

Standards Body · Foundation paper, public edition · Released July 17, 2026

Canonical record: <https://standardsbody.ai/library/foundation-paper/incentives-and-prestige/>

Standards Body is an independent research and institutional-design project. It is not currently a regulator, accreditation body, certification body, or governmental authority. This document is research; it is not an adopted standard.

FOUNDATION_07_INCENTIVES_AND_PRESTIGE.md

Foundation 7: Incentives and Prestige

Series: Foundations for Frontier AI Evaluation Infrastructure

Version: 1.0

Status: Canonical working white paper

Project: Standards Body

Primary domain: standardsbody.ai

Core line: Foundations for Frontier AI

Research basis reviewed through: July 16, 2026

Document owner: Standards Body

Review cycle: Annual review, with event-triggered revision after material changes in frontier AI markets, evaluation practice, professional incentives, standards adoption, or institutional behavior

Document Purpose

This paper defines the Standards Body position on incentives, recognition, reputation, and prestige within frontier AI evaluation and standards development.

It is intended to serve as:

- A first-principles explanation of why technical quality depends on institutional incentives
- A framework for aligning the interests of developers, evaluators, researchers, standards bodies, purchasers, insurers, governments, and open communities
- A guide to responsible use of prizes, grants, rankings, public recognition, procurement preference, certification, disclosure, career credit, and reputational signals
- A model for rewarding evidence, correction, transparency, and public-interest contribution without encouraging metric gaming or prestige capture
- A bridge between voluntary practice and progressively stronger standards
- A reference for future partnership programs, contributor systems, evaluator recognition, public reporting, research funding, and institutional design
- A durable source document from which shorter articles, operational policies, and incentive-program specifications can be developed

This paper is not a marketing plan.

It does not propose that safety or standards compliance should be reduced to public rankings.

It does not assume that financial rewards are always the strongest motivators.

It establishes the conditions under which incentives and prestige can strengthen frontier AI evaluation rather than distort it.

Executive Summary

Frontier AI evaluation is usually discussed as a technical problem.

It is also an incentive problem.

A benchmark can be well designed and still fail if developers benefit more from optimizing the visible score than from improving the underlying capability or safeguard.

An independent evaluator can be technically competent and still soften conclusions if future access, contracts, prestige, or publication depend on pleasing the system developer.

A researcher can identify an important weakness and still avoid publishing it if the work is difficult to credit, risky to disclose, or less prestigious than building a new model.

A company can invest substantially in safety while receiving little market benefit because purchasers, users, investors, insurers, and governments cannot distinguish serious practice from public relations.

A standards organization can become more interested in adoption numbers, institutional status, or sponsorship than in whether its standards remain valid.

An open-source contributor can provide critical evaluation infrastructure while receiving less recognition than a highly visible but less rigorous public commentator.

A regulator can impose reporting requirements that reward the production of documents rather than the reduction of risk.

These are not secondary concerns.

They shape which evidence is produced, which failures are revealed, which methods receive resources, which standards are adopted, which experts remain in the field, and which institutions gain authority.

The central proposition of this foundation is:

Frontier AI evaluation infrastructure will be no stronger than the incentives surrounding it. Institutions should deliberately reward accurate evidence, independent challenge, correction, reproducibility, safeguard improvement, responsible disclosure, and public-interest contribution, while limiting rewards for superficial compliance, selective transparency, score optimization, prestige accumulation, and unsupported claims.

Incentives operate through more than money.

Relevant incentive channels include:

- Revenue
- market access
- procurement
- insurance
- investment
- legal exposure
- professional advancement
- publication
- model access
- data access
- compute access
- grants
- prizes
- awards
- standards participation
- certification
- public recognition
- peer respect
- institutional legitimacy
- reputation
- moral commitment
- mission
- belonging
- autonomy
- curiosity
- fear of sanction
- fear of exclusion
- desire for influence

Prestige is particularly important in frontier AI because the field is concentrated, technically specialized, rapidly changing, and highly visible.

Prestige can function as:

- A signal of competence
- A reward for public contribution
- A recruiting advantage
- A source of institutional authority
- A gateway to model access
- A substitute for formal credentials
- A mechanism for coordinating attention
- A defense against commercial pressure

Prestige can also become dangerous.

It can reward:

- Dramatic claims
- exclusive access
- association with powerful laboratories
- benchmark leadership without validity
- public certainty under uncertainty
- institutional proximity
- selective disclosure
- fashionable topics
- consensus signaling
- self-promotion

A serious institution should not attempt to eliminate prestige.

Prestige is unavoidable wherever communities allocate attention, trust, and opportunity.

The task is to make prestige more evidence-sensitive.

Standards Body therefore proposes an incentive architecture based on six principles.

1. Reward the underlying contribution, not merely the visible artifact

A high benchmark score should not be rewarded more than a valid benchmark.

A polished report should not be rewarded more than an accurate one.

A certificate should not be rewarded more than sustained conformity.

2. Reward correction as well as initial performance

Institutions should receive credit for:

- Disclosing error
- correcting evidence
- withdrawing invalid claims
- improving safeguards
- reporting incidents
- revising standards

without making failure consequence-free.

3. Separate recognition from authority

An award, designation, or public profile should not automatically grant decision power, accreditation, or certification authority.

4. Use multiple incentive channels

Financial rewards alone can crowd out intrinsic motivation, distort priorities, or favor measurable outputs. A resilient system combines:

- Funding
- career recognition
- procurement
- public status
- access
- governance participation
- mission
- community respect

5. Anticipate gaming

Every important metric or reward can become a target.

Incentive design should include:

- Adversarial review
- multiple measures
- verification
- uncertainty
- rotation
- expiration
- anti-manipulation controls
- retrospective evaluation

6. Preserve plural pathways to contribution

Frontier AI standards should not reward only large organizations, elite universities, or highly visible individuals.

Valuable contribution can come from:

- Small laboratories
- independent researchers
- security practitioners
- domain experts
- open-source communities
- public-interest organizations
- affected users
- standards professionals
- auditors
- maintainers
- technical writers

- incident reporters

A mature incentive ecosystem should align the interests of several actor groups.

For developers, it should make rigorous evaluation and safeguards commercially, institutionally, and reputationally valuable.

For evaluators, it should reward independence, accuracy, security, correction, and methodological contribution rather than favorable client outcomes.

For researchers, it should reward replication, negative results, benchmark maintenance, incident analysis, and shared infrastructure.

For standards bodies, it should reward durability and effectiveness rather than document production.

For purchasers and insurers, it should make credible assurance decision-relevant.

For governments, it should create incentives for evidence-based procurement, research funding, disclosure protection, and international interoperability.

For open communities, it should provide recognition, access, funding, and governance participation without requiring institutional prestige in advance.

The most dangerous incentive mistake is to confuse visibility with value.

The seventh foundation of Standards Body is therefore the deliberate design of the motivational and reputational environment around frontier AI evaluation.

1. Foundational Proposition

1.1 Core Thesis

Frontier AI evaluation and standards will succeed only when the institutions and people involved have meaningful reasons to produce, reveal, verify, and act upon accurate evidence.

1.2 Prestige Thesis

Prestige is a form of institutional currency. It should be attached to demonstrated contribution, methodological integrity, correction, and public value rather than proximity, publicity, or unsupported certainty.

1.3 Multi-Incentive Thesis

No single incentive mechanism is sufficient. Durable alignment requires financial, professional, reputational, access-based, governance, and mission-driven incentives.

1.4 Anti-Gaming Thesis

Every incentive changes behavior. Incentives should be designed with an explicit theory of gaming, substitution, crowding out, capture, and unintended effects.

1.5 Correction Thesis

Institutions should face consequences for preventable failure while receiving meaningful credit for timely disclosure, remediation, and evidence-based revision.

1.6 Public-Goods Thesis

Evaluation methods, reference infrastructure, incident databases, open tools, and standards maintenance create public goods that markets may underfund without deliberate support.

1.7 Distribution Thesis

Incentive systems should not reserve recognition and opportunity for actors who already possess money, status, access, or institutional affiliation.

2. Scope and Boundaries

2.1 What This Foundation Covers

This paper covers incentives affecting:

- Frontier AI developers
- deployers
- evaluators
- auditors
- accreditation bodies
- researchers
- standards organizations
- governments
- purchasers
- insurers

- investors
- open-source communities
- public-interest organizations
- individual contributors
- affected users

It covers:

- Financial rewards
- market incentives
- procurement
- insurance
- grants
- prizes
- reputation
- awards
- public recognition
- rankings
- disclosure
- professional status
- career advancement
- access
- governance participation
- sanctions
- liability
- certification
- standards adoption

2.2 What This Foundation Does Not Fully Cover

This paper does not fully specify:

- Tax policy
- securities law
- antitrust enforcement
- compensation law
- academic tenure systems
- insurance regulation
- prize-contract law
- complete labor-market design
- political campaign incentives
- macroeconomic policy

2.3 Incentives Versus Values

Incentives influence behavior.

They do not replace:

- Ethics
- professional identity
- institutional mission
- personal judgment
- public duty

A system built only around rewards and penalties may undermine the motivations it needs.

2.4 Prestige Versus Public Relations

Prestige is durable social recognition within a community.

Public relations seeks favorable perception.

They can overlap.

They should not be treated as equivalent.

2.5 Reward Versus Authority

Recognition should not automatically create:

- Accreditation
- certification authority
- governance control
- regulatory status
- scientific truth

2.6 Positive and Negative Incentives

Positive incentives include:

- Funding
- recognition
- access
- procurement preference
- career credit

Negative incentives include:

- Sanctions
- loss of access
- liability
- public correction
- certificate withdrawal

Both require safeguards.

3. Canonical Definitions

3.1 Incentive

An incentive is a condition that changes the expected benefit, cost, status, opportunity, or consequence associated with an action.

3.2 Intrinsic Motivation

Motivation arising from interest, purpose, mastery, curiosity, identity, or satisfaction inherent in the activity.

3.3 Extrinsic Motivation

Motivation arising from external rewards, sanctions, status, access, or requirements.

3.4 Reputational Incentive

A benefit or cost created by how others perceive an actor's competence, reliability, integrity, or social value.

3.5 Prestige

Relatively durable esteem or status granted by a relevant community or institution.

3.6 Recognition

Formal or informal acknowledgment of contribution, competence, or achievement.

3.7 Credential

Evidence presented as proof of qualification, competence, status, or completion.

3.8 Award

A formal recognition granted according to stated criteria.

3.9 Prize

A reward, often financial or reputational, offered for achieving a defined objective.

3.10 Challenge Competition

A structured competition inviting participants to solve a defined problem under specified rules and judging criteria.

3.11 Bounty

A reward offered for identifying a vulnerability, failure, misuse pathway, benchmark flaw, or other specified finding.

3.12 Grant

Funding provided to support research, infrastructure, training, public-interest work, or institutional capacity.

3.13 Procurement Preference

An advantage given in purchasing decisions to actors meeting defined practices or assurance conditions.

3.14 Insurance Incentive

A change in coverage, premium, deductible, or underwriting treatment linked to evidence or controls.

3.15 Access Incentive

Access to models, data, compute, tools, events, or decision processes granted as a reward or qualification.

3.16 Governance Incentive

Influence, voting rights, committee participation, or advisory roles granted for contribution or competence.

3.17 Career Incentive

Professional advancement, hiring, promotion, tenure, publication, or reputation linked to behavior.

3.18 Public-Goods Contribution

Work that creates broadly usable value but may not be fully compensated by the direct beneficiaries.

3.19 Free Rider

An actor that benefits from shared infrastructure or risk reduction without contributing proportionately.

3.20 Moral Hazard

A condition in which protection from consequences changes behavior in a way that increases risk.

3.21 Principal-Agent Problem

A problem arising when an agent makes decisions on behalf of a principal but has different information or incentives.

3.22 Multi-Tasking Distortion

A condition in which rewarding one measurable task causes neglect of other important but less measurable tasks.

3.23 Crowding Out

Reduction of intrinsic or prosocial motivation after external rewards or controls are introduced.

3.24 Goodhart Effect

Degradation of a measure's value when it becomes a target for optimization.

3.25 Campbell Effect

Corruption pressure created when a quantitative indicator is used heavily for consequential decisions.

3.26 Signaling

Behavior intended to communicate quality, commitment, competence, or alignment to others.

3.27 Assurance Signal

A report, certification, evaluation result, or disclosure intended to reduce information asymmetry.

3.28 Prestige Capture

Control of recognition systems by actors able to convert existing status into further authority or rewards.

3.29 Recognition Inflation

Decline in the meaning of a designation as awards, badges, certificates, or titles proliferate.

3.30 Perverse Incentive

An incentive that predictably encourages behavior contrary to the intended objective.

3.31 Reward Hacking

Behavior that maximizes the measured reward while avoiding or undermining the intended goal.

3.32 Contribution Credit

Attribution of value to the people or organizations responsible for an outcome.

3.33 Corrective Credit

Recognition given for timely error disclosure, remediation, withdrawal, or improvement.

3.34 Prestige Decay

Loss of recognition value when evidence becomes stale, performance declines, or the designation is overused.

3.35 Incentive Compatibility

A condition in which participants benefit from acting in ways aligned with the intended rules or objective.

4. Why Incentives Are Foundational

4.1 Evidence Production Is Costly

High-quality evaluation requires:

- Expertise
- compute
- model access
- task development
- validation
- security
- documentation

- independent review
- maintenance

If the benefits flow broadly while costs are concentrated, underinvestment is likely.

4.2 Bad News Is Often Costly

An organization that reveals:

- A capability threshold
- a safeguard failure
- a security incident
- an invalid benchmark
- an incorrect claim

may face:

- Reputational loss
- legal exposure
- delayed deployment
- competitor advantage
- loss of access
- public criticism

Without a credible disclosure environment, problems remain hidden.

4.3 Visible Outputs Dominate Invisible Work

The field often rewards:

- New model releases
- benchmark scores
- influential papers
- dramatic findings
- product launches

More than:

- Dataset cleaning
- benchmark maintenance
- replication
- negative results
- documentation
- incident analysis
- standards editing
- security operations
- evaluator calibration

The less visible work is often essential infrastructure.

4.4 Access Is a Powerful Incentive

Access to frontier models can shape:

- Research agendas
- public commentary
- institutional relationships
- evaluator independence
- career opportunities

4.5 Reputation Influences Adoption

Organizations may adopt standards partly because they:

- Signal seriousness
- satisfy buyers
- attract talent
- reduce uncertainty
- improve legitimacy

4.6 Sanctions Shape Reporting

Excessively punitive systems can reduce voluntary disclosure.

Weak sanctions can reward negligence.

4.7 Institutional Survival Shapes Standards

Standards organizations may seek:

- Membership
- sponsorship
- adoption
- document sales
- relevance
- recognition

These incentives can affect technical judgment.

4.8 Incentives Affect Who Participates

Unpaid committee work and expensive travel favor:

- Large firms
- wealthy institutions

- established professionals
- actors with organizational support

4.9 Public Attention Is Scarce

Attention influences:

- Funding
- policy
- talent
- prestige
- research priorities

Incentive design must account for attention markets.

5. Actor Incentive Map

5.1 Frontier Developers

Potential incentives:

- Revenue
- market leadership
- talent
- investment
- access to compute
- regulatory influence
- reputation
- faster deployment
- lower liability
- procurement eligibility

Potential distortions:

- Release pressure
- selective evaluation
- benchmark optimization
- strategic disclosure
- favorable auditor selection
- underinvestment in public goods

5.2 Deployers

Potential incentives:

- Productivity
- cost reduction
- competitive advantage
- customer satisfaction
- legal compliance

Potential distortions:

- Underreporting incidents
- relying on vendor claims
- weak monitoring
- unsafe automation
- transferring responsibility

5.3 Evaluators and Auditors

Potential incentives:

- Fees
- repeat clients
- model access
- public recognition
- publication
- government contracts
- accreditation

Potential distortions:

- Client capture
- overclaiming
- proprietary lock-in
- favorable conclusions
- reluctance to correct
- prestige competition

5.4 Researchers

Potential incentives:

- Publication
- citations
- grants
- career advancement
- model access

- novelty
- public attention

Potential distortions:

- Sensationalism
- neglect of replication
- benchmark proliferation
- weak maintenance
- selective results
- access dependence

5.5 Standards Bodies

Potential incentives:

- Adoption
- membership
- sponsorship
- institutional relevance
- government recognition

Potential distortions:

- Slow retirement
- consensus dilution
- incumbent influence
- document production over impact

5.6 Governments

Potential incentives:

- Public safety
- economic growth
- national advantage
- political credit
- administrative control
- international leadership

Potential distortions:

- National favoritism
- symbolic regulation
- secrecy
- procurement capture
- short political time horizons

5.7 Purchasers

Potential incentives:

- Reliable performance
- reduced risk
- cost control
- supplier accountability

Potential distortions:

- Lowest-price selection
- checkbox assurance
- shifting responsibility to vendors
- proprietary requirements

5.8 Insurers

Potential incentives:

- Accurate risk pricing
- loss prevention
- market growth

Potential distortions:

- Opaque underwriting
- exclusion rather than improvement
- excessive reliance on certificates
- short claims history

5.9 Investors and Lenders

Potential incentives:

- Growth
- risk control
- governance quality
- legal predictability

Potential distortions:

- Short-term performance
- pressure for deployment
- superficial environmental, social, and governance signals

5.10 Open-Source Communities

Potential incentives:

- Technical contribution
- reputation
- public benefit
- autonomy
- learning
- adoption

Potential distortions:

- Visibility competition
- weak maintenance incentives
- ideological polarization
- decentralized responsibility gaps

5.11 Public-Interest Organizations

Potential incentives:

- Mission
- influence
- funding
- public attention

Potential distortions:

- Funder dependence
- issue amplification
- ideological commitment
- weak technical capacity

5.12 Individual Contributors

Potential incentives:

- Mastery
- credit
- career
- belonging
- mission
- access
- income

Potential distortions:

- Burnout
 - unpaid labor
 - status competition
 - retaliation fear
-

6. The Incentive Stack

A durable system should combine several layers.

6.1 Intrinsic Layer

Supports:

- Curiosity
- mission
- professional pride
- public service
- craftsmanship
- truth-seeking

Institutional design should protect autonomy and meaning.

6.2 Peer Layer

Supports:

- Respect
- recognition
- belonging
- professional norms
- constructive competition

6.3 Career Layer

Supports:

- Hiring
- promotion
- tenure
- leadership
- credentialing
- project opportunity

6.4 Access Layer

Supports access to:

- Models
- data
- compute
- evaluation environments
- working groups
- conferences
- institutional networks

6.5 Financial Layer

Supports:

- Salaries
- grants
- prizes
- bounties
- contracts
- insurance benefits
- procurement revenue

6.6 Market Layer

Supports:

- Customer preference
- investor confidence
- lower transaction cost
- market access
- certification value

6.7 Governance Layer

Supports:

- Voting
- committee participation
- advisory roles
- standard-setting influence
- public consultation

6.8 Enforcement Layer

Creates consequences through:

- Corrective action
- loss of recognition
- contract remedies
- penalties
- liability
- access restriction

6.9 Interaction

An incentive can be weak alone but strong in combination.

Example:

A developer may invest in independent evaluation because it provides:

- Procurement eligibility
 - lower insurance burden
 - public recognition
 - evaluator access
 - lower regulatory uncertainty
 - internal confidence
-

7. Intrinsic Motivation and Crowding Out

7.1 Why Intrinsic Motivation Matters

Many high-value contributors are motivated by:

- Scientific truth
- public safety
- professional duty
- open knowledge
- craftsmanship
- community

7.2 Crowding-Out Risk

External rewards can shift attention from:

- Quality
- mission
- judgment

toward:

- Payment
- score
- compliance
- status

Research in motivation psychology and economics has documented conditions under which controlling external rewards can reduce intrinsic motivation or prosocial behavior.[^deci-meta] [^frey-jegen]

7.3 Implications

Do not pay only for:

- Number of vulnerabilities
- number of papers
- benchmark points
- certificates issued
- standards published
- incidents closed

7.4 Autonomy-Supporting Incentives

Provide:

- Choice
- recognition of mastery
- meaningful purpose
- fair compensation
- peer respect
- ownership
- feedback

7.5 Mission Without Exploitation

Intrinsic motivation should not justify unpaid or underpaid labor.

7.6 Professional Norms

Codes, peer review, and public responsibility can reinforce intrinsic standards.

7.7 Balanced Design

Use external incentives to:

- Remove barriers
- fund public goods
- reward effort
- recognize contribution

without attempting to control every behavior.

8. Goodhart, Campbell, and Metric Gaming

8.1 The Core Problem

When a measure becomes consequential, actors optimize for the measure.

The connection between the measure and the goal can weaken.

8.2 Frontier Examples

- Benchmark optimization without generalization
- maximizing refusal rates at the expense of usefulness
- reducing reported incidents rather than incidents
- maximizing audit completion rather than control effectiveness
- publishing many standards rather than maintaining good ones
- rewarding vulnerability volume rather than severity or novelty
- ranking companies by disclosure length
- counting external reviewers rather than review independence

8.3 Multi-Metric Defense

Use several evidence types.

8.4 Hidden and Rotating Measures

Held-out and dynamic evaluation can reduce direct gaming.

8.5 Outcome Review

Connect metrics to real outcomes.

8.6 Qualitative Judgment

Do not eliminate expert judgment merely because it is harder to standardize.

8.7 Anti-Optimization Reserve

Some criteria should remain:

- Rotating
- sampled
- independently selected
- incident-driven

8.8 Metric Expiration

Retire measures that no longer discriminate or predict.

8.9 Reward Caps

Avoid unlimited reward for one metric.

8.10 Retrospective Gaming Review

Ask:

- What behavior changed?
 - What was neglected?
 - What became easier to hide?
 - Who benefited?
 - Did the metric preserve meaning?
-

9. Prestige as Institutional Currency

9.1 Why Prestige Matters

Prestige affects:

- Hiring
- funding
- access
- adoption
- authority
- media attention
- collaboration

- standards influence

9.2 Sources of Prestige

- Scientific contribution
- technical competence
- institutional affiliation
- model access
- public communication
- awards
- leadership
- successful predictions
- benchmark creation
- standards authorship

9.3 Legitimate Prestige

Prestige is useful when it helps identify:

- Reliable contributors
- competent evaluators
- durable institutions
- public-goods providers
- trusted stewards

9.4 Prestige Failure

Prestige becomes harmful when it:

- Replaces evidence
- protects institutions from criticism
- compounds access inequality
- rewards certainty
- creates celebrity experts
- suppresses junior dissent
- obscures conflicts
- becomes self-perpetuating

9.5 Prestige Portability

Recognition should identify the domain and contribution.

An expert in one domain should not receive automatic authority in another.

9.6 Prestige Expiration

Some designations should expire unless contribution continues.

9.7 Prestige Diversification

Recognize:

- Technical breakthroughs
- maintenance
- replication
- correction
- mentoring
- security
- standards service
- open tools
- public-interest contribution

9.8 Prestige Separation

Separate:

- Contribution recognition
- evaluator qualification
- certification authority
- governance authority

9.9 Prestige Audit

Institutions should examine:

- Who receives recognition?
- Which work is ignored?
- Does status predict competence?
- Are the same organizations repeatedly rewarded?
- Can newcomers gain recognition?

10. Recognition Architecture

Standards Body should build recognition around contribution classes.

10.1 Evidence Contribution

Recognition for:

- Valid evaluation
- important negative result
- replication
- benchmark correction
- uncertainty improvement

10.2 Infrastructure Contribution

Recognition for:

- Open tools
- task banks
- secure environments
- documentation
- reference assets
- long-term maintenance

10.3 Safety Improvement

Recognition for:

- Safeguard improvement
- incident reduction
- disclosure
- remediation
- independent verification

10.4 Institutional Contribution

Recognition for:

- Standards development
- governance
- interoperability
- evaluator training
- public consultation

10.5 Public-Interest Contribution

Recognition for:

- Accessibility
- small-actor support

- affected-party research
- transparency
- education

10.6 Corrective Contribution

Recognition for:

- Retraction
- corrected standard
- incident disclosure
- result withdrawal
- improved process

10.7 Mentorship and Community

Recognition for:

- Reviewer development
- open-source stewardship
- domain translation
- inclusive participation

10.8 Recognition Levels

Possible levels:

- Acknowledgment
- verified contribution
- distinguished contribution
- institutional stewardship

These should not imply accreditation.

10.9 Evidence Requirements

Every recognition should have:

- Criteria
- evidence
- reviewers
- conflicts
- decision
- expiry or permanence
- correction process

11. Awards and Honors

11.1 Purpose

Awards can direct attention toward neglected work.

11.2 Appropriate Award Categories

- Evaluation validity
- public infrastructure
- responsible disclosure
- replication
- safeguard innovation
- standards stewardship
- open-source evaluation
- interdisciplinary review
- early-career contribution
- institutional correction

11.3 Award Risks

- Popularity contests
- sponsor influence
- prestige concentration
- self-nomination bias
- publicity over substance
- winner-take-all dynamics

11.4 Selection

Use:

- Public criteria
- qualified judges
- conflict disclosure
- evidence review
- minority views
- public rationale

11.5 No Endorsement Spillover

Award language should state scope.

11.6 Team Credit

Recognize maintainers, reviewers, data contributors, and operational staff.

11.7 Post-Award Review

Serious error may require:

- Correction
- clarification
- withdrawal

11.8 Award Diversity

Avoid creating too many designations.

Recognition inflation reduces meaning.

12. Prizes and Challenge Competitions

12.1 Why Use Prizes

Prizes can attract:

- New participants
- unconventional approaches
- cross-disciplinary teams
- measurable progress
- public attention

NIST has used open innovation prize challenges, crowdsourcing, hackathons, and related incentive mechanisms for well-defined public-safety problems.[^nist-prizes]

12.2 Good Prize Problems

A prize is appropriate when:

- Objective is clear
- performance can be measured
- multiple approaches are possible
- participation can broaden
- solution has public value

12.3 Poor Prize Problems

Avoid when:

- Outcome is not measurable
- safety risk is high
- speed is rewarded over validity
- solution requires hidden institutional context
- winner-take-all structure discourages sharing

12.4 Prize Types

- Best performance
- milestone
- grand challenge
- open methods
- replication
- robustness
- safety
- infrastructure
- negative result
- community contribution

12.5 Prize Design

Specify:

- Goal
- eligibility
- data
- security
- evaluation
- intellectual property
- disclosure
- fairness
- conflicts
- award
- post-prize use

12.6 Milestone Prizes

Can reward partial progress and reduce winner-take-all risk.

12.7 Shared-Value Prizes

Reward several complementary contributions.

12.8 Open Infrastructure Condition

Some prizes can require:

- Open tools
- documentation
- reproducibility
- maintenance plan

12.9 Post-Competition Validation

Winning a competition does not establish production readiness.

12.10 Standards Body Use

Potential challenges:

- Dynamic benchmark renewal
 - held-out task generation
 - evaluator proficiency
 - safeguard stress testing
 - interoperability
 - model identity
 - incident taxonomy
-

13. Grants and Public-Goods Funding

13.1 Why Grants Matter

Markets may underfund:

- Benchmark maintenance
- negative results
- standards work
- replication
- open tools
- evaluator training
- public-interest analysis
- incident databases

13.2 Grant Categories

- Research
- infrastructure
- maintenance
- capacity
- early-career
- domain translation
- open-source
- international participation
- public-interest evaluation

13.3 Selection Risks

- Prestige bias
- institutional concentration
- proposal-writing advantage
- fashionable topics
- reviewer networks

13.4 Grant Design

Use:

- Clear objectives
- diverse panels
- conflict controls
- milestone review
- flexible methods
- open outputs where safe
- maintenance funding
- impact evaluation

13.5 Maintenance Grants

Infrastructure requires ongoing support.

Do not fund creation without stewardship.

13.6 Microgrants

Can broaden participation and support independent contributors.

13.7 Fellowship Programs

Can build:

- Evaluation science
- domain expertise
- standards competence
- auditor capacity
- public-interest leadership

13.8 Funding Independence

Diversify funders to reduce agenda control.

13.9 Failed Research

Allow publication of well-conducted negative or unsuccessful work.

14. Bounties and Responsible Disclosure

14.1 Bounty Applications

Bounties can reward discovery of:

- Safeguard bypasses
- benchmark flaws
- evaluation leakage
- model vulnerabilities
- security weaknesses
- documentation errors
- standard ambiguities

14.2 Vulnerability-Disclosure Lessons

NIST SP 800-216 provides guidance for receiving, assessing, managing, and communicating vulnerability disclosures within federal systems.[^nist-vdp]

Frontier AI disclosure programs can learn from:

- Defined scope
- safe harbor
- acknowledgment
- triage
- remediation
- communication

- timelines

14.3 Bounty Risks

- Quantity over quality
- duplicate findings
- unsafe disclosure
- adversarial relationships
- underpayment
- exploitation of unpaid researchers
- reward for creating the problem

14.4 Severity-Based Reward

Reward should consider:

- Impact
- novelty
- reproducibility
- actionability
- safe handling
- quality of evidence

14.5 Nonfinancial Recognition

Offer:

- Public credit
- private credit
- access
- collaboration
- contribution record

according to researcher preference.

14.6 Safe Harbor

Good-faith researchers should know:

- Permitted activity
- prohibited activity
- reporting channel
- legal treatment
- disclosure process

14.7 Standards Bounty

Standards Body could reward:

- Contradictions
- missing definitions
- invalid claims
- implementation conflicts
- unanticipated gaming
- international incompatibility

14.8 Corrective Loop

Bounty findings should update:

- Protocol
 - failure database
 - version history
 - evaluator guidance
 - research agenda
-

15. Publication and Academic Incentives

15.1 Current Distortions

Academic systems often reward:

- Novelty
- positive results
- citations
- first authorship
- venue prestige

more than:

- Replication
- maintenance
- datasets
- standards
- negative findings
- long-term validation

15.2 Evaluation Science Needs Different Credit

Recognize:

- Benchmark stewardship
- task validation
- environment engineering
- scoring audits
- contamination analysis
- replication
- longitudinal studies

15.3 Contributor Taxonomy

Use structured contributor roles.

Possible roles:

- Conceptualization
- methodology
- software
- validation
- data curation
- investigation
- security
- governance
- writing
- maintenance
- project administration

15.4 Registered Reports

Precommitted research designs can reduce selective publication.

15.5 Negative Results

Publication venues and funders should value valid negative evidence.

15.6 Replication Credit

Independent replication should carry professional value.

15.7 Maintenance Citations

Tools and datasets should have stable identifiers and citation guidance.

15.8 Reviewer Credit

Peer and standards review should receive documented professional credit without compromising confidentiality.

15.9 Access Independence

Research prestige should not depend excessively on exclusive frontier-model access.

16. Developer Incentives

16.1 Desired Behaviors

Developers should benefit from:

- Strong evaluation
- independent review
- safeguard improvement
- incident disclosure
- standards participation
- secure external access
- correction
- public-goods contribution

16.2 Market Incentives

Possible benefits:

- Procurement eligibility
- insurer recognition
- investor confidence
- trusted-user programs
- lower duplicated assurance
- faster partner due diligence
- certification within narrow scope

16.3 Reputational Incentives

Recognition for:

- Transparent evidence
- external review
- remediation
- responsible restraint

- reproducibility
- support for evaluator access

16.4 Regulatory Incentives

Possible mechanisms:

- Safe harbor
- reduced reporting duplication
- recognition of credible standards
- expedited review
- lower inspection frequency

These should not remove responsibility.

16.5 Access Incentives

Developers providing high-quality external access may receive:

- Recognition
- research partnerships
- procurement preference
- reduced uncertainty

16.6 Incident Incentives

Design so that:

- Concealment is costly
- timely disclosure receives credit
- negligence still has consequence
- remediation is visible

16.7 Framework Commitments

OpenAI, Anthropic, and Google DeepMind have published voluntary capability-linked safety frameworks. These can create internal and reputational incentives, especially when commitments are specific, externally reviewable, and difficult to revise opportunistically.[^openai-pf][^anthropic-rsp][^deepmind-fsf]

16.8 Avoiding Safety Marketing

Recognition should be tied to evidence, not language.

17. Evaluator and Auditor Incentives

17.1 Desired Behaviors

Evaluators should benefit from:

- Accurate findings
- independence
- secure operations
- correction
- methodological innovation
- proficiency
- transparent limitations
- public contribution

17.2 Harmful Incentives

- Favorable client outcomes
- report volume
- certificate volume
- exclusive methods
- sensational findings
- access preservation
- government alignment

17.3 Payment Design

Use:

- Fixed fees
- milestone payments
- no result-dependent compensation
- pooled funding
- transparent client concentration

17.4 Prestige Design

Recognize:

- Error correction
- replication
- robust dissent
- methodological transparency
- secure handling
- cross-evaluator consistency

17.5 Accreditation Incentives

Accreditation can create:

- Market access
- credibility
- procurement eligibility

It can also encourage minimal compliance.

17.6 Report Quality

Purchasers should reward:

- Evidence
- clarity
- uncertainty
- actionability

not length.

17.7 Independence Protection

Provide alternative funding and access so unfavorable conclusions do not end an evaluator's viability.

17.8 Failure Consequences

Material evaluator failure should affect:

- Scope
- recognition
- registry status
- professional standing

17.9 Correction Credit

An evaluator that identifies and corrects its own error promptly should be distinguished from one that conceals it.

18. Standards-Organization Incentives

18.1 Desired Behaviors

Standards organizations should benefit from:

- Validity
- adoption
- maintenance
- interoperability
- participation
- retirement of obsolete standards
- public value

18.2 Harmful Incentives

- Document volume
- membership revenue
- sponsor influence
- permanent committees
- slow retirement
- prestige through exclusivity
- paywall dependence

18.3 Maintenance Funding

Standards require:

- Editors
- evidence review
- testing
- translations
- public consultation
- versioning

18.4 Adoption Versus Quality

A widely adopted weak standard can be more harmful than a less adopted strong one.

18.5 Participation Incentives

Support smaller actors through:

- Travel funding
- remote access

- compensation
- grants
- open drafts
- transparent contribution credit

ISO research on standards and innovation notes the importance of supporting research institutions and small and medium-sized organizations in standardization participation.[^iso-innovation]

18.6 Retirement Incentive

Organizations should receive legitimacy for withdrawing outdated standards.

18.7 Public Access

Where standards support law or public-interest requirements, access should be considered part of legitimacy.

18.8 Institutional Scorecard

Standards bodies should be evaluated for outcomes, not only publication.

19. Procurement Incentives

19.1 Why Procurement Matters

Purchasers can create immediate demand for:

- Evaluation
- documentation
- independent assurance
- incident reporting
- monitoring
- interoperability

19.2 Public Procurement

Government purchasing can shape markets.

NIST's AI RMF identifies acquisition and procurement actors as part of the AI lifecycle and risk-management ecosystem.[^nist-rmf]

19.3 Procurement Criteria

Reward:

- Evidence quality
- qualified evaluation
- safeguards
- incident readiness
- system identity
- update control
- transparency

19.4 Avoid Lowest Price Only

Low price can hide:

- Weak evaluation
- externalized risk
- poor monitoring
- vendor lock-in

19.5 Outcome-Based Procurement

Specify outcomes while accepting equivalent methods.

19.6 Small-Business Access

Avoid requirements that only incumbent suppliers can satisfy.

19.7 Procurement Preference

Can reward:

- Standards adoption
- independent assurance
- open interfaces
- responsible disclosure
- evaluator access

19.8 Risks

- Checkbox compliance
- proprietary certification
- fragmented buyer requirements
- slow procurement
- incumbent advantage

19.9 Contractual Update

Require re-evaluation after material system change.

20. Insurance and Financial Incentives

20.1 Insurance Role

Insurance can translate risk evidence into:

- Premium
- deductible
- coverage
- exclusions
- control requirements

20.2 Potential Benefits

- Market discipline
- loss data
- control incentives
- independent review demand

20.3 Limitations

- Limited claims history
- correlated risk
- systemic events
- model opacity
- rapid change
- weak actuarial data

20.4 Evidence Requirements

Insurers should distinguish:

- Certification
- capability evaluation
- safeguard testing
- organizational audit
- deployment monitoring

20.5 Moral Hazard

Coverage should not reduce care.

20.6 Investor Incentives

Investors may reward:

- Governance
- evaluation maturity
- security
- transparent risk
- durable market access

20.7 Financial Disclosure

Avoid unverified claims.

20.8 Long-Term Capital

Patient funding can support:

- Safety research
 - infrastructure
 - standards
 - evaluator capacity
-

21. Public Reporting and Reputation

21.1 Transparency as Incentive

Reporting can create reputational pressure and peer comparison.

21.2 Hiroshima AI Process

The HAIP Reporting Framework provides a common voluntary structure for organizations to disclose advanced AI governance and risk-management practices.^[^haip]

21.3 Reporting Risks

- Disclosure length as proxy for quality
- selective answers
- unverifiable claims
- branding

- confidentiality abuse
- comparability failure

21.4 Evidence-Backed Reporting

Report:

- Practice
- evidence
- version
- scope
- independent review
- incidents
- limitations
- changes

21.5 Correction

Maintain visible corrections.

21.6 Reputation Recovery

An institution should be able to recover through:

- Disclosure
- remediation
- external verification
- sustained improvement

21.7 No Reputation Laundering

One award, partnership, or certificate should not erase contradictory evidence.

21.8 Report Quality Recognition

Recognize candor and specificity, not only favorable content.

22. Rankings and Scorecards

22.1 Benefits

Rankings can:

- Simplify information

- create competition
- attract attention
- identify leaders
- encourage adoption

22.2 Risks

- Goodhart effects
- false precision
- missing context
- reputational harm
- gaming
- winner-take-all status
- methodology disputes
- pressure to simplify uncertainty

22.3 When Rankings Are Appropriate

Only when:

- Construct is sufficiently stable
- data are comparable
- uncertainty is reported
- methodology is public
- conflicts are controlled
- appeal exists
- result expires

22.4 Alternatives

- Profiles
- tiers
- badges by contribution
- multidimensional scorecards
- evidence summaries
- maturity levels

22.5 No Universal Safety Ranking

Organizations and systems vary across domains.

22.6 Ranking Governance

Include:

- Method review
- anti-gaming
- correction
- versioning
- conflict disclosure
- appeals
- retirement

22.7 Standards Body Position

Prefer evidence profiles and contribution recognition over a single ordinal ranking.

23. Certification and Prestige

23.1 Certification as Market Signal

Certification can reduce information asymmetry.

23.2 Prestige Spillover

A narrow certificate can create broad reputational benefit.

23.3 Claim Controls

Require:

- System
- version
- scope
- scheme
- evaluator
- date
- expiry
- limitations

23.4 Certification Inflation

Too many schemes reduce trust.

23.5 Recognition Versus Certification

Standards Body awards or contributor recognition should not be confused with conformity certification.

23.6 Withdrawal

Certificate and associated prestige should be withdrawable.

23.7 Continuous Performance

Long-term prestige should depend on ongoing conduct.

24. Access as Incentive and Leverage

24.1 Model Access

Frontier-model access is valuable.

It can reward:

- Qualified evaluation
- responsible security
- public-interest research
- reproducibility

24.2 Access Capture

Developers may favor:

- Friendly researchers
- prestigious institutions
- supportive commentators
- repeat partners

24.3 Access Governance

Use:

- Public criteria
- independent selection
- conflict disclosure
- security requirements
- appeals
- rotation

24.4 Data and Compute Access

Grants of compute and data can expand participation.

24.5 Standards Participation

Committee seats and governance access are incentives.

They should be earned through contribution and balanced representation.

24.6 Access Revocation

Grounds should be clear.

Evidence-based criticism should not be a ground.

24.7 Access Portability

Qualified status should support access across more than one developer where possible.

24.8 Public Access Report

Organizations should report how access decisions are made.

25. Career and Professional Incentives

25.1 Professional Pathways

Frontier evaluation needs recognized careers in:

- Evaluation science
- audit
- standards
- model security
- domain assessment
- incident analysis
- benchmark stewardship
- public-interest technology

25.2 Credentials

Credentials can help but risk:

- Exclusion

- credential inflation
- outdated knowledge
- institutional monopoly

25.3 Portfolio-Based Recognition

Use evidence of:

- Methods
- evaluations
- open tools
- reviews
- incident work
- proficiency
- maintenance

25.4 Continuing Competence

Recognition should require updated practice.

25.5 Promotion Criteria

Organizations should reward:

- Identifying problems
- stopping unsafe work
- correcting claims
- improving controls
- mentoring
- documentation

25.6 Internal Dissent

Employees should not be punished for responsible, evidence-based safety concerns.

25.7 Whistleblower Pathways

Confidential channels and anti-retaliation protections are essential.

25.8 Standards Service

Committee and editorial work should receive professional credit.

25.9 Invisible Labor

Recognize:

- Operations
 - security
 - project management
 - data curation
 - moderation
 - documentation
-

26. Open-Source and Community Incentives

26.1 Public-Goods Nature

Open communities produce:

- Models
- evaluation tools
- datasets
- benchmarks
- documentation
- reproductions
- audits

26.2 Sustainability Problem

Maintainers face:

- Burnout
- uncompensated labor
- support burden
- security responsibility
- corporate free riding

26.3 Funding

Use:

- Grants
- sponsorship
- maintenance contracts
- public procurement
- bounty programs
- fellowships

26.4 Credit

Recognize:

- Contributors
- maintainers
- reviewers
- security responders
- translators
- educators

26.5 Governance Participation

Meaningful contributors should have paths into standards and policy discussions.

26.6 Security Incentives

Open projects need:

- Disclosure channels
- incident support
- signed releases
- model provenance
- reproducible evaluation

26.7 Avoid Corporate Capture

Sponsorship should not purchase community control.

26.8 Avoid Ideological Recognition

Do not reward openness or closedness as identities detached from evidence and context.

26.9 Contribution Portability

A contributor's verified work should remain legible across projects and institutions.

27. Incident Disclosure and Corrective Credit

27.1 The Disclosure Dilemma

Disclosure can create:

- Reputational cost

- liability
- regulatory scrutiny
- competitor advantage

Concealment can create greater public risk.

27.2 Corrective Credit Model

Give credit for:

- Timely discovery
- prompt containment
- complete reporting
- independent review
- remediation
- prevention
- public learning

27.3 No Failure Immunity

Corrective credit should not erase:

- Negligence
- repeated failure
- concealment
- preventable harm
- bad faith

27.4 Incident Quality

Assess:

- Detection time
- notification time
- evidence
- impact
- root cause
- corrective action
- recurrence
- disclosure

27.5 Near Misses

Recognize high-quality near-miss reporting.

27.6 Shared Learning

De-identified incident databases can create public value.

27.7 Procurement and Insurance

Responsible disclosure can receive favorable treatment.

27.8 Prestige for Correction

Institutional culture should admire accurate correction more than confident persistence.

28. Sanctions and Negative Incentives

28.1 Purpose

Negative incentives deter:

- Concealment
- fraud
- scope overreach
- security failure
- false certification
- repeated nonconformity
- retaliation

28.2 Sanction Types

- Correction
- warning
- enhanced monitoring
- loss of award
- registry notice
- scope suspension
- certificate withdrawal
- loss of access
- contract remedy
- financial penalty
- legal liability

28.3 Proportionality

Consider:

- Intent
- severity
- preventability
- recurrence
- cooperation
- remediation
- impact

28.4 Chilling Effects

Excessive penalties can reduce:

- Disclosure
- experimentation
- participation
- honest uncertainty

28.5 Restoration

Provide a path to recover status through verified improvement.

28.6 Public Notice

Material sanctions should be visible where others rely on the status.

28.7 No Social-Media Enforcement

Public attention should not substitute for evidence and due process.

29. Incentive Compatibility in Standards Development

29.1 Participation Incentives

Contributors need reasons to invest time.

29.2 Contribution Credit

Standards should list meaningful contributions.

29.3 Compensation

Pay:

- Editors
- technical leads
- small-actor representatives
- public-interest contributors
- domain experts

where feasible.

29.4 Conflict Risk

Payment should not purchase conclusions.

29.5 Open Drafts

Public drafts can broaden participation.

29.6 Comment Quality

Recognize substantive comments, not volume.

29.7 Implementation Feedback

Organizations testing standards should receive:

- Credit
- early learning
- influence within conflict controls

29.8 Retirement Incentive

Committees should not depend on a standard remaining active for status.

29.9 Standards Body Contributor System

A future contributor framework should record:

- Role
- evidence
- conflicts
- work
- review
- maintenance

- correction
-

30. Designing Incentive Programs

Every program should begin with a theory of change.

30.1 Objective

What behavior or outcome should change?

30.2 Actor

Who controls the behavior?

30.3 Baseline

What happens without the incentive?

30.4 Mechanism

Why would the incentive change behavior?

30.5 Signal

How is performance observed?

30.6 Verification

Who verifies the signal?

30.7 Reward

What is offered?

30.8 Timing

When is the reward granted?

30.9 Duration

Is the reward one-time or ongoing?

30.10 Gaming

How could actors maximize the reward without achieving the objective?

30.11 Distribution

Who can participate?

30.12 Crowding Out

Could the mechanism weaken intrinsic motivation?

30.13 Market Effects

Could it increase concentration or dependency?

30.14 Correction

Can awards or status be revised?

30.15 Evaluation

How will program impact be measured?

30.16 Sunset

When should the program end?

31. Incentive Portfolios by Maturity Stage

Stage 0: Research

Use:

- Grants
- fellowships
- open challenges
- publication credit
- model access

Avoid:

- Premature certification
- public rankings

Stage 1: Recommended Practice

Use:

- Recognition
- implementation support
- pilot grants
- contributor credit

Stage 2: Voluntary Framework

Use:

- Reporting
- peer recognition
- procurement signals
- access incentives

Stage 3: Standard

Use:

- Training
- certification readiness
- procurement
- insurer interest
- standards participation credit

Stage 4: Independent Assurance

Use:

- Accredited evaluator market
- public registry
- purchaser reliance
- professional status

Stage 5: Formal Requirement

Use:

- Enforcement
- safe harbor
- remediation credit
- public accountability

Stage 6: Continuous Assurance

Use:

- Ongoing market access
 - reduced duplication
 - adaptive oversight
 - sustained recognition
-

32. Anti-Gaming Architecture

32.1 Pre-Mortem

Before launch, ask how the incentive will be gamed.

32.2 Multiple Measures

Combine:

- Quantitative
- qualitative
- held-out
- operational
- incident
- peer evidence

32.3 Independent Verification

Do not rely solely on self-report.

32.4 Random Audit

Use risk-based and random review.

32.5 Metric Rotation

Rotate measures where gaming is likely.

32.6 Reward Delay

Some rewards should depend on sustained performance.

32.7 Clawback

Allow withdrawal after:

- Fraud
- material error
- concealed conflict
- later invalidation

32.8 Counter-Metric

Monitor undesirable substitution.

Example:

If rewarding incident reporting, also monitor incident severity and preventability.

32.9 Adversarial Participation

Include critics in program review.

32.10 Public Rationale

Explain why an award or designation was granted.

32.11 Appeal

Allow evidence-based challenge.

32.12 Program Audit

Evaluate the incentive system itself.

33. Distribution, Equity, and Access

33.1 Existing Advantage

Prestigious institutions have:

- Networks
- proposal capacity
- model access
- media attention
- funding

- committee access

33.2 Corrective Design

Use:

- Open nominations
- microgrants
- travel support
- remote participation
- paid committee service
- blinded review
- regional programs
- mentorship

33.3 Blinding

Blinded review can reduce status bias in some contexts.

It may be impractical when contribution identity is part of the evidence.

33.4 Geographic Diversity

Support regions with limited frontier access.

33.5 Language

Provide translation and multilingual evaluation recognition.

33.6 Early-Career Contributors

Avoid requiring status to earn status.

33.7 Independent Researchers

Create pathways without institutional affiliation.

33.8 Accessibility

Programs should accommodate disability and caregiving constraints.

33.9 Distribution Audit

Report who receives:

- Funding
 - awards
 - access
 - governance roles
 - procurement
 - recognition
-

34. International Incentive Alignment

34.1 Global Competition

National competition can encourage:

- Investment
- talent
- innovation

It can also discourage:

- Disclosure
- restraint
- shared standards
- external access

34.2 Shared Recognition

International recognition can reward:

- Interoperable evaluation
- incident reporting
- evaluator competence
- open public goods
- mutual recognition

34.3 Procurement Across Borders

Shared evidence can increase market reward for credible assurance.

34.4 Avoiding Prestige Nationalism

Technical recognition should not become geopolitical branding.

34.5 Developing-Economy Participation

Provide:

- Funding
- training
- evaluator capacity
- standards access
- infrastructure

34.6 International Awards

Use diverse panels and transparent criteria.

34.7 Voluntary Reporting

The HAIP framework demonstrates how shared reporting can create reputational incentives across organizations and jurisdictions.[^haip]

34.8 Mutual Recognition

Foundation 8 develops global interoperability fully.

35. Governance of Incentive Systems

35.1 Separation of Roles

Separate where feasible:

- Program sponsor
- criteria designer
- applicant
- evaluator
- decision-maker
- funder
- appeals body

35.2 Conflicts

Disclose:

- Funding
- affiliations
- competitor relationships

- prior collaboration
- ideological commitments
- financial interest

35.3 Decision Records

Record:

- Criteria
- evidence
- deliberation
- decision
- dissent
- conditions

35.4 Public Oversight

High-profile awards and rankings should permit scrutiny.

35.5 Appeals

Applicants and affected parties need defined challenge mechanisms.

35.6 Confidentiality

Protect:

- Sensitive findings
- personal data
- security
- reviewer deliberation

without hiding criteria.

35.7 Term Limits

Recognition-panel membership should rotate.

35.8 Sponsor Limits

Sponsors should not control winners.

35.9 Program Review

Review:

- Outcomes
- gaming
- distribution
- costs
- reputation
- capture

35.10 Retirement

End programs that no longer improve behavior.

36. Maturity Model

Level 0: Accidental Incentives

Characteristics:

- Rewards emerge informally
- attention dominates
- no theory of change
- no gaming analysis

Level 1: Explicit Recognition

Characteristics:

- Defined criteria
- contribution credit
- conflict disclosure
- narrow awards

Level 2: Verified Incentive Programs

Characteristics:

- Evidence
- independent review
- correction
- distribution analysis
- program metrics

Level 3: Integrated Market and Professional Incentives

Characteristics:

- Procurement
- grants
- career credit
- access
- insurance
- assurance
- public reporting

Level 4: Adaptive International Incentive Ecosystem

Characteristics:

- Cross-border recognition
 - public-goods funding
 - continuous impact review
 - anti-gaming
 - equitable access
 - institutional correction
 - interoperable evidence
-

37. Implementation Pathway

Phase 1: Incentive Audit

Map current rewards and penalties across the ecosystem.

Phase 2: Identify Underrewarded Contributions

Examples:

- Maintenance
- replication
- disclosure
- correction
- open infrastructure
- standards service

Phase 3: Define Desired Behaviors

Tie each program to an outcome.

Phase 4: Select Mechanisms

Choose:

- Grant
- prize
- recognition
- access
- procurement
- governance
- sanction

Phase 5: Conduct Gaming Analysis

Use adversarial review.

Phase 6: Pilot Small

Limit scope and duration.

Phase 7: Verify

Use independent assessment.

Phase 8: Publish Evidence

Explain criteria, winners, limitations, and conflicts.

Phase 9: Measure Behavior

Assess whether incentives changed the intended outcome.

Phase 10: Adjust

Modify reward, metrics, access, and eligibility.

Phase 11: Scale

Expand only after evidence.

Phase 12: Sunset or Institutionalize

End weak programs and maintain effective ones.

38. Proposed Standards Body Pilot

38.1 Pilot Name

Frontier Evaluation Public-Goods and Integrity Recognition Program

38.2 Purpose

Test whether structured recognition and modest funding can increase high-value contributions that existing academic and commercial incentives underreward.

38.3 Contribution Categories

Evaluation Integrity

- Benchmark correction
- contamination discovery
- scoring validation
- replication

Public Infrastructure

- Open evaluation tools
- task environments
- reference assets
- documentation
- maintenance

Responsible Disclosure

- Safeguard bypass
- evaluator security flaw
- protocol vulnerability
- standard ambiguity

Institutional Integrity

- Transparent correction
- incident learning
- conflict disclosure
- minority report

Open and Public-Interest Contribution

- Small-actor tooling
- multilingual evaluation
- accessibility

- open-source stewardship

38.4 Recognition Types

- Verified contribution record
- small grant
- maintenance award
- model-access nomination
- standards working-group invitation
- annual distinguished contribution

38.5 Selection

Use:

- Open nominations
- evidence package
- qualified reviewers
- conflict register
- public rationale
- appeal

38.6 Anti-Gaming

- No award based only on count
- severity and evidence required
- duplicate screening
- delayed verification
- clawback
- sponsor separation
- annual program audit

38.7 Distribution

Reserve pathways for:

- Independent researchers
- early-career contributors
- open-source maintainers
- underrepresented regions
- small evaluator organizations

38.8 Corrective Credit

Create a category recognizing organizations that:

- Disclose material error
- withdraw invalid claims
- remediate
- submit to external verification

38.9 Outputs

- Contribution registry
- annual evidence report
- incentive impact analysis
- distribution report
- program corrections
- future funding recommendations

38.10 Success Criteria

The pilot succeeds if it:

- Produces verified public goods
 - increases maintenance and replication
 - supports contributors outside elite institutions
 - rewards correction without excusing negligence
 - avoids popularity-based selection
 - leads to measurable reuse or risk reduction
 - remains financially and institutionally sustainable
-

39. Metrics for Evaluating Incentive Programs

39.1 Behavior Change

- New participation
- increased disclosure
- increased replication
- maintenance activity
- evaluator improvement
- standards adoption

39.2 Output Quality

- Validity

- reuse
- reproducibility
- security
- decision value
- longevity

39.3 Gaming

- Duplicate claims
- low-quality submissions
- metric manipulation
- strategic timing
- false disclosure
- prestige signaling

39.4 Distribution

- Organization size
- geography
- career stage
- affiliation
- open-source participation
- demographic access where appropriate

39.5 Crowding Out

- Changes in intrinsic participation
- unpaid contribution
- community trust
- topic diversity

39.6 Market Effects

- Concentration
- supplier behavior
- evaluator pricing
- procurement access
- insurance response

39.7 Reputation

- Public understanding
- recognition meaning
- status inflation

- trust
- correction behavior

39.8 Cost

- Administration
- verification
- reward
- applicant burden
- reviewer burden

39.9 Institutional Integrity

- Conflicts
- appeals
- corrections
- clawbacks
- sponsor influence
- decision quality

39.10 Long-Term Impact

- Risk reduction
- safeguard improvement
- public infrastructure
- talent retention
- standards quality
- international reuse

40. Failure Modes and Safeguards

40.1 Rewarding Visibility

Failure: Publicly visible work receives more credit than durable infrastructure.

Safeguard: Contribution categories and evidence-based review.

40.2 Metric Gaming

Failure: Participants optimize the reward measure.

Safeguard: Multiple measures, rotation, verification, outcome review.

40.3 Prestige Capture

Failure: Established institutions repeatedly select one another.

Safeguard: open nomination, rotation, blinded stages, distribution audit.

40.4 Sponsor Capture

Failure: Funders influence awards or standards.

Safeguard: sponsor separation and conflict disclosure.

40.5 Recognition Inflation

Failure: Too many badges reduce meaning.

Safeguard: limited designations and clear evidence thresholds.

40.6 Award Spillover

Failure: A narrow contribution award becomes a broad safety endorsement.

Safeguard: scope-specific language.

40.7 Quantity Over Quality

Failure: Counts dominate severity, validity, or impact.

Safeguard: weighted evidence and qualitative review.

40.8 Crowding Out

Failure: External rewards weaken mission or cooperation.

Safeguard: autonomy-supportive design and community review.

40.9 Winner-Take-All

Failure: One winner captures status while many valuable contributors receive nothing.

Safeguard: milestone, category, and shared recognition.

40.10 Bounty Abuse

Failure: Participants flood programs or create unsafe findings.

Safeguard: scope, triage, safe harbor, severity, disclosure controls.

40.11 Corrective Credit Abuse

Failure: Organizations manufacture or repeatedly cause problems to receive recognition for fixing them.

Safeguard: negligence analysis, recurrence tracking, no immunity.

40.12 Access Patronage

Failure: Model access rewards loyalty rather than competence.

Safeguard: public criteria and independent selection.

40.13 Procurement Capture

Failure: Recognition becomes a required proprietary badge.

Safeguard: equivalent evidence and competition review.

40.14 Career Exclusion

Failure: Credentials become barriers to new experts.

Safeguard: portfolio pathways and supervised entry.

40.15 Open-Source Exploitation

Failure: Companies benefit from unpaid maintainers without support.

Safeguard: maintenance funding and contribution credit.

40.16 Incident Suppression

Failure: Sanctions discourage disclosure.

Safeguard: corrective credit, safe channels, proportionate response.

40.17 Safety Marketing

Failure: Incentive programs become public-relations tools.

Safeguard: independent governance and evidence.

40.18 Institutional Permanence

Failure: Program continues because staff or sponsors depend on it.

Safeguard: sunset and external impact review.

40.19 Geographic Concentration

Failure: Recognition remains concentrated in a few regions.

Safeguard: regional access, translation, funding, diverse panels.

40.20 Prestige Without Accountability

Failure: High-status contributors resist correction.

Safeguard: visible corrections, expiry, appeal, withdrawal.

41. Serious Objections

Objection 1: Safety Should Be a Duty, Not a Rewarded Behavior

Some practices should be basic obligations.

Response:

Duties and incentives can coexist. Rewards are especially useful for work beyond minimum obligations and for public goods.

Residual concern:

Recognition can imply that ordinary compliance is exceptional.

Objection 2: Financial Incentives Corrupt Scientific Work

They can.

Response:

Use fixed funding, independent review, multiple incentives, and transparency.

Residual concern:

No design fully removes financial influence.

Objection 3: Prestige Is Inherently Elitist

Prestige can reinforce hierarchy.

Response:

Make recognition evidence-based, plural, scoped, and accessible.

Residual concern:

Status systems tend to accumulate.

Objection 4: Rankings Create Competition and Improvement

Sometimes.

Response:

Use rankings only where constructs and data support them.

Residual concern:

Even valid rankings can narrow behavior.

Objection 5: Market Incentives Are Enough

Markets often underprovide public goods and fail to price systemic or unobservable risk.

Objection 6: Regulation Is a Better Incentive

Law can create minimums.

It may not reward:

- Innovation
- public goods
- correction
- research
- voluntary leadership

A mixed system is stronger.

Objection 7: Corrective Credit Rewards Failure

Response:

Credit should reward disclosure and remediation while preserving consequences for negligence and harm.

Objection 8: Bounties Encourage Attack

They can increase attention to vulnerabilities.

Response:

Use bounded scope, safe harbor, controlled disclosure, and triage.

Objection 9: Awards Are Mostly Symbolic

Symbolic rewards can matter where prestige, career, and access are important.

They should not replace funding or institutional reform.

Objection 10: Procurement Preferences Entrench Standards

Correct.

Response:

Allow equivalent evidence, version review, and competition analysis.

Objection 11: Open-Source Communities Resist Formal Recognition

Some do.

Participation should remain voluntary and avoid centralizing community legitimacy.

Objection 12: Incentive Design Is Too Contextual for Standards Body

Context matters.

Standards Body can still define principles, failure modes, and evaluation requirements for incentive systems.

42. Evidence Gaps

42.1 Developer Behavior

Which incentives most effectively increase rigorous external evaluation and disclosure?

42.2 Evaluator Independence

How do payment, access, accreditation, and prestige affect findings?

42.3 Corrective Credit

Can institutions reward disclosure without creating moral hazard?

42.4 Prestige

Which recognition systems predict real competence and contribution?

42.5 Public Reporting

Does reporting create substantive change or reputational compliance?

42.6 Procurement

Do assurance preferences improve outcomes without excessive concentration?

42.7 Insurance

Can underwriters create useful AI risk incentives with limited loss data?

42.8 Academic Credit

Which reforms increase replication, maintenance, and negative results?

42.9 Prizes

When do challenge competitions produce durable evaluation infrastructure?

42.10 Bounties

Which reward structures produce high-quality AI vulnerability disclosure?

42.11 Crowding Out

When do external rewards reduce intrinsic safety motivation?

42.12 Open-Source Sustainability

Which funding models preserve community autonomy and maintenance?

42.13 Distribution

Which interventions broaden participation without weakening competence?

42.14 International Prestige

How do national and institutional status incentives affect cooperation?

42.15 Long-Term Impact

Do incentive programs produce sustained behavior after rewards end?

43. Research Agenda

Priority 1: Ecosystem Incentive Mapping

Map rewards, penalties, dependencies, and information flows.

Priority 2: Developer Disclosure

Test safe-harbor, procurement, and reputational mechanisms.

Priority 3: Evaluator Incentives

Study client concentration, access dependence, payment, and correction.

Priority 4: Prestige Measurement

Compare reputation with actual competence and impact.

Priority 5: Public-Goods Funding

Pilot maintenance grants, microgrants, and shared infrastructure.

Priority 6: Corrective Credit

Design and test credit for disclosure and remediation.

Priority 7: Bounties

Develop AI-specific vulnerability and evaluation-integrity programs.

Priority 8: Academic Reform

Test contributor credit, registered reports, replication grants, and maintenance citations.

Priority 9: Procurement

Measure effects on supplier behavior and market concentration.

Priority 10: Insurance

Develop evidence requirements and monitor moral hazard.

Priority 11: Access Governance

Study independent allocation of model, data, and compute access.

Priority 12: Anti-Gaming

Create adversarial review methods for incentive programs.

Priority 13: Distribution

Measure who receives funding, prestige, access, and governance roles.

Priority 14: International Alignment

Study shared recognition and cross-border public-goods funding.

Priority 15: Program Retirement

Develop criteria for ending ineffective incentive systems.

44. Near-Term Experiments

Experiment 1: Contribution Recognition

Compare expert review with popularity-based voting for evaluation contributions.

Experiment 2: Corrective Credit

Test how organizations respond to recognition for transparent correction.

Experiment 3: Maintenance Grant

Fund benchmark maintenance and measure reliability, reuse, and contributor retention.

Experiment 4: Replication Prize

Reward independent replication rather than novel results.

Experiment 5: Standards Bounty

Offer rewards for identifying contradictions, ambiguity, and gaming paths.

Experiment 6: Access Allocation

Compare developer-selected and independently selected external researchers.

Experiment 7: Procurement Preference

Test whether credible assurance changes supplier investment.

Experiment 8: Recognition Expiry

Compare permanent and renewable contribution designations.

Experiment 9: Multi-Metric Scorecard

Compare behavior under single ranking and multidimensional profile.

Experiment 10: Open-Source Sustainability

Pilot maintenance contracts for critical evaluation infrastructure.

Experiment 11: Geographic Access

Fund contributors from underrepresented regions and measure program quality.

Experiment 12: Program Red Team

Have independent reviewers attempt to game a proposed award or bounty.

45. Implications for Future Standards

A future standard or institutional policy for incentive programs could require:

45.1 Objective

Defined behavior or public outcome.

45.2 Actor Model

Who can act and what motivates them.

45.3 Baseline

Behavior without intervention.

45.4 Mechanism

How the incentive is expected to work.

45.5 Eligibility

Clear, fair, and accessible criteria.

45.6 Evidence

Proof required for reward or status.

45.7 Verification

Independent assessment and conflict control.

45.8 Reward

Financial, reputational, access, governance, or market benefit.

45.9 Anti-Gaming

Threat model, counter-metrics, audit, rotation, and clawback.

45.10 Distribution

Small-actor, open-source, geographic, and career-stage access.

45.11 Crowding Out

Assessment of intrinsic and community effects.

45.12 Correction

Appeal, correction, withdrawal, and restoration.

45.13 Transparency

Criteria, decisions, sponsors, conflicts, and rationale.

45.14 Impact Evaluation

Behavior, quality, burden, market effects, and long-term outcomes.

45.15 Sunset

Expiry or renewal based on evidence.

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

46. Relationship to the Other Foundations

Foundation 1: Dynamic Evaluation Protocols

Incentives should reward evaluation validity and maintenance rather than fixed benchmark scores.

Foundation 2: Held-Out Evaluations

Access, security, and disclosure incentives determine whether protected evidence remains credible.

Foundation 3: High-Stakes Capability Evaluation

Developers need meaningful reasons to test, disclose, and mitigate consequential capabilities.

Foundation 4: Independent Expert Review

Reviewer independence depends partly on funding, access, career, and prestige incentives.

Foundation 5: Third-Party Auditor Ecosystem

Evaluator markets require incentives for competence, impartiality, proficiency, and correction.

Foundation 6: Progressive Standards and Requirements

Voluntary stages rely heavily on recognition and market incentives before formal requirements emerge.

Foundation 8: Global Interoperability

International recognition and mutual acceptance can reward compatible evidence and standards.

47. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Incentive design is core infrastructure for frontier AI evaluation.
2. Technical standards should include analysis of the incentives they create.
3. Prestige should be treated as a real institutional resource.
4. Recognition should be attached to evidence, scope, contribution, and current validity.
5. Awards and public recognition should not confer evaluator, accreditation, certification, or regulatory authority.
6. High benchmark scores should not receive greater prestige than valid and well-maintained benchmarks.
7. Replication, negative results, maintenance, correction, incident analysis, and documentation deserve explicit professional credit.
8. Financial rewards should support public goods without attempting to replace intrinsic motivation.
9. Result-dependent evaluator compensation should be prohibited.
10. Model access should be allocated through legible competence and security criteria rather than loyalty or institutional prestige alone.
11. Responsible disclosure should receive protection, acknowledgment, and timely response.
12. Corrective credit should reward timely disclosure and remediation without excusing negligence, concealment, or repeated failure.
13. Recognition programs should include anti-gaming analysis, verification, appeal, correction, and sunset.
14. Rankings should be used only when construct validity, comparability, uncertainty, and governance support them.
15. Standards Body should prefer multidimensional evidence profiles over universal safety rankings.
16. Procurement and insurance can create powerful incentives but should accept equivalent evidence and avoid proprietary lock-in.
17. Certification prestige should remain limited to the system, version, scope, scheme, and period assessed.
18. Contributor recognition should include operational, maintenance, security, and community work.

19. Open-source communities should receive sustainable funding and meaningful credit without forced centralization.
 20. Standards participation should be financially and logistically accessible to smaller and public-interest actors.
 21. Awards and grants should disclose funders, reviewers, conflicts, criteria, and rationale.
 22. Existing prestige should not be a prerequisite for earning recognition.
 23. Recognition should be withdrawable after fraud, concealed conflict, or material invalidation.
 24. Sanctions should deter concealment and misconduct without suppressing good-faith disclosure.
 25. Incentive programs should be evaluated for crowding out, market concentration, gaming, and distribution.
 26. Institutions should receive legitimacy for retiring failed incentives and obsolete standards.
 27. National prestige should not override international technical cooperation.
 28. Public reporting should reward candor and evidence rather than volume and favorable language.
 29. The most valuable incentive is not always financial. Access, credit, authority, mission, and peer respect can be equally consequential.
 30. The incentive system itself should be treated as an object of continuous evaluation.
-

48. Decision Rules

An incentive program should be created only when:

- A meaningful behavior or public good is underprovided
- the responsible actor is identifiable
- the mechanism is plausible
- performance can be verified
- gaming risk can be managed
- distribution and crowding-out effects are considered
- the program can be reviewed and ended

A prize should be used when:

- The objective is sufficiently clear
- multiple solution paths are desirable
- performance can be tested
- participation can broaden
- winner selection will not create unsafe pressure

A bounty should be used when:

- The target and permitted methods are defined
- safe disclosure exists
- triage and remediation capacity exist
- reward reflects severity and quality
- legal treatment is clear

Recognition should be granted only when:

- The contribution is evidenced
- scope is explicit
- conflicts are managed
- independent review occurs
- correction and withdrawal are possible

A ranking should not be used when:

- The construct is unstable
- data are not comparable
- uncertainty is large
- one metric hides important tradeoffs
- the ranking will create severe gaming
- appeal is unavailable

Corrective credit should be granted when:

- Disclosure is timely
- evidence is complete
- remediation is meaningful
- recurrence prevention is verified
- there was no concealment or manufactured failure

A program should be suspended or retired when:

- Gaming dominates
- behavior does not improve
- burden exceeds value
- prestige capture becomes persistent
- distribution is unacceptably narrow
- superior mechanisms exist
- sponsor or institutional conflicts cannot be controlled

49. Incentive Program Design Template

A. Identity

- Program name
- owner
- version
- status
- duration

B. Objective

- Desired behavior
- public outcome
- problem

C. Actor

- Eligible participants
- decision control
- motivation

D. Baseline

- Current behavior
- evidence
- gap

E. Mechanism

- Why the incentive should work

F. Reward

- Financial
- prestige
- access
- governance
- market
- career

G. Evidence

- Submission

- validation
- outcome
- maintenance

H. Verification

- Reviewers
- conflicts
- methods
- security

I. Anti-Gaming

- Threat model
- counter-metrics
- audit
- clawback

J. Distribution

- Small actors
- open source
- geography
- career stage
- accessibility

K. Crowding Out

- Intrinsic motivation
- community effects
- collaboration

L. Decision

- Criteria
- deliberation
- dissent
- rationale

M. Correction and Appeal

N. Impact Evaluation

O. Sunset

50. Contribution Recognition Template

Contribution:

Contributor or team:

Category:

Date:

Contribution Description

Public Value

Evidence

Validation

Reuse or Impact

Maintenance

Conflicts

Reviewers

Recognition Scope

Expiration or Permanence

Conditions

Correction History

Decision

- Acknowledgment
- verified contribution
- distinguished contribution
- stewardship recognition
- not recognized

51. Responsible Disclosure Reward Template

Finding identifier:

Reporter:

Affected system or protocol:

Date:

Scope Compliance

Finding

Evidence

Severity

Novelty

Reproducibility

Potential Harm

Disclosure Handling

Remediation

Verification

Reporter Preference

- Public credit
- private credit
- anonymous
- financial reward
- access or collaboration

Reward Decision

Publication Timing

Follow-Up

52. Corrective Credit Template

Organization:

Incident or error:

Date identified:

Preventability

Detection

Disclosure Timing

Evidence Quality

Containment

Remediation

Independent Verification

Recurrence Prevention

Harm

Prior History

Concealment or Bad Faith

Credit Decision

- No credit
- acknowledgment
- verified corrective action
- distinguished institutional correction

Remaining Consequences

Review Date

53. Recognition and Prestige Scorecard

Dimension	Core Question
Objective	Is the desired behavior or public good clear?
Actor	Does the program target the actor who can change behavior?
Mechanism	Is there a plausible incentive pathway?
Evidence	Is the contribution or outcome demonstrated?
Verification	Is review independent and competent?
Scope	Is recognition limited to the actual contribution?
Metric validity	Does the measure represent the intended goal?
Anti-gaming	Are reward hacking and substitution addressed?
Crowding out	Could the program weaken intrinsic motivation?
Distribution	Can new, small, open, and geographically diverse actors participate?
Conflict	Are funder, reviewer, and sponsor interests controlled?
Transparency	Are criteria, evidence, and rationale legible?
Correction	Can error be corrected publicly?
Clawback	Can status or reward be withdrawn after serious invalidation?
Maintenance	Is long-term stewardship rewarded?
Public goods	Does the program support broadly usable infrastructure?
Market effects	Does it create concentration, lock-in, or exclusion?
Access	Are model, data, compute, and governance opportunities allocated fairly?
Prestige integrity	Does status track competence and contribution?
Decision separation	Is recognition kept separate from formal authority?
International fit	Can recognition travel without national or institutional favoritism?
Impact	Did behavior or outcomes improve?
Cost	Is administration proportionate?
Sunset	Can the program end when it no longer works?

54. Final Perspective

Frontier AI institutions will often claim that they value safety, rigor, transparency, and public benefit.

The more important question is what they reward.

What receives funding?

What leads to promotion?

What earns access?

What appears in public rankings?

What wins awards?

What attracts purchasers?

What improves insurance terms?

What creates committee authority?

What happens when someone reveals bad news?

What happens when an organization corrects itself?

What happens when a benchmark maintainer spends years doing work that produces no dramatic headline?

The answers reveal the real operating system of the field.

A system that praises independent review while punishing unfavorable reviewers does not value independence.

A system that praises transparency while imposing disproportionate cost on disclosure does not value transparency.

A system that praises open science while underfunding maintenance does not value open infrastructure.

A system that praises standards while rewarding adoption over validity does not value standards.

A system that praises safety while granting prestige only for capability progress will continue to underproduce safety work.

The solution is not to monetize every good behavior.

Nor is it to create a badge for every contribution.

Incentives can corrupt.

Prestige can concentrate.

Awards can become marketing.

Rankings can destroy the measures they use.

Certification can produce false reassurance.

Sanctions can suppress disclosure.

The solution is deliberate design.

That design should recognize that people and institutions act for mixed reasons.

They seek:

- Purpose
- mastery
- income
- status
- access
- belonging
- influence
- security
- public benefit

A strong system aligns several of these motives without allowing one metric or reward to dominate.

It gives prestige to difficult, durable work.

It funds public goods.

It protects responsible disclosure.

It rewards correction.

It creates consequences for concealment.

It broadens access to contribution.

It separates recognition from authority.

It allows recognition to expire.

It audits its own incentives.

The seventh foundation of Standards Body is therefore incentive alignment with epistemic integrity.

The field should make it more rewarding to know the truth, reveal the truth, correct the record, improve the system, and build infrastructure others can trust.

References and Research Basis

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

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

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

[^haip-overview]: OECD, **HAIP Reporting Framework Overview**. <https://oecd.ai/en/transparency/overview>

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

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

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

[^nist-prizes]: National Institute of Standards and Technology, **Open Innovation Prize Challenges**. <https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges>

[^challenge-gov]: United States General Services Administration, **Challenge.gov**. <https://www.challenge.gov/>

[^nist-vdp]: National Institute of Standards and Technology, **SP 800-216, Recommendations for Federal Vulnerability Disclosure Guidelines**, 2023. <https://csrc.nist.gov/pubs/sp/800/216/final>

[^nist-vdp-project]: National Institute of Standards and Technology, **Vulnerability Disclosure Guidelines**. <https://csrc.nist.gov/projects/vulnerability-disclosure-guidelines/iso-pubs>

[^iso-public-policy]: International Organization for Standardization, **Standards and Public Policy: A Toolkit for National Standards Bodies**, 2023. https://www.iso.org/files/live/sites/isoorg/files/publications/en/ISO_Public-Policy-Toolkit.pdf

[^iso-innovation]: International Organization for Standardization, **Standards and Innovation**, 2021. <https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100466.pdf>

[^iso-benefits]: International Organization for Standardization, **Repository of Studies on Benefits of Standards**. <https://www.iso.org/sites/materials/benefits-of-standards/>

[^iso-referencing]: International Organization for Standardization, **Using and Referencing IEC and ISO Standards to Support Public Policy**. <https://www.iso.org/iso/pub100358.pdf>

[^goodhart]: Charles A. E. Goodhart, **Problems of Monetary Management: The U.K. Experience**, 1975, in *Papers in Monetary Economics*.

[^campbell]: Donald T. Campbell, **Assessing the Impact of Planned Social Change, Evaluation and Program Planning**, 1979. [https://doi.org/10.1016/0149-7189\(79\)90048-X](https://doi.org/10.1016/0149-7189(79)90048-X)

[^holmstrom-milgrom]: Bengt Holmström and Paul Milgrom, **Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design**, *Journal of Law, Economics, and Organization*, 1991. https://doi.org/10.1093/jleo/7.special_issue.24

[^deci-meta]: Edward L. Deci, Richard Koestner, and Richard M. Ryan, **A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation**, *Psychological Bulletin*, 1999. <https://doi.org/10.1037/0033-2909.125.6.627>

[^frey-jegen]: Bruno S. Frey and Reto Jegen, **Motivation Crowding Theory**, *Journal of Economic Surveys*, 2001. <https://doi.org/10.1111/1467-6419.00150>

[^ostrom]: Elinor Ostrom, **Beyond Markets and States: Polycentric Governance of Complex Economic Systems**, Nobel Prize Lecture, 2009. https://www.nobelprize.org/uploads/2018/06/ostrom_lecture.pdf

[^lerner-tirole]: Josh Lerner and Jean Tirole, **Some Simple Economics of Open Source**, *Journal of Industrial Economics*, 2002. <https://doi.org/10.1111/1467-6451.00174>

[^kremer-williams]: Michael Kremer and Heidi Williams, **Incentivizing Innovation: Adding to the Toolkit**, *Innovation Policy and the Economy*, 2010. <https://doi.org/10.1086/605852>

[^merton]: Robert K. Merton, **The Matthew Effect in Science**, *Science*, 1968. <https://doi.org/10.1126/science.159.3810.56>

[^credit-taxonomy]: Allen et al., **Credit Where Credit Is Due**, *Nature*, 2014, describing the CRediT contributor-role taxonomy. <https://doi.org/10.1038/508312a>

Revision Record

Version 1.0

Date: July 16, 2026

Change type: Complete foundational edition

Summary: Establishes the fully developed canonical working white paper for Foundation 7. Defines the incentive and prestige problem, actor incentives, the incentive stack, intrinsic and extrinsic motivation, Goodhart and Campbell effects, prestige governance, recognition architecture, awards, prizes, grants, bounties, publication incentives, developer and evaluator incentives, standards-organization incentives, procurement, insurance, public reporting, rankings, certification, access, professional careers, open-source sustainability, corrective credit, sanctions, standards participation, program design, anti-gaming, distribution, international alignment, governance, maturity, implementation, a Standards Body pilot, metrics, failure analysis, objections, evidence gaps, research agenda, standards implications, operational templates, scorecard, and primary-source research basis.

Status: Ready for internal review and future expert critique.