Standards Body · Foundation paper, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/foundation-paper/independent-expert-review/

FOUNDATION_04_INDEPENDENT_EXPERT_REVIEW.md

Foundation 4: Independent Expert Review

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 evaluator access, review practice, institutional governance, or frontier AI capabilities

---

Document Purpose

This paper defines the Standards Body position on independent expert review for frontier artificial intelligence.

It is intended to serve as:

- A first-principles explanation of why developer evaluation should be complemented by external scrutiny
- A framework for defining meaningful independence rather than relying on the label alone
- A guide to selecting, supporting, governing, and evaluating expert reviewers
- A bridge between technical evaluation, institutional legitimacy, public accountability, and secure access
- A reference for future standards, review panels, evaluator guidance, contributor systems, and institutional partnerships
- A durable source document from which shorter articles, procedural manuals, and technical specifications can be developed

This paper is not an accreditation standard for evaluator organizations. That subject belongs primarily in `FOUNDATION_05_THIRD_PARTY_AUDITOR_ECOSYSTEM.md` and `EVALUATOR_ACCREDITATION_FRAMEWORK.md`.

This paper is not a claim that external reviewers are inherently correct, neutral, or free from incentives.

It establishes the conditions under which expert review can add credible, decision-relevant evidence to frontier AI evaluation.

Executive Summary

Frontier AI systems are commonly developed, configured, evaluated, secured, and released by the same organizations.

Those organizations possess knowledge that outside reviewers often do not.

They understand:

- The model architecture
- Training and post-training methods
- Internal evaluation results
- System prompts and safeguards
- Known incidents and failure modes
- Deployment constraints
- Model-access infrastructure
- Product timelines
- Security considerations
- Commercial and strategic context

Developer participation is therefore indispensable.

But developer expertise and developer independence are different properties.

The same organization may be responsible for:

1. Defining which risks deserve evaluation
2. Selecting the evaluation methods
3. Configuring the model
4. Interpreting ambiguous results
5. Determining whether a capability threshold has been crossed
6. Judging whether safeguards are adequate
7. Deciding what evidence becomes public
8. Choosing whether deployment proceeds

Even highly responsible organizations operate under incentives that can affect these judgments.

Relevant pressures can include:

- Product deadlines
- Competitive dynamics
- Investor expectations

- Government relationships
- Employee commitments
- Public reputation
- Access to capital
- Legal exposure
- Institutional optimism
- Familiarity with internal assumptions
- Desire to avoid releasing sensitive information
- Dependence on a planned deployment

These pressures do not prove misconduct.

They create a structural reason for independent challenge.

Independent expert review adds value when qualified people or institutions can:

- Examine claims made by system developers
- Challenge evaluation scope
- Test alternative threat models
- assess whether model access was sufficient
- identify missing evidence
- reproduce or independently interpret results
- scrutinize safeguards
- surface domain-specific concerns
- record disagreement
- provide additional confidence to decision-makers and the public

Independent review should not be understood as a ceremonial external endorsement.

A reviewer is not meaningfully independent merely because they:

- Work for a different legal entity
- Sign a conflict disclosure
- Receive a private demonstration
- Comment on a finished report
- Are described as an expert
- Are allowed to publish a quotation
- Lack direct employment by the developer

Independence is multidimensional.

It includes at least:

1. **Organizational independence**
   The reviewer is not governed by the organization whose claims are under review.

2. **Financial independence**
   The reviewer is not so dependent on the reviewed party that unfavorable conclusions threaten its survival or future access.

3. **Methodological independence**
   The reviewer can challenge the evaluation design, propose additional tests, and reject unsupported interpretations.

4. **Informational independence**
   The reviewer has enough access to form a judgment rather than merely validating a curated presentation.

5. **Publication independence**
   The reviewer can communicate material conclusions under agreed security and confidentiality constraints.

6. **Decision independence**
   The reviewer is not required to align its conclusion with the developer's preferred outcome.

7. **Intellectual independence**
   The reviewer is willing to revise their own assumptions and is not selected only because their prior views are convenient.

8. **Security independence**
   The reviewer can protect sensitive material without relying entirely on controls administered by the reviewed party.

No reviewer will maximize every dimension.

The objective is not perfect detachment.

The objective is sufficient independence for the claim and decision at issue.

Independence is also insufficient on its own.

An external reviewer may be:

- Technically underqualified
- unfamiliar with the domain
- denied adequate access
- rushed
- poorly resourced
- biased
- captured by ideology
- careless with security
- unable to reproduce results
- selected for prestige rather than competence
- unwilling to record uncertainty
- overconfident outside their expertise

Independent review must therefore be paired with:

- Competence
- access

- time
- resources
- method
- security
- conflict management
- documentation
- procedural fairness
- accountability
- quality assurance

Standards Body adopts the following core position:

> **No organization should be the sole authority over the evaluation of its own most consequential AI systems. Developer evidence should be complemented by independent expert review that is sufficiently competent, informed, resourced, secure, contestable, and free from controlling conflicts to improve the quality of consequential decisions.**

A mature review ecosystem should recognize multiple forms of review:

- Independent protocol review
- Independent task review
- Independent model evaluation
- Safeguard review
- Risk-report review
- Deployment-readiness review
- Incident review
- Standards review
- Replication
- Red teaming
- Public-interest review
- Open-source and open-weight review
- International peer review

These functions should not be collapsed into one generic category.

The level of review should increase with:

- Potential consequence
- Uncertainty
- novelty
- developer conflict
- restricted information
- threshold significance
- deployment scale
- model autonomy
- irreversibility
- public reliance on the result

The strongest review systems should also preserve dissent.

A review is less credible when every reviewer is expected to produce consensus.

Unresolved disagreement can be decision-relevant evidence.

The purpose of independent review is not to create a new class of unquestionable authorities.

It is to create a disciplined system in which important claims face informed challenge before institutions, markets, governments, or the public rely on them.

---

1. Foundational Proposition

1.1 Core Thesis

> **Developer evaluation is necessary, but consequential frontier AI claims should not depend solely on developer judgment.**

1.2 Independence Thesis

> **Independence should be assessed across multiple dimensions and relative to the specific claim, access arrangement, funding model, and decision.**

1.3 Competence Thesis

> **Independent but unqualified review can reduce rather than increase trust.**

1.4 Access Thesis

> **Review quality is constrained by what reviewers can inspect, test, question, and reproduce.**

1.5 Accountability Thesis

> **Independent reviewers should themselves be subject to transparent methods, conflict controls, quality review, correction, and appeal.**

1.6 Pluralism Thesis

> **No single professional community, institution, laboratory, government, or worldview should define the complete evaluation agenda for frontier AI.**

1.7 Proportionality Thesis

> **The depth and formality of review should be proportionate to the potential consequence and evidentiary uncertainty of the decision.**

---

2. Scope and Boundaries

2.1 What This Foundation Covers

This paper covers independent expert review of:

- Evaluation protocols
- Test tasks and datasets
- Capability findings
- Safeguard effectiveness
- Risk reports
- Safety cases
- Deployment plans
- Security claims
- Model access arrangements
- Threshold determinations
- Evaluation incidents
- Standards proposals
- Evaluator methods
- Post-deployment evidence

2.2 What This Foundation Does Not Fully Cover

This paper does not fully specify:

- Accreditation of evaluator organizations
- Certification schemes
- Legal audit requirements
- Formal inspection powers
- Regulatory enforcement
- Procurement law
- Professional licensing
- Complete institutional governance

Those topics are addressed elsewhere in the Standards Body architecture.

2.3 Expert Review Versus Public Consultation

Expert review evaluates technical or institutional evidence using relevant competence.

Public consultation enables broader participation, legitimacy, local knowledge, and affected-party input.

Both matter.

Neither should substitute for the other.

2.4 Expert Review Versus Audit

A review may:

- Challenge reasoning
- assess evidence
- recommend additional work
- interpret findings

An audit normally assesses conformity with defined criteria through a more formal evidence process.

Some activities combine both.

The term used should match the function.

2.5 Expert Review Versus Red Teaming

Red teaming seeks to identify weaknesses through adversarial testing.

Independent review can include red teaming but also covers:

- Methodology
- evidence
- governance
- scope
- interpretation
- conflicts
- decision linkage

2.6 Expert Review Versus Certification

Certification is a third-party attestation that specified requirements are fulfilled.

Independent review may inform certification but should not be described as certification unless the full scheme supports that claim.

2.7 Expert Review Versus Peer Review

Academic peer review normally evaluates research for publication.

Frontier AI review may require:

- Secure model access
- confidential evidence
- operational testing
- incident data
- time-sensitive decisions
- security controls
- deployment context

Traditional peer review alone is not enough.

---

3. Canonical Definitions

3.1 Expert

An expert is a person with demonstrable knowledge, skill, judgment, or experience relevant to the specific review question.

Prestige, title, institutional affiliation, or public visibility is not sufficient evidence of expertise.

3.2 Reviewer

A reviewer is a person or body responsible for examining evidence, methods, claims, or decisions and producing a reasoned judgment.

3.3 Independent Reviewer

An independent reviewer is sufficiently free from controlling interests and organizational authority to form and communicate a genuine judgment.

3.4 External Reviewer

An external reviewer is outside the reviewed organization.

Externality is one component of independence, not proof of it.

3.5 Third Party

A third party is separate from the supplier or developer and the immediate user or purchaser whose interests are directly represented.

In conformity-assessment practice, first-party, second-party, and third-party activities are distinguished according to the relationship of the assessor to the object and interested parties. [^iso-casco-bodies]

3.6 Subject-Matter Expert

A subject-matter expert possesses specialized domain competence, such as:

- Cybersecurity
- biology
- critical infrastructure
- statistics
- model evaluation
- safety engineering
- human factors
- governance
- security

3.7 Review Panel

A review panel is a group assembled to provide complementary expertise, deliberation, and judgment.

3.8 Conflict of Interest

A conflict of interest is a relationship, incentive, obligation, or commitment that could impair, or reasonably appear to impair, impartial judgment.

3.9 Impartiality

Impartiality is the presence of objectivity and the management of conflicts so that decisions are not improperly influenced.

3.10 Competence

Competence is the demonstrated ability to apply knowledge and skills to achieve intended results.

3.11 Review Mandate

A review mandate defines the question, scope, authority, access, outputs, constraints, timeline, and decision relationship of a review.

3.12 Review Independence

Review independence is the degree to which the reviewer can select methods, interpret evidence, communicate findings, and resist improper influence.

3.13 Review Access

Review access is the combination of model access, system information, artifacts, personnel, time, tools, and permissions available to the reviewer.

3.14 Review Finding

A review finding is a supported conclusion concerning evidence, process, conformance, risk, capability, safeguard effectiveness, or uncertainty.

3.15 Dissent

Dissent is a reasoned disagreement with the majority or final institutional conclusion.

3.16 Minority Report

A minority report is a documented dissenting assessment issued alongside or in connection with the primary review conclusion.

3.17 Right of Reply

The right of reply allows the reviewed party to respond to factual or procedural findings before publication without obtaining veto power over the conclusion.

3.18 Reviewer Capture

Reviewer capture occurs when a reviewer becomes materially influenced by the organization, funder, professional community, government, or ideology it is expected to scrutinize.

3.19 Access Asymmetry

Access asymmetry occurs when one reviewer or participant receives materially different information, model access, time, or support from another.

3.20 Epistemic Diversity

Epistemic diversity is the inclusion of different forms of relevant knowledge, methods, assumptions, and professional experience.

---

4. Why Independent Review Is Necessary

4.1 Information Concentration

Frontier developers possess essential private information.

This creates a tension.

Outside reviewers need access to evaluate claims, but developers often control access.

Independent review helps convert private evidence into more credible public or institutional assurance.

4.2 Internal Bias

Internal teams can be affected by:

- Familiarity with assumptions
- Shared incentives
- organizational hierarchy
- sunk costs
- release pressure
- selection of favorable metrics
- normalization of known failures
- internal language that obscures external interpretation

NIST's AI Risk Management Framework notes that independent review can improve testing effectiveness and reduce internal bias and conflicts of interest.[^nist-rmf-core]

4.3 Scope Selection

The choice of what to evaluate can be as important as the result.

Independent reviewers can ask:

- Which risks were excluded?
- Which system configuration was tested?
- Which users were represented?
- Which deployment conditions were ignored?
- Which threshold was selected?
- Which evidence was treated as decisive?

4.4 Alternative Threat Models

Developers may optimize for known threat models.

Independent experts can introduce:

- Different adversaries
- different deployment assumptions
- different misuse pathways
- different system boundaries
- different failure scenarios
- different institutional concerns

4.5 Verification

External testing can confirm, refine, or challenge internal findings.

AISI has described third-party evaluation as a source of independent verification that can complement developer testing and reduce real or perceived conflicts.[^aisi-lessons]

4.6 Public Confidence

Trust should not depend solely on self-description.

Independent scrutiny can increase confidence when:

- Reviewers are competent
- access is adequate
- methods are visible
- conflicts are managed
- findings are not censored
- limitations are clear

4.7 Decision Quality

Independent review can improve:

- Deployment decisions
- safeguard design
- standards
- procurement
- governance
- incident response
- research priorities

4.8 Error Correction

External reviewers may identify:

- Invalid tasks
- scoring errors

- weak elicitation
- contaminated benchmarks
- missing controls
- unsafe procedures
- overclaims
- underclaims

4.9 Institutional Legitimacy

Institutions gain legitimacy when important judgments can be challenged by qualified parties outside the immediate decision chain.

4.10 Limits

Independent review cannot guarantee:

- Truth
- neutrality
- complete access
- future safety
- absence of capture
- correct policy
- public agreement

It is one layer of assurance.

---

5. Independence Is Multidimensional

5.1 Organizational Independence

Questions:

- Is the reviewer legally separate?
- Can the reviewed party remove leadership?
- Does the reviewed party control the reviewer's mandate?
- Does the reviewer depend on the reviewed party for facilities or personnel?

5.2 Financial Independence

Questions:

- What percentage of revenue comes from the reviewed party?
- Is payment contingent on completion, result, or publication?
- Will an unfavorable finding threaten future contracts?
- Are reviewers compensated directly by the developer?

- Is funding pooled or mediated?

Financial ties do not automatically invalidate review.

They should be disclosed and managed.

5.3 Governance Independence

Questions:

- Who appoints reviewers?
- Who can dismiss them?
- Who approves scope?
- Who adjudicates disputes?
- Who controls publication?

5.4 Methodological Independence

Questions:

- Can reviewers propose additional tests?
- Can they challenge thresholds?
- Can they reject developer-defined interpretations?
- Can they access raw evidence?
- Can they choose independent tools?

5.5 Informational Independence

Questions:

- Does the reviewer receive full or curated evidence?
- Are internal incidents disclosed?
- Are model limitations known?
- Can the reviewer interview relevant personnel?
- Are omissions visible?

5.6 Operational Independence

Questions:

- Can reviewers run tests themselves?
- Can they verify model identity?
- Can they control the environment?
- Can they preserve logs?
- Can they rerun disputed tests?

5.7 Publication Independence

Questions:

- Can reviewers publish findings?
- What redactions are permitted?
- Who decides security concerns?
- Can the developer delay release?
- Can dissent be published?

5.8 Intellectual Independence

Questions:

- Was the reviewer selected because they are likely to agree?
- Can the reviewer revise prior beliefs?
- Does the panel include methodological disagreement?
- Are critics excluded categorically?

5.9 Professional Independence

Questions:

- Does the reviewer seek employment, investment, prestige, or access from the reviewed party?
- Is the reviewer embedded in the same professional network?
- Could reputational incentives distort judgment?

5.10 Political Independence

Questions:

- Is the review controlled by a government with strategic interests?
- Are national-security priorities shaping evidence interpretation?
- Can international reviewers participate?
- Are political conclusions separated from technical findings?

5.11 Security Independence

Questions:

- Can the reviewer protect sensitive evidence?
- Are logs controlled by the developer?
- Can access be revoked during disagreement?
- Is evidence preserved independently?

5.12 Independence Profile

Every material review should include an independence profile.

The profile should state:

- Strengths
- limitations
- conflicts
- access dependencies
- publication constraints
- residual concerns

---

6. Independence Is Not Isolation

A credible review system requires cooperation.

6.1 Developer Knowledge

Developers may need to provide:

- Model access
- architecture information
- system prompts
- safeguards
- known incidents
- training cutoff
- internal findings
- deployment plans
- technical support

6.2 Reviewer Challenge

Reviewers need enough separation to question that information.

6.3 Productive Relationship

The desired relationship is neither:

- Developer control
- permanent hostility
- passive endorsement
- unrestricted disclosure

It is structured cooperation under independent judgment.

6.4 Adversarial and Collaborative Modes

Some review tasks benefit from collaboration:

- Integration
- elicitation
- factual clarification
- remediation

Others benefit from adversarial independence:

- Red teaming
- safeguard challenge
- threshold review
- incident investigation

A review mandate should identify the mode.

6.5 Trust but Verify

"Trust but verify" is incomplete when verification is weak.

The relevant principle is:

> Cooperate enough to understand the system, verify enough to form an independent judgment.

---

7. Competence Framework

7.1 General Competencies

All reviewers should demonstrate:

- Analytical reasoning
- evidence assessment
- uncertainty communication
- conflict awareness
- documentation
- confidentiality
- professional judgment
- ability to distinguish fact from inference

7.2 Technical Evaluation Competence

May include:

- Machine learning
- benchmark design
- statistics
- experimental design
- model elicitation
- agent evaluation
- scoring
- reproducibility
- contamination analysis

7.3 Domain Competence

Examples:

- Cyber operations
- biology
- chemistry
- critical infrastructure
- finance
- persuasion
- human factors
- AI research and development
- security engineering

7.4 Institutional Competence

May include:

- Governance
- standards
- risk management
- safety cases
- conformity assessment
- audit
- regulation
- incident response
- organizational incentives

7.5 Sociotechnical Competence

May include:

- User behavior
- labor
- civil rights
- accessibility
- public-interest technology
- deployment context
- organizational practice
- community impact

7.6 Security Competence

Reviewers handling sensitive systems may need:

- Threat modeling
- secure operations
- access control
- incident response
- dual-use awareness
- protected disclosure

7.7 Review Competence

Expertise in the subject is not the same as expertise in reviewing.

Reviewers should know how to:

- Follow a mandate
- sample evidence
- document findings
- avoid overclaiming
- handle disagreement
- preserve independence
- distinguish major and minor issues

7.8 Team Competence

A panel can possess collective competence that no single member has.

The mandate should identify:

- Required disciplines
- coverage gaps
- role of each member

- integration process
- final decision method

7.9 Competence Evidence

Possible evidence:

- Relevant research
- professional experience
- demonstrated evaluations
- peer recognition
- proficiency testing
- work samples
- training
- incident experience
- references

Titles should not substitute for evidence.

7.10 Competence Limits

Every reviewer should state:

- Areas of expertise
- areas outside expertise
- reliance on others
- unresolved knowledge gaps

---

8. Reviewer Roles

8.1 Technical Protocol Reviewer

Examines:

- Construct
- tasks
- scoring
- reliability
- validity
- versioning
- limitations

8.2 Domain Reviewer

Examines:

- Realism
- operational relevance
- actor models
- consequences
- safeguards
- domain assumptions

8.3 Security Reviewer

Examines:

- Access
- threat model
- sensitive content
- infrastructure
- incident response
- model or data protection

8.4 Elicitation Reviewer

Examines whether capability was sufficiently revealed.

8.5 Safeguard Reviewer

Examines controls against representative and adaptive attacks.

8.6 Governance Reviewer

Examines:

- Authority
- conflicts
- thresholds
- decision process
- appeals
- transparency

8.7 Public-Interest Reviewer

Examines:

- Social consequence

- affected parties
- access
- concentration
- rights
- legitimacy
- distributional impact

8.8 Open-Source and Open-Weight Reviewer

Examines:

- Reproducibility
- community governance
- downstream modification
- weight access
- decentralized deployment
- smaller-actor participation
- benefit and risk asymmetries

8.9 Statistical Reviewer

Examines:

- Sampling
- uncertainty
- thresholds
- aggregation
- reliability
- false-positive and false-negative risk

8.10 Incident Reviewer

Examines:

- Timeline
- root cause
- controls
- disclosure
- corrective action
- recurrence

8.11 Review Chair

Coordinates scope, process, deliberation, dissent, and reporting.

The chair should not dominate technical conclusions outside their competence.

8.12 Observer

An observer may monitor process without voting or accessing all content.

---

9. Types of Independent Review

9.1 Protocol Review

Occurs before active evaluation.

Questions:

- Is the construct valid?
- Are tasks representative?
- Is scoring defensible?
- Are change and expiration rules adequate?

9.2 Pre-Administration Review

Examines readiness:

- Model access
- environment
- security
- configuration
- evaluator resources
- task integrity

9.3 Concurrent Review

Reviewer observes or participates during testing.

Useful for:

- Detecting deviations
- understanding anomalies
- preserving evidence

9.4 Result Review

Examines:

- Raw outputs
- scoring
- uncertainty

- interpretation
- claims

9.5 Independent Replication

A separate party repeats the evaluation.

9.6 Risk-Report Review

Examines the developer's integrated reasoning, evidence, assumptions, and mitigation decisions.

Anthropic's Responsible Scaling Policy now authorizes and, in specified circumstances, requires external review of risk reports under governance arrangements involving its Long-Term Benefit Trust.[^anthropic-rsp]

9.7 Safety-Case Review

Examines whether a structured argument adequately supports a defined deployment.

9.8 Safeguard Review

Examines whether controls reduce risk under representative attack.

9.9 Deployment-Readiness Review

Examines:

- Capability
- safeguards
- monitoring
- incident response
- scale
- access
- rollback

9.10 Post-Deployment Review

Examines real-world evidence and drift.

9.11 Incident Review

Triggered by:

- Misuse
- unexpected capability

- safeguard failure
- security compromise
- evaluation failure

9.12 Standards Review

Examines proposed technical or institutional standards.

9.13 Compliance Review

Assesses adherence to a defined framework or requirement.

9.14 Meta-Review

Examines the quality and consistency of earlier reviews.

9.15 Continuous Review

Ongoing access and evidence collection support repeated assurance.

This is more resource-intensive and belongs at higher maturity levels.

---

10. Selecting Reviewers

10.1 Selection Principle

Select reviewers for the mandate, not for prestige.

10.2 Required Criteria

- Relevant competence
- independence profile
- security capability
- availability
- review experience
- ability to communicate uncertainty
- willingness to record dissent
- absence of disqualifying conflicts

10.3 Selection Body

Potential selectors:

- Independent secretariat
- governing board committee
- multi-stakeholder panel
- public institute
- accredited organization
- random or rotating expert pool

The reviewed party may propose reviewers but should not control final selection for high-consequence review.

10.4 Reviewer Pool

Maintain a pool with metadata on:

- Expertise
- jurisdiction
- language
- conflicts
- security clearance
- prior reviews
- performance
- availability
- open-source experience
- public-interest experience

10.5 Rotation

Rotation can reduce:

- Familiarity bias
- dependence
- capture
- intellectual monoculture

Excessive rotation can lose institutional memory.

Use staggered terms.

10.6 Repeat Engagement

Repeat reviewers can gain valuable system knowledge.

Risks should be managed through:

- Revenue limits
- periodic independent members
- peer review
- rotation
- public disclosure
- performance review

10.7 Reviewer Diversity

Diversity should include relevant differences in:

- Discipline
- institution
- methodology
- geography
- language
- professional background
- open and closed model experience
- public and private sector
- risk perspective

Diversity should not be reduced to symbolic demographic representation.

10.8 Critics

Serious critics can improve review.

They should not be excluded merely for prior disagreement.

They should also be held to evidence and professional standards.

10.9 Developer Objection

The reviewed party may identify:

- Conflict
- lack of competence
- security concern
- prior misconduct

Objection should be adjudicated independently.

It should not become veto power over rigorous reviewers.

10.10 Public Nomination

For some panels, public nomination can expand the pool.

Final selection should remain competence-based.

---

11. Open-Source and Open-Weight Representation

11.1 Why Representation Matters

Open and closed development create different:

- Access models
- governance structures
- security assumptions
- reproducibility conditions
- downstream modifications
- benefit pathways
- misuse pathways
- market effects

A review system dominated by frontier laboratories may misunderstand decentralized development.

A review system dominated by open-source advocates may underestimate security and deployment concerns.

Both perspectives should face evidence.

11.2 Meaningful Participation

Open-source representation should include authority to:

- Review constructs
- challenge assumptions
- examine access implications
- assess downstream modification
- contribute tasks
- record dissent
- review public claims

11.3 Selection

Representatives should demonstrate:

- Technical contribution

- community knowledge
- relevant evaluation competence
- security awareness
- independence

Public visibility alone is insufficient.

11.4 Avoiding Tokenism

One representative should not be expected to speak for all open communities.

11.5 Conflicts

Open-source reviewers can also have conflicts:

- Project affiliation
- ideology
- commercial services
- competitive interest
- grant dependence

11.6 Confidential Access

Some open-source experts may require controlled access to review closed systems.

The process should not assume that community affiliation is incompatible with security.

11.7 Contribution to Standards

Open-source expertise is particularly valuable for:

- Reproducibility
- portable evaluation
- model modification
- decentralized deployment
- smaller evaluator access
- public infrastructure
- international participation

---

12. Public-Interest and Affected-Party Participation

12.1 Technical Expertise Is Not Complete Expertise

People affected by deployment may understand:

- Workflow failures
- institutional incentives
- practical harms
- accessibility
- implementation gaps
- cultural context

12.2 Appropriate Roles

Affected parties can contribute through:

- Issue identification
- scenario review
- deployment-context review
- public consultation
- post-deployment evidence
- governance oversight

12.3 Limits

Affected-party participation should not be asked to certify technical claims outside competence.

12.4 Confidentiality

Participation processes should avoid exposing individuals to:

- Retaliation
- privacy loss
- sensitive model content
- unnecessary security risk

12.5 Compensation

Community expertise should not be extracted without fair compensation where resources allow.

---

13. Review Mandate

Every review should begin with a written mandate.

13.1 Identity

- Review name
- reviewed object
- version
- sponsor
- reviewed party
- reviewers
- date

13.2 Question

What must the review determine?

13.3 Scope

What is included and excluded?

13.4 Authority

Can reviewers:

- Request evidence
- run tests
- interview personnel
- challenge scope
- publish findings
- issue dissent
- recommend suspension

13.5 Access

Specify:

- Model access
- system information
- artifacts
- logs
- people
- tools
- time
- compute

13.6 Standard of Evidence

What evidence is sufficient?

13.7 Outputs

Possible outputs:

- Advisory note
- technical report
- public summary
- restricted annex
- conformance finding
- dissent
- remediation plan

13.8 Decision Relationship

Who receives the review?

What decision can it influence?

13.9 Constraints

- Security
- legal
- privacy
- export control
- time
- publication
- access

13.10 Appeal

Define grounds and process.

13.11 Review Expiration

State how long the conclusion remains current.

---

14. Access

Review quality depends on access.

14.1 Model Access

Possible levels:

- Public interface
- API
- controlled endpoint
- evaluator-operated deployment
- weights
- fine-tuning
- internal tools

14.2 Information Access

Possible information:

- Architecture
- training methods
- training cutoff
- safety tuning
- system prompt
- internal evaluations
- incidents
- deployment plans
- risk reports
- model changes

14.3 Time Access

Reviewers need time for:

- Integration
- task development
- elicitation
- debugging
- analysis
- replication
- writing
- developer response

14.4 Personnel Access

Reviewers may need interviews with:

- Evaluation leads
- model developers

- safety teams
- security teams
- deployment owners
- governance leaders
- incident responders

14.5 Artifact Access

May include:

- Raw logs
- transcripts
- scoring code
- task provenance
- model cards
- internal reports
- safeguard tests
- change history

14.6 Environment Access

Reviewers may need:

- Sandboxes
- tools
- network
- reference systems
- monitoring
- controlled data

14.7 Access Profile

Every review report should state its access profile.

Research on external frontier-model evaluation proposes separating model access, model information, and evaluation timeframe because each changes what conclusions are justified. [^external-access]

14.8 Access Limitations

If access is insufficient, the conclusion should be narrowed.

14.9 Security Tradeoffs

Greater access can improve rigor while increasing:

- Model theft risk
- task leakage
- sensitive information exposure
- misuse
- operational burden

Use proportional controls rather than automatic denial.

14.10 Access Equality

Different reviewers may have different access.

Material asymmetries should be disclosed.

---

15. Review Methodology

15.1 Question Formulation

Translate the mandate into review questions.

15.2 Evidence Map

Identify:

- Claims
- required evidence
- sources
- missing evidence
- conflicts
- uncertainty

15.3 Sampling

Select representative:

- Tasks
- records
- incidents
- systems
- personnel
- configurations

- time periods

15.4 Triangulation

Compare:

- Developer evidence
- independent tests
- external research
- operational data
- expert judgment
- incidents
- prior reviews

15.5 Reperformance

Where possible, rerun or reproduce critical evidence.

15.6 Challenge Testing

Attempt to falsify important claims.

15.7 Interviews

Use structured interviews and preserve notes appropriately.

15.8 Configuration Verification

Confirm that the reviewed evidence applies to the actual system.

15.9 Uncertainty

Record:

- Unknowns
- access limits
- disagreement
- methodological limits
- time constraints

15.10 Findings

Classify findings by:

- Severity

- confidence
- decision relevance
- evidence
- remediation urgency

15.11 Review Trail

Preserve:

- Evidence requests
- responses
- tests
- deviations
- conflicts
- decisions
- dissent
- changes

15.12 Method Independence

The reviewer should not simply restate the developer's framework.

They should evaluate whether that framework is itself adequate.

---

16. Evidence Standards

16.1 Evidence Hierarchy

No rigid hierarchy fits every question.

Reviewers should distinguish:

- Direct test evidence
- operational evidence
- incident evidence
- internal analysis
- expert judgment
- simulation
- public research
- assertion

16.2 Raw Evidence

Where practical, review:

- Transcripts
- logs
- task-level scores
- failure examples
- system configurations
- version records

16.3 Developer Assertions

Assertions should be supported by verifiable evidence.

16.4 Confidential Evidence

Confidentiality does not reduce the need to assess:

- Source
- completeness
- conflict
- validity
- chain of custody

16.5 Negative Evidence

Absence of a finding may reflect:

- Insufficient search
- weak access
- short timeframe
- immature method

16.6 Confidence

Each major conclusion should include confidence and rationale.

16.7 Corroboration

High-consequence conclusions should ideally receive more than one independent line of evidence.

16.8 Review Independence From Evidence Production

Where the reviewer participated in creating the evidence, that dual role should be disclosed.

17. Developer Interaction

17.1 Initial Briefing

The developer explains:

- System
- scope
- known limitations
- access
- security
- internal evidence

17.2 Reviewer Questions

Reviewers should be able to request:

- Clarification
- raw evidence
- additional tests
- personnel interviews
- configuration changes
- known incidents

17.3 Integration Support

Developer support may be needed to avoid false negatives caused by poor integration.

Support should not become task coaching.

17.4 Factual Review

Before publication, the developer may review for:

- Factual errors
- security
- confidential information
- misunderstood configuration

17.5 No Conclusion Veto

The developer should not suppress an unfavorable but supported conclusion.

17.6 Disagreement

The final report may include:

- Reviewer conclusion
- developer response
- unresolved disagreement
- evidence needed to resolve it

17.7 Remediation

Reviewers can assess corrective actions without becoming responsible for operating them.

17.8 Ongoing Access

Repeated review can improve understanding.

It can also increase dependence.

Manage through rotation, disclosure, and governance.

---

18. Conflicts of Interest

18.1 Conflict Categories

Financial

- Fees
- grants
- investment
- equity
- employment
- vendor relationship

Professional

- Collaboration
- future employment
- board membership
- shared institution
- publication dependence

Personal

- Close relationship
- dispute

- loyalty
- hostility

Intellectual

- Public commitment
- proprietary method
- ideological identity
- reputational stake

Political

- Government agenda
- national competition
- regulatory advocacy

Competitive

- Rival product
- rival evaluation provider
- litigation

18.2 Disclosure

Disclosure should be specific enough to evaluate materiality.

18.3 Management Options

- No action
- additional disclosure
- role limitation
- independent co-review
- recusal
- replacement
- public explanation

18.4 Disqualifying Conflicts

Examples may include:

- Compensation tied to favorable result
- undisclosed material equity
- direct authority over reviewed decision
- review of one's own work without independent control
- active litigation materially related to the finding
- inability to publish because of private dependency

18.5 Institutional Conflict

An organization can be conflicted even if individual reviewers are not.

18.6 Appearance

Perceived conflict matters because legitimacy depends on justified public interpretation.

Appearance alone should not become a tool for arbitrary exclusion.

18.7 Conflict Register

Maintain:

- Interest
- assessment
- decision
- mitigation
- reviewer
- date
- update

---

19. Funding and Compensation

19.1 Reviewers Should Be Paid

Serious review requires time and expertise.

Unpaid systems can:

- Exclude less wealthy experts
- reduce quality
- favor institutions with independent funding
- create hidden incentives
- delay work

19.2 Payment Risks

Direct developer payment can create dependence.

19.3 Funding Models

Direct Fee

Developer pays reviewer.

Strength:

- Simple
- scalable

Risk:

- Client dependence

Pooled Fund

Developers contribute to a fund administered independently.

Strength:

- Reduces direct leverage

Risk:

- Collective industry capture

Public Funding

Government funds review.

Strength:

- Public mandate

Risk:

- Political influence
- procurement delay

Philanthropic Funding

Strength:

- Supports public-interest work

Risk:

- Funder agenda
- instability

Mixed Funding

Can reduce dependence if governed carefully.

19.4 Fee Structure

Payment should not depend on:

- Passing
- favorable conclusion
- publication outcome

19.5 Revenue Concentration

Reviewer organizations should disclose material client concentration.

19.6 Access as Compensation

Early model access, prestige, and relationships can be nonfinancial benefits that influence reviewers.

19.7 Compensation Transparency

Disclose:

- Payer
- fee structure
- nonfinancial benefits
- publication rights
- repeat engagement

---

20. Security and Confidentiality

20.1 Sensitive Inputs

Review may involve:

- Model weights
- vulnerabilities
- held-out tasks
- incidents
- user data
- security architecture
- risk reports
- proprietary methods

20.2 Reviewer Security Requirements

May include:

- Identity assurance
- secure devices
- access control
- encryption
- logging
- data retention
- incident response
- compartmentalization
- secure facilities

20.3 Need-to-Know

Not every panel member requires access to every artifact.

20.4 Confidential Deliberation

Some deliberation may remain private to support candor.

Material reasoning should still be represented in the report.

20.5 Redaction

Redaction should protect legitimate interests without removing the basis of the conclusion.

20.6 Security Review of Publication

A separate security reviewer can assess proposed disclosure.

The reviewed party may raise concerns but should not be the sole decision-maker.

20.7 Incident Duties

Reviewers should report:

- Leakage
- unauthorized access
- unsafe system behavior
- compromised evidence
- conflicts discovered late

20.8 Secure Access Infrastructure

OpenAI's 2026 third-party evaluation playbook highlights the importance of early model access, appropriate technical integration, clear communication, and secure handling for credible frontier evaluations.[^openai-playbook]

20.9 Confidentiality Limits

Reviewers should not promise confidentiality beyond their technical, legal, and institutional capacity.

---

21. Deliberation and Dissent

21.1 Consensus Is Not Always the Goal

Forced consensus can hide:

- Uncertainty
- methodological disagreement
- value disagreement
- evidence gaps
- minority expertise

21.2 Deliberation Process

A panel should:

1. Review evidence individually
2. share preliminary findings
3. identify disagreements
4. test reasons
5. seek additional evidence
6. record unresolved differences
7. issue majority and minority conclusions where needed

21.3 Dissent Categories

- Factual
- methodological
- interpretive
- threshold
- policy
- procedural
- security

21.4 Minority Report

A minority report should include:

- Disputed claim
- evidence
- reasoning
- consequence
- what would resolve disagreement

21.5 Anonymous Dissent

Anonymous dissent may protect individuals but weaken accountability.

Use only with justified safeguards.

21.6 No Retaliation

Reviewers should not lose access or future eligibility merely for evidence-based dissent.

21.7 Review Chair Duty

The chair should ensure dissent is represented fairly.

---

22. Appeals and Corrections

22.1 Grounds for Appeal

- Factual error
- scoring error
- invalid task
- wrong configuration
- undisclosed conflict
- procedural deviation
- insufficient access
- evidence omission
- security compromise
- overbroad interpretation

22.2 Non-Grounds

- Dislike of the conclusion
- commercial inconvenience
- political pressure

- unfavorable publicity

22.3 Appeal Body

Should include qualified people not responsible for the original decision.

22.4 Appeal Outcomes

- Uphold
- clarify
- correct
- narrow
- re-evaluate
- withdraw
- replace reviewer
- reopen evidence

22.5 Publication During Appeal

Material findings may be:

- Published with appeal status
- temporarily held
- corrected later

The approach should reflect consequence and urgency.

22.6 Corrections

Maintain:

- Original finding
- correction
- reason
- date
- impact
- reviewer
- affected decisions

22.7 Learning

Appeals should improve methods, not merely resolve individual disputes.

---

23. Transparency and Reporting

23.1 Public Reporting Minimum

A public review should disclose:

- Mandate
- scope
- reviewers or selection process
- expertise
- independence profile
- funding
- access
- methodology
- major findings
- confidence
- limitations
- dissent
- developer response
- result expiration

23.2 Restricted Annex

May include:

- Raw evidence
- vulnerabilities
- private tasks
- sensitive incidents
- system details
- reviewer deliberation

23.3 Review Access Statement

The report should state what reviewers could and could not access.

23.4 Independence Statement

Include:

- Relationships
- payment
- constraints
- publication rights
- selection process

23.5 Claims Boundary

State what the review does not establish.

23.6 Developer Response

Publish a concise response where appropriate.

23.7 Review Status

- Draft
- final
- under appeal
- corrected
- superseded
- withdrawn
- expired

23.8 Public Legibility

Avoid language that implies:

- Government approval
- certification
- complete safety
- unanimous expert consensus
- authority beyond the mandate

---

24. Quality Assurance for Reviewers

24.1 Reviewer Performance

Independent reviewers need evaluation.

24.2 Quality Indicators

- Accuracy
- completeness
- reproducibility
- timeliness
- security
- conflict handling
- evidence quality

- uncertainty
- useful recommendations
- correction behavior

24.3 Peer Review of Review

A second independent party can assess the review process.

24.4 Calibration Exercises

Reviewers can assess common cases and compare judgments.

24.5 Proficiency Testing

Where methods are standardized, reviewers can demonstrate performance on known or controlled problems.

24.6 Feedback

Collect feedback from:

- Reviewed party
- decision-maker
- peer reviewers
- affected stakeholders
- security staff

Feedback should not become client control.

24.7 Reviewer Record

Maintain:

- Mandates
- conflicts
- findings
- corrections
- security incidents
- appeals
- performance reviews

24.8 Suspension

Reviewers may be suspended for:

- Serious misconduct

- security breach
- undisclosed conflict
- repeated methodological failure
- fabrication
- retaliation
- inability to maintain competence

24.9 Foundation 5 Boundary

Formal accreditation, surveillance, and recognition systems will be developed in Foundation 5 and the accreditation framework.

---

25. Institutional Models

25.1 Ad Hoc Expert Panel

Strengths:

- Fast
- flexible
- specialized

Risks:

- Inconsistent
- weak continuity
- unclear accountability

25.2 Academic Consortium

Strengths:

- Research competence
- disciplinary depth

Risks:

- Slow
- publication incentives
- limited security
- uneven operational access

25.3 Government AI Institute

Strengths:

- Public mandate
- secure access
- continuity
- national coordination

Risks:

- political influence
- jurisdiction limits
- procurement constraints

The UK AI Security Institute has conducted frontier-model evaluations and published lessons about third-party evaluation, access, and the evolving science of testing.[^aisi-lessons][^aisi-trends]

25.4 Independent Nonprofit

Strengths:

- Mission focus
- public-interest orientation
- potential flexibility

Risks:

- funding dependence
- access dependence
- capacity limits

25.5 Commercial Evaluator

Strengths:

- Scalability
- professional operations
- repeatable service

Risks:

- client dependence
- proprietary methods
- market concentration

25.6 Multi-Stakeholder Review Body

Strengths:

- Diverse legitimacy
- broad expertise

Risks:

- slow deliberation
- political bargaining
- diffuse accountability

25.7 Standards-Organization Working Group

Strengths:

- Process
- international participation
- consensus infrastructure

Risks:

- slower than frontier change
- incumbent influence
- limited direct model access

25.8 Distributed Peer Network

Strengths:

- pluralism
- resilience
- open participation

Risks:

- inconsistent competence
- weak security
- coordination

25.9 Hybrid Model

A strong ecosystem may combine:

- Government institutes
- independent nonprofits
- commercial evaluators
- academics

- standards bodies
- open-source experts
- public-interest organizations

---

26. Conformity-Assessment Lessons

Frontier AI is not identical to established conformity-assessment domains.

It can still learn from them.

26.1 Competence, Consistency, and Impartiality

ISO/CASCO standards emphasize competence, consistent operation, and impartiality across testing, inspection, validation, verification, and certification activities.[^iso-casco-bodies][^iso-building-trust]

26.2 First, Second, and Third Party

The relationship of the assessor to the object matters.

26.3 Recognition

Conformity-assessment competence can be recognized through:

- Accreditation
- government recognition
- peer assessment

ISO/CASCO describes several recognition routes for conformity-assessment bodies.[^iso-recognition]

26.4 Applicability to Frontier AI

Useful lessons include:

- Define the scheme
- define competence
- manage conflicts
- preserve records
- calibrate assessors
- conduct surveillance
- support appeals
- separate standard setting from certification where possible

26.5 Limits of Analogy

Frontier AI differs because:

- Systems change rapidly
- constructs remain immature
- models may behave strategically
- evidence is security-sensitive
- evaluation access is concentrated
- standards are incomplete
- global consensus is limited

---

27. Review Levels

Standards Body proposes five review levels.

Level 0: Internal Peer Challenge

Characteristics:

- Different internal team
- documented challenge
- no external claim of independence

Use:

- Early development

Level 1: External Advisory Review

Characteristics:

- External experts
- limited access
- advisory findings
- narrow scope

Use:

- Protocol refinement
- low-consequence claims

Level 2: Independent Technical Review

Characteristics:

- Formal mandate

- sufficient model and evidence access
- conflict controls
- independent findings
- public summary
- appeal

Use:

- Material capability and safeguard claims

Level 3: Multi-Disciplinary High-Stakes Review

Characteristics:

- Multiple independent disciplines
- held-out testing
- security
- deployment context
- dissent
- external replication
- governance review

Use:

- High-stakes deployment or threshold decisions

Level 4: Continuous Multi-Institution Assurance

Characteristics:

- Ongoing access
- repeated review
- independent audits
- incident feedback
- international participation
- evaluator oversight
- continuous monitoring

Use:

- Critical systems and mature institutional regimes

The review level should match the claim and consequence.

---

28. Review Lifecycle

Stage 1: Trigger

A review begins because of:

- New model
- protocol change
- threshold finding
- deployment
- incident
- standard
- disputed claim
- scheduled review

Stage 2: Mandate

Define question, scope, authority, access, and outputs.

Stage 3: Reviewer Selection

Assess competence, independence, security, and conflicts.

Stage 4: Access Agreement

Define technical, informational, personnel, and time access.

Stage 5: Evidence Plan

Map claims to evidence.

Stage 6: Orientation

Developer and reviewers establish shared factual understanding.

Stage 7: Independent Analysis

Reviewers test, inspect, reproduce, interview, and challenge.

Stage 8: Deliberation

Reviewers compare findings and identify dissent.

Stage 9: Draft Report

Include evidence, confidence, limitations, and recommendations.

Stage 10: Factual and Security Review

Reviewed party identifies errors and security concerns.

Stage 11: Finalization

Reviewer retains conclusion authority within mandate.

Stage 12: Publication or Controlled Delivery

Issue public and restricted outputs.

Stage 13: Appeal

Resolve material challenges.

Stage 14: Remediation Review

Assess corrective action where relevant.

Stage 15: Expiration and Renewal

Re-review after system or evidence change.

---

29. Implementation Pathway

Phase 1: Define Reviewable Claims

List claims that should not rely solely on developer judgment.

Phase 2: Create Reviewer Taxonomy

Identify required roles and competence.

Phase 3: Establish Independence Criteria

Define conflicts, funding, selection, and publication rights.

Phase 4: Build Reviewer Pool

Recruit across:

- Technical
- domain
- security
- governance
- public interest
- open source
- international communities

Phase 5: Create Mandate Templates

Standardize scope, access, evidence, and outputs.

Phase 6: Pilot Access Profiles

Test different model, information, and time access arrangements.

Phase 7: Pilot Review Levels

Compare advisory and high-stakes reviews.

Phase 8: Build Secure Infrastructure

Support controlled evidence and model access.

Phase 9: Establish Dissent and Appeal

Make disagreement operational.

Phase 10: Evaluate Reviewers

Measure quality and revise process.

Phase 11: Publish Review Methodology

Make rigor legible.

Phase 12: Connect to Accreditation

Use lessons to develop Foundation 5.

---

30. Proposed Standards Body Pilot

30.1 Pilot Name

Independent Expert Review of a Dynamic Held-Out Evaluation Protocol

30.2 Reviewed Object

A pilot protocol developed under:

- `FOUNDATION_01_DYNAMIC_EVALUATION_PROTOCOLS.md`
- `FOUNDATION_02_HELD_OUT_EVALUATIONS.md`
- `FOUNDATION_03_HIGH_STAKES_CAPABILITY_EVALUATION.md`

The preferred initial domain is autonomous cyber capability.

30.3 Review Panel

Include:

- Evaluation scientist
- cybersecurity expert
- statistician
- security engineer
- governance expert
- open-source or open-weight expert
- public-interest technology reviewer

30.4 Mandate

Review:

- Capability graph
- task validity
- held-out integrity
- elicitation
- scoring
- thresholds
- safeguards
- deployment relevance
- governance
- public claims

30.5 Access

Provide:

- Protocol
- active task samples under control
- scoring code
- reference results
- environment
- task provenance
- threat model
- change history
- evaluator interviews

30.6 Independence Controls

- Reviewer selection by independent committee
- conflict register
- fixed compensation
- no result-dependent payment
- publication rights
- minority report
- appeal

30.7 Outputs

- Public review report
- restricted technical annex
- independence profile
- access statement
- dissent
- remediation recommendations
- process evaluation

30.8 Success Criteria

The pilot succeeds if it:

- Finds material issues not visible internally
- improves the protocol
- preserves security
- supports reasoned dissent
- produces a legible public account
- demonstrates adequate access
- avoids developer or reviewer control
- provides evidence for future accreditation requirements

31. Metrics for Evaluating Independent Review

31.1 Review Quality

- Factual accuracy
- evidence quality
- scope coverage
- reproducibility
- uncertainty
- usefulness
- correction rate

31.2 Independence

- Funding concentration
- selection control
- publication constraints
- conflicts
- developer influence
- reviewer dependence

31.3 Access

- Model access
- information access
- timeframe
- personnel access
- artifact access
- unresolved requests

31.4 Competence

- Required disciplines covered
- reviewer performance
- calibration
- domain adequacy
- security performance

31.5 Deliberation

- Dissent recorded
- evidence exchanged
- minority views

- unresolved issues
- chair neutrality

31.6 Operational Performance

- Cost
- duration
- integration delays
- security incidents
- reviewer burden
- developer burden

31.7 Decision Utility

- Decisions changed
- safeguards improved
- claims corrected
- false confidence reduced
- incidents anticipated
- public understanding

31.8 Ecosystem Quality

- Reviewer diversity
- open-source participation
- geographic participation
- repeat-client dependence
- access to smaller organizations

---

32. Failure Modes and Safeguards

32.1 Independence Theater

Failure: External names are used to legitimize a developer-controlled process.

Safeguard: Independence profile, selection transparency, publication rights, access statement.

32.2 Prestige Substitution

Failure: Famous reviewers are selected without relevant competence.

Safeguard: Mandate-specific competence evidence.

32.3 Underaccessed Review

Failure: Reviewers validate curated evidence without sufficient model or information access.

Safeguard: Access profile and narrowed claims.

32.4 Client Capture

Failure: Reviewer depends financially on repeated developer contracts.

Safeguard: Revenue disclosure, pooled funding, rotation, peer review.

32.5 Government Capture

Failure: National interests distort technical findings.

Safeguard: international participation, technical-policy separation, dissent.

32.6 Ideological Capture

Failure: Reviewers approach evidence with fixed pro- or anti-AI conclusions.

Safeguard: methodological pluralism and evidence standards.

32.7 Reviewer Monoculture

Failure: Everyone shares the same discipline, institution, or assumptions.

Safeguard: complementary panel design.

32.8 Token Representation

Failure: Open-source, public-interest, or affected-party members are present without authority.

Safeguard: defined role, access, vote, and dissent rights.

32.9 Forced Consensus

Failure: Disagreement is removed to create a clean conclusion.

Safeguard: minority reports and unresolved-issue register.

32.10 Unlimited Developer Review

Failure: Factual review becomes conclusion negotiation.

Safeguard: time limit, issue log, reviewer final authority.

32.11 Confidentiality Abuse

Failure: Security is used to conceal weak evidence.

Safeguard: controlled independent scrutiny and public methods.

32.12 Reviewer Overclaiming

Failure: Reviewer speaks beyond mandate or competence.

Safeguard: claims boundary and peer review.

32.13 Reviewer Security Failure

Failure: Sensitive models, tasks, or incidents leak.

Safeguard: security qualification, least privilege, incident response.

32.14 Excessive Formality

Failure: Review becomes slow, costly, and disconnected from capability pace.

Safeguard: tiered review levels and clear triggers.

32.15 Insufficient Formality

Failure: High-stakes decisions rely on informal consultation.

Safeguard: mandatory mandate, records, conflicts, and report.

32.16 Access Revocation Pressure

Failure: Reviewer fears losing future model access.

Safeguard: access agreements, pooled infrastructure, public disclosure.

32.17 Review Shopping

Failure: A developer seeks reviewers until receiving a favorable finding.

Safeguard: reviewer registry, disclosure of prior reviews, governance selection.

32.18 Fragmented Findings

Failure: Multiple reviews produce incompatible conclusions without explanation.

Safeguard: shared metadata, meta-review, explicit methodological differences.

32.19 Review as Certification

Failure: Advisory findings are marketed as proof of safety.

Safeguard: precise terminology and public claims controls.

32.20 Stale Review

Failure: Old findings are applied to changed systems.

Safeguard: expiration and event-triggered renewal.

---

33. Serious Objections

Objection 1: External Reviewers Lack Developer Knowledge

Often true.

Response:

- Require developer cooperation
- provide access
- include developer evidence
- allow factual review
- use complementary internal and external work

Residual concern:

Some knowledge remains tacit and difficult to transfer.

Objection 2: Independent Review Slows Deployment

It can.

Response:

- Use proportional review levels
- begin access early
- maintain standing reviewer pools
- use reusable infrastructure
- predefine mandates

Residual concern:

High-quality review requires time.

Objection 3: External Review Creates Security Risk

It can expose:

- Model weights
- vulnerabilities
- tasks
- incidents

Response:

- Security qualification
- controlled access
- least privilege
- secure environments
- restricted annexes

Residual concern:

Risk cannot be eliminated.

Objection 4: Reviewers Can Be Biased or Political

Correct.

Response:

- Independence profiles
- plural panels
- evidence standards
- dissent
- quality review
- appeals

Residual concern:

No reviewer is value-free.

Objection 5: The Developer Pays, So the Reviewer Is Not Independent

Direct payment creates a conflict, but it does not automatically invalidate review.

Response:

- Fixed fees
- result-independent payment
- disclosure
- pooled funding
- revenue limits

- external selection

Residual concern:

Future business can still influence behavior.

Objection 6: Government Review Is More Legitimate

Government review can carry public authority.

It can also face:

- Political influence
- national interests
- capacity limits
- opacity

A mixed ecosystem is preferable.

Objection 7: Open-Source Representation Weakens Security

Community affiliation does not imply security incapacity.

Use competence and controls rather than categorical exclusion.

Objection 8: Too Many Stakeholders Dilute Technical Quality

Response:

Assign roles according to competence.

Not every participant should decide every technical question.

Objection 9: Independent Review Will Become a Closed Profession

This is a serious risk.

Response:

- Transparent competence pathways
- small-evaluator access
- open tools
- supervised participation
- rotation
- public-interest funding

Objection 10: Reviewers Cannot Test Unknown Unknowns

Correct.

Independent review broadens challenge but cannot guarantee completeness.

Objection 11: Public Dissent Reduces Trust

It may reduce superficial confidence.

It can increase justified trust by making uncertainty visible.

Objection 12: Existing Lab Frameworks Already Include External Testing

Several frontier developers use external testing or external review.

This is valuable.

It does not eliminate the need for common standards governing independence, access, method, and claims.[^openai-external][^anthropic-rsp]

---

34. Evidence Gaps

34.1 Independence Metrics

Which measures predict genuinely independent reviewer behavior?

34.2 Access

What minimum access is required for different claim types?

34.3 Reviewer Selection

Which selection models reduce capture while preserving competence and speed?

34.4 Funding

Which funding models best balance sustainability and independence?

34.5 Panel Composition

How does disciplinary and institutional diversity affect findings?

34.6 Dissent

When does formal dissent improve decision quality?

34.7 Review Effectiveness

How often does external review identify material issues missed internally?

34.8 Reviewer Calibration

Can reviewers be meaningfully calibrated across frontier evaluations?

34.9 Security

Which controls enable deep access without unacceptable leakage risk?

34.10 Open-Source Participation

Which mechanisms support meaningful and secure participation?

34.11 International Review

How should competing national interests be managed?

34.12 Decision Impact

Do external reviews change deployment and safeguards in beneficial ways?

34.13 Review Expiration

How long should findings remain current?

34.14 Strategic Models

How should reviewers adapt if models can recognize and manipulate evaluation?

---

35. Research Agenda

Priority 1: Independence Profiles

Develop a standardized multidimensional independence disclosure.

Priority 2: Access Levels

Validate access categories across real evaluations.

Priority 3: Reviewer Competence

Develop task-specific competence models and proficiency exercises.

Priority 4: Funding Models

Compare direct fees, pooled funds, public funding, and mixed structures.

Priority 5: Selection Governance

Pilot independent committees, random selection, and rotating pools.

Priority 6: Review Methodology

Develop common procedures for evidence mapping, reperformance, challenge, and dissent.

Priority 7: Open-Source Participation

Design secure, meaningful pathways for open and decentralized communities.

Priority 8: Public-Interest Review

Clarify how sociotechnical evidence should connect to technical evaluation.

Priority 9: Reviewer Security

Develop secure access environments and reviewer security standards.

Priority 10: Meta-Review

Evaluate consistency and rigor across independent reviews.

Priority 11: Review Effectiveness

Track whether reviews detect problems and improve decisions.

Priority 12: Continuous Assurance

Study ongoing rather than one-time review.

Priority 13: Strategic Behavior

Develop review methods resistant to model manipulation and sabotage.

Priority 14: International Recognition

Create requirements for cross-border recognition of review findings.

---

36. Near-Term Experiments

Experiment 1: Access Comparison

Have reviewers assess the same claim under limited, moderate, and deep access.

Experiment 2: Reviewer Composition

Compare single-discipline and multidisciplinary panels.

Experiment 3: Independent Selection

Compare developer-selected and independently selected reviewers.

Experiment 4: Funding Disclosure

Test how different payment structures affect perceived and actual independence.

Experiment 5: Dissent Protocol

Run a panel with a formal minority-report mechanism.

Experiment 6: Open-Source Participation

Include qualified open-weight experts in a controlled review and evaluate their distinct contributions.

Experiment 7: Factual Review Boundary

Test a structured developer response process that separates correction from conclusion negotiation.

Experiment 8: Meta-Review

Have a second panel assess the first panel's methodology.

Experiment 9: Continuous Review

Compare one-time review with ongoing evidence access.

Experiment 10: Reviewer Calibration

Give reviewers common cases and measure judgment consistency.

Experiment 11: Security Drill

Simulate sensitive-evidence compromise during review.

Experiment 12: Review Impact

Track which recommendations are accepted and whether outcomes improve.

---

37. Implications for Future Standards

A future independent expert review standard could require:

37.1 Mandate

A written question, scope, authority, and decision relationship.

37.2 Competence

Documented reviewer and panel competence.

37.3 Independence Profile

Organizational, financial, methodological, informational, publication, and intellectual independence.

37.4 Conflict Management

Disclosure, assessment, mitigation, recusal, and register.

37.5 Selection

Transparent and proportionate reviewer-selection process.

37.6 Access

Defined model, information, personnel, artifact, and time access.

37.7 Method

Evidence map, sampling, testing, reperformance, interviews, and uncertainty.

37.8 Security

Controls for sensitive models, tasks, incidents, and reports.

37.9 Deliberation

Rules for panel judgment and dissent.

37.10 Developer Response

Factual and security review without conclusion veto.

37.11 Reporting

Public mandate, access, methods, findings, limitations, funding, conflicts, and status.

37.12 Appeal

Independent challenge and correction process.

37.13 Quality Assurance

Reviewer performance, meta-review, and suspension.

37.14 Expiration

Time and event triggers for renewal.

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

---

38. Relationship to the Other Foundations

Foundation 1: Dynamic Evaluation Protocols

Independent reviewers help determine when evaluations should change, expire, or retire.

Foundation 2: Held-Out Evaluations

Protected content requires independent scrutiny so confidentiality does not shield weak methods.

Foundation 3: High-Stakes Capability Evaluation

The consequence of the capability should determine review depth and panel competence.

Foundation 5: Third-Party Auditor Ecosystem

Foundation 4 defines the review function and principles. Foundation 5 will define how organizations scale, qualify, compete, and remain accountable.

Foundation 6: Progressive Standards and Requirements

Independent review can mature from voluntary challenge into procurement, certification, or formal assurance.

Foundation 7: Incentives and Prestige

Recognition should reward rigorous, honest review, including correction and dissent.

Foundation 8: Global Interoperability

Review findings need compatible metadata, competence, access, and recognition across jurisdictions.

---

39. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Developer evaluation is essential but insufficient for the most consequential claims.
2. No organization should be the sole authority over the evaluation of its own most consequential systems.
3. Externality is not the same as independence.
4. Independence is multidimensional and claim-specific.
5. Independence without competence does not create trustworthy review.
6. Competence without sufficient access does not support broad conclusions.
7. Every material review should disclose its mandate, access, funding, conflicts, and limitations.

8. Review depth should increase with potential consequence and uncertainty.
9. Reviewer selection should be based on mandate-specific competence rather than prestige.
10. The reviewed party may contribute expertise and correct factual errors but should not control the final conclusion.
11. Reviewers should be paid fairly, but payment should not depend on a favorable outcome.
12. Repeat client relationships require additional safeguards.
13. Independent reviewers should themselves be accountable, correctable, and reviewable.
14. Dissent is legitimate evidence and should not be erased for appearance of consensus.
15. Open-source and open-weight expertise should receive meaningful representation where relevant.
16. Public-interest and affected-party knowledge should complement technical review without replacing technical competence.
17. Security restrictions should be proportionate and should not prevent qualified scrutiny.
18. A review should state what reviewers could not access.
19. Advisory review should not be marketed as certification or proof of safety.
20. Reviews should expire after material system, protocol, or deployment changes.
21. Multiple independent institutions are preferable to one permanent review monopoly.
22. International recognition should depend on competence, independence, access, method, and accountability.
23. Public transparency alone cannot replace secure expert review of legitimately confidential evidence.
24. Confidential evidence alone cannot replace public accountability.
25. The review ecosystem should be evaluated for capture, concentration, exclusion, and effectiveness.

---

40. Decision Rules

Independent review should be required when:

- A capability finding may trigger major safeguards or deployment action
- A developer makes a consequential public safety claim
- Evaluation evidence is materially confidential
- A threshold is disputed or near a boundary
- The model has unusually high autonomy or access

- A major incident occurs
- A protocol changes substantially
- Evaluator access is contested
- The decision may affect large populations or critical systems
- Internal teams have material conflicts
- A standard or certification claim is proposed

A review should not be described as independent when:

- The reviewed party controls reviewer selection without oversight
- Payment depends on a favorable result
- Reviewers cannot challenge scope
- Access is too narrow for the claim
- Findings cannot be communicated
- conflicts are undisclosed
- the reviewer is validating their own work without independent control

A reviewer should be replaced or supplemented when:

- Required competence is missing
- a material conflict cannot be mitigated
- security requirements cannot be met
- repeated methodological failure occurs
- access limitations make the mandate impossible
- public claims exceed evidence
- retaliation or bad-faith conduct occurs

A review should be renewed when:

- Model version changes
- system prompt changes materially
- tools or scaffolds change
- deployment expands
- safeguards change
- new incidents occur
- protocol changes
- evidence expires
- reviewer access improves materially

---

41. Independent Review Mandate Template

A. Identity

- Review name
- identifier
- reviewed object

- version
- sponsor
- reviewed party
- date

B. Review Question

C. Decision Relationship

- Decision
- decision-maker
- consequence
- timeline

D. Scope

- Included
- excluded
- dependencies

E. Reviewers

- Names or selection process
- roles
- competence
- jurisdiction
- security status

F. Independence Profile

- Organizational
- financial
- governance
- methodological
- informational
- publication
- intellectual
- security

G. Conflicts

- Interest
- assessment
- mitigation

- recusal

H. Access

- Model
- information
- artifacts
- personnel
- tools
- environment
- time
- compute

I. Methods

- Evidence review
- testing
- sampling
- interviews
- reperformance
- challenge
- deliberation

J. Evidence Standard

K. Security and Confidentiality

L. Outputs

- Public report
- restricted annex
- dissent
- remediation

M. Developer Response

N. Appeal

O. Expiration

---

42. Independence Profile Template

Reviewer or organization:
Reviewed party:
Mandate:
Date:

Organizational Independence

Financial Relationships

Governance Relationships

Professional Relationships

Intellectual Commitments

Political or National Interests

Selection Process

Methodological Authority

Information Access

Operational Access

Publication Rights

Security Dependencies

Future Engagements

Mitigations

Residual Concerns

Overall Independence Assessment

- Strong
- adequate with controls
- limited
- insufficient

---

43. Reviewer Conflict Disclosure Template

Reviewer:
Organization:
Review:

Employment and Consulting

Funding and Grants

Equity and Investments

Board or Advisory Roles

Research Collaborations

Competitive Interests

Personal Relationships

Public Commitments

Litigation or Disputes

Government or Political Interests

Other Material Interests

Proposed Mitigation

Decision

- No material conflict
- disclose
- role limitation
- co-review
- recusal
- disqualify

---

44. Review Finding Template

Finding identifier:
Review:
Category:
Severity:
Confidence:

Claim

Evidence

Method

Access Limitations

Alternative Explanations

Developer Response

Reviewer Assessment

Dissent

Decision Relevance

Recommendation

Remediation

Verification

Expiration

---

45. Minority Report Template

Review:
Minority reviewer or group:
Date:

Majority Conclusion in Dispute

Minority Conclusion

Evidence

Methodological Difference

Interpretation Difference

Consequence

What Would Resolve the Disagreement

Recommended Decision Treatment

---

46. Independent Review Scorecard

Dimension	Core Question
Mandate	Is the question, scope, and authority explicit?
Competence	Does the reviewer or panel cover the required expertise?
Organizational independence	Is the reviewer outside the reviewed authority chain?
Financial independence	Are payment and client dependence controlled?
Methodological independence	Can reviewers challenge methods and add tests?
Informational independence	Is access sufficient and not excessively curated?
Operational independence	Can reviewers verify and run relevant evidence?
Publication independence	Can material findings be communicated?
Intellectual independence	Is the panel capable of genuine disagreement?
Selection	Was reviewer choice legitimate and transparent?
Conflicts	Are interests disclosed and managed?
Access	Are model, information, artifacts, personnel, and time adequate?

Dimension	Core Question
Evidence	Are findings supported by appropriate evidence?
Reperformance	Can critical evidence be reproduced or verified?
Security	Are sensitive materials handled competently?
Deliberation	Is panel reasoning documented?
Dissent	Can minority views be recorded?
Developer response	Can factual errors be corrected without conclusion veto?
Reporting	Are methods, access, funding, limitations, and findings legible?
Appeal	Can material error be challenged independently?
Quality assurance	Is reviewer performance evaluated?
Open-source participation	Is relevant decentralized expertise included meaningfully?
Public interest	Are deployment and affected-party considerations represented?
Expiration	Is the review tied to time and system version?
Decision utility	Does the review improve an actual decision?

---

47. Final Perspective

Independent expert review is often described as a trust mechanism.

That description is incomplete.

Trust should not be produced by the presence of an external name on a report.

It should be produced by a process that makes scrutiny real.

A credible reviewer needs:

- The right expertise
- The right question
- The right access
- The right amount of time
- The right to challenge assumptions
- The right to record uncertainty
- The right to disagree
- The ability to protect sensitive evidence
- The freedom to communicate supported conclusions
- Accountability for error

Frontier AI creates a difficult asymmetry.

Developers possess the strongest access.

Outsiders may possess greater independence.

Neither access without independence nor independence without access is enough.

The institutional task is to combine them.

This requires cooperation without control.

It requires confidentiality without opacity.

It requires expertise without prestige theater.

It requires compensation without purchased conclusions.

It requires pluralism without loss of technical rigor.

It requires public accountability without unsafe disclosure.

It also requires humility.

Independent reviewers can be wrong.

Panels can be captured.

Governments can be political.

Nonprofits can depend on funders.

Academics can chase publication.

Commercial evaluators can chase contracts.

Critics can become ideological.

Experts can overreach.

For that reason, independent review should not create a new unquestionable authority.

It should create a more contestable evidence system.

The fourth foundation of Standards Body is therefore not external approval.

It is informed challenge under conditions that make judgment genuinely independent.

References and Research Basis

[^nist-rmf-core]: National Institute of Standards and Technology, **AI RMF Core**, including guidance on independent review and internal bias. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[^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-conformance]: National Institute of Standards and Technology, **Overview of Conformance Testing**. https://www.nist.gov/itl/ai/applied-ai-research-group/overview-conformance-testing

[^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

[^aisi-lessons]: UK AI Security Institute, **Early Lessons from Evaluating Frontier AI Systems**, October 24, 2024. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems

[^aisi-trends]: UK AI Security Institute, **Frontier AI Trends Report**. https://www.aisi.gov.uk/frontier-ai-trends-report

[^aisi-qa]: UK AI Security Institute, **Early Insights from Developing Question-Answer Evaluations for Frontier AI**, September 23, 2024. https://www.aisi.gov.uk/blog/early-insights-from-developing-question-answer-evaluations-for-frontier-ai

[^aisi-research]: UK AI Security Institute, **Research Agenda**. https://www.aisi.gov.uk/research-agenda

[^openai-external]: OpenAI, **Strengthening Our Safety Ecosystem with External Testing**, November 19, 2025. https://openai.com/index/strengthening-safety-with-external-testing/

[^openai-playbook]: OpenAI, **A Shared Playbook for Trustworthy Third-Party Evaluations**, May 29, 2026. https://openai.com/index/trustworthy-third-party-evaluations-foundations/

[^openai-early]: OpenAI, **Early Access for Safety Testing**, December 20, 2024. https://openai.com/index/early-access-for-safety-testing/

[^anthropic-rsp]: Anthropic, **Responsible Scaling Policy**, Version 3.2, April 29, 2026. https://www.anthropic.com/responsible-scaling-policy

[^anthropic-rsp3]: Anthropic, **Responsible Scaling Policy Version 3.0**, February 24, 2026. https://www.anthropic.com/news/responsible-scaling-policy-v3

[^deepmind-warning]: Google DeepMind, **An Early Warning System for Novel AI Risks**, 2023. https://deepmind.google/blog/an-early-warning-system-for-novel-ai-risks/

[^external-access]: Jacob Charnock, Alejandro Tlaie, Kyle O'Brien, Stephen Casper, and Aidan Homewood, **Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations**, 2026. https://arxiv.org/abs/2601.11916

[^frontier-auditing]: Miles Brundage et al., **Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies**, 2026. https://arxiv.org/abs/2601.11699

[^compliance-review]: Aidan Homewood et al., **Third-Party Compliance Reviews for Frontier AI Safety Frameworks**, 2025. https://arxiv.org/abs/2505.01643

[^dangerous-capabilities]: Mary Phuong et al., **Evaluating Frontier Models for Dangerous Capabilities**, 2024. https://arxiv.org/abs/2403.13793

[^iso-casco-bodies]: International Organization for Standardization, **CASCO Toolbox: Bodies**. https://casco.iso.org/bodies.html

[^iso-recognition]: International Organization for Standardization, **CASCO Toolbox: Recognition of Conformity Assessment Bodies**. https://casco.iso.org/recognition-of-cabs.html

[^iso-building-trust]: International Organization for Standardization, **Building Trust: The Conformity Assessment Toolbox**. https://www.iso.org/iso/casco_building-trust.pdf

[^iaf-accreditation]: International Accreditation Forum, **What Is Accreditation?** https://iaf.nu/en/faq/what-is-accreditation/

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 4. Defines the rationale for independent expert review, multidimensional independence, competence, reviewer roles, review types, reviewer selection, open-source and public-interest participation, mandates, access, methodology, evidence, developer interaction, conflicts, funding, security, dissent, appeals, transparency, reviewer quality assurance, institutional models, conformity-assessment lessons, review levels, lifecycle, implementation pathway, a Standards Body pilot, metrics, failure analysis, objections, evidence gaps, research agenda, standards implications, operational templates, scorecard, and primary-source research basis.

Status: Ready for internal review and future expert critique.