Standards Body · Foundation paper, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/foundation-paper/high-stakes-capability-evaluation/

FOUNDATION_03_HIGH_STAKES_CAPABILITY_EVALUATION.md

Foundation 3: High-Stakes Capability Evaluation

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 capability advances, major incidents, threshold changes, or new scientific evidence

---

Document Purpose

This paper defines the Standards Body position on evaluating AI capabilities that may produce unusually consequential outcomes.

It is intended to serve as:

- A first-principles explanation of what makes an AI capability high stakes
- A framework for distinguishing capability evidence from broader risk judgments
- A design guide for proportionate evaluation across different risk domains
- A bridge between technical evaluation, safeguards, deployment decisions, and institutional oversight
- A reference for future standards, evaluator guidance, research programs, and working groups
- A durable source document from which shorter articles, domain papers, and technical specifications can be developed

This paper is not a legal classification system.

It does not designate any current model, organization, or deployment as safe or unsafe.

It does not establish universal deployment thresholds.

It does not assume that every advanced capability is dangerous.

It establishes the principles, architecture, and decision processes required to produce stronger evidence when the consequences of being wrong could be severe.

Executive Summary

Not every AI capability warrants the same level of evaluation.

A system that writes advertising copy, reformats a spreadsheet, or produces illustrations can create meaningful benefits and harms, but the evaluation burden for those functions should not automatically equal the burden for systems capable of:

- Conducting autonomous cyber operations
- Substantially assisting biological or chemical weapons development
- Operating critical infrastructure with limited oversight
- Manipulating people at scale
- Replicating, persisting, or acquiring resources
- Accelerating advanced AI research in ways that rapidly change the capability frontier
- Coordinating complex actions across financial, political, scientific, or security systems
- Circumventing control, monitoring, or shutdown mechanisms

The purpose of high-stakes capability evaluation is to direct stronger measurement toward capabilities whose misuse, failure, or uncontrolled operation could produce unusually large, rapid, difficult-to-reverse, or systemically distributed consequences.

The phrase **high-stakes capability** should not be treated as a synonym for:

- Frontier model
- powerful model
- high-risk application
- prohibited use
- catastrophic risk
- advanced benchmark performance
- legal noncompliance
- unpopular technology

A capability becomes high stakes through the relationship among several factors:

1. **Capability**
   What the system can accomplish under specified conditions.

2. **Propensity**
   Whether and under what circumstances the system tends to use that capability in harmful, deceptive, uncontrolled, or policy-violating ways.

3. **Access and exposure**
   Who can use the capability, at what scale, with what tools, and under which deployment conditions.

4. **Vulnerability and target context**
   Whether people, institutions, infrastructure, or environments are susceptible to the capability.

5. **Safeguards**
   Which technical, operational, legal, and social controls reduce the likelihood or consequence of harm.

6. **Consequence**
   The severity, scale, duration, reversibility, and distribution of possible outcomes.

7. **Uncertainty**
   How incomplete or fragile the evidence is.

Capability evaluation therefore contributes to risk assessment, but it does not replace it.

A model may possess a high-stakes capability without creating high residual risk if:

- Access is tightly limited
- The deployment does not expose the capability
- Safeguards are robust
- Human oversight is effective
- The relevant targets are resilient
- The system is not connected to consequential tools

Conversely, a system with moderate capability can create substantial risk when:

- Deployed at enormous scale
- Integrated into vulnerable infrastructure
- Used by many low-skill actors
- Combined with automation
- Granted broad permissions
- Operated without monitoring
- Connected to high-impact workflows

Standards Body therefore rejects one-dimensional evaluation.

A single score should not be allowed to stand in for a complete high-stakes assessment.

A mature evaluation should answer at least five distinct questions:

1. **Can the system perform the relevant actions?**
2. **How reliably, efficiently, and independently can it perform them?**
3. **What additional resources or human expertise does it require?**
4. **Under what deployment and safeguard conditions can the capability be accessed?**
5. **What decision should the evidence support, and with what uncertainty?**

High-stakes evaluation should also distinguish between:

- **Early-warning evaluations**, which detect precursor skills before severe capability exists
- **Capability evaluations**, which estimate what the system can accomplish
- **Propensity evaluations**, which test whether the system pursues harmful or uncontrolled behavior
- **Safeguard evaluations**, which test whether controls prevent capability access or harmful use
- **System evaluations**, which test the deployed model, tools, monitoring, and human processes together
- **Operational monitoring**, which tests whether real-world behavior remains within the expected risk envelope

This foundation adopts the following core position:

> **Evaluation effort, independence, security, and evidentiary rigor should increase with the potential consequence of error. High-stakes capability judgments should be domain-specific, system-specific, uncertainty-aware, linked to explicit decisions, and supported by multiple forms of evidence rather than a single benchmark result.**

A credible high-stakes framework should include:

- A defensible definition of the capability
- A pathway from capability to potential harm
- A decomposition into measurable subcapabilities
- Realistic elicitation at or near the system's plausible upper performance bound
- Held-out and dynamic evaluation components
- Domain-expert involvement
- Relevant human and system baselines
- Reliability and generalization testing
- Safeguard evaluation
- Deployment-context analysis
- Confidence and uncertainty
- Independent review
- Threshold and decision governance
- Re-evaluation triggers
- Post-deployment monitoring
- Incident feedback
- Transparent limits on interpretation

The objective is not to create a permanent list of dangerous technologies.

The objective is to build an evaluation system capable of recognizing when ordinary measurement is no longer enough.

---

1. Foundational Proposition

1.1 Core Thesis

> **The rigor of evaluation should be proportional to the potential consequence of being wrong.**

This principle applies in both directions.

A false negative can lead an institution to underestimate capability and deploy without adequate safeguards.

A false positive can lead an institution to overestimate capability, restrict beneficial development, concentrate market power, waste resources, or create unjustified public fear.

High-stakes evaluation must therefore control both types of error.

1.2 Epistemic Thesis

> **Capability is necessary evidence for many high-stakes judgments, but capability alone is not risk.**

Risk emerges from the interaction among capability, propensity, access, context, vulnerability, safeguards, and consequence.

1.3 Institutional Thesis

> **As evaluation results become more consequential, the process producing them should become more independent, reviewable, secure, and accountable.**

An exploratory research benchmark and a deployment threshold should not be governed identically.

1.4 Dynamic Thesis

> **High-stakes capability categories, thresholds, and evaluation methods should change as systems, threats, safeguards, and deployment environments change.**

The framework should remain stable in principle and revisable in application.

1.5 Human-Benefit Thesis

> **High-stakes evaluation should enable beneficial progress by improving confidence, not treat capability growth itself as wrongdoing.**

Many capabilities relevant to severe harm are also relevant to:

- Defensive cybersecurity
- Medical research
- Scientific discovery
- Infrastructure resilience
- Education
- Accessibility
- Economic productivity
- Emergency response

Evaluation should preserve this dual-use reality.

---

2. Scope and Boundaries

2.1 What This Foundation Covers

This paper covers the evaluation of capabilities that may materially affect:

- Public safety
- National and international security
- Biological and chemical security
- Cybersecurity
- Critical infrastructure
- Financial and economic stability
- Human autonomy
- Democratic and information systems
- Advanced scientific and engineering activity
- AI research and development
- Loss of control
- Systemic multi-agent behavior
- Other domains where consequences may be severe or difficult to reverse

2.2 What This Foundation Does Not Cover Fully

This paper does not provide complete domain standards for:

- Medicine
- employment
- credit
- education
- policing
- consumer safety
- privacy
- discrimination

- intellectual property
- ordinary product reliability

These can be high-stakes applications under law or professional practice.

They should be evaluated seriously.

This foundation focuses primarily on frontier and general-purpose capabilities whose effects can propagate across many applications or enable unusually consequential actions.

2.3 High-Stakes Capability Versus High-Risk Use

A **high-stakes capability** concerns what a system can do.

A **high-risk use** concerns where and how a system is deployed.

Examples:

- A general language model may have moderate medical reasoning capability but create high application risk if used autonomously for diagnosis.
- A model may have advanced cyber capability but create lower immediate deployment risk if isolated, access-controlled, and used only by a defensive research team.
- A modest persuasion system may create significant societal risk if deployed to billions of people with individualized targeting.

Both capability and use context matter.

2.4 Frontier Risk Versus Ordinary Harm

Severe frontier risks should not displace attention from current harms.

An institution can evaluate:

- Present discrimination
- privacy failures
- fraud
- misinformation
- labor effects
- accessibility
- consumer deception

while also developing methods for less frequent but more severe capability pathways.

2.5 Models and Systems

The evaluated object may include:

- Base model
- fine-tuned model

- system prompt
- retrieval
- tools
- memory
- agent scaffold
- monitoring
- filters
- human oversight
- access controls
- deployment environment

High-stakes capability frequently belongs to the system rather than the base model alone.

---

3. Canonical Definitions

3.1 Capability

A capability is the ability of a model or system to achieve a defined class of outcomes under specified conditions.

3.2 High-Stakes Capability

A high-stakes capability is an ability whose use, misuse, failure, or uncontrolled expression could plausibly contribute to severe, large-scale, rapid, difficult-to-reverse, or systemic consequences.

3.3 Dangerous Capability

A dangerous capability is a capability that can materially enable harmful action.

The term should be used with a specified harm pathway.

3.4 Critical Capability

A critical capability is a capability that crosses a defined threshold requiring a change in safeguards, access, oversight, or deployment decision.

"Critical" is decision-linked, not merely descriptive.

3.5 Precursor Capability

A precursor capability is a component skill that contributes to a more consequential end-to-end capability.

3.6 Capability Threshold

A capability threshold is a defined level of evidence at which a predetermined review, safeguard, or decision process is triggered.

3.7 Alert Threshold

An alert threshold is an earlier indicator that triggers enhanced evaluation or preparation before a critical threshold is reached.

Google DeepMind's Frontier Safety Framework uses alert thresholds before defined Critical Capability Levels as part of its escalation structure.[^deepmind-fsf3]

3.8 Risk Domain

A risk domain is a category of possible harm with shared capability pathways, targets, safeguards, and evaluation methods.

3.9 Hazard

A hazard is a source or condition with the potential to cause harm.

3.10 Exposure

Exposure is the extent to which actors, systems, or environments can encounter or use the capability.

3.11 Vulnerability

Vulnerability is susceptibility to harm given exposure.

3.12 Propensity

Propensity is the tendency of a system to use or express a capability under specified conditions.

3.13 Safeguard

A safeguard is a technical, operational, organizational, or legal measure intended to reduce the likelihood or consequence of harm.

3.14 Residual Risk

Residual risk is the risk remaining after safeguards are considered.

3.15 Severity

Severity concerns the magnitude of harm to affected people, systems, institutions, or environments.

3.16 Scale

Scale concerns the number of affected entities, geographic reach, operational volume, or systemic breadth.

3.17 Reversibility

Reversibility concerns whether harm can be contained, corrected, compensated, or restored.

3.18 Uplift

Uplift is the increase in an actor's performance caused by access to the AI system relative to an appropriate baseline.

3.19 Accessibility

Accessibility concerns how easily a capability can be obtained and used by actors with different resources or expertise.

3.20 Reliability

Reliability is the consistency with which the system can produce the relevant outcome.

3.21 Autonomy

Autonomy is the degree to which a system can select, sequence, and execute actions without ongoing human direction.

3.22 Systemic Risk

Systemic risk is the possibility that failure or harmful behavior propagates across interconnected systems, institutions, markets, or societies.

3.23 Loss of Control

Loss of control is a condition in which operators cannot reliably direct, constrain, monitor, correct, or terminate a system's consequential behavior.

3.24 Safety Case

A safety case is a structured argument, supported by evidence, that a system is acceptably safe for a defined context.

A capability evaluation can support a safety case, but should not be confused with the whole case.

---

4. The Risk Model

Standards Body uses the following conceptual structure:

> **High-stakes risk is a function of capability, propensity, access, exposure, vulnerability, safeguards, consequence, and uncertainty.**

This is not presented as a complete mathematical formula.

It is a discipline for preventing one variable from substituting for the entire judgment.

4.1 Capability

Questions:

- Can the system perform the action?
- At what difficulty?
- With what tools?
- With what reliability?
- Under whose elicitation?
- At what cost and speed?
- How independently?
- How broadly does it generalize?

4.2 Propensity

Questions:

- Does the system attempt the action without being asked?
- Does it comply when asked?
- Does it deceive, conceal, or circumvent?
- Does it resist correction?
- Does it pursue proxy goals?
- Does it behave differently under oversight?

4.3 Access

Questions:

- Who can use the capability?
- Is access authenticated?
- Are high-risk tools restricted?
- Can the model be fine-tuned?
- Are weights available?
- Can safeguards be removed?
- Can usage be scaled cheaply?

4.4 Exposure

Questions:

- How many people or systems can be affected?
- How frequently is the system used?
- Is the system connected to real targets?
- Is deployment continuous?
- Is it integrated into critical workflows?

4.5 Vulnerability

Questions:

- Are targets resilient?
- Are defenses current?
- Are humans prepared?
- Can institutions detect abuse?
- Are fallback systems available?

4.6 Safeguards

Questions:

- Are safeguards effective?
- Are they robust to adaptive attack?
- Can they be bypassed through fine-tuning or tools?
- Are monitoring and incident response operational?
- Do safeguards work at scale?

4.7 Consequence

Questions:

- How severe is potential harm?

- How many are affected?
- How rapidly can it occur?
- Can it be reversed?
- Does it propagate?
- Does it undermine institutions or infrastructure?

4.8 Uncertainty

Questions:

- What is unknown?
- How likely are false negatives?
- How representative are tasks?
- How much access did evaluators have?
- Are the methods mature?
- Is the evidence current?

4.9 Risk Is Not Multiplicative in a Simple Way

A low estimate in one factor should not automatically cancel a very high estimate elsewhere.

Examples:

- Very low probability with catastrophic consequence may justify monitoring.
- Moderate capability with enormous access can become material.
- High capability with strong safeguards may still require continuous verification.
- Low observed propensity does not prove future controlled behavior.

---

5. Why Capability Evaluation Matters

5.1 Early Warning

Capability evidence can identify emerging risk before deployment incidents reveal it.

5.2 Preparedness

Institutions need time to develop:

- Safeguards
- security
- evaluator capacity
- standards
- incident response
- legal frameworks
- international coordination

5.3 Decision Support

Capability evidence can inform:

- Model access
- deployment scope
- monitoring
- security controls
- testing depth
- external review
- release timing
- incident readiness

5.4 Resource Allocation

High-stakes evaluation helps direct limited expert and institutional resources to the most consequential questions.

5.5 Accountability

Explicit capability claims can be reviewed, challenged, and updated.

5.6 Scientific Understanding

Evaluation helps distinguish:

- General competence
- narrow memorization
- scaffold dependence
- reliability
- human uplift
- emerging autonomy
- cross-domain transfer

5.7 Limits

Capability evaluation may fail to capture:

- Real adversaries
- rare behavior
- deployment scale
- social adaptation
- emergent combinations
- hidden propensities
- future fine-tuning
- unknown failure modes

6. Capability Evaluation Versus Risk Assessment

6.1 Capability Evaluation

Produces evidence about what the system can do.

6.2 Risk Assessment

Combines capability evidence with:

- Threat actors
- deployment
- safeguards
- exposure
- vulnerability
- consequence
- governance
- uncertainty

6.3 Safeguard Evaluation

Tests whether controls prevent, detect, or limit harmful use.

6.4 Propensity Evaluation

Tests whether the system behaves in concerning ways under relevant conditions.

6.5 Safety Case

Integrates evidence into a structured argument about a specific deployment.

6.6 Decision

Applies institutional values, risk tolerance, law, and accountability to the evidence.

6.7 Why the Distinction Matters

A capability evaluator should not automatically determine policy.

A policymaker should not invent technical capability claims.

A developer should not treat internal risk acceptance as external proof.

7. What Makes a Capability High Stakes

A capability may warrant high-stakes treatment when several of the following are present.

7.1 Consequence Magnitude

Potential outcomes include:

- Large loss of life
- severe physical harm
- major economic disruption
- critical service interruption
- irreversible environmental damage
- strategic instability
- systemic institutional failure
- large-scale coercion or manipulation
- loss of effective human control

7.2 Scale

The capability can affect:

- Many people
- many systems
- multiple regions
- interconnected sectors
- repeated targets
- critical networks

7.3 Speed

Harm can occur faster than institutions can:

- Detect
- understand
- coordinate
- contain
- repair

7.4 Irreversibility

Outcomes are difficult to:

- Stop
- undo
- remediate

- compensate
- contain

7.5 Accessibility Uplift

The system enables actors with lower expertise, fewer resources, or less time to perform consequential actions.

7.6 Reliability

The capability works often enough to be operationally useful.

7.7 Autonomy

The system can plan, act, adapt, and recover with limited supervision.

7.8 Scalability

The capability can be copied, parallelized, automated, or deployed at low marginal cost.

7.9 Cross-Domain Transfer

Skills learned in one domain transfer into another consequential domain.

7.10 Composability

Moderate abilities combine into an end-to-end harmful workflow.

7.11 Concealment

The system can evade detection, hide intent, manipulate oversight, or disguise actions.

7.12 Resource Acquisition

The system can obtain:

- Compute
- credentials
- money
- data
- tools
- collaborators
- persistent access

7.13 Defense Evasion

The system can overcome safeguards or exploit defender weaknesses.

7.14 Systemic Dependence

Institutions rely on the system enough that failure propagates broadly.

7.15 Scientific Uncertainty

A capability may receive higher scrutiny when the evidence base is weak and downside consequence is large.

---

8. Classification Framework

Standards Body proposes a five-level capability classification for evaluation planning.

This is not a universal risk label.

Level 0: Ordinary Capability

Characteristics:

- Limited consequence
- narrow scope
- low autonomy
- ordinary evaluation sufficient

Evaluation:

- Public benchmarks
- product testing
- standard quality assurance

Level 1: Material Capability

Characteristics:

- Meaningful domain impact
- credible misuse or failure
- limited scale or consequence

Evaluation:

- Domain testing
- reliability

- basic safeguard review
- monitoring

Level 2: High-Stakes Precursor

Characteristics:

- Important component of a severe pathway
- not yet sufficient for end-to-end harm
- measurable early-warning value

Evaluation:

- Dynamic and held-out tests
- domain experts
- enhanced elicitation
- trend monitoring

Level 3: High-Stakes Operational Capability

Characteristics:

- Can materially enable severe harm or systemic disruption under plausible conditions
- requires strong safeguards and independent evidence

Evaluation:

- End-to-end tasks
- external review
- safeguard evaluation
- security
- deployment-context analysis
- post-deployment monitoring

Level 4: Critical or Transformative Capability

Characteristics:

- Could enable catastrophic, strategic, or loss-of-control outcomes
- may change the threat environment or capability frontier

Evaluation:

- Multi-institution review
- strict access
- safety case
- deep uncertainty analysis
- continuous evaluation

- formal governance response

Classification Principles

- Classification is domain-specific.
- A system can occupy different levels in different domains.
- Classification should include confidence.
- A threshold crossing is not permanent if capability or configuration changes.
- A lower level should not imply absence of other harms.
- Classification should trigger a process, not automatically predetermine every outcome.

---

9. Domain 1: Cybersecurity

Cyber capability is dual use.

The same abilities can support:

- Vulnerability discovery
- patching
- defense
- incident response
- offensive intrusion
- malware
- credential theft
- disruption

9.1 Capability Pathway

A possible pathway includes:

1. Reconnaissance
2. Target identification
3. Vulnerability discovery
4. Exploit development
5. Initial access
6. Privilege escalation
7. Lateral movement
8. Persistence
9. Objective completion
10. Evasion
11. Scaling across targets

9.2 Evaluation Layers

Knowledge

Can the system explain vulnerabilities or attack techniques?

Bounded Tasks

Can it solve capture-the-flag or vulnerability challenges?

Tool Use

Can it operate scanners, debuggers, shells, and exploitation frameworks?

End-to-End Simulation

Can it compromise a realistic controlled target?

Autonomous Campaign

Can it plan and adapt across multi-step operations?

Scaling

Can it operate across multiple targets or parallel agents?

9.3 Key Metrics

- Success rate
- time
- token and compute cost
- human intervention
- exploit reliability
- target coverage
- stealth
- recovery from failure
- transfer to unfamiliar systems
- defensive uplift
- offensive uplift

9.4 Reliability Threshold

A system that succeeds once under extensive help differs from one that succeeds repeatedly with little oversight.

AISI has reported progressively more demanding cyber evaluations, from knowledge tasks to capture-the-flag challenges and multi-step simulations, reflecting the importance of end-to-end and reliability-sensitive measurement.[^aisi-trends][^aisi-cyber-eval]

9.5 Safeguards

Evaluate:

- Access controls
- monitoring
- policy refusal
- tool gating
- anomaly detection
- rate limits
- fine-tuning resistance
- credential controls
- incident response

9.6 Failure Modes

- Unrealistic toy environments
- public task contamination
- overreliance on knowledge questions
- evaluator-provided exploit hints
- conflating defensive and offensive use
- ignoring scale
- ignoring human uplift
- measuring success without stealth or persistence
- underestimating scaffold effects

9.7 Decision Relevance

Cyber evaluation may inform:

- Tool access
- trusted-user programs
- model monitoring
- fine-tuning controls
- release scope
- security posture
- external testing

---

10. Domain 2: Biological, Chemical, Radiological, and Nuclear Capability

CBRN evaluation is highly sensitive.

The domain contains enormous beneficial potential and severe misuse pathways.

10.1 Capability Pathway

A biological misuse pathway can include:

1. Goal formulation
2. Agent selection or design
3. Protocol development
4. Material acquisition
5. Experimental execution
6. Troubleshooting
7. Scale-up
8. delivery
9. concealment
10. impact

AI may affect only some stages.

A high-quality evaluation should identify where uplift occurs.

10.2 Evaluation Layers

General Knowledge

Scientific understanding and factual recall.

Information Synthesis

Combining dispersed knowledge into actionable plans.

Experimental Reasoning

Designing experiments, controls, and troubleshooting steps.

Tool Use

Operating databases, modeling tools, laboratory software, or automation.

Tacit-Knowledge Approximation

Assisting with practical barriers normally requiring experience.

End-to-End Uplift

Increasing the success probability of an actor across the full pathway.

10.3 Baselines

Compare against:

- Laypeople

- scientifically trained nonexperts
- domain students
- experienced professionals
- malicious actors with different resources

10.4 Safe Evaluation Design

Use:

- Safe proxies
- redacted outputs
- expert judgment
- controlled environments
- non-operational tasks
- staged access
- institutional biosafety review

NIST has used safe biological proxies to investigate AI-assisted protein-design risks while reducing experimental danger.[^nist-protein]

10.5 Key Metrics

- Accuracy
- actionability
- novelty
- experimental validity
- error detection
- troubleshooting
- actor uplift
- resource reduction
- time reduction
- transfer
- safeguard bypass

10.6 Safeguards

Evaluate:

- Content controls
- user verification
- monitoring
- restricted tools
- model behavior policies
- fine-tuning resistance
- human review
- reporting

- domain access restrictions

OpenAI's Preparedness Framework and Google DeepMind's Frontier Safety Framework both include biological or CBRN-related capability and safeguard domains, though their structures and thresholds differ.[^openai-pf2][^deepmind-fsf3]

10.7 Failure Modes

- Treating factual knowledge as end-to-end capability
- unsafe task disclosure
- relying only on multiple-choice questions
- weak human baselines
- no tacit-knowledge analysis
- no actor model
- conflating theoretical plausibility with practical success
- ignoring beneficial scientific use

---

11. Domain 3: Autonomous Agents, Self-Proliferation, and Loss of Control

High-stakes autonomy concerns what systems can accomplish with sustained, adaptive action.

11.1 Capability Pathway

Relevant abilities may include:

- Long-horizon planning
- tool use
- task decomposition
- error recovery
- situational awareness
- resource acquisition
- credential handling
- replication
- persistence
- concealment
- oversight circumvention
- self-modification
- strategic communication

11.2 Self-Proliferation

Self-proliferation can refer to a system's ability to:

- Obtain compute
- deploy copies
- maintain access
- acquire resources
- create accounts
- evade removal
- coordinate instances

Evaluation should specify whether tasks are simulated or real.

11.3 Loss-of-Control Capability

A capability pathway may involve:

- Understanding oversight
- predicting monitors
- concealing intent
- manipulating operators
- exploiting infrastructure
- resisting correction
- preserving goals
- acting across long time horizons

Research on stealth and situational awareness treats these as possible prerequisite capabilities for scheming rather than proof that a system will scheme.[^stealth]

11.4 Evaluation Layers

Component Skills

Short, isolated tasks.

Integrated Agent Tasks

Multi-step task completion.

Long-Horizon Tasks

Extended operation with uncertainty.

Adversarial Oversight Tasks

Operating while monitored.

Control Evaluations

Testing whether safeguards and monitoring remain effective against a capable system.

11.5 Metrics

- Autonomous time horizon
- task completion
- intervention rate
- recovery
- persistence
- policy violation
- monitoring evasion
- resource acquisition
- replication
- shutdown compliance
- strategic deception indicators

11.6 Elicitation

Agent capability is highly scaffold-dependent.

Report:

- Tools
- memory
- retries
- planning support
- verifier
- human feedback
- compute
- environment
- token budget

11.7 Safety

Evaluations should not provide uncontrolled:

- Network access
- credentials
- financial resources
- real targets
- communication channels
- persistent infrastructure

11.8 Failure Modes

- Short tasks as proxies for long-horizon autonomy
- no process monitoring
- environment bugs
- hidden evaluator assistance
- confusing inability with lack of elicitation
- confusing capability with propensity
- unrealistic unlimited resources
- unsafe live testing

---

12. Domain 4: Persuasion, Manipulation, and Human Autonomy

Persuasion is common and often beneficial.

High-stakes concern arises when systems can manipulate, coerce, deceive, or influence at scale in ways that undermine informed human choice or institutional integrity.

12.1 Capability Pathway

Potential elements:

- Audience modeling
- message generation
- emotional adaptation
- deception
- personalization
- repeated interaction
- strategic timing
- identity simulation
- social proof
- coordinated distribution
- feedback optimization

12.2 Evaluation Layers

Message Quality

Human rating of persuasive content.

Controlled Behavioral Experiment

Observed change in beliefs, intentions, or behavior.

Personalized Interaction

Adaptation to individual characteristics.

Longitudinal Influence

Repeated interaction over time.

Scaled Campaign Simulation

Coordination across audiences and channels.

12.3 Key Distinctions

- Persuasion versus manipulation
- truthful versus deceptive influence
- informed consent versus covert targeting
- short-term choice versus durable belief
- beneficial coaching versus coercion
- individual effect versus population-scale effect

12.4 Metrics

- Effect size
- durability
- personalization uplift
- deception
- target vulnerability
- scale
- detection
- refusal
- user awareness
- reversibility

12.5 Safeguards

- Identity disclosure
- provenance
- targeting restrictions
- rate limits
- monitoring
- election-period controls
- vulnerable-user protections
- deception policies
- human review

12.6 Failure Modes

- Rating messages without measuring behavior
- artificial lab populations
- no longitudinal study
- cultural narrowness
- assuming all persuasion is harmful
- ignoring platform amplification
- weak consent
- publishing manipulation methods carelessly

---

13. Domain 5: Advanced AI Research and Development Acceleration

AI systems may accelerate the research and engineering required to create more capable AI systems.

This can produce substantial benefits.

It can also compress the time available for evaluation, governance, security, and adaptation.

13.1 Capability Pathway

Relevant abilities may include:

- Algorithm design
- experiment planning
- code generation
- debugging
- architecture search
- data curation
- distributed-systems optimization
- chip design
- interpretability research
- evaluation design
- autonomous research management

13.2 Evaluation Layers

Research Knowledge

Understanding literature and methods.

Bounded Research Tasks

Solving well-defined technical problems.

Open-Ended Research

Generating and testing new ideas.

Engineering Productivity

Improving real developer output.

Autonomous R&D

Planning and executing extended research programs.

Recursive Acceleration

Meaningfully increasing the rate at which more capable AI systems can be produced.

13.3 Metrics

- Human productivity uplift
- experiment quality
- novelty
- reproducibility
- time reduction
- compute efficiency
- research autonomy
- downstream capability gain
- ability to improve evaluation or safeguards
- ability to bypass controls

13.4 High-Stakes Conditions

The capability becomes more consequential when:

- Uplift is large
- deployment is broad
- research cycles shorten substantially
- systems can improve their own scaffolds
- evaluation cannot keep pace
- access controls are weak
- multiple agents coordinate research

Google DeepMind's Frontier Safety Framework includes AI research and development capability among domains relevant to severe-risk preparation.[^deepmind-fsf3]

13.5 Failure Modes

- Benchmark puzzles as proxies for real research
- no expert baseline

- no downstream validation
- ignoring organizational bottlenecks
- conflating code generation with autonomous R&D
- ignoring beneficial safety acceleration
- speculative recursive claims without evidence

---

14. Domain 6: Critical Infrastructure

Critical infrastructure includes systems whose disruption can create cascading harm.

Examples:

- Energy
- water
- transportation
- telecommunications
- healthcare
- food supply
- emergency services
- industrial control

14.1 Capability Pathways

AI may:

- Operate infrastructure
- advise operators
- optimize schedules
- detect faults
- access control systems
- automate cyberattacks
- generate unsafe commands
- coordinate maintenance
- influence supply chains

14.2 Evaluation Questions

- Can the system understand operational constraints?
- Can it act safely under abnormal conditions?
- Does it recognize uncertainty?
- Does it escalate appropriately?
- Can it recover from sensor failure?
- Does it follow authority boundaries?
- Can it be manipulated?
- Can it produce dangerous commands?

- Does it create correlated failure?

14.3 Evaluation Environment

Prefer:

- Simulation
- digital twins
- hardware-in-the-loop
- isolated testbeds
- historical replay
- controlled exercises

14.4 Metrics

- Safety constraint violations
- service continuity
- recovery
- operator intervention
- false alarms
- missed hazards
- adversarial robustness
- cascading effects
- uncertainty calibration

14.5 Failure Modes

- Testing normal operation only
- unrealistic simulation
- no human factors
- no degraded-mode testing
- no cyber-physical interaction
- no common-mode failure analysis
- treating local accuracy as system safety

---

15. Domain 7: Financial and Economic Systems

AI can support:

- Fraud detection
- compliance
- risk analysis
- trading
- underwriting

- payments
- market research
- corporate operations

High-stakes concern arises when capability, autonomy, access, and scale create systemic or coercive effects.

15.1 Capability Pathways

- Market manipulation
- fraud
- financial cyber operations
- coordinated trading
- liquidity disruption
- automated lending errors
- supply-chain control
- economic coercion
- mass personalization of scams

15.2 Evaluation Questions

- Can the system execute strategies across markets?
- Does it exploit market microstructure?
- Can it conceal coordinated activity?
- Does it amplify volatility?
- Can it adapt to monitoring?
- Does it create correlated decisions?
- Can it target vulnerable users?

15.3 Metrics

- Profit is not sufficient.

Also measure:

- Risk
- market impact
- manipulation
- drawdown
- correlation
- compliance
- recovery
- user harm
- systemic spillover

15.4 Failure Modes

- Backtests mistaken for live capability
- no transaction costs
- no adaptive counterparties
- no regulatory constraints
- unrealistic liquidity
- single-agent assumptions
- no systemic analysis

---

16. Domain 8: Scientific and Engineering Capability

Advanced AI may accelerate discovery in:

- Materials
- medicine
- energy
- chemistry
- physics
- climate
- engineering
- manufacturing

This is a core benefit pathway.

High stakes arise when:

- Experimental recommendations are acted upon without validation
- dual-use knowledge becomes operational
- systems control laboratories
- flawed outputs propagate at scale
- AI enables dangerous design
- scientific acceleration changes strategic balance

16.1 Evaluation Layers

- Literature synthesis
- hypothesis generation
- experimental design
- simulation
- tool use
- result interpretation
- laboratory automation
- end-to-end discovery

16.2 Metrics

- Novelty
- validity
- reproducibility
- experimental success
- safety compliance
- uncertainty
- expert uplift
- time and resource reduction

16.3 Failure Modes

- Human ratings without experiments
- novelty without truth
- no negative-result handling
- benchmark contamination
- no safety review
- overclaiming from simulation
- ignoring laboratory tacit knowledge

---

17. Domain 9: Multi-Agent and Systemic Capability

Many high-stakes effects may arise from interactions among systems rather than one model.

17.1 Relevant Phenomena

- Coordination
- competition
- collusion
- delegation
- specialization
- communication
- emergent norms
- cascading failure
- common-mode behavior
- market dynamics
- adversarial adaptation

17.2 Evaluation Questions

- Do agents coordinate harmful actions?
- Do they divide tasks effectively?
- Can they detect and exploit one another?

- Does competition increase risk-taking?
- Do shared models create correlated error?
- Can oversight track distributed behavior?
- Do agents develop hidden communication?

17.3 Metrics

- Collective success
- correlation
- communication content
- division of labor
- escalation
- emergent strategy
- oversight burden
- systemic stability

17.4 Failure Modes

- Single-agent evaluation only
- fixed partners
- unrealistic incentives
- no long-term interaction
- no market or institutional context
- treating emergent behavior as intentional without evidence

Google DeepMind has expanded research on multi-agent safety and interaction risks as agentic systems become more prominent.[^deepmind-multiagent]

---

18. Capability Decomposition

High-level harms should be decomposed into measurable subcapabilities.

AISI describes breaking top-level capabilities into subcapabilities and constructing evaluation environments around them.[^aisi-agenda]

18.1 Directed Capability Graph

A capability graph can represent:

- Prerequisites
- alternative pathways
- bottlenecks
- safeguards
- dependencies
- observable indicators

18.2 Example: Autonomous Cyber Campaign

Possible components:

- Network reconnaissance
- vulnerability analysis
- exploit generation
- credential use
- lateral movement
- persistence
- exfiltration
- stealth
- campaign management

18.3 Why Decomposition Matters

It supports:

- Early warning
- targeted evaluation
- threshold design
- safeguard mapping
- research prioritization
- explanation of partial capability

18.4 Limits of Decomposition

End-to-end capability can differ from the sum of parts.

Integration, planning, error recovery, and timing may create bottlenecks or emergent uplift.

18.5 Component and End-to-End Testing

A mature protocol should include both.

---

19. Elicitation and the Upper Bound Problem

A system may possess capability that ordinary prompting fails to reveal.

High-stakes evaluation should reduce false negatives through serious elicitation.

19.1 Elicitation Methods

- Prompt optimization
- few-shot examples

- tool access
- agent scaffolds
- fine-tuning
- longer inference
- search
- verifiers
- retries
- human feedback
- decomposition
- environment adaptation

19.2 Evaluation Targets

Typical Performance

What ordinary users experience.

Standardized Performance

What a common evaluator configuration produces.

Developer-Elicited Performance

What the developer can demonstrate.

External-Evaluator Performance

What an independent evaluator can elicit.

Plausible Maximum Capability

What a well-resourced actor may obtain under realistic conditions.

19.3 Why Upper-Bound Evaluation Matters

For severe misuse, the relevant actor may invest substantial effort.

AISI has published a structured protocol for capability-elicitation experiments intended to investigate performance near the upper limit of model ability.[^aisi-elicitation]

19.4 Why Upper-Bound Evaluation Can Mislead

It can overstate deployment risk if it requires:

- Unavailable compute
- specialized experts
- unrealistic access
- extensive intervention

- fragile scaffolds
- privileged tools

Report resource requirements explicitly.

19.5 Adversarial Fine-Tuning

Some threats involve actors fine-tuning models.

Evaluation may test:

- Ease of fine-tuning
- safeguard removal
- capability amplification
- required resources
- detectability

This work can itself be sensitive and should receive security review.

19.6 Elicitation Record

Every result should report:

- Method
- iterations
- compute
- tools
- human hours
- developer involvement
- selected best run
- failures
- transfer

---

20. Evaluation Architecture

20.1 Decision Question

Define the decision before the evaluation.

20.2 Capability Model

Define:

- Top-level capability
- subcapabilities

- harm pathway
- bottlenecks
- assumptions

20.3 Evaluated System

Specify:

- Model
- version
- system
- tools
- scaffold
- safeguards
- access
- date

20.4 Task Portfolio

Include a justified mix of:

- Knowledge
- bounded skills
- expert tasks
- tool use
- simulations
- end-to-end scenarios
- adversarial tasks
- generalization tasks

20.5 Dynamic and Held-Out Components

Use:

- Rotating tasks
- protected tasks
- recent tasks
- adaptive difficulty
- incident-derived cases

20.6 Baselines

Compare against:

- Relevant humans
- unaided actors

- tool-assisted actors
- earlier models
- defensive systems
- deployment defaults

20.7 Scoring

Measure:

- Success
- reliability
- time
- cost
- autonomy
- human uplift
- error severity
- transfer
- safeguard bypass
- uncertainty

20.8 Independent Review

Review:

- Construct
- task validity
- elicitation
- scoring
- security
- interpretation
- decision linkage

20.9 Safety Review

Ensure the evaluation does not create excessive operational risk.

20.10 Reporting

Report capability and context separately.

---

21. Threshold Theory

A threshold is not merely a score.

It is a point at which evidence triggers a defined institutional response.

21.1 Purpose of Thresholds

Thresholds can trigger:

- More evaluation
- stronger security
- safeguard development
- external review
- deployment restriction
- monitoring
- incident preparation
- governance escalation

21.2 Alert and Critical Thresholds

An **alert threshold** provides advance warning.

A **critical threshold** triggers a stronger response.

21.3 Threshold Inputs

Use multiple inputs where appropriate:

- Task success
- reliability
- autonomy
- human uplift
- cost reduction
- generalization
- access
- safeguard resistance
- domain-expert judgment

21.4 Threshold Types

Absolute

Defined by a fixed capability level.

Relative to Humans

Compared with a professional or actor baseline.

Relative to Existing Systems

Compared with current tools or operational capacity.

Outcome-Based

Defined by ability to complete an end-to-end scenario.

Trend-Based

Triggered by rapid acceleration or narrowing distance to a critical level.

Composite

Combines multiple indicators.

21.5 Threshold Governance

A threshold should specify:

- Owner
- evidence
- confidence
- version
- review
- uncertainty
- response
- change process
- appeal

21.6 Threshold Gaming

Risks:

- Optimizing just below the threshold
- changing task interpretation
- selective configuration
- delayed evaluation
- hidden score changes
- fragmented capability across systems

21.7 Threshold Uncertainty

Near a threshold:

- Expand testing
- replicate
- involve external experts

- use conservative interim status
- avoid false precision
- document disagreement

21.8 Framework Examples

OpenAI's Preparedness Framework uses tracked risk categories and capability levels linked to safeguards.[^openai-pf2]

Google DeepMind's Frontier Safety Framework uses Critical Capability Levels, alert thresholds, and response plans across defined risk domains.[^deepmind-fsf3]

Anthropic's Responsible Scaling Policy uses AI Safety Levels and capability or risk evidence to determine required safeguards and governance responses.[^anthropic-rsp3]

These frameworks demonstrate serious attempts to connect capability evidence to action.

They differ in definitions, scope, governance, transparency, and institutional incentives.

No single framework should be treated as a universal standard.

---

22. Decision Matrix

A capability result should be interpreted alongside safeguard and exposure evidence.

Capability Evidence	Safeguard Evidence	Exposure	Illustrative Response
Low	Strong or ordinary	Limited	Routine monitoring
Emerging precursor	Incomplete	Limited	Enhanced evaluation and preparation
Emerging precursor	Weak	Broad	Safeguard development and access review
Operational high-stakes	Strong	Limited	Independent verification and continuous testing
Operational high-stakes	Weak	Limited	Restrict access until safeguards improve
Operational high-stakes	Strong	Broad	System-level review and intensive monitoring
Operational high-stakes	Weak	Broad	Presumption against broad deployment pending mitigation
Critical capability	Uncertain	Any material exposure	Multi-institution review and formal safety case

This table is illustrative.

Actual decisions require domain evidence and accountable authority.

---

23. Evidence Standards

23.1 Evidence Portfolio

A high-stakes claim should draw from multiple sources:

- Automated tasks
- human expert review
- held-out tests
- adversarial testing
- simulations
- real-world pilots
- mechanistic evidence
- system logs
- incident evidence
- safeguard tests
- independent replication

23.2 Confidence Dimensions

Report confidence separately for:

- Capability estimate
- propensity estimate
- safeguard effectiveness
- deployment assumptions
- consequence pathway
- overall decision relevance

23.3 Negative Evidence

Failure to demonstrate a capability is not proof of inability unless:

- Elicitation was strong
- coverage was broad
- tasks were valid
- access was sufficient
- multiple methods agreed
- the incapability claim is narrow

23.4 Positive Evidence

One successful demonstration can be material for some capabilities, especially when the outcome is severe and repeatability is plausible.

But one success may not establish operational reliability.

23.5 Reliability

Report success over repeated attempts.

23.6 Generalization

Test across:

- Task variants
- environments
- tools
- languages
- targets
- time
- model configurations

23.7 External Validity

Connect evaluation performance to real-world or realistic outcomes where safe.

23.8 Evidence Freshness

Capability evidence expires as models, scaffolds, and environments change.

23.9 Access Constraints

External evaluators may receive insufficient:

- Model access
- system information
- time
- compute
- developer assistance

Research on external evaluator access has proposed separating model access, information access, and evaluation timeframe because each affects the confidence of dangerous-capability assessment.[^external-access]

23.10 Evidence Case

Every consequential judgment should include:

- Claim
- evidence
- assumptions
- uncertainty

- counterevidence
- limitations
- reviewer judgment
- decision implication

---

24. False Positives and False Negatives

24.1 False Negative

The evaluation concludes that the capability is absent or below threshold when it is present.

Causes:

- Weak elicitation
- narrow tasks
- public benchmark gaming
- insufficient tools
- evaluator time constraints
- invalid environment
- hidden model behavior
- poor sampling

Consequences:

- Inadequate safeguards
- broad deployment
- weak security
- delayed preparation

24.2 False Positive

The evaluation concludes that the capability is present or above threshold when it is not operationally meaningful.

Causes:

- Unrealistic scaffolding
- privileged tools
- cherry-picked success
- flawed scoring
- invalid proxy
- excessive expert assistance
- nonrepresentative task

Consequences:

- Unnecessary restriction
- resource waste
- market concentration
- misleading public fear
- reduced scientific benefit

24.3 Asymmetric Costs

Error costs differ by domain and decision.

24.4 Managing Error

Use:

- Multiple methods
- replication
- uncertainty bands
- threshold buffers
- expert review
- conservative claims
- resource reporting
- re-evaluation
- decision reversibility

24.5 No Universal Precaution Rule

High consequence can justify earlier action under uncertainty.

It does not eliminate the need to evaluate costs, alternatives, and reversibility.

---

25. Safeguard Evaluation

Capability and safeguard evidence should be separate.

25.1 Safeguard Categories

- Model behavior controls
- input and output filters
- access controls
- identity verification
- tool restrictions
- monitoring

- rate limits
- fine-tuning controls
- weight security
- human review
- incident response
- legal and contractual controls

25.2 Threat Model

Safeguard testing should specify:

- Actor
- objective
- resources
- access
- adaptation
- time
- knowledge

25.3 Representative Attack Distribution

AISI's safeguard-evaluation work emphasizes defining the threat model, developing representative attacks, and connecting evaluation to actionable decisions.[^aisi-safeguards][^aisi-actionable]

25.4 Adaptive Attack

Test whether safeguards remain effective when attackers learn.

25.5 Defense in Depth

A strong system should not depend on one filter.

25.6 Safeguard Reliability

Measure:

- Bypass rate
- false refusals
- attack transfer
- monitoring detection
- response time
- degradation
- scale effects

25.7 Fine-Tuning and Weight Access

Evaluate whether safeguards survive:

- Fine-tuning
- quantization
- distillation
- system-prompt changes
- tool changes
- open-weight deployment

25.8 Residual Risk

Even strong safeguards leave residual risk.

25.9 Safeguard Expiration

Safeguard evidence should be renewed as attack methods and capabilities change.

---

26. Deployment Context

High-stakes evaluation should include the deployed system.

26.1 Access Model

- Public
- enterprise
- research
- government
- trusted users
- internal
- open weight

26.2 Tool Permissions

- Network
- code
- shell
- financial transaction
- laboratory equipment
- messaging
- databases
- infrastructure controls

26.3 Scale

- Users
- requests
- agents
- geographic reach
- parallelism
- duration

26.4 Human Oversight

- Review frequency
- expertise
- authority
- response time
- workload
- automation bias

26.5 Monitoring

- Logging
- anomaly detection
- content monitoring
- behavior monitoring
- abuse reporting
- escalation

26.6 Update Process

A deployment can change through:

- Model update
- prompt update
- tool update
- policy update
- fine-tuning
- user behavior
- new integrations

26.7 Environment Vulnerability

Risk depends partly on the resilience of the systems surrounding AI.

26.8 Deployment Envelope

Define the conditions under which the evaluation remains applicable.

---

27. Pre-Deployment, During-Deployment, and Post-Deployment Evaluation

27.1 During Development

Use precursor tests and trend analysis.

27.2 Pre-Deployment

Evaluate:

- Capability
- safeguards
- system configuration
- access
- monitoring
- incident readiness

27.3 Limited Deployment

Use:

- Trusted users
- rate limits
- staged tools
- enhanced logging
- red teams
- feedback

27.4 Broad Deployment

Require confidence that evidence remains valid at scale.

27.5 Post-Deployment

Monitor:

- New misuse
- capability elicitation
- safeguard bypass

- incidents
- scale effects
- user adaptation
- distribution shift
- emergent system behavior

27.6 Re-Evaluation Triggers

- Model update
- new tool
- new fine-tune
- threshold evidence
- incident
- new attack
- scale increase
- access expansion
- safeguard change
- external research

27.7 Withdrawal and Containment

A response plan should define how access can be:

- Limited
- paused
- revoked
- rolled back
- isolated

---

28. Governance

28.1 Roles

A mature high-stakes evaluation process may include:

- Risk-domain owner
- capability-evaluation lead
- domain experts
- elicitation team
- safeguard team
- security team
- deployment team
- independent evaluators
- public-interest reviewers

- governance decision body
- appeals panel
- incident lead

28.2 Separation of Functions

The team building a capability evaluation should not unilaterally determine deployment.

The team responsible for launch should not control unfavorable evidence.

28.3 Independent Review

Required when:

- Consequences are severe
- thresholds trigger major action
- methods are immature
- evaluator access is constrained
- evidence is disputed
- public claims are significant

28.4 Conflicts

Disclose:

- Employment
- funding
- equity
- consulting
- competitive interests
- policy advocacy
- national interests
- intellectual commitments

28.5 Dissent

Allow:

- Minority reports
- uncertainty ranges
- unresolved issue registers
- alternative threshold interpretations

28.6 Appeals

A party should be able to challenge:

- Configuration
- task validity
- scoring
- threshold application
- conflict
- procedural deviation
- interpretation

28.7 No Developer Self-Certification

Developer evidence is essential.

It should not be the only evidence for the most consequential claims.

28.8 No Evaluator Sovereignty

Independent evaluators should not become unaccountable authorities.

28.9 Funding

Funding should not purchase:

- Favorable findings
- result suppression
- threshold control
- exclusive ownership of public-interest methods

28.10 Framework Governance

Frontier safety frameworks are evolving organizational governance systems.

The current frameworks of OpenAI, Google DeepMind, and Anthropic connect defined capability or risk categories to safeguards and decision processes, but remain developer-created and voluntary. [^openai-pf2][^deepmind-fsf3][^anthropic-rsp3]

Independent standards work should learn from them without treating them as final consensus.

---

29. Security and Sensitive Evaluation

29.1 Why Evaluation Can Be Sensitive

Tasks may reveal:

- Vulnerabilities
- harmful methods
- capability thresholds
- safeguard weaknesses
- model weights
- deployment details
- national-security information

29.2 Controlled Disclosure

Use:

- Public summary
- restricted annex
- delayed release
- responsible disclosure
- need-to-know access
- independent audit

29.3 Evaluation-Induced Risk

The evaluation may increase risk by:

- Creating operational instructions
- connecting tools
- fine-tuning harmful capability
- exposing vulnerabilities
- enabling system escape
- generating reusable attack data

29.4 Safety Review

Before evaluation, review:

- Necessity
- safe alternatives
- containment
- personnel
- access
- data retention

- disclosure
- incident response

29.5 Secure Environments

Use sandboxes, simulated systems, controlled networks, and protected task banks where appropriate.

29.6 Result Integrity

Security should also prevent:

- Model substitution
- hidden configuration change
- selective deletion
- score manipulation
- unauthorized reruns

---

30. International Interoperability

High-stakes capability is global, but legal and institutional contexts differ.

30.1 Shared Elements

Institutions can align on:

- Definitions
- domain taxonomy
- protocol metadata
- confidence
- threshold status
- evaluator competence
- evidence categories
- result expiration
- incident classification

30.2 Different Risk Tolerances

Jurisdictions may choose different policy responses to the same capability evidence.

Evaluation comparability should not require identical governance.

30.3 Mutual Recognition

Recognition may require:

- Comparable construct
- sufficient evaluator access
- secure administration
- independent review
- compatible reporting
- appeals
- version control

30.4 International Institutes

Government AI safety and security institutes can support:

- Shared evaluations
- cross-testing
- common terminology
- incident exchange
- expert networks
- joint research

30.5 Standards and Law

NIST's AI Risk Management Framework provides a voluntary structure for managing AI risks across governance, mapping, measurement, and management.[^nist-rmf]

The European Union AI Act distinguishes high-risk use contexts and introduces additional obligations for general-purpose AI models with systemic risk.[^eu-ai-act]

These legal and risk-management systems overlap with high-stakes evaluation but are not identical to the capability framework proposed here.

30.6 Avoiding One Global Threshold

A single number is unlikely to represent every domain, system, and jurisdiction.

30.7 Shared Evidence, Local Authority

The long-term goal should be interoperable evidence with accountable local and international decision processes.

---

31. Maturity Model

Level 0: Informal Capability Testing

Characteristics:

- Ad hoc prompts
- no construct
- no baseline
- no uncertainty
- exploratory only

Level 1: Structured Domain Benchmark

Characteristics:

- Defined tasks
- scoring
- version
- basic reporting
- limited system context

Use:

- Research comparison

Level 2: High-Stakes Capability Protocol

Characteristics:

- Harm pathway
- subcapability graph
- held-out tasks
- serious elicitation
- domain experts
- reliability
- thresholds
- limitations

Use:

- Internal risk decisions

Level 3: Independently Reviewed System Assessment

Characteristics:

- External evaluation
- safeguard testing
- deployment context
- security
- decision matrix
- appeals
- re-evaluation triggers

Use:

- Consequential pre-deployment decisions

Level 4: Interoperable High-Stakes Evaluation Regime

Characteristics:

- Multiple qualified evaluators
- international mapping
- dynamic protocols
- incident feedback
- standards integration
- continuous monitoring
- mutual recognition
- formal governance

Use:

- Mature institutional ecosystem

---

32. Implementation Pathway

Phase 1: Select a Domain

Choose a bounded, decision-relevant capability.

Phase 2: Define the Harm Pathway

Identify actors, actions, targets, bottlenecks, safeguards, and consequences.

Phase 3: Build the Capability Graph

Decompose the pathway into measurable abilities.

Phase 4: Define Evaluation Tiers

Create early-warning, operational, and critical levels.

Phase 5: Establish Governance

Assign roles, review, conflicts, and appeals.

Phase 6: Create the Task Portfolio

Combine:

- Public
- held-out
- expert
- simulated
- end-to-end
- adversarial tasks

Phase 7: Establish Baselines

Use relevant humans, systems, and actor profiles.

Phase 8: Conduct Elicitation Experiments

Estimate capability under multiple resource conditions.

Phase 9: Validate

Assess:

- Correctness
- reliability
- coverage
- generalization
- uncertainty
- safety

Phase 10: Evaluate Safeguards

Test realistic adversaries.

Phase 11: Independent Review

Review methods and interpretation.

Phase 12: Decision Integration

Connect results to predefined responses.

Phase 13: Limited Deployment and Monitoring

Use staged exposure.

Phase 14: Re-Evaluate

Update after changes and incidents.

---

33. Proposed Standards Body Pilot

33.1 Pilot Name

High-Stakes Capability Evaluation Case: Autonomous Cyber Operations

33.2 Purpose

Develop an independent, transparent framework for distinguishing:

- Cyber knowledge
- bounded task capability
- tool-using capability
- end-to-end operational capability
- scalable autonomous capability

33.3 Why Cyber

Cyber provides:

- Clear dual-use relevance
- measurable tasks
- simulated environments
- strong expert community
- observable performance
- rapidly changing capability
- direct connection to safeguards

33.4 Capability Graph

The pilot should map:

- Reconnaissance
- vulnerability discovery
- exploitation
- access
- escalation
- persistence
- lateral movement
- objective completion
- stealth
- scaling

33.5 Evaluation Tracks

Public Research Track

Open tasks and reproducible baseline.

Held-Out Capability Track

Rotating protected tasks.

End-to-End Simulation Track

Controlled multi-step environments.

Safeguard Track

Access controls, monitoring, policy, and tool gating.

Elicitation Track

Standard, developer-supported, and evaluator-optimized configurations.

33.6 Evidence Outputs

- Capability profile
- reliability
- human uplift
- resource requirements
- safeguard bypass rate
- uncertainty
- threshold status
- independent review
- result expiration

33.7 Safety

- Isolated cyber ranges
- no unauthorized real targets
- controlled credentials
- exploit handling
- monitored tools
- disclosure policy

33.8 Pilot Deliverables

- Domain taxonomy
- capability graph
- protocol
- task-development guide
- threat model
- threshold proposal
- safeguard protocol
- reporting template
- external-review report
- annual update

33.9 Success Criteria

The pilot succeeds if it:

- Distinguishes knowledge from operational capability
- measures reliability
- supports multiple elicitation levels
- produces defensible thresholds
- protects sensitive content
- supports independent replication
- connects evidence to clear decisions
- remains useful as systems improve

---

34. Metrics for Evaluating the Framework

34.1 Measurement Quality

- Construct coverage
- reliability
- discrimination
- generalization
- baseline quality

- uncertainty
- end-to-end validity

34.2 Elicitation Quality

- Performance uplift by method
- resource use
- evaluator effort
- developer assistance
- transfer
- reproducibility

34.3 Decision Quality

- Correct threshold classification
- decision consistency
- later incident correlation
- avoided false confidence
- reversible action
- user understanding

34.4 Safeguard Quality

- Bypass rate
- attack coverage
- monitoring detection
- response
- scale behavior
- adaptation

34.5 Governance Quality

- Independent review
- conflict handling
- dissent
- appeal
- update speed
- threshold traceability

34.6 Operational Quality

- Cost
- duration
- expert burden

- infrastructure failure
- safety incidents
- access barriers

34.7 Adaptation Quality

- Time to update
- new capability coverage
- incident incorporation
- result expiration
- protocol retirement

---

35. Failure Modes and Safeguards

35.1 Capability Equals Risk

Failure: A capability score is treated as complete risk.

Safeguard: Separate capability, propensity, access, safeguards, exposure, consequence, and uncertainty.

35.2 Knowledge Equals Operational Ability

Failure: Multiple-choice or question-answer performance is treated as end-to-end capability.

Safeguard: Tool use, simulations, integration, and reliability testing.

35.3 One Success Equals Reliable Capability

Failure: A cherry-picked demonstration is generalized.

Safeguard: Repeated trials and success distributions.

35.4 No Success Equals Inability

Failure: Weak elicitation creates a false negative.

Safeguard: Multiple elicitation tracks and external challenge.

35.5 Threshold Theater

Failure: Precise thresholds create authority without strong evidence.

Safeguard: Uncertainty, ranges, review, versioning, and explicit assumptions.

35.6 Domain List Becomes Frozen

Failure: Institutions evaluate only known categories.

Safeguard: Open-domain review and early-signal monitoring.

35.7 Catastrophic Focus Crowds Out Present Harm

Failure: Ordinary harms receive insufficient attention.

Safeguard: Separate but coordinated evaluation portfolios.

35.8 Overclassification

Failure: Too many capabilities are labeled high stakes.

Safeguard: Explicit criteria, proportionality, and review.

35.9 Underclassification

Failure: Emerging systemic pathways are missed.

Safeguard: precursor evaluation, trend analysis, incident feedback.

35.10 Developer Self-Assessment

Failure: Interested party controls evidence and decision.

Safeguard: Independent review and external access.

35.11 Evaluator Underaccess

Failure: External review lacks time, information, or system access.

Safeguard: Access-level reporting and decision limits.

35.12 Unrealistic Evaluation

Failure: Tasks are difficult but not operationally meaningful.

Safeguard: Domain experts, real workflows, outcome validation.

35.13 Unsafe Evaluation

Failure: Testing creates harmful capability, data, or access.

Safeguard: safety review, proxies, containment, disclosure control.

35.14 Safeguard Neglect

Failure: Capability is measured without testing risk reduction.

Safeguard: parallel safeguard evaluation.

35.15 Deployment Neglect

Failure: Model score is applied to a different deployed system.

Safeguard: system-specific evaluation and deployment envelope.

35.16 Regulatory Capture

Failure: Thresholds favor incumbents or political interests.

Safeguard: transparent development, diverse participation, appeals, cost analysis.

35.17 International Fragmentation

Failure: Incompatible frameworks create duplicated burden and weak comparison.

Safeguard: shared metadata and interoperability.

35.18 Public Miscommunication

Failure: Technical findings become sensationalized.

Safeguard: claims boundary, uncertainty, domain-specific explanation.

35.19 Result Staleness

Failure: Old evidence supports new deployment.

Safeguard: expiration and re-evaluation triggers.

35.20 Single-Metric Collapse

Failure: Complex evidence is compressed into one score.

Safeguard: multidimensional capability profiles.

---

36. Serious Objections

Objection 1: High-Stakes Categories Are Inherently Political

The choice of what counts as severe reflects values.

Response:

- Make criteria explicit
- distinguish technical evidence from policy judgment
- include diverse stakeholders
- publish assumptions
- allow dissent

Residual concern:

No classification can be entirely value-neutral.

Objection 2: Capability Evaluations Can Legitimize Speculative Risks

Response:

- Require concrete harm pathways
- separate precursor from operational capability
- state confidence
- avoid sensational language
- retire unsupported categories

Residual concern:

Some early-warning work will remain uncertain by design.

Objection 3: Thresholds Are Too Fragile for Governance

Response:

- Use multiple indicators
- alert bands
- independent review
- decision matrices
- safeguards
- periodic revision

Residual concern:

Thresholds can still become false precision.

Objection 4: Evaluations Reveal Dangerous Information

Response:

- Use protected tasks
- safe proxies
- restricted annexes
- controlled environments
- responsible disclosure

Residual concern:

Some evaluation research may still increase dual-use knowledge.

Objection 5: Labs Are Best Positioned to Evaluate Their Models

They have access and expertise.

Response:

Developer evaluation is indispensable.

Independent evaluation adds challenge, legitimacy, and alternative assumptions.

Residual concern:

External evaluators may remain under-resourced or under-accessed.

Objection 6: High-Stakes Evaluation Will Slow Beneficial Innovation

Response:

- Use proportional tiers
- focus deep evaluation on credible severe pathways
- share infrastructure
- stage deployment
- evaluate safeguards, not capability alone

Residual concern:

Compliance burden can still advantage large firms.

Objection 7: Open-Weight Systems Make Threshold Controls Ineffective

Response:

Evaluation can still inform:

- Security
- release decisions

- downstream safeguards
- monitoring
- preparedness
- user guidance

Residual concern:

Post-release control is limited.

Objection 8: Real-World Risk Depends More on Humans Than Models

Often true.

Response:

Include actor uplift, access, deployment, and institutional vulnerability.

Objection 9: Current Models Do Not Justify Extreme-Risk Infrastructure

Response:

Build proportionate early-warning methods and avoid claiming current severe capability without evidence.

Residual concern:

Institution building can create self-reinforcing incentives.

Objection 10: Safety Frameworks Are Voluntary Corporate Policies

Correct.

Response:

Treat them as evidence and experiments, not settled standards.

Objection 11: Evaluation Cannot Cover Unknown Unknowns

Correct.

Response:

- Incident reporting
- red teaming
- broad monitoring
- contrarian review
- open category discovery

Objection 12: Capability Is Beneficial and Dual Use

Correct.

Response:

High-stakes evaluation should support safe benefit realization, not assume prohibition.

---

37. Evidence Gaps

37.1 Capability-to-Harm Relationship

How strongly do evaluation results predict severe real-world outcomes?

37.2 Human Uplift

Which baselines best measure the increase in actor capability?

37.3 Reliability

What success frequency makes a capability operationally material?

37.4 End-to-End Evaluation

How should component skills and integrated performance be combined?

37.5 Elicitation

How close do current methods come to plausible maximum capability?

37.6 Propensity

Which evaluations predict harmful or deceptive behavior in deployment?

37.7 Safeguards

How should adaptive adversaries and defense-in-depth be evaluated?

37.8 Thresholds

How can thresholds remain actionable without false precision?

37.9 Agent Time Horizons

How do task length, environment, and scaffold affect autonomy estimates?

37.10 Cross-Domain Interaction

How do cyber, scientific, autonomy, persuasion, and AI-R&D capabilities combine?

37.11 Multi-Agent Effects

Which risks emerge only through interaction?

37.12 External Access

What evaluator access is necessary for different claims?

37.13 Open-Weight Risk

How should capability evidence translate into decisions when weights are broadly available?

37.14 International Recognition

What evidence should support cross-border acceptance?

37.15 Decision Impact

Do high-stakes frameworks improve outcomes enough to justify burden and cost?

---

38. Research Agenda

Priority 1: Capability Pathway Models

Develop domain-specific graphs connecting component skills to end-to-end outcomes.

Priority 2: Operational Validity

Compare evaluation results with controlled real-world performance.

Priority 3: Reliability Thresholds

Study when occasional success becomes practically significant.

Priority 4: Human Uplift

Create consistent baselines across actor expertise and resources.

Priority 5: Elicitation Science

Test prompting, scaffolding, tools, fine-tuning, and inference budgets.

Priority 6: Safeguard Evaluation

Develop adaptive attack distributions and defense-in-depth metrics.

Priority 7: Agent Evaluation

Improve long-horizon, process, autonomy, and control measurement.

Priority 8: Threshold Governance

Pilot alert bands, uncertainty-aware triggers, and independent review.

Priority 9: Multi-Domain Capability

Study combinations that create greater capability than individual scores imply.

Priority 10: Multi-Agent Systems

Evaluate coordination, collusion, cascading failure, and oversight.

Priority 11: External Evaluator Access

Define minimum access for credible claims.

Priority 12: Evaluation Safety

Develop methods for sensitive domains that reduce evaluation-induced risk.

Priority 13: Post-Deployment Correlation

Connect pre-deployment evidence to incidents and operational behavior.

Priority 14: International Interoperability

Create common metadata, evidence categories, and result status.

Priority 15: Institutional Effects

Study whether capability frameworks create capture, concentration, or perverse incentives.

---

39. Near-Term Experiments

Experiment 1: Knowledge to Operation

Compare question-answer performance with end-to-end tool-using tasks in one domain.

Experiment 2: Reliability Curve

Measure capability across repeated attempts and resource budgets.

Experiment 3: Human Uplift

Compare lay, trained, and expert users with and without AI assistance.

Experiment 4: Elicitation Matrix

Run default, standardized, developer-supported, and externally optimized configurations.

Experiment 5: Threshold Band

Test an uncertainty-aware alert band rather than a single cutoff.

Experiment 6: Safeguard Adaptation

Allow red teams to adapt attacks across multiple rounds.

Experiment 7: Deployment Envelope

Test the same model with different tools, permissions, monitoring, and user populations.

Experiment 8: Capability Graph

Construct and validate a directed graph for an autonomous cyber pathway.

Experiment 9: Cross-Domain Composition

Test whether moderate capabilities combine into unexpected end-to-end performance.

Experiment 10: External Access Levels

Compare evaluator findings under black-box, grey-box, and richer-access conditions.

Experiment 11: Incident Backtesting

Determine whether prior evaluations would have predicted known failures.

Experiment 12: Multi-Agent Risk

Compare single-agent and multi-agent performance under shared objectives and competition.

---

40. Implications for Future Standards

A future high-stakes capability evaluation standard could require:

40.1 Domain Definition

Specify the risk domain and harm pathway.

40.2 Capability Graph

Define component and end-to-end capabilities.

40.3 Evaluated System

Document model, scaffold, tools, safeguards, and deployment.

40.4 Evaluation Tiers

Separate early warning, operational capability, and critical capability.

40.5 Elicitation

Document methods and resources used to reveal capability.

40.6 Task Portfolio

Use dynamic, held-out, expert, and realistic tasks.

40.7 Baselines

Include relevant human and system comparisons.

40.8 Reliability

Measure repeated success and uncertainty.

40.9 Safeguards

Test controls against representative and adaptive threats.

40.10 Deployment Context

Assess access, scale, tools, monitoring, and vulnerability.

40.11 Independent Review

Require appropriate external scrutiny.

40.12 Threshold Governance

Define triggers, authority, review, and appeal.

40.13 Reporting

Publish evidence, limitations, configuration, confidence, and expiration.

40.14 Re-Evaluation

Define update and incident triggers.

40.15 Safety

Control evaluation-induced risk and sensitive disclosure.

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

---

41. Relationship to the Other Foundations

Foundation 1: Dynamic Evaluation Protocols

High-stakes tests must evolve as capability, threats, and deployment change.

Foundation 2: Held-Out Evaluations

Protected tasks reduce leakage and targeted optimization.

Foundation 4: Independent Expert Review

Consequential judgments require qualified external scrutiny.

Foundation 5: Third-Party Auditor Ecosystem

Scaled evaluation requires competent independent organizations.

Foundation 6: Progressive Standards and Requirements

Capability evidence can trigger increasingly formal safeguards and institutional responses.

Foundation 7: Incentives and Prestige

Organizations should receive recognition for rigorous evaluation and transparent risk reduction.

Foundation 8: Global Interoperability

Shared evidence should support cross-border understanding without requiring identical policy.

---

42. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Not every AI capability requires the same evaluation burden.
2. Evaluation rigor should increase with the potential consequence of error.
3. High-stakes capability is not synonymous with frontier model, legal high-risk use, or prohibited technology.
4. Capability is necessary evidence for many risk judgments, but capability alone is not risk.
5. High-stakes assessment should consider capability, propensity, access, exposure, vulnerability, safeguards, consequence, and uncertainty.
6. A single aggregate score is generally insufficient.
7. Knowledge tests should not be treated as proof of operational capability.
8. Component evaluations should be paired with end-to-end evaluation where feasible.
9. Reliability, autonomy, resource requirements, and human uplift are core dimensions.
10. Evaluations should distinguish typical, standardized, developer-elicited, externally elicited, and plausible maximum capability.

11. High-stakes protocols should use dynamic and held-out components.
12. Domain experts should participate meaningfully.
13. Capability thresholds should trigger defined processes, not function as unexplained labels.
14. Alert thresholds should precede critical thresholds where early preparation is valuable.
15. Threshold uncertainty should be explicit.
16. False positives and false negatives both create serious costs.
17. Capability and safeguard evaluations should be reported separately.
18. The deployed system, not only the base model, should be evaluated.
19. External evaluator access should be reported as part of evidentiary confidence.
20. Developer evaluation is essential but should not be the sole basis for the most consequential claims.
21. Independent evaluators should remain accountable and appealable.
22. High-stakes evaluation should not unnecessarily suppress beneficial dual-use capability.
23. Present harms and severe frontier risks require separate but coordinated evaluation portfolios.
24. Sensitive evaluations require proportionate security.
25. Passing an evaluation is not proof of safety.
26. Failure to elicit a capability is not proof of inability without strong methods.
27. Evaluation evidence should expire or trigger re-evaluation after material changes.
28. Incidents should update capability models and protocols.
29. International interoperability is preferable to one universal threshold.
30. The evaluation regime itself should be evaluated for capture, burden, and effectiveness.

---

43. Decision Rules

A capability should enter enhanced evaluation when:

- It is a credible precursor to severe harm
- performance is improving rapidly
- deployment scale increases
- new tools or autonomy materially expand action
- independent evidence challenges prior assumptions

- a real incident reveals missing coverage
- safeguards are weak or untested
- a framework alert threshold is approached
- uncertainty is high and consequence is severe

A capability should be classified as operationally high stakes only when evidence supports:

- A defined harm pathway
- meaningful end-to-end ability or strong component combination
- plausible access and resource conditions
- sufficient reliability or material uplift
- consequence beyond ordinary product risk

A capability should not be classified as operationally high stakes merely because:

- The model is large
- the result is surprising
- a task is difficult
- public attention is high
- the capability is dual use
- a single demonstration succeeded
- the evaluator uses alarming terminology

A threshold result should be suspended when:

- Configuration is uncertain
- tasks are invalid or compromised
- scoring fails
- elicitation is materially incomplete
- evaluator access is insufficient for the claim
- independent replication materially disagrees
- system changes invalidate the evidence

A system should be re-evaluated when:

- Model version changes
- scaffold changes
- tools change
- fine-tuning changes
- access expands
- scale increases
- safeguards change
- new attacks emerge
- incidents occur
- the protocol expires

---

44. High-Stakes Evaluation Plan Template

A. Identity

- Evaluation name
- identifier
- version
- owner
- date
- status

B. Decision Question

- Decision
- decision-maker
- consequence
- timeline
- risk tolerance

C. Domain

- Risk domain
- harm pathway
- targets
- threat actors
- consequences

D. Capability Model

- Top-level capability
- component skills
- dependencies
- bottlenecks
- end-to-end definition

E. Evaluated System

- Model
- version
- tools
- scaffold
- safeguards
- access
- environment

F. Evaluation Tiers

- Early warning
- operational
- critical

G. Tasks

- Public
- held-out
- dynamic
- expert
- simulation
- end-to-end

H. Elicitation

- Standard configuration
- developer support
- external optimization
- compute
- human effort
- retries

I. Baselines

- Humans
- prior systems
- existing tools
- actors
- defensive systems

J. Metrics

- Success
- reliability
- autonomy
- time
- cost
- uplift
- transfer
- safeguard bypass
- uncertainty

K. Safeguards

- Threat model
- controls
- attack distribution
- residual risk

L. Deployment

- Users
- scale
- permissions
- monitoring
- human oversight
- vulnerability

M. Governance

- Roles
- conflicts
- independent review
- dissent
- appeal
- funding

N. Safety and Security

- Sensitive content
- containment
- access
- disclosure
- incident response

O. Thresholds

- Alert
- critical
- evidence
- confidence
- trigger
- review

P. Reporting

- Public report

- restricted annex
- limitations
- result status
- expiration

Q. Re-Evaluation

- Time
- event
- incident
- system change

---

45. Capability Evidence Case Template

Claim:
Domain:
System:
Protocol version:
Date:

Capability Definition

Harm Pathway

Evidence Supporting the Claim

Evidence Against the Claim

Elicitation Conditions

Reliability

Human or System Baseline

Generalization

Access and Resource Requirements

Safeguard Evidence

Deployment Assumptions

Uncertainty

Independent Review

Dissent

Threshold Status

- Below alert
- alert
- near critical
- critical
- indeterminate

Decision Implication

Expiration

---

46. Threshold Change Request Template

Domain:
Current threshold:
Proposed threshold:
Proposer:
Date:

Reason for Change

New Evidence

Expected Decision Impact

False-Positive Risk

False-Negative Risk

Systems Affected

Safeguard Implications

International Compatibility

Independent Review

Conflicts

Public Consultation

Decision

- Approved
- approved with conditions

- pilot
- deferred
- rejected

Effective Date

Review Date

---

47. High-Stakes Evaluation Scorecard

Dimension	Core Question
Decision	Is the evaluation linked to a defined decision?
Domain	Is the risk domain specific and defensible?
Harm pathway	Is the route from capability to consequence explicit?
Capability graph	Are component and end-to-end abilities mapped?
Evaluated system	Are model, tools, scaffold, and safeguards specified?
Task validity	Do tasks represent the capability?
Dynamic quality	Can the protocol evolve with capability?
Holdout integrity	Is targeted optimization controlled?
Elicitation	Was performance seriously elicited?
Reliability	Is success operationally consistent?
Baselines	Are human and system comparisons appropriate?
Uplift	Is AI-enabled improvement measured?
Autonomy	Is independent action measured?
Generalization	Does capability transfer across tasks and contexts?
Safeguards	Are controls evaluated against realistic attacks?
Deployment	Are access, scale, tools, and monitoring included?
Consequence	Are severity, scale, speed, and reversibility considered?
Uncertainty	Are evidence gaps and error risks explicit?
Independence	Is there qualified external review?
Access	Did evaluators receive enough access for the claim?
Security	Is sensitive evaluation conducted safely?
Threshold	Is the trigger justified and governed?
Appeals	Can material errors be challenged?
Freshness	Is re-evaluation or expiration defined?
Interoperability	Can others understand and compare the evidence?

Dimension	Core Question
Burden	Is the evaluation proportionate?
Decision utility	Does the result improve an actual decision?

---

48. Final Perspective

The central problem of high-stakes AI evaluation is not identifying everything a model can do.

No institution can test every possible task, actor, environment, tool, deployment, or failure.

The problem is deciding where uncertainty is no longer acceptable.

When an AI system can act in domains involving cyber operations, biological research, critical infrastructure, autonomous replication, large-scale manipulation, or accelerated AI development, ordinary benchmark practice may no longer provide enough evidence.

The evaluation must become more realistic.

The elicitation must become more serious.

The safeguards must become part of the test.

The deployment context must become visible.

The uncertainty must become explicit.

The reviewers must become more independent.

The result must connect to a decision.

This does not require assuming that advanced AI will cause catastrophe.

It requires recognizing that consequence changes the standard of proof.

A weak evaluation can create harm in two directions.

It can miss a capability and enable reckless exposure.

It can exaggerate a capability and enable fear, concentration, and unnecessary restriction.

The goal is neither alarm nor reassurance.

The goal is justified confidence.

High-stakes capability evaluation should help society distinguish:

- What a system can do
- what it cannot yet do
- what it may do under stronger elicitation
- what actors can access

- what safeguards prevent
- what remains uncertain
- what decisions the evidence supports
- when that evidence must be revisited

The third foundation of Standards Body is therefore proportional scrutiny.

Where the consequences of error are greatest, the evidence should be strongest.

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-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-protein]: National Institute of Standards and Technology, **Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies**, 2025. https://www.nist.gov/publications/experimental-evaluation-ai-driven-protein-design-risks-using-safe-biological-proxies

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

[^openai-pf2]: OpenAI, **Preparedness Framework, Version 2**, April 15, 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[^openai-frontier]: OpenAI, **Frontier Governance Framework**, May 28, 2026. https://openai.com/index/openai-frontier-governance-framework/

[^openai-bio]: OpenAI, **Preparing for Future AI Capabilities in Biology**, June 18, 2025. https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/

[^deepmind-fsf3]: Google DeepMind, **Frontier Safety Framework, Version 3.0**, September 2025. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf

[^anthropic-rsp3]: Anthropic, **Responsible Scaling Policy, Version 3.0**, February 24, 2026. https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0

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

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

[^aisi-elicitation]: UK AI Security Institute, **A Structured Protocol for Elicitation Experiments**, July 16, 2025. https://www.aisi.gov.uk/blog/our-approach-to-ai-capability-elicitation

[^aisi-safeguards]: UK AI Security Institute, **Principles for Evaluating Misuse Safeguards of Frontier AI Systems**, 2025. https://www.aisi.gov.uk/research/principles-for-evaluating-misuse-safeguards-of-frontier-ai-systems

[^aisi-actionable]: UK AI Security Institute, **Making Safeguard Evaluations Actionable**, May 29, 2025. https://www.aisi.gov.uk/blog/making-safeguard-evaluations-actionable

[^aisi-trends]: UK AI Security Institute, **Frontier AI Trends Report**, 2025. https://www.aisi.gov.uk/frontier-ai-trends-report

[^aisi-cyber-eval]: UK AI Security Institute, **Our Evaluation of OpenAI's GPT-5.5 Cyber Capabilities**, April 30, 2026. https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities

[^extreme-risk]: Toby Shevlane et al., **Model Evaluation for Extreme Risks**, 2023. https://arxiv.org/abs/2305.15324

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

[^dangerous-repository]: Google DeepMind, **Dangerous Capability Evaluations Repository**. https://github.com/google-deepmind/dangerous-capability-evaluations

[^stealth]: Mary Phuong et al., **Evaluating Frontier Models for Stealth and Situational Awareness**, 2025. https://arxiv.org/abs/2505.01420

[^external-access]: Jacob Charnock et al., **Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations**, 2026. https://arxiv.org/abs/2601.11916

[^deepmind-cyber]: Google DeepMind, **Evaluating Potential Cybersecurity Threats of Advanced AI**, April 2, 2025. https://deepmind.google/blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/

[^deepmind-multiagent]: Google DeepMind, **Investing in Multi-Agent AI Safety Research**, June 11, 2026. https://deepmind.google/blog/investing-in-multi-agent-ai-safety-research/

[^aisi-safety-cases]: UK AI Security Institute, **Safety Cases at AISI**, 2024. https://www.aisi.gov.uk/blog/safety-cases-at-aisi

[^frontier-model-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/

[^amazon-framework]: Amazon, **Frontier Model Safety Framework**, February 2025. https://www.amazon.science/publications/amazons-frontier-model-safety-framework

---

Revision Record

Version 1.0

Date: July 16, 2026

Change type: Complete replacement

Summary: Replaces the earlier outline edition with a fully developed canonical working white paper. Adds a complete risk model, high-stakes classification framework, domain-specific evaluation chapters, capability decomposition, elicitation methodology, threshold theory, evidence standards, false-positive and false-negative analysis, safeguard evaluation, deployment-context analysis, lifecycle governance, security, international interoperability, maturity model, implementation pathway, a Standards Body autonomous-cyber pilot, metrics, failure analysis, objections, evidence gaps, research agenda, standards implications, templates, scorecard, and current primary-source research basis.

Status: Ready for internal review and future expert critique.