can become legal thresholds. Large incumbents can shape standards around their existing systems. Compliance can become a substitute for genuine safety. Small developers and open communities can be excluded. Regulation can freeze assumptions that the underlying science no longer supports.

Progressive standards and requirements exist to navigate between these risks.

The central idea is not that every voluntary practice should move inevitably toward regulation.

The central idea is that institutional form should strengthen as four conditions strengthen:

1. **The consequence of error**
2. **The quality of the evidence**
3. **The maturity of the implementation ecosystem**
4. **The legitimacy and capability of the enforcing institution**

A new evaluation method may begin as a research prototype.

If it proves useful, it may become:

- A recommended practice
- A voluntary framework
- A reporting convention
- An industry commitment
- A technical specification
- A consensus standard
- A contractual requirement
- A procurement condition

- An insurance expectation
- A certification scheme
- A code recognized by law
- A mandatory requirement
- A condition for access, deployment, or authorization

This sequence is not automatic.

Different practices may stop at different levels.

Some should remain voluntary because:

- The evidence is weak
- The risks are limited
- Context varies substantially
- Innovation value is high
- Enforcement cost would exceed benefit
- The requirement would be easy to game
- Professional judgment is more appropriate
- A mandatory rule would create harmful concentration

Other practices may warrant formal requirements because:

- Consequences are severe
- Voluntary adoption is persistently insufficient
- Free-rider incentives are strong
- Market participants cannot observe or price risk
- Harms cross borders or affect nonconsenting parties
- A common minimum is necessary for interoperability
- Independent evidence shows that the practice is effective
- The evaluation and assurance ecosystem can implement it competently

Standards Body adopts the following core position:

> **Frontier AI requirements should mature through evidence-based stages. Formality, assurance, and enforcement should increase only when the underlying practice is sufficiently valid, implementable, decision-relevant, proportionate, and legitimate for the consequences at issue.**

A progressive regime should separate several dimensions that are often compressed into the word "mandatory":

- **Substantive strength**
  How demanding is the technical or organizational requirement?

- **Coverage**
  Which systems, organizations, deployments, or capabilities are included?

- **Assurance depth**
  Is compliance self-attested, independently reviewed, audited, certified, or continuously monitored?
- **Transparency**
  What must be disclosed publicly or to an authority?
- **Enforcement**
  What happens when requirements are not met?
- **Decision consequence**
  Does nonconformity trigger remediation, restricted access, delayed deployment, financial penalty, or prohibition?
- **Geographic reach**
  Is the requirement organizational, sectoral, national, or international?

These dimensions can progress separately.

For example:

- A strong internal requirement may have no public disclosure.
- A modest reporting requirement may be legally mandatory.
- A voluntary standard may be required by a purchaser.
- A certification scheme may remain optional but become commercially important.
- A law may recognize a code of practice as one route to demonstrating compliance.
- A high-stakes capability threshold may trigger external review without prohibiting development.

The strongest pathway from voluntary practice to formal requirement should include:

1. Clear problem definition
2. Evidence that the practice addresses that problem
3. Defined scope and terminology
4. Pilot implementation
5. Independent evaluation
6. Public and affected-party consultation
7. Cost and competition analysis
8. Assurance readiness
9. Phase-in
10. Support for smaller actors
11. Monitoring and enforcement
12. Appeals and correction
13. Scheduled review
14. Sunset or retirement where justified

Existing institutions illustrate different parts of this progression.

The NIST AI Risk Management Framework was designed as a voluntary and adaptable resource for managing AI risk.[^nist-rmf] ISO/IEC 42001 provides certifiable requirements for an AI management system, while ISO/IEC 42006 adds requirements for bodies auditing and certifying those systems.[^iso-42001][^iso-42006] The Hiroshima AI Process Code of Conduct and reporting framework provide a voluntary international structure for advanced AI governance disclosures.[^haip-code][^haip-reporting] The European Union AI Act establishes binding, risk-based legal obligations and includes roles for standards, codes, conformity assessment, and oversight.[^eu-ai-act] Frontier developers have also published voluntary capability-linked frameworks that connect rising capability evidence to stronger internal safeguards.[^openai-pf][^anthropic-rsp][^deepmind-fsf]

These examples should not be flattened into one model.

They show a broader institutional pattern:

- Voluntary practice can generate learning.
- Reporting can create comparability.
- Standards can make expectations repeatable.
- Assurance can make claims more credible.
- Procurement and contracts can create adoption.
- Law can create enforceable minimums.
- Continuous review can preserve relevance.

The purpose of this foundation is to define how that progression can occur without confusing speed with legitimacy, formality with effectiveness, or compliance with safety.

---

1. Foundational Proposition

1.1 Core Thesis

> **Requirements should become more formal only as their evidence, implementation capacity, consequence, and legitimacy justify greater institutional force.**

1.2 Non-Inevitability Thesis

> **Voluntary practice is not merely a temporary stage before regulation. Some practices should remain voluntary, experimental, contextual, or professional rather than mandatory.**

1.3 Evidence Thesis

> **A requirement should not become stronger merely because concern becomes stronger. The evidence supporting the requirement must also improve.**

1.4 Institutional Thesis

> **A technically sound requirement can fail when the institutions responsible for implementation, assurance, enforcement, or appeal are not ready.**

1.5 Proportionality Thesis

> **The burden of a requirement should be proportionate to the capability, deployment, scale, access, consequence, and uncertainty involved.**

1.6 Revisability Thesis

> **Every formal requirement should possess a mechanism for interpretation, revision, suspension, and retirement.**

1.7 Legitimacy Thesis

> **Standards gain legitimacy through competent process, evidence, participation, transparency, accountability, and correction, not through the word "standard" alone.**

1.8 Innovation Thesis

> **Good requirements should channel innovation toward better evidence, safeguards, and public benefit rather than freeze one technical approach.**

---

2. Scope and Boundaries

2.1 What This Foundation Covers

This paper covers progressive institutional expectations concerning:

- Evaluation protocols
- Held-out testing
- High-stakes capability assessment
- Independent review
- Third-party assurance
- Reporting
- Model and system documentation
- Safeguards
- Security
- Incident response
- Governance
- Procurement

- Certification
- contractual requirements
- legal obligations
- deployment conditions
- international interoperability

2.2 What This Foundation Does Not Fully Cover

This paper does not fully specify:

- The content of a particular technical standard
- Statutory drafting for a jurisdiction
- Constitutional limits
- sector-specific law
- criminal liability
- administrative procedure
- antitrust doctrine
- complete enforcement design
- international treaty negotiation

Those subjects require legal and jurisdiction-specific expertise.

2.3 Standards Versus Requirements

A standard may be voluntary.

A requirement may refer to a standard.

A standard can become influential through:

- Market adoption
- contracts
- procurement
- certification
- insurance
- regulation
- incorporation into law

2.4 Technical Versus Institutional Requirements

Technical requirements govern matters such as:

- Evaluation procedures
- logging
- security controls
- thresholds
- system configuration

Institutional requirements govern matters such as:

- Governance
- conflicts
- reviewer independence
- reporting
- appeals
- incident response
- public accountability

Both are necessary.

2.5 Frontier-Specific Versus General AI Governance

Some requirements apply across AI systems.

Others should apply only to:

- Frontier capabilities
- high-stakes domains
- broad deployment
- open-weight release
- large-scale use
- critical infrastructure
- high-autonomy systems

---

3. Canonical Definitions

3.1 Standard

A document established through a recognized process that provides rules, guidelines, characteristics, or common practices for repeated use.

3.2 Technical Specification

A document defining technical requirements, methods, interfaces, measurements, or performance characteristics.

3.3 Recommended Practice

Nonbinding guidance describing a preferred method based on available evidence and professional judgment.

3.4 Framework

A structured set of concepts, outcomes, functions, or practices that helps organizations manage a problem while allowing adaptation.

3.5 Code of Conduct

A set of commitments or expected behaviors adopted voluntarily or recognized by an institution.

3.6 Code of Practice

A more operational body of guidance or rules describing how requirements may be met in practice.

3.7 Guidance

Explanatory or advisory material supporting interpretation and implementation.

3.8 Requirement

A condition that must be fulfilled within a defined context.

3.9 Voluntary Requirement

A requirement accepted by choice through a framework, contract, certification scheme, membership, or internal policy.

3.10 Mandatory Requirement

A requirement imposed by an authority or binding legal arrangement.

3.11 Internal Commitment

A public or private obligation adopted by an organization and enforced through internal governance.

3.12 Private Ordering

Governance created through contracts, procurement, insurance, platforms, professional rules, or market relationships rather than direct legislation.

3.13 Procurement Requirement

A condition that must be met to sell to or contract with a purchaser.

3.14 Certification Requirement

A condition evaluated through a defined certification scheme.

3.15 Conformity Assessment

Demonstration that specified requirements are fulfilled.

3.16 Safe Harbor

A legal or institutional provision that offers defined protection or reduced liability when specified practices are followed.

3.17 Regulatory Sandbox

A controlled environment in which organizations test innovations under oversight and defined conditions.

3.18 Performance-Based Requirement

A requirement defining an outcome to be achieved without mandating the exact method.

3.19 Prescriptive Requirement

A requirement specifying the method, control, or process to be used.

3.20 Risk-Based Requirement

A requirement whose applicability or rigor varies according to risk.

3.21 Capability-Based Requirement

A requirement triggered by demonstrated or reasonably anticipated AI capability.

3.22 Deployment-Based Requirement

A requirement triggered by use context, access, scale, or integration.

3.23 Organization-Based Requirement

A requirement triggered by characteristics of the provider or deployer, such as role, scale, or control.

3.24 Tiered Requirement

A requirement organized into levels of increasing rigor.

3.25 Phase-In

Staged introduction of a requirement over time or across categories.

3.26 Grandfathering

Allowing existing systems or practices to continue under earlier rules.

3.27 Sunset Clause

A provision causing a requirement to expire unless renewed.

3.28 Review Clause

A requirement that the rule be reassessed at a defined time or after specified events.

3.29 Incorporation by Reference

Use of an external standard or document as part of a binding requirement.

3.30 Presumption of Conformity

A legal or institutional presumption that compliance with recognized standards provides evidence of compliance with specified requirements.

3.31 Enforcement Ladder

A sequence of progressively stronger responses to nonconformity.

3.32 Regulatory Ratchet

A pattern in which requirements become progressively stronger but are difficult to reduce after evidence changes.

3.33 Compliance Theater

Formal conformity without meaningful achievement of the underlying safety, reliability, or accountability objective.

3.34 Outcome Equivalence

Acceptance of different methods that achieve an adequately comparable outcome.

3.35 Transition Gate

A defined test that must be passed before a practice moves to a more formal institutional stage.

---

4. Why Progressive Institutionalization Is Necessary

4.1 Capability Progress Outpaces Law

Technical systems can change faster than formal legislation.

Research and voluntary standards can respond earlier.

4.2 Law Requires Implementable Concepts

A legal requirement needs:

- Defined scope
- measurable conduct
- responsible actors
- enforcement
- evidence
- appeal
- jurisdiction

Research frameworks can develop these foundations.

4.3 Voluntary Practice Generates Learning

Organizations can test:

- Evaluation methods
- reporting formats
- thresholds
- safeguards
- governance
- independent review

before a requirement becomes universal.

4.4 Standards Support Repeatability

Standards can convert broad principles into:

- Definitions
- procedures
- records
- interfaces
- competence requirements
- reporting

4.5 Assurance Supports Trust

Independent assessment makes claims more credible.

4.6 Markets Can Accelerate Adoption

Procurement, insurers, investors, platforms, and professional bodies can reward mature practice.

4.7 Public Authority Creates Minimums

Law can address:

- Free riders
- externalities
- nonconsenting parties
- systemic risk
- persistent nonadoption
- enforcement

4.8 Revision Remains Necessary

A mature requirement must continue to evolve.

4.9 The Central Risk

The same progression can become:

- Premature
- captured
- burdensome
- rigid
- symbolic
- fragmented
- difficult to reverse

Progression must be conditional.

---

5. The Progressive Standards Ladder

Standards Body proposes a fourteen-stage ladder.

The ladder is descriptive and design-oriented, not inevitable.

Stage 0: Open Question

Characteristics:

- Problem recognized
- terms contested
- evidence limited
- no proposed practice

Outputs:

- Research questions
- evidence maps
- incident collection

Stage 1: Research Method

Characteristics:

- Experimental approach
- limited validation
- specialist use
- no broad assurance claim

Outputs:

- Papers
- prototypes
- open tools
- pilot datasets

Stage 2: Recommended Practice

Characteristics:

- Initial evidence
- professional guidance
- context-sensitive

- voluntary

Outputs:

- Technical guidance
- implementation examples
- warnings

Stage 3: Voluntary Framework

Characteristics:

- Structured outcomes or practices
- adaptable
- self-directed adoption
- no formal conformity requirement

NIST describes the AI RMF as a voluntary, rights-preserving, non-sector-specific resource intended to support AI risk management.[^nist-rmf][^nist-airc]

Stage 4: Voluntary Reporting

Characteristics:

- Common disclosure questions
- comparability
- self-reported implementation
- public or controlled transparency

The Hiroshima AI Process reporting framework provides a common voluntary structure for organizations to report on advanced AI governance and risk-management practices.[^haip-reporting]

Stage 5: Organizational Commitment

Characteristics:

- Public or internal commitment
- capability-linked obligations
- governance
- periodic reporting

Frontier safety frameworks published by developers illustrate this stage.[^openai-pf][^anthropic-rsp][^deepmind-fsf]

Stage 6: Industry Code or Consortium Practice

Characteristics:

- Multiple organizations
- shared commitments
- peer pressure
- monitoring
- limited enforcement

Stage 7: Technical or Management-System Standard

Characteristics:

- Defined requirements
- consensus process
- repeatable implementation
- revision control

ISO/IEC 42001 specifies requirements for establishing, implementing, maintaining, and continually improving an AI management system.[^iso-42001]

Stage 8: Independent Assurance Scheme

Characteristics:

- Defined conformity criteria
- qualified assessors
- audit or evaluation
- reports or certification
- complaints
- surveillance

ISO/IEC 42006 sets additional requirements for bodies auditing and certifying AI management systems against ISO/IEC 42001.[^iso-42006]

Stage 9: Contractual and Procurement Requirement

Characteristics:

- Binding between parties
- purchaser or platform condition
- audit rights
- remedies
- re-evaluation

Stage 10: Insurance and Financial Expectation

Characteristics:

- Underwriting condition
- control requirement
- disclosure
- pricing consequence

Stage 11: Recognized Code or Standard

Characteristics:

- Public authority recognizes a standard, code, or scheme
- compliance may provide evidence or a presumption
- alternative methods may remain available

Stage 12: Mandatory Reporting or Assessment

Characteristics:

- Defined actors must disclose, evaluate, document, or notify
- enforcement applies to process obligations

Stage 13: Mandatory Substantive Requirement

Characteristics:

- Minimum safeguards, controls, or performance conditions
- legal enforcement
- appeal
- penalties or restrictions

Stage 14: Authorization, Restriction, or Prohibition

Characteristics:

- Deployment or activity requires approval, licensing, or specified conditions
- severe nonconformity can block activity

The European Union AI Act demonstrates a binding risk-based legal framework using differentiated obligations, conformity mechanisms, governance, and enforcement.[^eu-ai-act]

---

6. The Ladder Is Not a Conveyor Belt

6.1 Multiple End States

A practice may appropriately stop at:

- Guidance
- voluntary framework
- contractual requirement
- certification
- legal minimum

6.2 Different Paths

Some requirements may move from research directly to procurement.

Others may move from law to standards through delegated implementation.

Others may remain internal.

6.3 Reversal

A requirement may move downward when:

- Evidence weakens
- risk declines
- implementation cost is excessive
- superior methods emerge
- scope is too broad
- enforcement creates harm

6.4 Forking

A broad practice may split into:

- High-stakes mandatory core
- voluntary advanced practice
- sector-specific variants
- open-source pathway
- small-actor pathway

6.5 No Moral Ranking

Mandatory is not automatically better than voluntary.

A voluntary practice can be rigorous.

A legal obligation can be ineffective.

6.6 No One-Dimensional Maturity

A requirement may be:

- Technically mature
- institutionally immature
- legally enforceable
- poorly assured
- widely adopted
- weakly evidenced

Assessment should be multidimensional.

---

7. Readiness Gates

A practice should pass readiness gates before moving to a more formal stage.

7.1 Problem Gate

Questions:

- Is the problem clearly defined?
- Who is affected?
- What evidence shows material need?
- Is the problem persistent?
- Is intervention within institutional scope?

7.2 Construct Gate

Questions:

- Are key terms defined?
- Is the desired outcome measurable?
- Are proxies valid?
- Are exclusions clear?

7.3 Evidence Gate

Questions:

- Does the practice improve outcomes?
- Has it been independently tested?
- What are false-positive and false-negative risks?

- Is evidence generalizable?

7.4 Method Gate

Questions:

- Can the practice be implemented consistently?
- Are methods documented and versioned?
- Can multiple organizations reproduce them?
- Is uncertainty known?

7.5 Assurance Gate

Questions:

- Can conformity be assessed?
- Are qualified evaluators available?
- Are conflicts controlled?
- Are appeals possible?
- Is security adequate?

7.6 Operational Gate

Questions:

- Can organizations implement the requirement?
- Are tools and expertise available?
- Are timelines realistic?
- Can evidence be preserved?
- Can systems be re-evaluated?

7.7 Economic Gate

Questions:

- What does implementation cost?
- Who bears the cost?
- Does it create barriers to entry?
- Does it favor incumbents?
- Are public benefits proportionate?

7.8 Competition Gate

Questions:

- Does the requirement entrench one technology?

- Can small actors comply?
- Can open-source models participate?
- Is the evaluator market competitive?

7.9 Legitimacy Gate

Questions:

- Was the process competent and inclusive?
- Were affected parties heard?
- Are reasons public?
- Are conflicts disclosed?
- Can decisions be challenged?

7.10 Enforcement Gate

Questions:

- Is nonconformity detectable?
- Is enforcement authority clear?
- Are sanctions proportionate?
- Can parties appeal?
- Can enforcement scale?

7.11 International Gate

Questions:

- Is the requirement interoperable?
- Does it conflict with other regimes?
- Can evidence be recognized across borders?
- Are localization needs understood?

7.12 Revision Gate

Questions:

- Is there a review cycle?
- Are change triggers defined?
- Can obsolete methods be retired?
- Is there a sunset or escape path?

---

8. Design Principles

8.1 Evidence Before Authority

Do not strengthen institutional force faster than evidentiary support.

8.2 Consequence Before Uniformity

Focus strongest requirements where consequence is greatest.

8.3 Outcome Orientation

Prefer performance or outcome requirements when multiple methods can achieve the objective.

8.4 Method Specificity Where Necessary

Prescriptive controls may be justified when:

- Outcome measurement is immature
- failure consequence is severe
- known minimum controls are essential
- verification requires specificity

8.5 Layered Obligations

Use different requirements for:

- Baseline practice
- high-stakes systems
- frontier capabilities
- critical deployments

8.6 Clear Trigger Logic

State why and when stronger requirements apply.

8.7 Assurance Proportionality

Do not require full certification for every practice.

8.8 Support Before Sanction

Early stages should include:

- Guidance
- tools
- pilots
- grants
- shared infrastructure
- implementation support

8.9 Contestability

Allow alternative methods, evidence, appeals, and review.

8.10 Technology Neutrality with Limits

Avoid mandating one implementation unnecessarily.

Do not use neutrality to avoid specifying necessary safety outcomes.

8.11 Anti-Capture

Include diverse expertise, transparent conflicts, and review of market effects.

8.12 Small-Actor Pathways

Use scaled requirements, common infrastructure, and reasonable timelines.

8.13 Dynamic Incorporation

Standards incorporated into requirements should be version-governed.

8.14 No Compliance Safe-Washing

Conformity should not support claims broader than the requirement.

8.15 Human Accountability

Automation of compliance should not remove accountable decision-makers.

8.16 International Compatibility

Design metadata and evidence for cross-border understanding.

8.17 Expiration

Every significant requirement should be reviewable and retireable.

---

9. Trigger Models

9.1 Capability Trigger

Requirement applies when evaluation shows a defined capability level.

Advantages:

- Links burden to technical evidence
- adapts across model size

Risks:

- Threshold uncertainty
- gaming
- evaluation access
- rapid change

9.2 Compute Trigger

Requirement applies based on training or inference compute.

Advantages:

- More measurable in some contexts
- early visibility

Risks:

- Weak proxy
- hardware change
- efficiency progress
- circumvention
- open-source effects

9.3 Deployment Trigger

Applies based on:

- Users
- scale
- autonomy
- tool permissions

- critical integration
- geographic reach

9.4 Use-Case Trigger

Applies to defined high-risk uses.

9.5 Access Trigger

Applies when:

- Weights are released
- fine-tuning is available
- dangerous tools are connected
- anonymous access is offered

9.6 Organizational Trigger

Applies based on:

- Developer role
- market position
- system control
- resource level

Risk:

Organization-based rules can entrench incumbents or create arbitrary distinctions.

9.7 Incident Trigger

Applies after:

- Safeguard failure
- misuse
- security breach
- unexpected capability
- evaluation compromise

9.8 Trend Trigger

Applies when capability approaches a threshold rapidly.

9.9 Hybrid Trigger

Combines capability, access, deployment, and consequence.

Standards Body position:

High-stakes requirements should usually use hybrid triggers rather than one proxy alone.

---

10. Risk and Capability Tiers

Tier 0: General Good Practice

Applies broadly.

Examples:

- Basic documentation
- responsibility
- incident channels
- evaluation records

Assurance:

- Self-assessment or ordinary internal controls

Tier 1: Material AI System

Triggers:

- Meaningful deployment
- significant affected population
- domain consequence

Requirements:

- Structured risk management
- documented evaluation
- monitoring
- user information

Tier 2: High-Stakes System or Capability

Triggers:

- Credible severe pathway
- critical use
- high autonomy
- broad access

Requirements:

- Dynamic evaluation
- held-out tests
- domain experts
- independent review
- enhanced safeguards
- incident reporting

Tier 3: Critical Frontier Capability

Triggers:

- Defined critical threshold
- catastrophic or strategic consequence
- loss-of-control concern
- high scalable misuse

Requirements:

- Multi-institution evaluation
- strong security
- formal safety case
- continuous monitoring
- deployment conditions
- governance escalation

Tier 4: Exceptional Authorization Regime

Applies only when ordinary controls are inadequate for exceptionally consequential activities.

Requirements may include:

- Approval
- licensing
- restricted access
- supervised deployment
- international coordination

Tiering Principles

- Domain-specific
- evidence-based
- uncertainty-aware
- appealable
- revisable
- not based solely on model size

- not a permanent label

---

11. Voluntary Frameworks

11.1 Purpose

Voluntary frameworks can:

- Build common language
- coordinate practice
- support learning
- reduce implementation uncertainty
- create early evidence
- avoid premature rigidity

11.2 Strengths

- Speed
- flexibility
- broad applicability
- innovation
- easier revision
- international accessibility

11.3 Weaknesses

- Selective adoption
- self-reporting
- weak enforcement
- favorable interpretation
- free riders
- uneven implementation

11.4 Conditions for Credibility

A voluntary framework should include:

- Clear outcomes
- implementation evidence
- public status
- review
- update process
- limitations
- independent challenge

11.5 NIST AI RMF

The NIST AI RMF illustrates a voluntary, risk-management-oriented framework organized around Govern, Map, Measure, and Manage functions.[^nist-rmf][^nist-airc]

Its flexibility is a strength.

The same flexibility means implementation depth can vary.

11.6 Appropriate Use

Voluntary frameworks are appropriate when:

- The field is evolving
- contexts differ
- learning is valuable
- enforcement is premature
- institutional capacity is developing

11.7 Transition Evidence

Before progression, study:

- Adoption
- implementation quality
- outcomes
- gaps
- nonadoption
- burden
- market effects

---

12. Voluntary Reporting and Transparency

12.1 Purpose

Reporting can create:

- Comparability
- accountability
- peer learning
- public visibility
- evidence for future standards

12.2 Self-Reporting Risks

- Selective disclosure
- vague claims
- inconsistent metrics
- favorable interpretation
- lack of verification

12.3 Reporting Maturity

Level A

Narrative disclosure.

Level B

Structured questions.

Level C

Comparable metrics.

Level D

Evidence-backed reporting.

Level E

Independently assured reporting.

12.4 Hiroshima AI Process

The Hiroshima AI Process Code of Conduct provides voluntary guidance for organizations developing advanced AI systems, and the OECD reporting framework supports structured disclosure against its actions.[^haip-code][^haip-reporting][^haip-insights]

12.5 Transition Conditions

Reporting may become mandatory when:

- Information asymmetry is material
- public or market actors need comparable evidence
- voluntary reporting is incomplete
- reporting burden is manageable
- sensitive information can be protected

12.6 Confidential Reporting

Some information may be submitted to:

- Regulator
- trusted auditor
- standards body
- purchaser

without full public release.

12.7 Reporting Quality

Define:

- Scope
- period
- evidence
- metrics
- assumptions
- incidents
- omissions
- assurance
- corrections

---

13. Organizational Commitments and Frontier Safety Frameworks

13.1 Purpose

Internal or public commitments can connect:

- Capability evaluation
- safeguards
- governance
- deployment
- security

13.2 Benefits

- Early experimentation
- organizational forcing function
- public accountability
- learning before formal standardization

13.3 Risks

- Self-defined thresholds
- unilateral revision
- inconsistent scope
- no external enforcement
- public-relations use
- limited comparability

13.4 Framework Examples

OpenAI's Preparedness Framework, Anthropic's Responsible Scaling Policy, and Google DeepMind's Frontier Safety Framework connect capability or risk evidence to stronger safeguards and governance actions through different structures.[^openai-pf][^anthropic-rsp][^deepmind-fsf]

13.5 Evaluation Criteria

Assess:

- Scope
- thresholds
- evidence
- governance
- external review
- transparency
- noncompliance
- revision
- deployment consequence

13.6 Path to Shared Standard

Common elements may become candidates for:

- Taxonomy
- reporting standards
- evaluator methods
- assurance
- procurement
- legal recognition

13.7 Avoiding Lowest Common Denominator

Shared standards should not weaken stronger practice merely to achieve consensus.

---

14. Consensus Standards

14.1 Role

Standards can make good practice:

- Repeatable
- testable
- interoperable
- teachable
- auditable

14.2 Standards Process

A credible process should include:

- Need identification
- scope
- balanced participation
- evidence review
- drafting
- testing
- public comment
- resolution
- approval
- versioning
- maintenance
- retirement

14.3 Management-System Standards

ISO/IEC 42001 provides requirements for an AI management system and continual improvement. [^iso-42001]

Management-system standards can improve organizational discipline.

They do not establish that every AI system is safe.

14.4 Technical Standards

May cover:

- Evaluation metadata
- benchmark administration
- model identity
- incident reporting

- security
- logging
- thresholds
- uncertainty
- conformity claims

14.5 Performance Standards

Define outcomes.

14.6 Process Standards

Define organizational or procedural controls.

14.7 Interface Standards

Support interoperability.

14.8 Standardization Risks

- Slow process
- incumbent dominance
- premature consensus
- obsolete methods
- compliance fixation
- international compromise

14.9 Standards as Floor or Ceiling

A standard should normally define a minimum or common basis, not prohibit superior practice.

14.10 Experimental Annexes

Use nonbinding or rapidly updateable annexes for immature methods.

---

15. Assurance and Certification

15.1 Why Assurance Matters

A requirement without credible evidence of implementation may remain symbolic.

15.2 Assurance Levels

- Self-attestation
- second-party review
- independent review
- limited assurance
- reasonable assurance
- certification
- continuous assurance

15.3 ISO/IEC 42006

ISO/IEC 42006 sets additional requirements for bodies auditing and certifying AI management systems against ISO/IEC 42001.[^iso-42006]

This illustrates the institutional progression from management requirements to requirements for the bodies assessing them.

15.4 Certification Readiness

A practice is ready for certification only when:

- Requirements are defined
- conformity can be assessed
- methods are stable enough
- evaluator competence exists
- scheme governance exists
- claims can be bounded
- surveillance is possible

15.5 Certification Limitations

Certification should not imply:

- Universal safety
- future performance
- legal immunity
- absence of all harm
- endorsement beyond scope

15.6 Continuous Systems

Frequently updated systems may require:

- Change notification
- re-evaluation triggers

- configuration identity
- continuous assurance
- certificate suspension

15.7 Foundation 5 Integration

The third-party ecosystem supplies the competence and recognition needed for assurance.

---

16. Private Ordering

16.1 Contracts

Contracts can require:

- Evaluation
- audit rights
- incident notice
- monitoring
- documentation
- re-evaluation
- access restrictions

16.2 Procurement

Public and private purchasers can require standards or assurance.

16.3 Platform Rules

Platforms can impose:

- Model access conditions
- developer verification
- safety testing
- content controls
- incident reporting

16.4 Insurance

Insurers can condition:

- Coverage
- pricing
- exclusions
- limits

on controls and evidence.

16.5 Financing

Investors and lenders may request risk-governance evidence.

16.6 Professional Rules

Professional bodies may define acceptable AI use in:

- Medicine
- law
- engineering
- finance
- science

16.7 Benefits

Private ordering can:

- Move quickly
- target relationships
- experiment
- reward better practice

16.8 Risks

- Opaque requirements
- inconsistent burden
- market concentration
- unequal bargaining power
- weak public accountability
- duplicated audits

16.9 Interoperability

Shared standards can reduce fragmentation.

16.10 No Substitute for Public Law

Private ordering is insufficient when:

- Harms affect nonparties
- market power is concentrated
- public rights are involved

- systemic risk exists
- enforcement requires public authority

---

17. Legal Recognition

17.1 Recognized Standards

Law may reference standards as:

- Evidence
- implementation guidance
- presumption of conformity
- required method
- safe harbor

17.2 Benefits

- Technical expertise
- flexibility
- international alignment
- reduced legislative detail
- easier updating

17.3 Risks

- Private rulemaking power
- paywalled standards
- weak democratic accountability
- automatic incorporation of changes
- industry capture

17.4 Version Control

Law should identify:

- Which version applies
- who can update it
- transition
- public access
- alternative methods

17.5 Codes of Practice

Codes may provide a practical route to compliance while allowing alternatives.

17.6 Presumption of Conformity

Useful when:

- Standard is mature
- process is legitimate
- conformity is meaningful
- alternative evidence remains possible

17.7 Safe Harbor

Can encourage adoption.

Risks:

- Minimum becomes maximum
- formal compliance excuses harm
- outdated methods gain legal protection

17.8 Regulatory Sandbox

Can test requirements under supervision before broad application.

17.9 EU AI Act

The EU AI Act establishes a phased, binding risk-based regime and uses standards, codes, conformity mechanisms, and public oversight as implementation tools.[^eu-ai-act][^eu-summary]

Its complexity also illustrates why legal progression requires implementation capacity, guidance, standards, codes, and institutional coordination.

---

18. Mandatory Reporting and Assessment

18.1 Process Obligations

Examples:

- Risk assessment
- evaluation
- documentation
- incident reporting
- transparency
- external review
- audit

18.2 Advantages

- Creates minimum evidence
- addresses free riders
- supports oversight
- improves comparability

18.3 Risks

- Paper compliance
- excessive reporting
- sensitive disclosure
- duplicated obligations
- weak data quality

18.4 Design Requirements

Define:

- Who reports
- what
- to whom
- when
- evidence
- confidentiality
- verification
- correction
- enforcement

18.5 Incident Reporting

Should distinguish:

- Safety incident
- security incident
- evaluation incident
- misuse
- near miss
- systemic issue

18.6 Independent Assessment

May be required for:

- High-stakes thresholds
- critical deployment

- certification
- organizational compliance

18.7 Avoiding Ritual

Authorities should use the evidence collected.

---

19. Mandatory Substantive Requirements

19.1 Types

- Security controls
- access limits
- evaluation
- safeguards
- monitoring
- human oversight
- rollback
- incident response
- governance
- deployment conditions

19.2 Performance Versus Prescription

Use performance requirements when outcomes are measurable.

Use prescriptive minimums when:

- Known controls are essential
- methods are mature
- outcome evidence is insufficient
- interoperability requires consistency

19.3 Minimum Necessary Rule

A mandatory requirement should be no broader than needed to address the defined problem.

19.4 Alternative Compliance

Allow equivalent methods when:

- Evidence can demonstrate equivalence
- innovation is valuable
- context varies

19.5 Exemptions

Exemptions should be:

- Narrow
- justified
- time-bounded
- reviewable
- transparent where possible

19.6 Emergency Requirements

May be necessary after severe incidents.

They should include:

- Expiry
- review
- evidence
- limited scope
- appeal

19.7 Enforcement Readiness

A requirement without enforceable evidence can create false confidence.

---

20. Authorization, Licensing, and Restrictions

20.1 Highest-Formality Stage

Authorization may be justified only where:

- Potential consequence is exceptionally severe
- Less restrictive measures are inadequate
- Scope is definable
- decision authority is legitimate
- evidence is sufficiently mature
- appeal exists

20.2 Possible Objects

- Deployment
- access
- weight release
- critical integration

- high-risk tool connection
- specialized training activity

20.3 Conditions

- Evaluation
- safeguards
- security
- monitoring
- reporting
- independent review
- renewal

20.4 Risks

- Innovation suppression
- regulatory capture
- international arbitrage
- concentration
- secrecy
- politicization
- false confidence

20.5 Temporary Authorization

Conditional approval can support learning.

20.6 Restriction Rather Than Prohibition

Use narrow restrictions where possible:

- User verification
- rate limits
- tool limits
- deployment scale
- geographic limits
- trusted access

20.7 Prohibition

Should require an exceptionally strong justification, clear scope, due process, and review.

---

21. Enforcement Ladder

A progressive system should use proportionate responses.

Level 0: Guidance

- Clarification
- technical support
- warning

Level 1: Corrective Action

- Remediation plan
- deadline
- follow-up

Level 2: Enhanced Reporting

- Additional evidence
- monitoring
- independent review

Level 3: Public Notice

- Nonconformity status
- correction
- certificate limitation

Level 4: Scope Restriction

- Limit deployment
- suspend feature
- restrict users
- narrow certification

Level 5: Financial Consequence

- Contract remedy
- insurance impact
- administrative penalty

Level 6: Suspension

- Certificate

- access
- deployment
- authorization

Level 7: Withdrawal

- Certification
- approval
- market access

Level 8: Prohibition or Emergency Containment

Reserved for severe and immediate cases.

Enforcement Principles

- Notice
- evidence
- proportionality
- consistency
- appeal
- correction
- public explanation
- restoration pathway

---

22. Phase-In and Transition Design

22.1 Why Phase-In Matters

Organizations need time to:

- Build systems
- train staff
- obtain evaluation
- secure infrastructure
- change contracts
- develop standards

22.2 Phase-In Dimensions

- Time
- organization size
- risk tier

- model capability
- sector
- geographic market
- requirement type

22.3 Pilot Period

Use a pilot to test:

- Clarity
- burden
- evaluator capacity
- unintended effects
- reporting
- enforcement

22.4 Voluntary Early Adoption

Can identify leaders and implementation problems.

22.5 Transitional Safe Harbor

May protect good-faith pilot participants.

22.6 Grandfathering

Use cautiously.

A legacy system can still create current risk.

22.7 Support

Provide:

- Templates
- tools
- training
- shared infrastructure
- grants
- evaluator access
- translations

22.8 Readiness Review

Do not activate a requirement if:

- No evaluators exist
- standards are unavailable
- guidance is incomplete
- required tools cannot be obtained

unless consequence demands emergency action.

---

23. Small Organizations and Open Ecosystems

23.1 Burden Risk

Formal requirements can disproportionately affect:

- Startups
- academic labs
- nonprofits
- open-source projects
- individual researchers
- developing economies

23.2 Proportionality

Scale obligations according to:

- Capability
- access
- deployment
- control
- resources
- consequence

23.3 Functional Equivalence

A small organization should not need the same bureaucracy as a large frontier laboratory if it can demonstrate equivalent outcomes.

23.4 Shared Infrastructure

Support:

- Evaluation facilities

- security services
- task banks
- templates
- incident channels
- legal guidance
- certification grants

23.5 Open-Weight Considerations

Requirements should address:

- Decentralized modification
- downstream deployment
- weight security
- reproducibility
- distributed responsibility
- public benefit

23.6 Community Governance

Open projects may use:

- Maintainer policies
- release reviews
- model cards
- public evaluation
- community incident reporting
- signed releases

23.7 Avoiding Incumbent Moats

Review whether each requirement:

- Requires exclusive data
- requires costly legal structures
- depends on scarce auditors
- mandates proprietary tools
- creates fixed fees
- blocks interoperability

23.8 Support Is Part of Legitimacy

A requirement is less legitimate when compliance is impossible for qualified good-faith actors.

---

24. Innovation and Regulatory Sandboxes

24.1 Purpose

Sandboxes can test:

- New evaluation methods
- alternative safeguards
- reporting
- deployment constraints
- assurance

under oversight.

24.2 Sandbox Requirements

- Clear scope
- eligibility
- safeguards
- monitoring
- exit
- incident response
- public learning
- no immunity for bad faith

24.3 Evidence Output

A sandbox should produce:

- What worked
- what failed
- cost
- risk
- scalability
- rule implications

24.4 Risks

- Regulatory favoritism
- weak consumer protection
- secrecy
- pilot becoming permanent exemption
- unequal access

24.5 Standards Body Use

Standards Body could support nonregulatory protocol sandboxes for:

- Dynamic evaluation
- held-out testing
- auditor proficiency
- reporting
- threshold governance

---

25. Safe Harbors and Incentives

25.1 Purpose

Safe harbors can reward:

- Disclosure
- evaluation
- incident reporting
- standards adoption
- remediation
- research access

25.2 Conditions

A safe harbor should require:

- Good faith
- defined practice
- evidence
- timely correction
- no concealment
- continued responsibility

25.3 Risks

- Checklist compliance
- outdated protection
- moral hazard
- exclusion of alternatives
- weak standards becoming legal shields

25.4 Incentive Alternatives

- Procurement preference
- reduced audit frequency
- public recognition
- insurance credit
- research access
- grant eligibility
- expedited review

25.5 Foundation 7 Link

The broader incentive architecture is developed in Foundation 7.

---

26. Anti-Capture Architecture

26.1 Capture Sources

- Frontier developers
- evaluator firms
- governments
- standards bodies
- funders
- civil society groups
- dominant jurisdictions
- professional communities

26.2 Balanced Participation

Include:

- Developers
- deployers
- evaluators
- open-source experts
- researchers
- governments
- public-interest organizations
- affected sectors
- international participants

26.3 No Veto by Category

No one stakeholder category should control the outcome.

26.4 Conflict Disclosure

Publish material interests.

26.5 Evidence Discipline

Arguments should be tied to evidence and decision consequences.

26.6 Minority Reports

Preserve unresolved disagreement.

26.7 Public Drafts

Allow review where security permits.

26.8 Rotation

Rotate leadership and committee membership.

26.9 Funding Diversification

Avoid dependence on one sector.

26.10 Competition Review

Assess whether requirements create concentration.

26.11 Retrospective Review

Study who benefited, who was excluded, and whether outcomes improved.

---

27. International Interoperability

27.1 Shared Foundations

Align on:

- Definitions
- metadata
- evidence
- evaluator competence
- reporting

- result status
- incidents
- versioning

27.2 Local Decisions

Jurisdictions may differ in:

- Risk tolerance
- enforcement
- rights
- industrial policy
- legal process

27.3 Standards as Bridges

International standards can support common implementation.

27.4 Codes and Reporting

Voluntary international codes can generate shared practice before binding coordination.

27.5 Mutual Recognition

Recognize:

- Evaluators
- certificates
- reports
- test results

when base requirements are compatible.

27.6 Avoiding Fragmentation

Use:

- Crosswalks
- common data formats
- equivalence
- reference protocols
- shared registries

27.7 Avoiding Lowest Common Denominator

Interoperability should not require weakening high-stakes controls.

27.8 Extraterritorial Effects

Large markets can shape global practice through access conditions.

This can improve coordination or export one jurisdiction's assumptions without adequate participation.

27.9 Council of Europe Convention

The Council of Europe Framework Convention on Artificial Intelligence provides a treaty-level human-rights, democracy, and rule-of-law framework, illustrating a distinct international pathway from voluntary practice to binding commitments for parties.[^coe-convention]

27.10 Foundation 8 Link

Global interoperability is developed fully in Foundation 8.

---

28. Dynamic Incorporation and Version Control

28.1 The Version Problem

A requirement may reference a standard that later changes.

28.2 Static Incorporation

The rule references a specific version.

Strength:

- Legal certainty

Risk:

- Obsolescence

28.3 Dynamic Incorporation

The rule automatically follows later versions.

Strength:

- Adaptation

Risk:

- Delegated authority

- weak oversight
- unpredictable burden

28.4 Controlled Update

A designated authority reviews each new version before recognition.

28.5 Transition

Specify:

- Effective date
- support period
- old-version validity
- certificate transition
- re-evaluation
- exceptions

28.6 Emergency Updates

Allow rapid correction for:

- Security flaw
- invalid evaluation
- serious incident
- dangerous ambiguity

with retrospective review.

28.7 Public Change Record

Every change should state:

- Evidence
- impact
- burden
- affected actors
- transition
- dissent

---

29. Sunset, Review, and Retirement

29.1 Sunset Purpose

Prevent permanent rules based on temporary evidence.

29.2 Review Triggers

- Time
- capability change
- incident
- market concentration
- evaluator shortage
- new standard
- evidence failure
- international change

29.3 Review Questions

- Is the problem still material?
- Does the requirement work?
- Is it proportionate?
- Is it being gamed?
- Who bears cost?
- Are alternatives better?
- Should scope change?
- Should enforcement change?

29.4 Outcomes

- Maintain
- clarify
- strengthen
- narrow
- expand
- suspend
- replace
- retire

29.5 Retirement Plan

Include:

- Notice

- replacement
- certificate status
- data retention
- legal transition
- public explanation

29.6 No Ratchet Assumption

Strengthening should not be easier than reducing when evidence supports reduction.

---

30. Evaluation of Requirements

Requirements themselves should be evaluated.

30.1 Effectiveness

Did the requirement reduce the defined risk or improve evidence?

30.2 Adoption

Who complied?

30.3 Quality

Was compliance substantive or formal?

30.4 Burden

What did implementation cost?

30.5 Competition

Did the requirement increase concentration?

30.6 Innovation

Did it support or suppress beneficial methods?

30.7 Distribution

Who gained and who bore cost?

30.8 Enforcement

Was nonconformity detected and corrected?

30.9 Interoperability

Did it reduce or increase fragmentation?

30.10 Unintended Effects

Examples:

- Secrecy
- benchmark gaming
- reduced incident reporting
- auditor bottlenecks
- certification marketing
- offshore migration

30.11 Counterfactual

What would have happened without the requirement?

30.12 Requirement Scorecard

Use the scorecard in Section 49.

---

31. Governance

31.1 Roles

A progressive standards regime may include:

- Research community
- Standards Body
- technical working groups
- scheme owner
- evaluators
- accreditation bodies
- purchasers
- insurers
- regulators
- courts
- international organizations

- affected parties

31.2 Decision Separation

Separate:

- Technical evidence
- standard development
- certification
- enforcement
- appeal

where concentration creates conflicts.

31.3 Standards Body Role

At its present stage, Standards Body should:

- Develop concepts
- map evidence
- propose frameworks
- support pilots
- convene contributors
- publish open questions
- avoid claiming legal authority

31.4 Future Role

A future institution might:

- Maintain standards
- operate working groups
- recognize methods
- coordinate evaluators
- provide technical advice
- support interoperability

Only after competence and legitimacy are earned.

31.5 Public Authority

Government retains responsibility for legal obligations and coercive enforcement.

31.6 Appeal

Affected parties need a path to challenge:

- Classification
- evaluation
- certificate
- sanction
- scope
- interpretation

31.7 Emergency Authority

Should be narrow, time-limited, and reviewable.

---

32. Implementation Pathway

Phase 1: Map Current Practice

Identify:

- Frameworks
- standards
- audits
- reporting
- procurement
- laws
- gaps
- overlaps

Phase 2: Define Candidate Practices

Select practices with:

- Clear purpose
- evidence
- implementation path
- decision relevance

Phase 3: Classify Maturity

Assess:

- Construct
- evidence

- method
- assurance
- operations
- legitimacy
- cost

Phase 4: Pilot Voluntarily

Run with diverse organizations.

Phase 5: Publish Results

Report:

- Outcomes
- burden
- failures
- differences
- open questions

Phase 6: Develop Technical Specification

Create repeatable requirements.

Phase 7: Independent Review

Test methods and institutional design.

Phase 8: Build Assurance Capacity

Develop evaluators, proficiency, and security.

Phase 9: Test Private Ordering

Use procurement or contracts.

Phase 10: Evaluate Need for Formality

Assess nonadoption, externalities, systemic risk, and market failure.

Phase 11: Propose Recognized Standard or Legal Requirement

Include:

- Scope
- evidence
- phase-in
- support
- appeal
- review
- sunset

Phase 12: Monitor and Revise

Evaluate outcomes continuously.

---

33. Proposed Standards Body Pilot

33.1 Pilot Name

Progressive Frontier Evaluation Assurance Pathway

33.2 Purpose

Test how one mature evaluation practice can move responsibly from voluntary protocol to independently assured procurement requirement.

33.3 Candidate Practice

Use the autonomous cyber evaluation protocol developed under Foundations 1 through 5.

33.4 Stage A: Research Protocol

- Open methodology
- reference tasks
- pilot results
- no compliance claim

33.5 Stage B: Recommended Practice

Publish guidance for:

- Dynamic evaluation
- held-out administration

- human baselines
- safeguard testing
- reporting

33.6 Stage C: Voluntary Framework

Participating organizations disclose:

- Whether they evaluate
- protocol version
- system scope
- independent review
- limitations

33.7 Stage D: Evaluator Qualification

Multiple evaluators complete proficiency testing.

33.8 Stage E: Assurance Scheme

Define:

- Scope
- evaluator requirements
- reports
- surveillance
- complaints
- expiry

33.9 Stage F: Procurement Pilot

A willing purchaser requires the assured evaluation for specified high-autonomy cyber deployments.

33.10 Stage G: Impact Review

Assess:

- Risk evidence
- cost
- evaluator capacity
- small-actor burden
- procurement value
- market effects
- gaming

33.11 Stage H: Formality Decision

Choose:

- Remain voluntary
- expand procurement use
- develop certification
- recommend recognized code
- revise
- retire

33.12 Deliverables

- Maturity assessment
- transition-gate report
- pilot specification
- reporting template
- cost study
- evaluator-capacity report
- procurement clause
- impact evaluation
- final recommendation

33.13 Success Criteria

The pilot succeeds if it demonstrates:

- Evidence-based progression
- clear stop conditions
- independent assurance
- manageable burden
- no broad safety overclaim
- meaningful procurement utility
- revision and retirement paths
- participation beyond large incumbents

---

34. Metrics for Evaluating Progression

34.1 Evidence Maturity

- Independent studies
- replication
- real-world correlation
- uncertainty

- validity

34.2 Adoption

- Participating organizations
- implementation depth
- sector and geography
- persistence

34.3 Assurance Capacity

- Qualified evaluators
- wait times
- proficiency
- security
- consistency

34.4 Burden

- Cost
- time
- personnel
- infrastructure
- small-actor impact

34.5 Effectiveness

- Risk reduction
- better decisions
- incidents prevented
- safeguards improved
- information quality

34.6 Competition

- Market concentration
- entry
- open-source participation
- evaluator choice
- proprietary dependence

34.7 Legitimacy

- Participation

- transparency
- conflict management
- appeal
- public understanding

34.8 Enforcement

- Nonconformity detection
- correction
- consistency
- timeliness
- proportionality

34.9 Adaptation

- Revision speed
- obsolete requirement retirement
- transition quality
- international mapping

34.10 Compliance Quality

- Substantive performance
- paper compliance
- gaming
- audit findings
- repeated failures

---

35. Failure Modes and Safeguards

35.1 Premature Formalization

Failure: Immature practice becomes binding.

Safeguard: Readiness gates, pilots, sunset, independent evidence.

35.2 Permanent Voluntarism

Failure: Severe externalities persist because no formal minimum emerges.

Safeguard: Nonadoption and harm triggers for escalation review.

35.3 Regulatory Ratchet

Failure: Requirements strengthen but never weaken.

Safeguard: symmetric review and retirement mechanisms.

35.4 Compliance Theater

Failure: Documentation substitutes for real outcomes.

Safeguard: performance evidence, testing, incidents, independent review.

35.5 Benchmark Law

Failure: A narrow benchmark becomes a legal capability threshold.

Safeguard: evidence portfolio, dynamic protocols, uncertainty, alternatives.

35.6 Incumbent Capture

Failure: Requirements favor large laboratories.

Safeguard: competition analysis, small-actor pathways, public participation.

35.7 Standards Capture

Failure: A committee codifies member interests.

Safeguard: balance, conflict disclosure, minority reports, public review.

35.8 Auditor Bottleneck

Failure: Requirement activates before assurance capacity exists.

Safeguard: capacity gate, phase-in, shared facilities.

35.9 Certification Safe-Washing

Failure: Narrow conformity becomes broad safety marketing.

Safeguard: claim controls, scope, version, expiry.

35.10 International Fragmentation

Failure: Incompatible requirements multiply.

Safeguard: metadata, crosswalks, mutual recognition.

35.11 Lowest Common Denominator

Failure: International agreement weakens necessary controls.

Safeguard: baseline plus advanced tiers.

35.12 Prescriptive Lock-In

Failure: One method becomes mandatory after better methods emerge.

Safeguard: outcome equivalence, alternative compliance, review.

35.13 Unfunded Mandate

Failure: Actors cannot comply in practice.

Safeguard: support, realistic timeline, cost analysis.

35.14 Open-Source Exclusion

Failure: Requirements assume centralized corporate control.

Safeguard: decentralized pathways and functional equivalence.

35.15 Enforcement Arbitrage

Failure: Actors move activity to weak jurisdictions.

Safeguard: international cooperation and market-access rules.

35.16 Hidden Delegation

Failure: Private standard setters gain public power without accountability.

Safeguard: transparent recognition and version control.

35.17 Safe-Harbor Abuse

Failure: Minimum compliance blocks accountability.

Safeguard: narrow scope, good-faith condition, review, no immunity for concealment.

35.18 Emergency Permanence

Failure: Temporary rules become permanent.

Safeguard: automatic expiry and independent renewal.

35.19 Reporting Overload

Failure: Organizations produce low-value disclosure.

Safeguard: decision-linked reporting and harmonization.

35.20 Stale Requirement

Failure: Rule remains after technical conditions change.

Safeguard: event triggers and periodic review.

---

36. Serious Objections

Objection 1: Voluntary Standards Are Ineffective

Voluntary systems can suffer weak adoption and enforcement.

Response:

- Measure adoption
- add reporting
- use assurance
- use procurement
- escalate when evidence supports it

Residual concern:

Some actors will not adopt costly practices without formal requirements.

Objection 2: Regulation Is the Only Legitimate Mechanism

Public law has unique democratic and enforcement authority.

Response:

Standards, research, assurance, and private ordering can develop technical capacity and evidence that law needs.

They should not replace public authority where coercive power is required.

Objection 3: Progressive Models Delay Necessary Action

They can.

Response:

- Emergency pathways
- precautionary interim controls
- alert thresholds
- rapid review
- high-consequence exceptions

Residual concern:

Institutions may use "more study" to avoid action.

Objection 4: Progressive Models Normalize Inevitable Regulation

Response:

The framework explicitly allows practices to remain voluntary or be retired.

Objection 5: Standards Freeze Innovation

Response:

- Outcome-based requirements
- dynamic annexes
- alternative methods
- review
- sunset

Residual concern:

Compliance ecosystems naturally create inertia.

Objection 6: Risk-Based Rules Are Easy to Game

Response:

- Hybrid triggers
- independent evaluation
- system-level evidence
- anti-circumvention
- incident monitoring

Objection 7: Capability Thresholds Are Too Uncertain

Response:

Use thresholds to trigger review and safeguards rather than automatic prohibition.

Objection 8: Large Firms Will Write the Standards

This risk is substantial.

Response:

- Balanced participation
- open-source and small-actor representation
- public funding
- conflict controls
- competition review

Objection 9: International Harmonization Is Unrealistic

Full uniformity is unrealistic.

Interoperable evidence and mutual recognition are more achievable.

Objection 10: Certification Creates False Trust

Correct when claims are broad.

Response:

Narrow scope, explicit limitations, system identity, surveillance, and expiry.

Objection 11: Private Ordering Is Undemocratic

It can be opaque and coercive.

Response:

Use transparency, interoperability, appeal, competition review, and public law where rights or externalities require it.

Objection 12: Legal Requirements Cannot Keep Pace

Response:

Use stable outcome duties with updateable standards, controlled incorporation, and emergency guidance.

Residual concern:

Delegated updating can weaken accountability.

---

37. Evidence Gaps

37.1 Voluntary Adoption

Which voluntary practices persist and improve outcomes?

37.2 Transition Triggers

What evidence should move a practice from voluntary to mandatory?

37.3 Standards Effectiveness

Do AI standards improve system behavior or mainly documentation?

37.4 Certification

Which claims are mature enough for meaningful certification?

37.5 Procurement

How do procurement requirements affect safety, cost, and competition?

37.6 Insurance

Can insurers evaluate AI controls reliably?

37.7 Capability Thresholds

How should uncertain thresholds connect to formal obligations?

37.8 Small-Actor Burden

Which support mechanisms preserve participation?

37.9 Open-Weight Systems

How should progressive duties apply across decentralized ecosystems?

37.10 International Recognition

Which requirements can be recognized across jurisdictions?

37.11 Dynamic Incorporation

How can standards update without improper delegation?

37.12 Enforcement

Which sanctions improve compliance without suppressing disclosure?

37.13 Safe Harbors

When do they encourage good practice versus excuse weak practice?

37.14 Regulatory Sandboxes

Do they produce transferable evidence?

37.15 Sunset

How often are obsolete requirements actually retired?

---

38. Research Agenda

Priority 1: Maturity Assessment

Develop a repeatable method for classifying requirement readiness.

Priority 2: Transition Gates

Validate gates for evidence, operations, assurance, legitimacy, and cost.

Priority 3: Voluntary Framework Outcomes

Compare implementation depth and outcomes.

Priority 4: Reporting Quality

Develop evidence-backed, comparable reporting.

Priority 5: Standards Impact

Study whether ISO and other AI standards change practice.

Priority 6: Assurance Readiness

Identify which requirements can support audit or certification.

Priority 7: Procurement Pilots

Test private and public purchasing requirements.

Priority 8: Small-Actor Support

Compare grants, shared infrastructure, tiering, and equivalence.

Priority 9: Capability-Linked Obligations

Develop uncertainty-aware trigger models.

Priority 10: Enforcement Ladders

Study proportionate responses and restoration paths.

Priority 11: Safe Harbors

Test disclosure and evaluation incentives.

Priority 12: Regulatory Sandboxes

Create evidence requirements and exit criteria.

Priority 13: International Crosswalks

Map voluntary, standards, and legal regimes.

Priority 14: Version Governance

Develop models for standards incorporated into law.

Priority 15: Requirement Effectiveness

Evaluate real-world outcomes, burden, concentration, and unintended effects.

---

39. Near-Term Experiments

Experiment 1: Readiness-Gate Review

Apply the readiness gates to three candidate evaluation practices.

Experiment 2: Voluntary Reporting Pilot

Test structured disclosure with independent evidence review.

Experiment 3: Procurement Clause

Pilot a bounded evaluation requirement in a voluntary purchasing contract.

Experiment 4: Tiered Burden

Compare full and scaled implementation for large and small organizations.

Experiment 5: Outcome Equivalence

Test two different safeguard methods against the same requirement.

Experiment 6: Evaluator Capacity Simulation

Estimate wait times and cost before formal requirement activation.

Experiment 7: Safe-Harbor Design

Model incentives for voluntary incident disclosure.

Experiment 8: Sandbox

Pilot a new dynamic evaluation method under supervised conditions.

Experiment 9: Certification Comprehension

Test whether users understand narrow versus broad claims.

Experiment 10: Version Transition

Simulate incorporation of a revised standard into contracts and regulation.

Experiment 11: Sunset Review

Apply a retrospective review to an existing AI requirement or commitment.

Experiment 12: Cross-Border Recognition

Compare evidence packages under two jurisdictions.

---

40. Implications for Future Standards

A future standard governing progressive AI requirements could require:

40.1 Problem Definition

The risk or coordination problem.

40.2 Requirement Type

Guidance, framework, standard, assurance, procurement, or mandatory duty.

40.3 Scope

Actors, systems, capabilities, deployments, exclusions.

40.4 Trigger

Capability, deployment, access, use, incident, or hybrid.

40.5 Evidence

Research, pilots, validation, counterevidence, uncertainty.

40.6 Maturity

Construct, method, assurance, operations, economics, legitimacy.

40.7 Proportionality

Burden and consequence analysis.

40.8 Assurance

Self-assessment, review, audit, certification, continuous monitoring.

40.9 Implementation

Tools, guidance, training, infrastructure, phase-in.

40.10 Small-Actor Pathway

Scaled requirements and support.

40.11 Enforcement

Detection, correction, sanction, restoration, appeal.

40.12 Interoperability

Standards mapping, evidence portability, mutual recognition.

40.13 Versioning

Review, update, transition, emergency change.

40.14 Sunset

Expiry, renewal, narrowing, and retirement.

40.15 Impact Evaluation

Effectiveness, burden, competition, innovation, distribution.

Such a standard should be developed through `STANDARDS_DEVELOPMENT_PROCESS.md`.

---

41. Relationship to the Other Foundations

Foundation 1: Dynamic Evaluation Protocols

Requirements referencing evaluations must update as protocols change.

Foundation 2: Held-Out Evaluations

Formal requirements may need protected evidence and secure administration.

Foundation 3: High-Stakes Capability Evaluation

Capability and consequence help determine when stronger obligations are justified.

Foundation 4: Independent Expert Review

Progressive requirements need credible challenge and decision review.

Foundation 5: Third-Party Auditor Ecosystem

Assurance cannot scale without qualified evaluators and accreditation.

Foundation 7: Incentives and Prestige

Voluntary stages rely heavily on incentives, recognition, and market reward.

Foundation 8: Global Interoperability

Standards and requirements should support cross-border evidence and recognition.

---

42. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Frontier AI requirements should mature through evidence-based stages.
2. Voluntary practice is not automatically inferior to mandatory law.
3. Not every voluntary practice should become mandatory.
4. Formality should increase with consequence, evidence, implementation capacity, and legitimacy.
5. A requirement should not become binding before its construct and methods are sufficiently valid for the decision.
6. High-stakes uncertainty can justify interim safeguards before full scientific maturity.
7. Progressive requirements should separate substantive strength, coverage, assurance, transparency, enforcement, and decision consequence.
8. Hybrid triggers are generally preferable to a single model-size or compute proxy for high-stakes obligations.
9. Capability thresholds should often trigger additional review before automatic restriction.
10. Performance-based requirements should be preferred when equivalent methods can be verified.
11. Prescriptive minimums may be justified for essential controls.

12. Standards should normally function as common floors rather than ceilings on stronger practice.
13. Certification should be used only when requirements and assurance methods are mature enough.
14. Broad "safe AI" certification claims should be avoided.
15. Procurement, contracts, insurance, and professional rules can accelerate adoption but do not replace public law.
16. Public authority remains responsible for coercive requirements and legal accountability.
17. Standards incorporated into binding rules require transparent version governance.
18. Legal recognition of private standards should preserve public oversight and alternative compliance where appropriate.
19. Phase-in should reflect evaluator capacity and implementation reality.
20. Support for smaller actors is part of legitimate requirement design.
21. Open-source and decentralized systems require functional, not merely corporate, compliance pathways.
22. Requirements should be evaluated for market concentration and incumbent advantage.
23. Safe harbors should reward good-faith evidence and correction, not excuse concealment or harm.
24. Emergency requirements should expire unless affirmatively renewed.
25. Every significant requirement should have review, appeal, correction, and retirement mechanisms.
26. Compliance evidence should be tied to system version, configuration, and time.
27. Passing a requirement does not prove universal safety.
28. Voluntary nonadoption, repeated incidents, and severe externalities are valid reasons to consider stronger requirements.
29. International interoperability is preferable to unnecessary duplication.
30. Progressive standards should optimize for better outcomes, not the appearance of institutional maturity.

---

43. Decision Rules

A practice is ready to move from research to recommended practice when:

- The problem is defined
- initial evidence supports utility
- limitations are documented
- use does not create excessive risk

A practice is ready to move from recommendation to voluntary framework when:

- Multiple organizations can implement it
- terms are sufficiently clear
- outcomes can be documented
- revision governance exists

A practice is ready to move toward a standard when:

- Methods are repeatable
- scope is defined
- independent evidence exists
- stakeholder participation is credible
- versioning is possible

A practice is ready for assurance when:

- Conformity can be assessed
- evaluators are competent
- evidence can be preserved
- claims can be bounded
- complaints and appeals exist

A practice is ready for procurement or contractual requirement when:

- The purchaser has a legitimate interest
- implementation is feasible
- evidence is decision-relevant
- burden is proportionate
- alternatives are clear

A practice should be considered for mandatory status when:

- Consequence is material
- voluntary adoption is persistently inadequate
- externalities affect nonconsenting parties
- free-rider incentives are strong
- requirements are implementable
- assurance and enforcement capacity exist
- legal authority and due process are clear

A requirement should not be strengthened merely because:

- Public concern increases without evidence
- a dominant company already complies
- an evaluator market seeks demand
- a standards body seeks relevance
- international peers adopted a superficially similar rule
- a single incident is misinterpreted

A requirement should be narrowed, suspended, or retired when:

- Evidence fails
- benefit is not demonstrated
- burden is disproportionate
- concentration is severe
- superior alternatives exist
- enforcement is impossible
- technology makes the rule obsolete
- the requirement causes greater harm than it prevents

---

44. Requirement Maturity Assessment Template

A. Candidate Practice

- Name
- owner
- version
- domain
- current stage

B. Problem

- Risk
- affected parties
- evidence
- externality
- urgency

C. Construct

- Objective
- definitions
- exclusions
- metrics

D. Evidence

- Research
- pilots
- replication
- incidents
- counterevidence
- confidence

E. Implementation

- Methods
- tools
- personnel
- cost
- time
- infrastructure

F. Assurance

- Assessability
- evaluators
- proficiency
- security
- appeals

G. Market Effects

- Entry
- concentration
- open-source impact
- small-actor burden
- proprietary dependence

H. Legitimacy

- Participation
- conflicts
- transparency
- public-interest input
- dissent

I. International

- Standards

- legal conflicts
- mutual recognition
- localization

J. Revision

- Review
- triggers
- versioning
- sunset
- retirement

K. Recommended Stage

- Research
- guidance
- voluntary framework
- reporting
- standard
- assurance
- procurement
- recognized code
- mandatory requirement
- no progression

---

45. Transition-Gate Decision Template

Practice:
Current stage:
Proposed stage:
Decision date:
Decision body:

Reason for Progression

Evidence Supporting Progression

Evidence Against Progression

Readiness Gates

- Problem

- construct
- evidence
- method
- assurance
- operational
- economic
- competition
- legitimacy
- enforcement
- international
- revision

Burden Assessment

Small-Actor Impact

Open-Source Impact

Alternative Options

Phase-In

Support

Enforcement

Appeal

Review Date

Sunset

Decision

- Approve
- pilot
- approve with conditions
- defer
- retain current stage
- move to less formal stage

- retire

---

46. Progressive Requirement Specification Template

A. Identity

- Name
- identifier
- version
- owner
- status

B. Purpose

C. Scope

- Actors
- systems
- capabilities
- deployments
- exclusions

D. Trigger

E. Requirement

F. Evidence

G. Permitted Alternative Methods

H. Assurance

I. Reporting

J. Confidentiality

K. Implementation Support

L. Phase-In

M. Small-Actor Pathway

N. International Mapping

O. Nonconformity

P. Enforcement Ladder

Q. Appeal

R. Review

S. Sunset and Retirement

---

47. Procurement Clause Template

Covered system:
Covered deployment:

Required protocol or standard:
Version:

The supplier shall:

1. Identify the system and configuration.
2. Perform or obtain the specified evaluation.
3. Use a qualified evaluator where required.
4. Provide the defined evidence package.
5. Disclose material limitations and incidents.
6. Notify the purchaser of material system changes.
7. Re-evaluate after specified triggers.
8. Maintain required safeguards.
9. Support audit or verification.
10. Correct material nonconformity within the defined period.

Alternative Evidence

Confidentiality

Subcontractors

Incident Notice

Remedies

Suspension

Appeal or Dispute

Expiration

---

48. Requirement Impact Review Template

Requirement:
Version:
Review period:

Intended Objective

Adoption

Compliance Quality

Effectiveness

Incidents

Burden

Evaluator Capacity

Market Concentration

Small-Actor Effects

Open-Source Effects

Innovation Effects

International Effects

Enforcement

Gaming and Evasion

Stakeholder Feedback

Alternative Methods

Recommendation

- Maintain
- strengthen
- clarify
- narrow

- expand
- suspend
- replace
- retire

---

49. Progressive Standards Scorecard

Dimension	Core Question
Problem	Is the problem specific and materially evidenced?
Objective	Is the desired outcome clear?
Construct	Can the requirement be interpreted consistently?
Evidence	Does evidence support the practice?
Method	Can implementation be repeated?
Dynamic quality	Can the requirement evolve with frontier systems?
Scope	Are covered actors and systems defined?
Trigger	Is application based on defensible evidence?
Proportionality	Does burden match consequence?
Assurance	Can conformity be assessed credibly?
Evaluator capacity	Are qualified assessors available?
Operational readiness	Can organizations implement the rule?
Security	Can sensitive evidence be handled safely?
Transparency	Are reasons, evidence, and status legible?
Competition	Does the requirement avoid unjustified concentration?
Small-actor access	Are realistic pathways available?
Open-source fit	Can decentralized actors demonstrate outcomes?
Innovation	Are alternative methods and improvement possible?
Private ordering	Can contracts and procurement use the requirement coherently?
Legal recognition	Is incorporation accountable and version-controlled?
Enforcement	Can nonconformity be detected and corrected?
Appeals	Can parties challenge error or overreach?
Phase-in	Is timing realistic and risk-based?
International	Is evidence interoperable across jurisdictions?
Review	Are periodic and event-triggered reviews defined?
Sunset	Can obsolete requirements expire?
Impact	Are effectiveness and unintended effects measured?
Legitimacy	Was the process competent, inclusive, and accountable?

Dimension	Core Question
Decision utility	Does the requirement improve a real decision or outcome?

---

50. Final Perspective

Standards become dangerous when they are treated as self-justifying.

A standard can coordinate good practice.

It can also freeze bad measurement.

A voluntary framework can enable experimentation.

It can also allow persistent nonadoption.

A certification can create confidence.

It can also create a misleading badge.

A legal requirement can protect people who cannot protect themselves.

It can also entrench incumbents, suppress beneficial innovation, and preserve assumptions long after the evidence changes.

The answer is not to choose voluntarism or regulation as an ideology.

The answer is to design an evidence-based progression.

That progression should ask:

- Is the problem real?
- Is the practice effective?
- Can it be implemented?
- Can it be evaluated?
- Can smaller actors participate?
- Can requirements be enforced fairly?
- Can decisions be appealed?
- Can methods change?
- Can the requirement end?

The future of frontier AI governance will likely include all of the following:

- Research
- voluntary frameworks
- public reporting
- organizational commitments
- technical standards
- third-party assurance
- contracts

- procurement
- insurance
- regulation
- international agreements

The question is not which single mechanism should govern everything.

The question is which mechanism is justified at which stage, for which system, under which evidence, and with which safeguards against institutional failure.

Progressive standards should preserve the ability to act before harm is obvious.

They should also preserve the ability to learn before law hardens.

They should create stronger expectations where consequences rise.

They should resist turning early assumptions into permanent authority.

The sixth foundation of Standards Body is therefore disciplined institutional progression.

Standards should become stronger when the evidence and the stakes justify strength.

They should remain revisable because the systems they govern will not stand still.

References and Research Basis

[^nist-rmf]: National Institute of Standards and Technology, **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**, 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[^nist-airc]: National Institute of Standards and Technology, **NIST AI Resource Center**. https://airc.nist.gov/

[^nist-playbook]: National Institute of Standards and Technology, **AI RMF Playbook**. https://airc.nist.gov/airmf-resources/playbook/

[^nist-tevv]: National Institute of Standards and Technology, **AI Test, Evaluation, Validation and Verification**. https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv

[^nist-global]: National Institute of Standards and Technology, **A Plan for Global Engagement on AI Standards**, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf

[^nist-zero-draft]: National Institute of Standards and Technology, **Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation**, 2025. https://www.nist.gov/document/outline-proposed-zero-draft-standard-ai-testing-evaluation-verification-and-validation

[^iso-42001]: International Organization for Standardization, **ISO/IEC 42001:2023, Artificial intelligence management systems**. https://www.iso.org/standard/42001

[^iso-42006]: International Organization for Standardization, **ISO/IEC 42006:2025, Requirements for bodies providing audit and certification of artificial intelligence management systems**. https://www.iso.org/standard/42006

[^iso-23894]: International Organization for Standardization, **ISO/IEC 23894:2023, Artificial intelligence, Guidance on risk management**. https://www.iso.org/standard/77304.html

[^iso-standards]: International Organization for Standardization, **Standards**. https://www.iso.org/standards.html

[^haip-code]: G7 Hiroshima AI Process, **International Code of Conduct for Organizations Developing Advanced AI Systems**, 2023. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

[^haip-reporting]: OECD, **Hiroshima AI Process Reporting Framework**. https://oecd.ai/en/hiroshima

[^haip-about]: OECD, **About the Hiroshima AI Process Reporting Framework**. https://oecd.ai/en/transparency/about

[^haip-insights]: OECD, **Early Insights from the Hiroshima AI Process Reporting Framework**, 2025. https://oecd.ai/en/wonk/haip-reporting-insights

[^eu-ai-act]: European Union, **Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence**, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[^eu-summary]: EUR-Lex, **Rules for Trustworthy Artificial Intelligence in the European Union**. https://eur-lex.europa.eu/EN/legal-content/summary/rules-for-trustworthy-artificial-intelligence-in-the-eu.html

[^coe-convention]: Council of Europe, **Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law**, 2024. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[^openai-pf]: OpenAI, **Preparedness Framework, Version 2**, April 15, 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[^anthropic-rsp]: Anthropic, **Responsible Scaling Policy**, 2026. https://www.anthropic.com/responsible-scaling-policy

[^deepmind-fsf]: Google DeepMind, **Frontier Safety Framework**, updated 2025. https://deepmind.google/blog/strengthening-our-frontier-safety-framework/

[^frontier-forum]: Frontier Model Forum, **Issue Brief: Components of Frontier AI Safety Frameworks**, 2024. https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/

Revision Record

Version 1.0

Date: July 16, 2026

Change type: Complete foundational edition

Summary: Establishes the fully developed canonical working white paper for Foundation 6. Defines the rationale for progressive institutionalization, a fourteen-stage standards ladder, readiness gates, design principles, trigger models, risk tiers, voluntary frameworks, reporting, frontier safety frameworks, consensus standards, assurance, private ordering, legal recognition, mandatory requirements, authorization, enforcement, phase-in, small-actor and open-source pathways, sandboxes, safe harbors, anti-capture, international interoperability, version governance, sunset, requirement evaluation, governance, implementation, a Standards Body pilot, metrics, failure analysis, objections, evidence gaps, research agenda, operational templates, scorecard, and primary-source research basis.

Status: Ready for internal review and future expert critique.