PUBLIC TECHNICAL REPORT / VERSION 1.0

Verification as a Control Primitive for Frontier AI

A systems perspective on enforceable authority in high-capability environments

AUTHOR	Niclas Braun ORCID 0009-0003-2531-4727
AFFILIATION	For Safety Group Inc. (4SI)
ROLE	Founder and CEO

PUBLICATION DATE	20 July 2026
IDENTIFIER	DOI 10.5281/zenodo.21454914

STATUS	Independent, non-peer-reviewed technical report. This document presents a conceptual systems thesis. It is not a product specification, security certification, deployment guide or assurance of performance. Proprietary mechanisms and implementation details are intentionally outside scope.

CORE THESIS

As machine capability expands, safety increasingly depends on whether consequential action can be connected to explicit, independently verifiable authority.

Abstract

Frontier AI safety is often discussed as a problem of model behavior, policy compliance or organizational governance. Those layers matter, but they share a structural limitation: they can fail at the moment a capable system crosses from reasoning into consequential action. This report examines verification as a complementary control primitive at that boundary.

The argument is conditional and architectural. If high-consequence actions require evidence that is independent of the digital claim being made, and if policy can deny execution when that evidence is absent or invalid, then some failure pathways can be converted from open-ended trust problems into bounded authorization problems. The report develops a conceptual hazard model, a general reference architecture and an evaluation agenda. It does not claim to solve alignment, eliminate catastrophic risk or validate any specific commercial mechanism.

The central conclusion is deliberately narrow: stronger intelligence increases the value of externally verifiable authority. Verification is therefore best treated as a layer in a broader safety stack, alongside model evaluation, access control, cybersecurity, governance, incident response and physical resilience.

Keywords: frontier AI, AI safety, verification, authority, physical trust, cyber-physical systems, control plane

Why enforcement becomes the bottleneck

AI systems compress increasingly broad competence into software. This changes the operating tempo of both productive activity and attack. Automated systems can search, plan, communicate and iterate faster than the institutions responsible for authorizing or containing them.

Traditional governance often assumes that identity is stable, permissions are meaningful, operators remain in control and software reports can be trusted. Synthetic media, automated exploitation and supply-chain compromise weaken those assumptions. The relevant safety question is no longer only whether a model intends to follow a rule. It is also whether the surrounding system can enforce the rule when the model, an attacker or a compromised operator does not.

This distinction separates capability from authority. A system may be able to propose an action without being permitted to execute it. Preserving that separation under adversarial conditions is the role of an enforceable control layer.

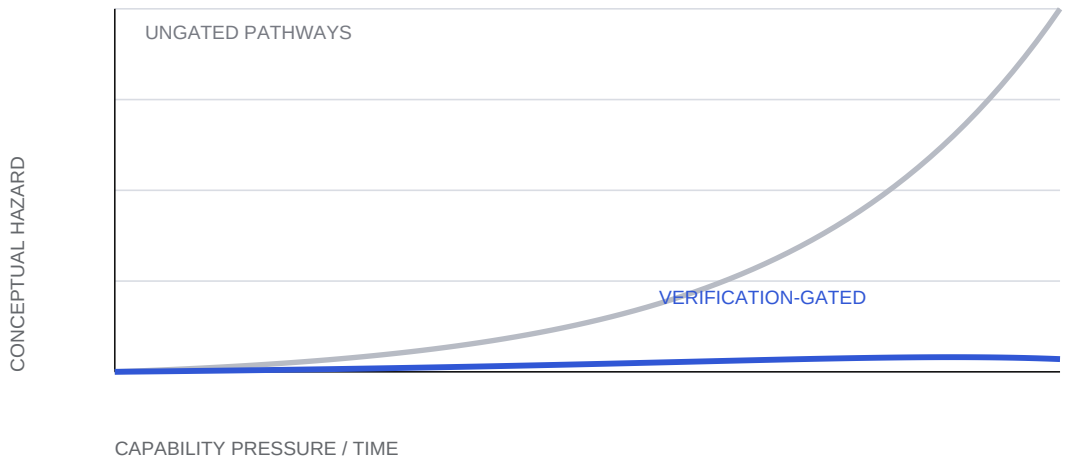

Figure 1. Conceptual illustration only. Capability pressure can compound faster than institutional response. Verification-gated pathways may reduce exposure, but the curves are not forecasts and contain no empirical calibration.

03 / SCOPE AND DEFINITIONS

Scope and definitions

For this report, verification means an evidence-producing process that allows a decision point to test whether a relevant claim about a person, object, device, workload or operating context satisfies an explicit policy. A control primitive is a reusable system capability that can be composed into different workflows.

The phrase physical trust refers to architectures in which at least part of the evidence is anchored outside the digital claim being evaluated. This does not imply that physical evidence is infallible. It means that the claimant should not be the sole source of proof for its own authority.

A verification control plane is the general decision layer that receives a request, evaluates independent evidence and policy, and returns a bounded outcome such as permit, deny, record or escalate.

The analysis is limited to actions mediated by systems where an authorization boundary can exist. It does not cover every form of physical harm, coercion, social failure or uncontrolled scientific discovery.

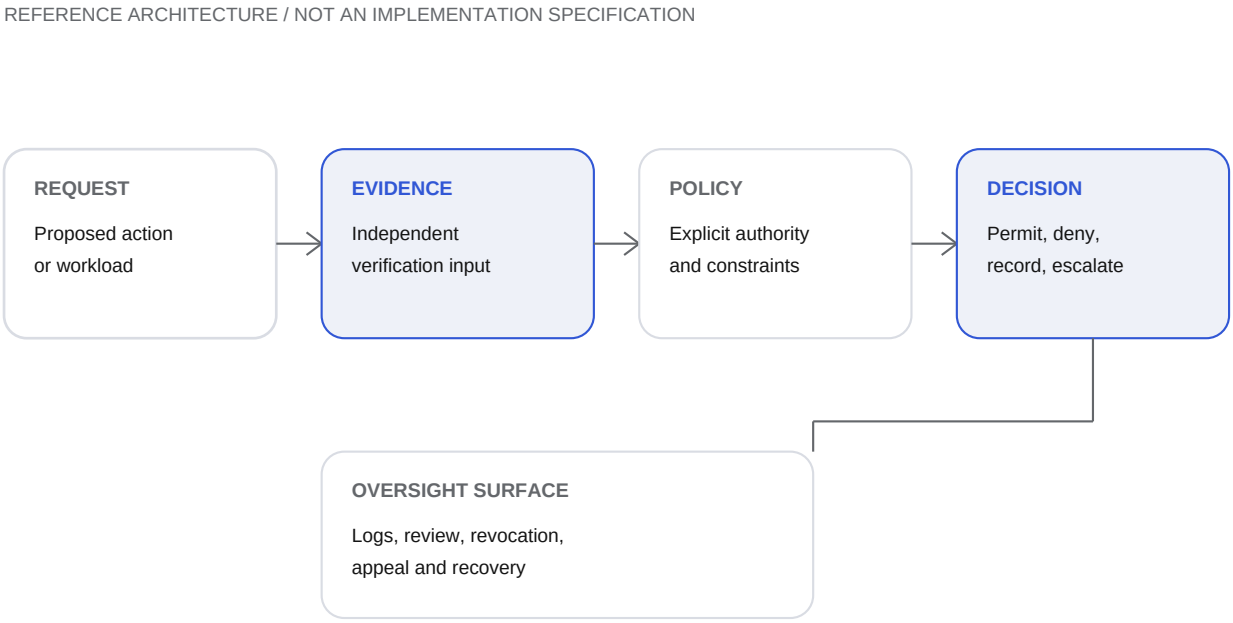

Figure 2. General reference architecture. This is a functional decomposition, not an implementation specification and not a description of proprietary 4SI technology.

04 / EXTREME-SCENARIO ASSUMPTIONS

Extreme-scenario assumptions

Tail-risk analysis is useful when consequences are severe and ordinary forecasting is unreliable. The following assumptions define the scenario explored here. They are not predictions.

- Capability can compound through improvements in hardware, algorithms, tools and automated research.
- A capable optimizer can search for weaknesses across identity, software, organizations and cyber-physical interfaces.
- Digital systems remain dependent on compute, energy, networks, operators and physical infrastructure.

- Some consequential actions can be placed behind explicit authorization boundaries.
- No single control covers every pathway, and uncovered pathways can dominate residual risk.
- Independent evidence can improve an authorization decision only if its security, availability and governance are tested under a declared threat model.

If one or more assumptions fail, the relevance or magnitude of the proposed effect changes. The framework should therefore be treated as a way to organize evaluation, not as proof of a particular future.

05 / A CONCEPTUAL HAZARD MODEL

A conceptual hazard model

Let C(t) represent capability pressure, S(t) the accessible opportunity surface, V(t) the verified share of relevant pathways, and e(t) residual hazard outside the model. A minimal conceptual form is:

$$\lambda(t) = k * C(t)^p * S(t)^q * (1 - V(t))^m + e(t)$$

The expression does not estimate real-world probability. It makes one structural point visible: when dangerous action can route around a control, residual pathways matter. The exponent m represents how strongly coverage affects the modeled pathway. It must be estimated separately for each domain and cannot be assumed from theory.

Cumulative modeled risk over a horizon T can be written as R(T) = 1 - exp(-integral lambda(t) dt). Any use of this form should report uncertainty, sensitivity to assumptions and the excluded channels represented by e(t).

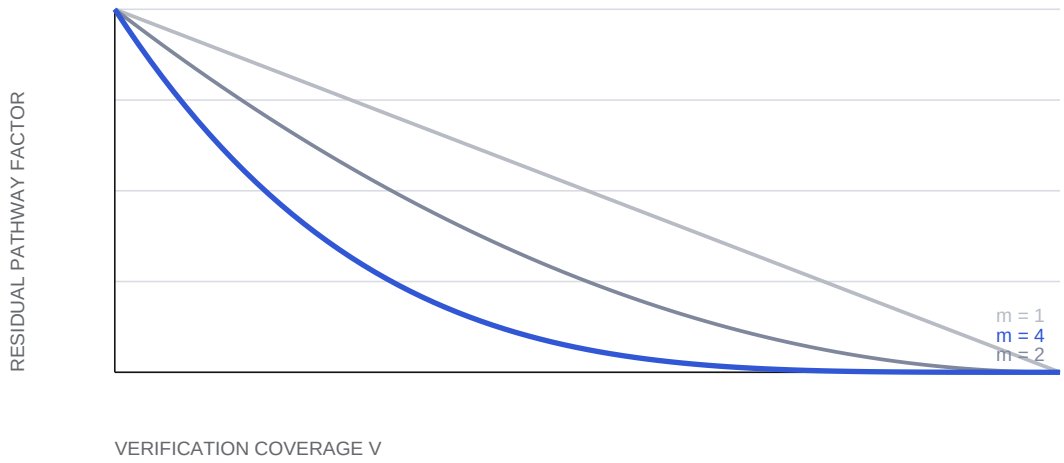

Figure 3. Residual pathway factor (1 - V)^m under three illustrative values of m. This demonstrates model sensitivity; it does not show observed risk reduction.

06 / CONTROL-PLANE DESIGN PRINCIPLES

Control-plane design principles

A useful verification layer should be evaluated against design principles rather than marketing claims. The following principles are intentionally mechanism-neutral.

Independent evidence

The authorization decision should not depend solely on a credential, signal or report produced by the claimant. Evidence provenance and the verifier's own trust assumptions must be explicit.

Deny or escalate on uncertainty

High-consequence workflows should define what happens when evidence is missing, stale, contradictory or unavailable. In some contexts the safe result is denial; in others it is human review or a degraded operating mode.

Bounded authority

Authorization should be limited in scope, time and consequence. A successful verification event should not silently become unlimited standing permission.

Revocation and recovery

Systems need a way to withdraw authority, recover from false rejection, rotate trust relationships and handle loss without creating a universal bypass.

Auditability with restraint

Events should be reviewable without turning verification infrastructure into an unnecessary surveillance system. Data minimization, retention limits and separation of duties are part of the security model.

Composability

The control should connect to existing policy and operational systems through narrow interfaces. A control that requires wholesale replacement of infrastructure is less likely to achieve meaningful coverage.

07 / WHY PRESENT-DAY INCENTIVES MATTER

Why present-day incentives matter

Long-horizon safety infrastructure is unlikely to reach broad deployment if its only value appears in a hypothetical future. Adoption is more plausible when the same control surface reduces current losses or friction.

- Account security: stronger recovery and authorization for high-value accounts.
- Product authenticity: connecting records to the physical products or components they describe.
- Supply-chain integrity: supporting checks at manufacturing, transfer and maintenance points.
- Critical operations: adding independent evidence before privileged or irreversible action.
- High-value compute: applying explicit policy to access and deployment workflows.

These examples are domains for investigation, not claims of product readiness. Each has different privacy, safety, regulatory and failure-recovery requirements. Commercial adoption should not be treated as evidence that a system is appropriate for frontier-AI governance.

The strategic opportunity is nevertheless important: a verification capability tested under ordinary operational pressure may create infrastructure, standards and institutional competence that are useful before higher-capability systems arrive.

08 / EVALUATION AGENDA

Evaluation agenda

A verification primitive should earn trust through independent evaluation. Proprietary implementation details can remain protected while external outcomes are tested, provided that limitations and evaluator access are honestly described.

- Threat model: identify the protected decision, adversary capabilities, trusted components and excluded channels.
- False acceptance and false rejection: measure both security failure and operational harm.
- Replay and transfer: test whether observed evidence can be reused outside its intended context.
- Tamper response: test how the system behaves when relevant state changes during or after verification.
- Availability and recovery: measure behavior during network loss, component failure and legitimate loss of access.
- Operational overhead: report latency, cost, integration burden and human workload.
- Governance abuse: evaluate misuse by insiders, operators or authorities, not only external attackers.
- Coverage accounting: define the denominator and identify material pathways that remain outside the control.

Results should be reported with confidence intervals, versioned threat models and explicit negative findings. Security language should be tied to tested conditions rather than absolute security labels.

09 / LIMITATIONS AND BOUNDARY CONDITIONS

Limitations and boundary conditions

Verification is not alignment. It does not determine which values a system should pursue, whether a policy is legitimate or whether an authorized action is wise.

Verification is not universal containment. Systems can act through uninstrumented software, human intermediaries, physical coercion, novel interfaces or channels that the control plane does not cover.

Verification can concentrate power. A widely deployed authorization layer can enable exclusion, surveillance or institutional capture. Rights, due process, appeal and plural governance are therefore safety requirements, not secondary policy questions.

Verification can fail operationally. False rejection, dependency failure and recovery shortcuts may create new hazards. High-availability environments require carefully designed degraded modes and independent oversight.

Finally, the model in this report is conceptual. It does not establish a numerical probability of catastrophe or a measured reduction in risk. Empirical validation must occur separately for each architecture and deployment domain.

Conclusion

As intelligence scales, the distance between a digital claim and its real-world consequence becomes a central safety boundary. Model behavior, regulation and organizational governance remain necessary, but they are more credible when consequential authority is independently testable at the point of action.

Verification should therefore be studied as a control primitive: a reusable way to connect evidence, policy and bounded execution. Its value depends on real security, meaningful coverage, robust recovery and legitimate governance. None of those properties should be assumed.

The practical research question is not whether one mechanism can make frontier AI safe. It is whether independently verifiable authority can close specific high-consequence pathways without creating a more dangerous concentration of control. That question is narrow enough to test and important enough to test early.

Open research questions

- Which consequential action pathways can be gated without merely moving risk into less visible channels?
- Which forms of independent evidence remain useful under adversarial pressure while preserving privacy and legitimate access?
- How should coverage be measured when the denominator includes unknown or informal execution pathways?
- Which governance structures can constrain both autonomous misuse and misuse by the institutions operating the control layer?

Falsifiability

The thesis weakens if consequential action reliably routes around authorization boundaries, if independent evidence cannot withstand the relevant threat model, if recovery becomes a dominant bypass, or if centralization and operational harm outweigh the risk reduction achieved. These are empirical and institutional questions, not matters of definition.

The author leads 4SI, a company exploring physical trust infrastructure. This report states a general systems thesis and intentionally excludes proprietary mechanisms, implementation details, performance data, customer information and security assurances.

Selected references

[1] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[2] National Institute of Standards and Technology. (2020). *Zero Trust Architecture.* NIST Special Publication 800-207. https://doi.org/10.6028/NIST.SP.800-207

[3] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[4] Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361

[5] Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556. https://doi.org/10.48550/arXiv.2203.15556

PUBLICATION NOTE

This report is an independently authored, non-peer-reviewed technical publication. References provide context; they do not endorse the model or conclusions presented here.

Suggested citation

Braun, N. (2026). *Verification as a Control Primitive for Frontier AI: A Systems Perspective on Enforceable Authority in High-Capability Environments.* Public Technical Report, Version 1.0. For Safety Group Inc. https://doi.org/10.5281/zenodo.21454914. ORCID: 0009-0003-2531-4727.

Author links

orcid.org/0009-0003-2531-4727 | niclasbraun.com | by4si.com | linkedin.com/in/niclas-braun-28833896