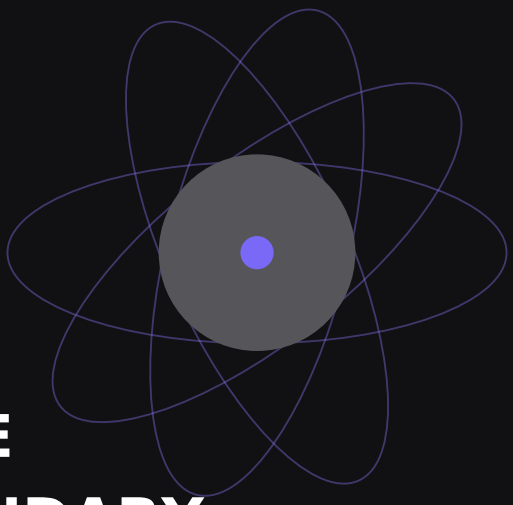
NICLAS BRAUN / RESEARCH REPORT

VERSION 1.0 / JULY 2026



PHYSICAL TRUST AT THE DIGITAL-PHYSICAL BOUNDARY

An Information-Theoretic Framework for Verification in Autonomous Action Systems

SCIENTIFIC REPORT

Information theory · decision theory · assurance architecture

Mechanism-neutral · no implementation-specific disclosure

NICLAS BRAUN · INDEPENDENT RESEARCHER

Contents

A research framework for physical trust assurance, interoperability, and institutional control.

Scope and disclosure statement

This article develops a conceptual and formal framework for reasoning about evidence at the boundary between digital representations and decision-relevant physical states. It is intentionally mechanism-neutral. It does **not** disclose or imply any specific composition, fabrication process, sensing modality, signal-acquisition method, feature representation, enrollment procedure, anti-tamper design, secure-processing topology, or other protected implementation detail.

The proposed Physical Trust Assurance Standard (PTAS) is an analytical construct and prospective standardization agenda. It is not a published consensus standard, a certification claim, an empirical validation of a product, or a statement of current deployment. All application profiles are hypothetical boundary analyses unless expressly stated otherwise.

Abstract

Digital security mechanisms establish properties of representations, communication channels, credentials, and computing environments. They do not, without an appropriate observation model, establish that a represented person, object, authority, or context presently has the physical state presupposed by a consequential action. This article characterizes that residual problem as the **physical trust gap** and develops a mechanism-neutral framework for its analysis. Physical trust is defined as the capacity of a relying institution to obtain and appraise evidence about a specified physical state through an observation path that is causally bound to the referent and adversarially distinct from the disputed digital claim, within a declared operating domain and threat model, before an action is authorized.

The contribution is fourfold. First, an input-equivalence lemma shows that no decision rule, deterministic or randomized, can condition its output on a physical distinction that induces no difference in the rule's accessible information. A same-domain corollary shows that adding evidence controlled within the disputed claim path does not close that gap when the joint inputs remain observationally equivalent under attack. The result is computationally agnostic and therefore applies independently of advances in artificial intelligence. Second, an information-theoretic account represents physical observation as an experiment that may reduce conditional uncertainty; Blackwell ordering and Bayesian decision risk distinguish observations that are merely correlated from those that are decision-relevant. Third, a reference architecture separates binding, observation, evidence, appraisal, authorization, and enforcement, thereby preventing verification results from being interpreted as unrestricted claims of truth. Fourth, the article derives candidate assurance levels, result semantics, evaluation requirements, and a governance pathway for a prospective Physical Trust Assurance Standard (PTAS).

The framework is positioned as complementary to digital identity, zero-trust architecture, remote attestation, verifiable credentials, content provenance, post-quantum cryptography, safety engineering, and institutional governance. It neither assumes that physical evidence is unforgeable nor treats physical presence as equivalent to authority. Its central claim is narrower: when policy depends on a present physical distinction, epistemically adequate authorization requires a causally connected observation path whose evidentiary source is not reducible to the disputed claim's control path and whose residual risk is commensurate with the consequence of error.

Keywords: physical trust; cyber-physical assurance; observation; authorization; autonomous systems; decision theory; information theory; remote attestation; AI governance

Principal propositions

1 **Representational validity does not entail physical correspondence.** Integrity, provenance, authentication, and truth conditions are analytically distinct properties.

2 **Computational capability and same-domain evidence do not remove observational non-identifiability.** If two physical states induce the same accessible joint input distribution under the declared adversary, neither additional computation nor additional evidence from that same compromise domain enables discrimination.

3 **Scaled action creates a control-capacity problem.** When the rate of machine-generated action proposals exceeds accountable review capacity, authorization predicates must become explicit, testable, and machine-consumable.

4 **Physical trust is a bounded assurance property.** It concerns a specified state variable, time interval, operating domain, decision boundary, and adversary model; it is not a general synonym for physical security.

5 **Interoperability should attach to semantics and assurance, not to a protected mechanism.** Claim profiles, result states, lifecycle rules, evaluation methods, and policy interfaces can be standardized while implementations remain heterogeneous. Interface compatibility does not imply equal security, equal independence, or mechanism fungibility.

6 **Physical and digital assurance are compositional but non-substitutable.** Cryptography protects defined information relationships; identity binds digital subjects and authenticators; attestation appraises computing state; physical trust addresses selected facts about embodied subjects and contexts.

1. Introduction: from information systems to action systems

Contemporary security architecture developed largely around systems that store information, transmit instructions, and mediate access. Increasingly, software also ranks alternatives, negotiates, purchases, routes, schedules, approves, and initiates changes in other systems or in the physical environment. The transition from information-processing systems to action-oriented systems changes the location of the relevant control boundary.

For an information system, a central question is whether data has been kept confidential and unchanged. For an action system, the decisive question is broader:

> What must be true in the physical world before this system is permitted to act?

This is not a rejection of digital security. It is a recognition of scope. Shannon's theory provided a mathematical foundation for transmitting information through noisy channels [1]. Modern cryptography provides confidentiality, integrity, authentication, and evidence that may support non-repudiation under explicit technical, procedural, and institutional assumptions. Digital identity systems establish subjects and authenticators. Zero trust architectures continually evaluate access to resources rather than granting trust based on network location [6]. Remote attestation can produce evidence about a device or execution environment [8]. Verifiable credentials can express tamper-evident claims by issuers about subjects [9]. Content provenance can make the history and binding of digital media inspectable [10].

These achievements are foundational. Yet each operates over a defined representational boundary. A signed record can be a faithfully preserved falsehood. A valid credential can be presented in an unintended physical context. A healthy device can be operated by the wrong person. A correct maintenance record can refer to a substituted component. A provenance chain can describe the history of an asset without proving that the asset depicts a true event; the C2PA explainer states this limitation directly [10]. The W3C Verifiable Credentials model similarly distinguishes

cryptographic verification from the business validation a verifier must apply before relying on a claim [9].

The gap is therefore not between security and insecurity in the abstract. It is between a **digital claim** and the **physical state of affairs** on which a consequential decision depends.

1.1 The physical trust gap

Consider five recurrent separations:

Digital property established	Physical fact still potentially unresolved
A credential is valid	The intended person is presently exercising the authority
A message signature verifies	The signer is acting in the expected physical context and under legitimate authority
A record is immutable	The record still corresponds to the physical object now presented
A device attests to an approved state	The device is connected to the expected asset, operator, location, or process
An AI action is policy-compliant in software	The physical prerequisite and accountable authority for execution are present

We call this residual uncertainty the **physical trust gap**. It is not present in every transaction, and it need not be closed everywhere. The gap matters where an incorrect acceptance produces material consequence: release of funds, privileged recovery, command execution, component installation, custody transfer, access to hazardous or classified operations, activation of autonomous equipment, or modification of critical infrastructure. Where policy makes the physical distinction a prerequisite, an institution must observe it through adequate evidence, redesign or constrain the action, or decline to act; cost or convenience cannot convert an unobserved prerequisite into a satisfied one.

1.2 Why the problem is becoming structural

Three trends make the gap structural rather than exceptional.

First, synthetic content and synthetic interaction reduce the cost of producing plausible claims. The marginal cost of generating messages, voices, images, identities, and procedural artifacts is falling. The relevant security assumption can no longer be that plausibility is expensive.

Second, software agents reduce the distance between claim and action. A generated instruction may be evaluated, routed, and executed without the latency that once allowed human institutions to notice contradictions. This does not require speculative artificial general intelligence. It follows from ordinary automation combined with increasing model capability and tool access.

Third, cryptographic transition creates long-lived heterogeneity. NIST finalized its first three post-quantum cryptography standards in 2024 [11-13] and subsequently emphasized crypto agility as a system property required for secure transition [14]. For assets and institutions with multi-decade lifecycles, trust architectures must operate across algorithm changes, mixed generations of equipment, intermittent connectivity, and evolving adversary capabilities. Physical trust cannot substitute for post-quantum cryptography, but it can provide an additional, differently rooted control at selected moments where digital authority becomes physical consequence.

1.3 Contribution and scope

This paper makes five contributions:

1 It defines physical trust as a bounded assurance property rather than a slogan or universal claim.

2. It formalizes the difference between a digital claim and evidence about physical state.
3. It identifies the invariant epistemic boundary that remains even as machine intelligence scales.
4. It proposes a mechanism-neutral reference architecture for converting physical evidence into policy-consumable results.
5. It outlines the contents and governance path of a future Physical Trust Assurance Standard.

The paper does not claim that physical mechanisms are unforgeable, that physical evidence is categorically superior to digital evidence, or that a single control can establish truth across contexts. Every observation channel is conditional on assumptions; every verifier has failure modes; and every enrollment or reference process may be corrupted. Physical trust is analytically useful only when the target claim is explicit, the inference boundary is narrow, and assurance is evaluated against a declared adversary and consequence model.

1.4 Research questions

The analysis is organized around four research questions:

1. **Identifiability:** Under what conditions can a decision system distinguish physical states that require different institutional actions?
2. **Decision value:** When does an additional physical observation reduce expected decision loss after acquisition, privacy, latency, and availability costs are included?
3. **Assurance architecture:** Which semantic and governance boundaries permit a physical verification result to be used without converting it into an unrestricted claim of truth?
4. **Standardization:** Which properties can be made interoperable and independently evaluable without disclosing or prescribing a particular physical mechanism?

1.5 Method and epistemic status

This is a theory-building and architecture paper rather than an empirical performance study. Its method combines: (i) conceptual analysis of claim types and decision boundaries; (ii) formalization using statistical decision theory and information theory; (iii) architectural synthesis across identity, attestation, provenance, and conformity-assessment frameworks; and (iv) adversarial analysis of the resulting control path. The formal results are deliberately limited to propositions that follow from stated information structures. The assurance levels and application profiles are design hypotheses to be tested, not findings inferred from deployment data.

Three validity constraints follow. **Construct validity** requires that each operational measure correspond to the physical claim actually used by policy. **Internal validity** requires that evaluation isolate relevant failure and attack mechanisms rather than report aggregate accuracy alone. **External validity** requires an explicit operating domain and prohibits extrapolation beyond tested populations, environments, and lifecycle states. No empirical claim about a specific implementation is made in this article.

2. The historical arc: representation, computation, and authority

2.1 From matter to symbols

Human institutions historically attached authority to physical presence, possession, ceremony, seals, witnesses, and controlled spaces. These mechanisms were imperfect, local, and expensive, but they made authority costly to separate from embodiment. Industrial bureaucracy converted many of these relationships into records. Networked computing converted records into transmissible symbols. Public-key cryptography made selected properties of those symbols verifiable at global scale.

This separation enabled information to travel independently of the physical object or person it described. It also created a persistent ambiguity: a representation may remain operationally authoritative after its physical referent, context, or controlling authority has changed.

This history can be summarized as four overlapping eras:

Era	Dominant scaling achievement	Primary trust question
Physical institutions	Stable local authority	Who or what is physically here?
Recorded institutions	Durable representation	What was documented?
Networked institutions	Global transmission and cryptographic integrity	Who controls the credential or key?
Intelligent action systems	Scaled inference and execution	What is physically true, and who is authorized, before action?

The fourth era does not supersede the others. It composes them. A robust action system may need a cryptographically authenticated message, an attested device, a validated institutional policy, an authentic physical object, and an accountable person. The engineering question is how to express and combine these claims without collapsing them into the undifferentiated word trust.

2.2 Computation is physical, but representations are not their referents

Landauer's analysis connected logically irreversible computation to physical dissipation [2]. The broader lesson is not that every security decision should be reduced to thermodynamics. It is that computation is instantiated in physical processes even when its semantic objects are abstract. A bit is represented physically; a credential is stored physically; a model executes on physical hardware. Yet the physical instantiation of a representation does not guarantee correspondence between the representation and the external thing it denotes.

Let the proposition encoded by a record be:

> Component C is installed in vehicle V.

The record's integrity may be perfect. Its storage hardware may be attested. Its signature may be post-quantum secure. None of those facts alone prevents component C from having been removed and replaced after the last trusted observation. The unresolved variable is not the integrity of the proposition but the current correspondence between proposition and world.

This is a familiar issue in measurement science. Observations are not reality itself; they are outputs of an interaction between a system and an instrument under assumptions. Physical trust therefore begins not with metaphysical certainty, but with disciplined measurement semantics:

- Which physical variable matters?
- At what time and location?
- Through which observation boundary?
- Under what adversary model?
- With what uncertainty and failure modes?
- For which permitted decision?

2.3 The compounding future, stated carefully

The term exponential is frequently used without a specified variable, interval, or generative mechanism. This article does not assume that every dimension of artificial intelligence improves exponentially, nor that empirical scaling relationships extrapolate indefinitely. Existing studies report regular power-law relationships between loss and scale over substantial experimental ranges [3] and systematic performance gains under coordinated scaling of model parameters and training data [4]. Those studies motivate, but do not empirically establish, the institutional scenario examined here: over a relevant planning interval, the rate at which software generates and evaluates action proposals may grow faster than the capacity for bespoke human review.

Let the arrival rate of machine-proposed consequential actions be

$$\lambda_A(t) = \lambda_0 e^{g t}, g > 0,$$

over a planning interval $[0,T]$. This is a local scenario model rather than a universal forecast. Let accountable review capacity satisfy

$$\mu_R(t) \leq \mu_0 + r t.$$

On an unbounded horizon, e^{gt}/t diverges to infinity, implying a finite t^* after which $\lambda_A(t) > \mu_0 + rt \geq \mu_R(t)$. On the bounded planning interval, this crossing occurs only if $T \geq t^*$. The model therefore specifies a conditional capacity scenario, not an unconditional forecast. Let $B(t)$ denote a nonnegative unreviewed backlog. Whenever the backlog is positive, a reflected fluid approximation gives

$$\frac{dB(t)}{dt} = \lambda_A(t) - \mu_R(t),$$

and at the boundary $B(t)=0$,

$$\frac{dB(t)}{dt} = \max(0, \lambda_A(t) - \mu_R(t)).$$

Once the arrival rate persistently exceeds service capacity, backlog, latency, or review compression must increase. The result does not prove that human review becomes irrelevant. It shows that an architecture in which every consequential action depends on unstructured, non-scalable review lacks a stable capacity model. Ashby's law of requisite variety provides a complementary systems interpretation: a regulator must possess sufficient discriminatory variety to constrain the disturbances relevant to its objective [16]. Here, that variety must be expressed in machine-evaluable policy predicates and evidence states rather than supplied exclusively by ad hoc human judgment.

Physical trust is one candidate precondition when the decision depends on present human authority, object authenticity, custody, or physical context. It is not needed for every action. It is most valuable where the consequence-weighted residual risk justifies an independent physical observation.

2.4 Relation to adjacent research traditions

The proposed framework intersects several mature literatures without being reducible to any one of them. Shannon provides a theory of information transmitted through a channel [1]; Landauer establishes the physical realization costs of computation [2]; Blackwell orders statistical experiments by their usefulness across decision problems [17]; and Pearl distinguishes statistical conditioning from causal intervention [18]. Security architecture contributes complete mediation, fail-safe defaults, separation of privilege, and open design [19]. Identity and attestation standards specify how digital subjects, authenticators, and computing-state evidence are established [6-9]. Conformity assessment and security evaluation address the relationship among claims, targets of evaluation, evidence, and evaluator judgment [15, 20].

The distinct object of analysis here is the correspondence relation at an institutional action boundary: whether the physical state presupposed by policy is sufficiently evidenced at the time an action becomes consequential. The novelty claimed is therefore integrative and formal rather than mechanistic. The article introduces a common model in which physical observation is treated simultaneously as an information channel, a decision experiment, an adversarial surface, and an assurance object.

---

3. Definitions and conceptual boundaries

Precise language is necessary because identity, authentication, presence, authenticity, attestation, authorization, and trust are frequently used as substitutes for one another.

3.1 Core definitions

Physical subject. A person, object, asset, component, document, location, or bounded physical system about which a claim is made.

Digital claim. A machine-readable or human-readable assertion about a subject, context, event, or authority.

Physical state. The subset of real-world properties relevant to a decision at a specified time and boundary.

Observation. Information produced by a causal interaction with a physical subject or its environment. Causal origin alone is insufficient for assurance if the observation can be substituted, replayed, relayed, or generated within the disputed claim's compromise domain.

Physical anchor. A governed reference to a physical subject whose observation is causally coupled to the claimed referent and remains outside the disputed claim's generative or control path under the declared adversary model. An anchor is not assumed to be unforgeable; its security is an evaluated property.

Claim-path independence. The degree to which an adversary capable of manipulating a disputed digital claim cannot, through the same capability, generate an accepted observation about the relevant physical state. Independence is relative to a stated threat model and may be partial rather than absolute.

Physical evidence. Observation-derived information bound to a declared physical anchor and packaged for appraisal, with explicit scope, provenance, freshness, independence assumptions, and uncertainty.

Verification result. A verifier's bounded conclusion about whether specified evidence satisfies an appraisal policy. A verification result is not the evidence itself and is not a universal statement of truth.

Physical trust event. A time-bounded event in which physical evidence is appraised and transformed into a result consumable by a relying system.

Relying system. The institutional system that uses a verification result as one input to a decision.

Decision boundary. The point immediately before a consequential action at which defined conditions must be satisfied.

Physical trust. The ability of a relying institution to obtain and evaluate evidence about a decision-relevant physical state through a governed physical anchor and a claim-path-independent observation process, under an explicit threat model and assurance profile, before permitting a consequential action.

Assurance. Justified confidence that a specified claim is supported within declared conditions and limits. Assurance is graded, contextual, and never absolute.

3.2 What physical trust is not

Physical trust is not:

- a claim that physical systems cannot be attacked;
- a replacement for cryptography, identity proofing, access control, safety engineering, or governance;
- a general proof that an action is wise, legal, ethical, or aligned;
- a guarantee that a verified person is acting voluntarily or without coercion;
- a guarantee that an authentic component is fit for purpose;
- a guarantee that a verified command is operationally correct;
- a universal surveillance layer;
- a synonym for possession of a token;
- a claim that a single measurement establishes permanent truth.

The scope discipline matters. If a verification result establishes only that an enrolled physical object was observed at a checkpoint, the relying system must not silently reinterpret it as proof of quality, lawful custody, continuous integrity, or correct use between checkpoints.

3.3 Four separate questions

A complete decision may require four different classes of assurance:

1. **Information integrity:** Has the data changed?
2. **Digital identity and system state:** Which credential, account, device, or execution environment is involved?
3. **Physical correspondence:** Is the expected person, object, or context physically present at the relevant boundary?
4. **Institutional authorization:** Given all evidence and policy, is the action permitted?

Security architectures fail when an answer to one question is silently treated as an answer to all four.

4. Formal model of the physical trust gap

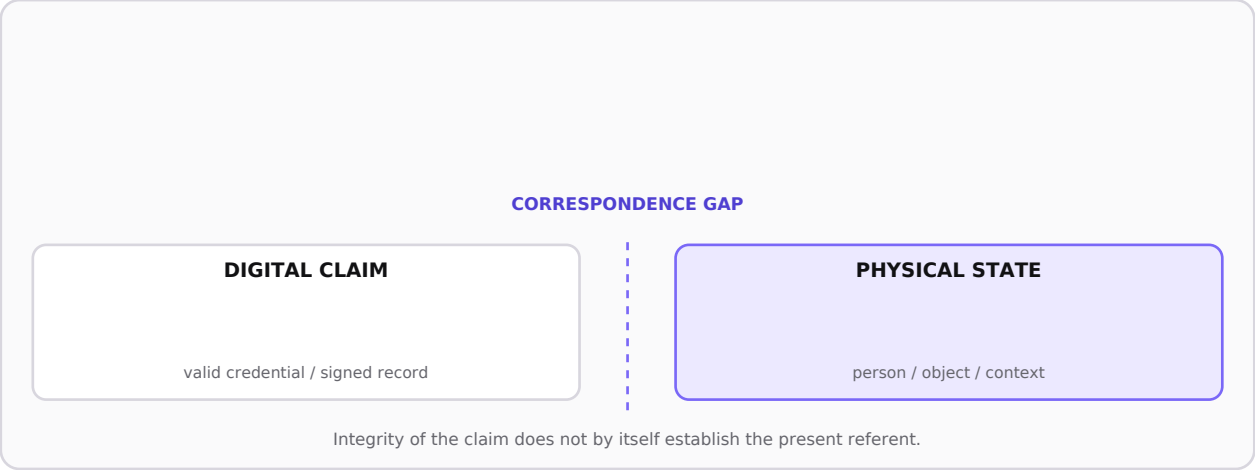


Figure 1. Representational validity and present physical correspondence occupy distinct inference domains.

4.1 State, claims, observations, and actions

Let:

- $X_t \in X$ be the decision-relevant physical state at time t;
- $D_t \in D$ be the available digital claims and credentials;
- $C_t \in C$ be institutional context, including policy, role, location class, mission phase, and consequence;
- $Y_t \in Y$ be an observation produced by a physical observation channel;
- $V_t \in V$ be a verification result;
- $A_t \in A$ be the institutional action.

The observation process is represented abstractly by

$$Y_t = H(X_t, C_t, \eta_t),$$

where H is an implementation-specific observation function and η_t represents noise, environmental variation, and unmodeled effects. This paper deliberately does not specify H.

A verifier applies an appraisal function

$$V_t = \varphi(Y_t, D_t, C_t, R_t),$$

where R_t contains reference values, lifecycle state, endorsements, or other policy-authorized material. A relying system then evaluates

$$A_t = \pi(D_t, V_t, C_t),$$

with a policy π whose output may include

$$A = \{permit, deny, escalate, record\}.$$

The separation between φ and π is substantive. The verifier appraises evidence under a declared appraisal policy; the relying institution applies authorization policy and remains accountable for the action. This distinction is consistent with the separation between evidence appraisal and relying-party policy in the IETF RATS architecture [8].

The formal analysis adopts four assumptions. **A1 (input closure):** the decision rule can depend only on variables made available through its defined interfaces. **A2 (policy separability):** at least two physical states require distinct policy outcomes. **A3 (channel specification):** the observation channel is described by a conditional distribution `p(y| x,d,c)` within a declared operating domain. **A4 (bounded inference):** conclusions are restricted to the variables and domain represented by the model. These assumptions do not assert that the observation channel is trustworthy; channel compromise is analyzed separately as part of the threat model.

4.2 Observational equivalence

Define an equivalence relation over physical states relative to a decision system's available inputs:

```
x sim_I x' <=> I(x) = I(x'),
```

where `I` maps a physical world to every input visible to the decision system. The equivalence class

```
[x]_I = x' ∈ X: I(x') = I(x)
```

contains all physical states the system cannot distinguish.

Suppose `x_a` represents an authorized person operating a valid credential and `x_u` represents an unauthorized person operating the same credential. If the system sees only the credential transcript and the transcripts are identical, then `x_a sim_I x_u`. No classifier, language model, or planning system operating only on that transcript can deterministically distinguish the two states. More compute may improve inference over differences present in the input; it cannot recover a distinction that the input channel does not contain.

Lemma 1: input-equivalence limitation

Let a possibly randomized decision system be represented by a Markov kernel `q(a| I)` over actions. If two physical states `x_1` and `x_2` induce the same accessible input, `I(x_1)=I(x_2)`, then

```
q(a| x_1)=q(a| I(x_1))=q(a| I(x_2))=q(a| x_2)
```

for every action `a`. Thus the complete output distributions are equal. If policy requires disjoint admissible action sets `A_1` and `A_2` in the two states, no rule operating only on `I` can select an admissible action with probability one in both states.

Proof. By A1, `X` influences the output only through `I`. Equality of `I(x_1)` and `I(x_2)` therefore induces equality of the conditional action kernels. If `A_1∩A_2=∅`, any common output distribution must assign positive error probability to at least one state unless it assigns zero mass to both admissible sets, which is itself erroneous. □

The lemma is an identifiability result, not a bound on intelligence or computational complexity. Additional computation may exploit distinctions encoded in `I`; it cannot make the output state-dependent when the relevant states induce the same accessible information distribution.

Corollary 1: same-domain evidence does not close the gap

Let D denote the disputed digital claim and let Z denote any additional credential, record, telemetry stream, attestation result, or inferred feature available to the decision system. Suppose that, under an adversary capability a^* and two policy-separable physical states x_1 and x_2,

$$p(d,z|\ x_1,a^*) = p(d,z|\ x_2,a^*).$$

Then every decision rule whose accessible input is (D,Z) has the same conditional output distribution in the two states and cannot select the correct policy outcome with probability one in both.

Proof. For a discrete joint input, the action distribution in physical state x_i is

$$p(a|\ x_i,a^*) = \Sigma_{(d,z)}\ q(a|\ d,z)p(d,z|\ x_i,a^*).$$

Equality of the joint input distributions therefore implies equality of the action distributions for every a; the general case follows by integration. This is the distributional form of Lemma 1 and holds irrespective of the number, complexity, or apparent consistency of the added digital artifacts. □

The corollary identifies the property a physical trust control must add. The output of a verifier will ultimately be digital, but its evidentiary source must break the relevant equivalence class under the declared adversary. A second assertion produced inside the same compromised account, device, database, or administrative domain may corroborate a narrative without establishing present physical correspondence. The security value of a physical anchor lies in creating a differently rooted observation path, not in attaching a physical label to another self-referential claim.

4.3 Observation as posterior refinement

Before a physical observation, a digital claim D leaves a posterior distribution over possible states $p(X|\ D)$. After an observation Y, the posterior becomes

$$p(X|\ D,Y) = (p(Y|\ X,D)p(X|\ D))/(p(Y|\ D)).$$

The average reduction in uncertainty is the conditional mutual information

$$I(X;Y|\ D) = H(X|\ D) - H(X|\ D,Y) >= 0.$$

Proposition 1: expected uncertainty cannot increase under conditioning

For discrete X with finite conditional entropy, adding an observation cannot increase expected conditional entropy:

$$H(X|\ D,Y) <= H(X|\ D).$$

Proof. Conditional mutual information is a conditional Kullback-Leibler divergence averaged over D and is therefore non-negative. Rearranging its entropy identity yields the inequality. □

The practical value depends on $I(X;Y|\ D)$, not on the mere existence of a physical signal. A noisy, spoofable, stale, redundant, or context-free observation may contribute negligible information. Moreover, information gain alone does not imply decision value: the observation may resolve

variables that do not alter the optimal action. In an adversarial setting, mutual information estimated under a routine distribution is also insufficient: the observation must preserve decision-relevant distinguishability under the declared attack capabilities.

Proposition 2: decision-risk monotonicity under Blackwell dominance

Let observations Y_1 and Y_2 be statistical experiments about X. For discrete observation spaces, if Y_2 Blackwell-dominates Y_1, then there exists a stochastic kernel K such that

$$p(y_1| x)=\Sigma_(y_2) K(y_1| y_2)p(y_2| x).$$

That is, Y_1 can be generated by garbling Y_2. For general observation spaces, the corresponding relation is expressed by integration with respect to the Markov kernel. For every prior over X, action space, and bounded loss function, the minimum Bayes risk achievable with Y_2 is no greater than the minimum Bayes risk achievable with Y_1 [17].

Proof sketch. Any decision rule based on Y_1 can be reproduced from Y_2 by first applying K and then applying the Y_1 rule. The feasible set of decision procedures available with Y_2 therefore contains all procedures available with Y_1. Minimization over a superset cannot yield higher risk. □

This result gives a precise meaning to the claim that one observation architecture is more decision-informative than another. It does not establish that Y_2 is cheaper, more private, more available, or harder to attack; those attributes enter the net decision criterion separately.

4.4 Decision risk and value of verification

Let $L(a,x)$ be the loss incurred by action a in physical state x. Without physical evidence, the Bayes action is

$$a_D^* = arg min_a E[L(a,X)| D].$$

With observation Y, the Bayes action becomes

$$a_D,Y^* = arg min_a E[L(a,X)| D,Y].$$

The net expected value of the physical observation is

$$VOI_phys = E[L(a_D^*,X)] - E[L(a_D,Y^*,X)] - K_Y,$$

where the expectations integrate over the institution's declared distributions for D, Y, and X, and K_Y includes acquisition cost, delay, availability loss, privacy burden, and integration cost. This is a value-of-information expression for an imperfect observation, not the classical value of perfect information. A physical control is justified when $VOI_phys>0$ under the institution's risk model. The framework therefore rejects both extremes: physical verification everywhere and physical verification nowhere.

This economic criterion must not be interpreted as permission to act on a missing prerequisite. If institutional policy requires a current physical fact and no alternative control establishes it at acceptable residual risk, the available choices are to obtain adequate evidence, redesign or constrain the action, or deny it. A negative value for one proposed observation method can justify choosing another method or eliminating the workflow; it cannot make the unobserved physical condition true.

4.5 Consequence-weighted verification

Let action class j have annual attempt rate λ_j, unauthorized acceptance probability p_j, conditional consequence c_j expressed in a declared loss unit, and recovery factor $r_j \in [0,1]$. A simplified annualized risk in that loss unit is

$$R = \Sigma_j \lambda_j p_j c_j (1-r_j).$$

If a physical trust control changes the false-accept probability to p'_j while adding annualized availability and operational cost k_j, monetized or utility-normalized in the same loss unit, then the approximate net risk reduction is

$$\Delta R = \Sigma_j \lambda_j (p_j-p'_j)c_j(1-r_j) - \Sigma_j k_j.$$

This expression clarifies why verification should be adaptive. An ordinary low-value action may not justify a physical checkpoint. An irreversible transfer, command-authority change, privileged recovery, or safety-critical activation may justify multiple independent checks. Assurance should scale with consequence, reversibility, observability, and adversary capability.

---

5. Computational capability and the physical observation boundary

5.1 The physical-causality thesis

The statement "AGI cannot change physics" conflates physical intervention with epistemic access. A sufficiently capable system may change physical states through machines, people, markets, laboratories, or infrastructure. It may discover improved theories, construct new instruments, exploit side channels, or manipulate the observation process. The invariant required by Lemma 1 is narrower:

> A decision process cannot condition on a physical distinction that has produced no decision-accessible difference through a causal observation channel.

Call this the **physical-causality thesis**. Within an ordinary causal model, epistemic access to a state variable requires that the variable, or a causally related proxy, affect an accessible observation. A proxy qualifies for assurance only if its relationship to the state and its distinguishability survive the declared adversary model; causal ancestry alone is not enough. Telemetry emitted by a compromised device, for example, may be physically caused yet remain controllable within the disputed claim path. If the current physical state is policy-relevant, a governed information path from that state to the decision process is therefore necessary. Pearl's distinction between observation and intervention is relevant here: conditioning on correlated data and actively establishing a measurement relation are not semantically interchangeable [18].

An intelligent system can infer latent variables from correlated evidence: a person's likely location from schedules, a machine state from telemetry, or component authenticity from supply-chain records. Such inference may be decision-relevant. Prediction and verification nevertheless have different assurance semantics. A prediction is a posterior distribution conditional on an observation

model. Verification, as defined here, is an appraisal that specified evidence satisfies a declared claim profile at a defined control point. A high posterior derived from records controlled by the same adversary does not become independent evidence of current correspondence. Neither prediction nor verification is infallible; the distinction concerns provenance, adversarial independence, temporal scope, and claim semantics rather than an intrinsic hierarchy of truth.

5.2 Intelligence changes both sides of the equation

Advanced AI may strengthen defenders through anomaly detection, policy analysis, simulation, and test generation. It may also strengthen attackers through scalable social interaction, vulnerability discovery, coordination, and adaptive search. A durable assurance argument should therefore avoid assumptions that reduce adversary capability to contemporary labor cost or cognitive bandwidth.

Let attacker search capacity be `B_A(t)` and defender review capacity be `B_D(t)`. If both are augmented by AI, the security objective is not to win a permanent intelligence race. It is to design boundaries whose security properties depend on constrained, independently observable facts and explicit policy, not solely on an adversary's inability to generate a convincing digital narrative.

This asymmetry motivates four design requirements:

1. **Assume synthesis.** Digital appearance and linguistic coherence are not sufficient evidence of authority.
2. **Bound action.** Capability to produce an instruction is not permission to execute it.
3. **Preserve an accountable root.** High-consequence action should remain connected to evidence that the institution can appraise and govern.
4. **Keep the anchor outside the claim.** Evidence used to establish present physical correspondence must not be generated solely by the disputed claim or by a compromise domain that the declared adversary already controls.

5.3 The observation channel is itself part of the threat model

Physical grounding does not eliminate adversarial uncertainty. An adversary may attempt to spoof, substitute, replay, relay, coerce, occlude, damage, clone, or re-contextualize the physical evidence source. The verifier, lifecycle database, policy engine, or integration path may be compromised. Enrollment may bind the wrong reference. A legitimate person may be compelled. An authentic component may be defective.

Accordingly, the physical-causality thesis establishes necessity, not sufficiency. A causal channel is required when physical distinction matters; the channel must then be engineered and evaluated against relevant attacks. The resulting assurance claim must state what was observed, what was not established, and how long the result remains valid.

6. A mechanism-neutral reference architecture

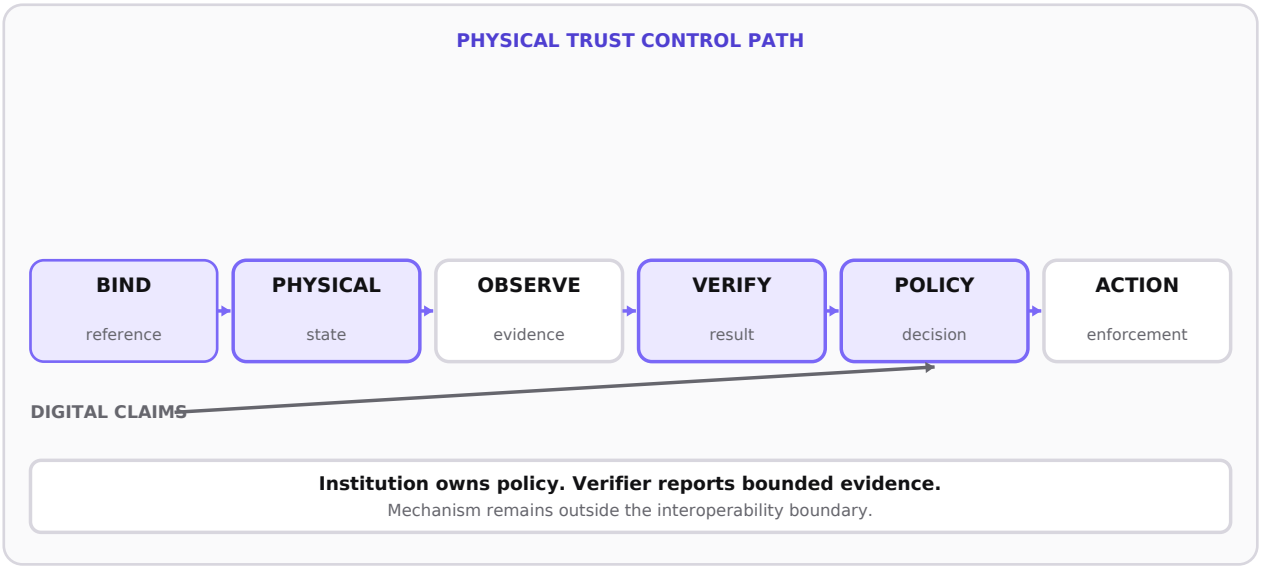


Figure 2. Logical control path separating governed binding, physical observation, evidence appraisal, institutional policy, and action enforcement.

The purpose of a public reference architecture is to make integration and assurance discussable without revealing the protected implementation that produces physical evidence. The architecture separates eight roles.

6.1 Roles

1. Physical subject

The person, object, component, asset, or bounded environment about which a physical claim is made.

2. Claiming context

The workflow that supplies the claim, requested action, role, mission or business context, and consequence class. Examples include an ERP release, a maintenance action, a treasury instruction, an agent execution request, or a command transition.

3. Binding and lifecycle authority

The governed function that establishes the initial correspondence between a physical subject and its scoped reference, and that manages activation, transfer, suspension, revocation, recovery, and retirement. A repeatable observation without a valid binding can consistently verify the wrong referent.

4. Observation subsystem

The implementation-specific subsystem that interacts with the physical subject and produces protected measurement-derived evidence. Its mechanism is outside the scope of this paper and should remain outside the core interoperability standard.

5. Physical verifier

The component or service that appraises evidence against reference material, freshness requirements, lifecycle status, and an assurance profile. It produces a bounded verification result.

6. Policy authority

The institutional owner of rules that map verification results and context to permitted outcomes. The policy authority determines, for example, whether one high-assurance human verification is sufficient, whether two-person control is required, or whether a component verification must be paired with an approved work order.

7. Relying system

The system that enforces the decision. It may be an identity platform, transaction engine, ERP system, maintenance system, access controller, agent orchestrator, mission system, or safety controller.

8. Evidence and audit service

The service that stores the minimum necessary event record, policy reference, result, lifecycle state, and audit metadata. Raw physical evidence should not be retained by default when a narrower result satisfies the relying purpose.

6.2 Logical flow

The reference flow is:

```
governed binding + current physical state → observation → evidence appraisal →
       verification result → institutional policy → controlled action.
```

Digital claims and system context enter at multiple points, but they do not replace the physical observation. The binding and observation path must remain distinguishable from the disputed claim path under the declared adversary. The final decision remains institution-owned.

An abstract event can be represented as

$$E = (s, q, c, \tau, \alpha, o, v, p, \ell\ell, z),$$

where:

- s is a privacy-preserving subject or object reference;
- q is the claim class being evaluated;
- c is the declared context and consequence class;
- τ is freshness and temporal scope;
- α is the assurance profile;
- o is the bounded outcome;
- v is the verifier reference and version;
- p is the appraisal-policy reference and version;
- $\ell\ell$ is lifecycle status;
- z contains optional audit commitments or integrity protection.

The public schema should not require disclosure of raw measurement features or protected mechanism parameters. Interoperability is achieved at the result and policy boundary.

6.3 Evidence, result, and decision must remain separate

Conflating these layers creates systemic risk.

- **Evidence** is observation-derived material submitted for appraisal.
- **Result** is a verifier's bounded conclusion under a named profile.

- **Decision** is the relying institution's policy outcome.

A verifier may produce `verified-under-profile-human-authority` while the relying system still denies the action because the work order has expired. Conversely, a relying system must not reinterpret `object-observed` as `object-safe-to-operate`. The standard should make such semantic overreach detectable.

6.4 Freshness and context

A result about physical state decays differently from a credential. The relevant subject may move, the object may be substituted, custody may change, or the environment may be altered. A physical trust event therefore requires a declared temporal and contextual scope.

Let `F(t)` represent an illustrative freshness factor for an observation made at time `t_0`. Under a constant-hazard assumption, a generic survival-style decay model is

```
F(t) = F_0 exp[-λ (t-t_0)], t >= t_0,
```

where λ is a profile-specific hazard parameter that may depend on subject mobility, custody controls, environment, and threat. This expression is an uncalibrated example, not a general law of assurance. Some events expire almost immediately; others remain valid until a controlled lifecycle transition. The standard should require an explicit, empirically justified freshness rule rather than impose one universal duration.

6.5 Composability

Physical trust is most useful when it composes with other evidence:

```
Decision assurance = f(identity, device state, physical evidence, policy, context).
```

Composition must not be treated as naive multiplication of confidence scores. Evidence sources may share failure modes. For example, a device credential and a physical verification result may depend on the same compromised enrollment authority. Assurance cases must document dependencies, common-cause failures, and trust roots.

---

7. Threat model and security properties

A standard that defines only the successful flow is incomplete. Physical trust must be specified relative to adversarial capabilities and institutional failure.

7.1 Threat classes

T1. Credential-control substitution

An attacker controls a valid account, token, session, or signing key but is not the physical subject expected by policy.

T2. Object substitution

A digital record refers to an expected object while a different physical object is presented, installed, transferred, or serviced.

T3. Replay

Previously valid evidence or a previous verification result is reused outside its freshness window or context.

T4. Relay and context displacement

A valid interaction is proxied from another location, subject, device, or operational context so that a relying system accepts the wrong physical boundary.

T5. Observation spoofing

The observation subsystem is exposed to crafted physical or environmental inputs intended to cause false acceptance.

T6. Enrollment corruption

The wrong subject or object is enrolled, reference material is altered, or lifecycle authority is compromised before routine verification begins.

T7. Verifier or policy compromise

The verifier, reference service, policy engine, or relying integration is modified to issue or accept unauthorized results.

T8. Insider misuse and coercion

An enrolled and correctly verified person acts outside legitimate intent, under coercion, or in collusion with another party.

T9. Denial and degradation

An attacker prevents verification, increases false rejects, blocks lifecycle data, or forces insecure fallback.

T10. Privacy extraction

Verification data is used to track people, infer sensitive behavior, correlate activity across domains, or reconstruct protected physical features beyond the relying purpose.

7.2 Adversary parameterization

A threat label is insufficient without an associated capability model. Let an adversary class be represented by

$$\Gamma=(K,Q,B,T,P),$$

where K denotes prior knowledge, Q interface and query access, B physical and computational budget, T available time, and P privileges or positions within the institutional system. Evaluation should state which components of Γ are fixed, measured, or excluded. An attack experiment against profile P_i estimates a conditional security error

$$\varepsilon_{sec}(P_i,\Gamma,\Omega)= P(\text{PERMIT}\mid \text{policy predicate false},\Gamma,\Omega),$$

within operating domain Ω. This quantity is not an intrinsic constant of a mechanism. It depends on the adversary, environment, integration, threshold, and lifecycle state. A conformance claim should therefore bind any reported security error to a versioned profile and test protocol.

The model also distinguishes false acceptance under the declared routine distribution from adversarial acceptance after adaptive search. The latter permits the attack strategy to depend on

previous observations and may be orders of magnitude larger than routine FAR. Reporting one as a substitute for the other is a category error.

7.3 Required security properties

A conformant assurance profile should specify which of the following properties it addresses:

- **Subject binding:** the result refers to the intended enrolled subject or object class.
- **Claim-path independence:** under the declared adversary, control of the disputed digital claim is insufficient by itself to generate an accepted physical observation.
- **Context binding:** the result is valid only for the declared workflow, relying party, consequence class, or operational boundary.
- **Freshness:** the result cannot be accepted beyond its time or event scope.
- **Anti-replay:** copied evidence or results cannot be reused as a new event.
- **Lifecycle validity:** revoked, retired, transferred, or superseded subjects are handled explicitly.
- **Policy integrity:** the rules applied to evidence are identified, versioned, and protected.
- **Result integrity:** relying systems can detect unauthorized modification of a result.
- **Failure safety:** unavailable or ambiguous evidence fails according to declared policy, not an implicit bypass.
- **Auditability:** an independent reviewer can determine which profile, policy, and lifecycle state governed a decision.
- **Minimal disclosure:** only data necessary for the relying purpose is exposed or retained.
- **Crypto agility:** cryptographic dependencies can transition without redefining the physical claim semantics.
- **Operational independence:** high-assurance profiles identify and reduce common-cause failures between digital credentials and physical verification.

7.4 Non-goals and residual risk

Physical verification cannot, by itself, establish mental intent, absence of coercion, legal authority, engineering fitness, or continuous custody between discrete observations. The institution must specify compensating controls. Examples include dual control, approved work orders, background checks, separation of duties, environmental monitoring, safety interlocks, and post-event review.

The standard should require a **residual-risk statement** for every conformance profile. A result without declared limitations invites unsafe semantic expansion.

8. Toward a Physical Trust Assurance Standard

Figure 3. Proposed disclosure and standardization boundary: public semantics, controlled assurance evidence, and a protected implementation core.

8.1 Why standardize

The functional value of standardization lies in enabling independent actors to specify, evaluate, procure, integrate, and compare systems without requiring access to a common proprietary mechanism.

A future **Physical Trust Assurance Standard (PTAS)** should make it possible for:

- an institution to state the physical claim required by a workflow;
- a verifier to return a result with unambiguous semantics;
- a relying system to consume that result without knowing the protected observation mechanism;
- an evaluator to test conformance against a declared assurance profile;
- multiple implementations to compete while preserving interoperability;
- policy owners to increase assurance as consequence increases;
- cryptographic and software components to evolve without changing the meaning of the physical claim.

ISO/IEC 17007:2026 describes drafting requirements for normative documents intended for conformity assessment, including clarity and demonstrability [15]. ISO/IEC 15408-1:2026 similarly separates a target of evaluation, its security properties, the evaluation context, and the audience for assurance claims [20]. PTAS should adopt this claim-evidence discipline from inception. Requirements that cannot be connected to inspection, test, analysis, or audit may supply vocabulary, but they do not supply assurance.

8.2 Standardization boundary

The central design decision is what **not** to standardize.

PTAS should standardize:

- terminology and claim classes;
- event and result semantics;
- assurance-profile requirements;
- context and freshness expression;
- lifecycle and revocation behavior;

- policy outcomes and error states;
- privacy and data-minimization requirements;
- audit commitments;
- crypto-agility expectations;
- conformance declarations and test categories;
- security and residual-risk documentation.

PTAS should not require public disclosure of:

- materials or physical construction;
- sensing modality or sensor configuration;
- protected feature representation;
- signal-processing or classification details;
- enrollment encoding;
- anti-cloning and anti-tamper mechanisms;
- manufacturing tolerances;
- secure-processing topology;
- implementation-specific thresholds that would weaken security.

This division follows a mature pattern: standardize the contract, not every means of satisfying it.

8.3 Design principles

P1. Claim specificity

Every result must identify the claim it supports. Generic labels such as `trusted` or `secure` are non-conformant because they conceal scope.

P2. Mechanism neutrality

The standard defines observable properties and evaluation requirements, not a mandatory physical implementation. Mechanism neutrality at the interface does not imply that implementations are interchangeable or security-equivalent; their physical binding, independence, operating domain, attack resistance, and evaluation evidence may differ materially.

P3. Claim-path independence

A conformant physical result must identify why its evidentiary source is not reducible to manipulation of the disputed claim under the declared adversary model. Adding factors inside one compromise domain is not equivalent to adding an independently rooted observation.

P4. Consequence proportionality

Assurance requirements increase with consequence, irreversibility, adversary capability, and recovery difficulty.

P5. Explicit uncertainty

False acceptance, false rejection, environmental limits, coverage, and unresolved threats must be stated rather than implied away.

P6. Separation of verification and authorization

The verifier reports. The institution decides.

P7. Context and freshness by construction

No result is context-free or indefinitely valid unless a profile explicitly and defensibly states so.

P8. Minimal disclosure

Verification should prove the narrowest fact necessary for the decision and avoid exposing reusable physical features or unrelated personal data.

P9. Composability without transitive overclaim

The standard supports composition with digital identity, attestation, and policy but forbids stronger claims than the combined evidence justifies.

P10. Lifecycle completeness

Enrollment, activation, transfer, suspension, revocation, recovery, retirement, and destruction are part of assurance.

P11. Failure-aware operation

Profiles must define behavior under outage, ambiguity, environmental degradation, and loss of upstream services.

P12. Cryptographic agility

The physical claim and policy semantics must survive algorithm transition. NIST's crypto-agility guidance emphasizes that protocols and systems should accommodate new algorithms and parameter sizes without brittle assumptions [14].

P13. Evaluability

Every normative requirement must map to an inspection, test, analysis, audit, or evidence artifact.

8.4 Candidate assurance levels

The following levels are a research proposal, not a certification scheme. They illustrate how consequence-sensitive profiles might be structured. The labels are not scalar security scores: a higher level evaluated in the wrong operating domain does not dominate a lower level validated for the actual deployment. Any future level assignment would require a profile-specific protection claim, evaluation target, and evidence package.

Level	Intended use	Minimum assurance characteristics	Illustrative decision
PTA-L0: Unverified	Baseline comparison	Digital claim only; no PTAS physical result	Informational display
PTA-L1: Contextual	Low-consequence workflow	Physical event tied to a declared anchor, subject class, context, and freshness window	Record or low-risk permit
PTA-L2: Controlled	Material enterprise action	Managed binding and lifecycle, documented claim-path separation, anti-replay, policy binding, audited verifier, quantified operating envelope	Release, access, or custody handoff
PTA-L3: High Assurance	High-value or safety-relevant action	Strong separation of duties, protected verification path, adversarial evaluation, secure failure behavior, independent assurance case	Privileged recovery, critical maintenance, exceptional transaction
PTA-L4: Mission Critical	Severe or strategic consequence	Redundant or diverse evidence, multi-party authority where appropriate, offline-continuity profile, independent conformity assessment, mission-specific red-team evidence	Command transition, strategic asset control, irreversible mission action

Higher levels should not merely require more factors. They should require stronger independence, lifecycle governance, failure containment, and evidence about the entire decision path.

8.5 Candidate claim profiles

Human Authority Profile

Establishes that the expected enrolled human authority satisfied a defined physical checkpoint for a specified action and context. It does not establish intent, competence, or freedom from coercion.

Object Authenticity Profile

Establishes that an observed physical object satisfies the enrolled reference and context requirements of a named profile. It does not establish fitness, lawful ownership, or uninterrupted custody unless separately claimed.

Custody Transition Profile

Combines object evidence, accountable parties, location or facility class, and transfer policy at a discrete handoff. It does not prove conditions between checkpoints.

Command Authority Profile

Requires named physical authority conditions before a command, credential transition, privileged recovery, or high-consequence configuration change is released.

Autonomous Action Profile

Provides a policy-consumable physical prerequisite before an autonomous system executes a consequential action. It does not establish that the model's proposed action is correct or aligned.

Physical Access Profile

Establishes a physical trust result as one input to entry or activation. It does not replace site security, supervision, or safety procedure.

8.6 Candidate result semantics

A relying system should receive explicit states rather than a single boolean:

- `SATISFIED`: evidence met the named profile within scope;
- `NOT_SATISFIED`: evidence failed a required condition;
- `INDETERMINATE`: evidence could not support a conclusion;
- `EXPIRED`: the freshness or lifecycle window has ended;
- `REVOKED`: the subject, object, verifier, or reference is no longer valid;
- `POLICY_MISMATCH`: the result does not satisfy the relying context;
- `SYSTEM_ERROR`: the verification path failed operationally.

Boolean interfaces tend to convert uncertainty into false certainty. `INDETERMINATE` must not be silently mapped to `SATISFIED`.

8.7 Conformance statement

A conformant implementation should declare:

1. standard version;
2. claim profile and assurance level;
3. supported subject classes and environments;
4. declared adversary model;
5. physical binding and lifecycle model;
6. claim-path-independence rationale and known common-cause dependencies;
7. freshness behavior;
8. cryptographic dependencies and transition plan;

9 privacy and retention policy;
10 evaluation method and sample domain;
11 residual-risk and non-goal statement;
12 implementation and evaluator identity;
13 deviations or extensions.

This is the minimum grammar of credible assurance.

9. Relationship to existing trust architectures

PTAS should be additive. It should reuse established concepts and avoid inventing new terminology where existing standards already provide interoperable semantics.

9.1 Digital identity

NIST SP 800-63-4 defines requirements for identity proofing, enrollment, authentication, authenticators, and federation in network interactions [7]. Physical trust can complement these functions at selected action boundaries. A person may be correctly proofed and authenticated while policy still requires evidence of present physical authority before a high-consequence act.

The distinction is temporal and contextual:

- identity proofing asks who was enrolled;
- authentication asks whether a claimant controls an accepted authenticator;
- physical authority verification asks whether the expected physical authority satisfies a defined checkpoint now;
- authorization asks whether the action is permitted in context.

9.2 Zero trust

NIST SP 800-207 rejects implicit trust based on network location and emphasizes resource-focused, granular authentication and authorization [6]. PTAS extends the same logic to selected physical claims: physical location, possession, or appearance should not be trusted implicitly; they should be evaluated as explicit evidence where relevant.

Physical trust is therefore not a return to perimeter security. It is a way to make physical prerequisites consumable by zero-trust policy engines.

9.3 Remote attestation

The IETF RATS architecture separates an Attester, Evidence, Verifier, Attestation Result, and Relying Party [8]. PTAS can reuse this role-separation logic for claims about bounded physical subjects and contexts beyond the computing environments ordinarily appraised in remote attestation. This is an architectural adaptation, not an assertion that PTAS changes or extends RFC 9334. The frameworks should interoperate but not be conflated. A device may attest that it is running approved code while the physical subject presented to the device remains wrong.

Where possible, PTAS result formats should reuse or profile existing attestation containers rather than create incompatible transport layers.

9.4 Verifiable credentials

W3C Verifiable Credentials 2.0 provides a data model for tamper-evident issuer claims and presentations [9]. A PTAS result could be conveyed as or referenced by a verifiable credential, provided the semantics remain bounded. The credential would prove who issued the result and protect it from tampering; the physical verifier and assurance profile would establish what the result means.

9.5 Content provenance

C2PA demonstrates how an ecosystem can standardize provenance claims, manifests, bindings, trust lists, and validation while distinguishing integrity from factual truth [10]. PTAS addresses a different object - decision-relevant physical state - but should adopt the same discipline of explicit scope and transparent non-goals.

9.6 Post-quantum cryptography and crypto agility

NIST FIPS 203 specifies a module-lattice-based key-encapsulation mechanism, while FIPS 204 and FIPS 205 specify post-quantum digital-signature standards [11-13]. When correctly integrated, these primitives support quantum-resistant key establishment and digital-signature security under their stated assumptions. They do not secure an entire communications system by themselves and do not establish physical correspondence.

PTAS must be cryptographically agile so that its results can be protected by current or future algorithms without changing the underlying claim semantics. The relationship is complementary:

```
PQC: support quantum-resistant key establishment and signatures
```

```
PTAS: define the physical claim entering that channel
```

9.7 AI risk management

NIST AI RMF 1.0 organizes AI risk work around governance, mapping, measurement, and management [5]. PTAS can provide one control family for AI-enabled workflows in which physical authority or object authenticity is a necessary condition for action. It does not evaluate model validity, bias, robustness, alignment, or the correctness of generated plans.

---

10. Application profiles

The following examples illustrate control boundaries. They are not statements of present deployment.

10.1 Consequential AI execution

Scenario

An autonomous or semi-autonomous system prepares an action that moves funds, changes privileged access, operates connected equipment, releases sensitive data, or alters a physical process.

Physical trust boundary

Immediately before execution, policy requires a named class of accountable physical authority. The verifier returns a profile-scoped result. The orchestration system evaluates the result together with model output, transaction context, and institutional policy.

Value

The control preserves a distinction between machine capability and institutional permission. It also makes escalation requirements machine-consumable as action volume grows.

Limit

The result does not prove that the model is aligned, that the proposed action is correct, or that the verified authority has made a wise decision.

10.2 High-value financial authority

Scenario

A transaction or administrative change exceeds normal value, reversibility, or anomaly thresholds. Ordinary digital authentication has succeeded.

Physical trust boundary

One or more designated officers must satisfy a Human Authority Profile before release. Policy may require separation of duties and independently governed verification paths.

Value

Credential compromise alone becomes insufficient to release the action. The control creates a physically grounded checkpoint for exceptional authority.

Limit

It does not establish beneficial ownership, sanctions compliance, economic legitimacy, or absence of coercion.

10.3 Component and supply-chain assurance

Scenario

A high-consequence component moves from supplier through receiving, controlled storage, maintenance, and installation. Digital records identify the expected item.

Physical trust boundary

An Object Authenticity Profile is evaluated at custody transfer and installation. Results are conveyed to ERP, MRO, asset-management, or quality systems. The institution may require the object result to be paired with a Human Authority Profile and approved work order.

Value

The digital lifecycle record remains connected to the physical item at the moments when substitution would become operational consequence.

Limit

Discrete verification does not prove continuous custody, engineering quality, or environmental compliance between checkpoints.

10.4 Critical infrastructure maintenance

Scenario

A technician is scheduled to replace a critical component or change a protected configuration. Accounts, devices, and work orders are digitally valid.

Physical trust boundary

Policy requires the expected technician and component to be verified in the declared facility and maintenance context before the control system accepts the change.

Value

The system gains evidence about the physical person-object pairing at the operational boundary, reducing dependence on credentials and records alone.

Limit

The verification does not replace lockout procedures, safety supervision, technical qualification, or change validation.

---

11. Evaluation science

A physical trust standard must be measurable. Evaluation should combine statistical performance, adversarial testing, systems assurance, human factors, privacy analysis, and operational evidence.

11.1 Statistical decision performance

For a binary claim under a specified operating domain, define:

- **True acceptance rate (TAR):** probability that an authorized condition is accepted;
- **False rejection rate (FRR):** probability that an authorized condition is rejected;
- **False acceptance rate (FAR):** probability that an unauthorized condition is accepted;
- **True rejection rate (TRR):** probability that an unauthorized condition is rejected.

If S is a verification score and threshold θ determines acceptance, then

$$FAR(\theta)=P(S>=\theta|\ U),$$

$$FRR(\theta)=P(S<\theta|\ A),$$

where U and A denote unauthorized and authorized conditions under the declared test distribution.

A reported FAR or FRR is not interpretable without the population, environment, sample size, confidence interval, attack model, and threshold-selection procedure. PTAS evaluations should report binomial confidence bounds and distinguish routine-condition testing from adaptive adversarial testing.

If zero false accepts are observed in n independent unauthorized trials, the "rule of three" gives an approximate 95 percent upper confidence bound of $3/n$ on the event probability [21]. Thus zero observed failures in 300 trials supports only an upper bound near 1 percent, not evidence of impossibility. High-assurance claims require sample sizes and test designs proportionate to their

consequences. More generally, performance testing should distinguish algorithmic, scenario, and operational evaluations and should report uncertainty rather than only point estimates [22].

11.2 Decision-weighted performance

Thresholds should be selected using consequence, not equal error rate alone. Let C_FA and C_FR be false-accept and false-reject costs, and let P(U) and P(A) be prior probabilities. The expected decision cost is

$$J(\theta)=C_{FA}P(U)FAR(\theta) +C_{FR}P(A)FRR(\theta).$$

A mission-critical command may rationally accept higher friction and false rejection to reduce false acceptance. An emergency workflow may require a different policy because denial itself carries severe consequence. PTAS should allow context-specific thresholds while requiring governance, versioning, and test evidence.

11.3 Environmental validity

Evaluations must define an operating domain Ω that includes relevant ranges of environment, handling, device variation, subject variation, age, wear, maintenance state, and connectivity. Claims outside Ω should return INDETERMINATE or another explicit out-of-domain result rather than silently extrapolate.

Let p_test(x) be the evaluation distribution and p_deploy(x) the deployment distribution. A low measured error under p_test does not establish low error under distribution shift. The standard should require:

- operating-domain declaration;
- out-of-domain detection or procedural controls;
- monitoring for drift;
- re-evaluation after material design, software, manufacturing, or policy change;
- separation between development and independent test evidence.

11.4 Adversarial evaluation

Routine accuracy testing does not measure security. Adversarial evaluation should include, where relevant:

- replay and relay attempts;
- subject or object substitution;
- crafted physical presentation;
- environmental manipulation;
- verifier and policy tampering;
- lifecycle inconsistency;
- downgrade and fallback attacks;
- integration-message modification;
- common-cause compromise;
- privacy inference and cross-context correlation;
- denial and recovery behavior.

The test authority should document attacker knowledge, access, budget, iteration count, and whether the test was black-box, gray-box, or white-box. Security claims should expire or be reviewed as attacker methods and implementations change.

11.5 End-to-end assurance

The correct unit of evaluation is the decision path, not only the physical verifier. Define component compromise events:

- E_o: observation subsystem fails;
- E_v: verifier fails;
- E_r: reference or lifecycle service fails;
- E_p: policy fails;
- E_i: integration or relying enforcement fails.

If the listed events cover every modeled path to an unsafe or incorrect system decision, then system failure satisfies

$$E_{sys} \subseteq E_o \cup E_v \cup E_r \cup E_p \cup E_i.$$

The inclusion may be strict because a component failure can be detected and contained without producing system-level failure. Without justified independence assumptions,

$$P(E_{sys}) \leq \Sigma_k P(E_k)$$

by the union bound. This conservative expression is often more honest than multiplying component reliabilities. Correlated failures - for example, one administrator controlling enrollment, policy, and audit - can dominate the system even when each technical component performs well.

High-assurance profiles should therefore evaluate:

- separation of duties;
- trust-root diversity;
- secure administration;
- update and rollback behavior;
- policy review and change control;
- audit integrity;
- offline and degraded modes;
- incident response and revocation latency.

11.6 Availability and human factors

A control that operators routinely bypass is not high assurance. Evaluation should measure:

- end-to-end latency distributions;
- completion and abandonment rates;
- failure recovery time;
- accessibility and usability across expected users;
- error under stress, protective equipment, poor connectivity, or constrained environments;
- frequency and governance of overrides;
- susceptibility to habituation and social engineering;
- training and operational burden.

The relationship between security and availability must be explicit. A secure failure policy may deny an action when verification is unavailable. For emergency or mission-continuity contexts, a governed break-glass path may be necessary. That path becomes part of the assurance boundary and must be tested as rigorously as the nominal path.

11.7 Privacy metrics

Physical evidence can be sensitive even when it is not conventionally biometric. A privacy evaluation should consider:

- whether results are linkable across relying systems;
- whether retained data can reconstruct protected physical features;
- whether verification reveals location, routine, role, health, or association;
- whether raw evidence is necessary after appraisal;
- whether selective disclosure or pairwise references can reduce correlation;
- retention duration and deletion verification;
- insider access and secondary-use controls.

Data minimization should be testable. If a relying system needs only `Human Authority Profile satisfied at time t for action q`, the default interface should not disclose reusable measurement data.

12. Governance and standardization pathway

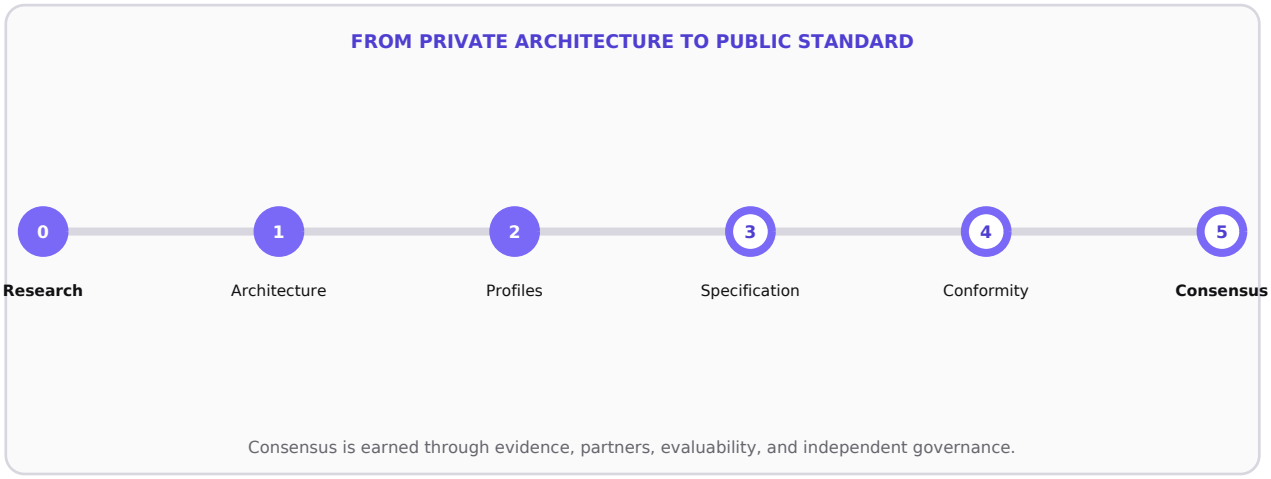


Figure 4. Evidence-dependent progression from a research framework to an externally governed consensus specification.

An author, research group, or commercial implementer may propose a specification, but no originator can establish ecosystem consensus unilaterally. The transition from analytical framework to recognized standard should therefore be staged and evidence-dependent.

12.1 Stage 0: research framework

Publish a non-normative conceptual model, glossary, threat model, and assurance principles. Invite criticism from security engineering, safety, privacy, identity, industrial, defense, space, and standards communities.

Output: this report and related critical research.

12.2 Stage 1: reference architecture

Define role separation, result semantics, lifecycle states, error conditions, and integration patterns. Validate the architecture through a small number of institution-owned workflows.

Output: PTAS Reference Architecture 0.x and implementation-neutral event schema.

12.3 Stage 2: founding-institution profiles

Work with institutions that own consequential workflows to define claim profiles and assurance requirements. Each pilot should begin with one decision boundary, one accountable policy owner, and one explicit residual-risk statement.

Output: Human Authority, Object Authenticity, Custody Transition, Command Authority, and Autonomous Action profiles.

12.4 Stage 3: public technical specification

Publish testable normative requirements using `shall`, `should`, and `may` consistently. Define extension rules, version negotiation, deprecation, crypto agility, privacy, and conformance statements. Establish an intellectual-property policy suitable for implementers.

Output: Proposed PTAS 1.0.

12.5 Stage 4: independent conformity program

Develop test methods, evaluator qualification, evidence requirements, surveillance, recertification triggers, vulnerability handling, and public registry rules. ISO/IEC 17007 provides useful drafting guidance for normative documents intended for conformity assessment [15].

Output: PTAS Conformity Assessment Scheme and accredited or otherwise independently governed evaluation pathways.

12.6 Stage 5: external standards venue

Transfer, submit, or jointly develop relevant components through an appropriate consensus organization or sector body. A credible standard should survive participation by competitors and independent experts.

Output: consensus specification, sector profiles, and interoperable implementations.

12.7 Governance principles

The governance model should include:

- transparent versioning and change records;
- security disclosure and correction processes;
- balanced representation of implementers, relying institutions, evaluators, and affected users;
- conflict-of-interest rules;
- published extension and deprecation policy;
- privacy and civil-liberties review;
- interoperability events;
- incident-driven profile revision;
- clear separation between standard ownership and proprietary product claims.

Independent governance is necessary if the assurance meaning is to remain distinguishable from the commercial interests of any originator or reference implementer. Conformance claims should therefore depend on published semantics and reproducible evidence, not on organizational provenance.

13. Economics of a physical trust layer

Standards become infrastructure when they reduce coordination cost across organizations. The economic case for PTAS has three components.

13.1 Reduction of semantic integration cost

Without shared semantics, every verifier-relying-system integration becomes bespoke. If `m` verification implementations connect to `n` relying platforms through pairwise integrations, the potential integration surface scales as

```
O(mn).
```

With a stable result and policy interface, each implementation and relying system can integrate once to the shared contract, reducing the idealized surface toward

```
O(m+n).
```

Real deployments remain more complex, but the direction matters. Standardized result semantics can turn a protected physical mechanism into an ecosystem component rather than a closed appliance.

13.2 Option value under technological transition

Long-lived institutions face uncertainty about AI capability, cryptographic algorithms, identity architectures, and infrastructure providers. A mechanism-neutral physical trust interface creates option value: observation mechanisms, cryptographic envelopes, or policy engines may change while the institution retains stable claim semantics. The interface does not make observation mechanisms assurance-equivalent; substitution remains constrained by the required claim profile, adversary model, operating domain, binding model, and independent evaluation evidence.

Let transition scenarios be `s∈ S`, with probability `p_s`, migration cost without abstraction `M_s`, and cost with PTAS-aligned abstraction `M'_s`. The expected option value is

```
V_option=Σ_(s∈ S) p_s(M_s-M'_s)-C_standard,
```

where `C_standard` includes specification, conformance, and governance cost. The value is highest where asset lifecycles are long and transitions are difficult: aerospace, defense, critical infrastructure, industrial control, regulated supply chains, and identity systems.

13.3 Risk concentration and common infrastructure

A shared standard can reduce integration cost while concentrating dependency. If many institutions accept one assurance profile, an error in its semantics or evaluation can propagate widely. Standardization therefore increases the importance of diversity, transparent limitations, and profile-specific risk ownership. The goal is interoperable meaning, not monoculture.

14. Limitations and open research questions

14.1 The oracle problem is relocated, not eliminated

Physical trust adds an observation channel; it does not create direct access to reality. The observation subsystem, enrollment process, reference values, verifier, and policy remain potential failure points. The architecture improves assurance only when these dependencies are more difficult to compromise jointly than the digital claim alone.

This relocation is not vacuous. It replaces an unobserved correspondence assumption inside a digital narrative with an explicit, bounded, challengeable, and independently evaluable control surface. Physical trust therefore does not abolish the oracle problem; it makes the relevant oracle and its failure modes governable.

14.2 Physical uniqueness is not sufficient

A physically distinctive feature may be stable but privacy-invasive, difficult to measure, environmentally sensitive, or transferable in unintended ways. Uniqueness, unclonability, usability, revocability, and institutional appropriateness are different properties.

14.3 Presence is not authority

A person can be physically present without possessing legitimate authority. An object can be authentic but unauthorized in the current workflow. PTAS must bind physical evidence to institutional policy rather than treat presence as permission.

14.4 Authority is not correctness

An authorized person or system can make a harmful decision. Physical trust gates who or what is present at a boundary; it does not validate the substantive correctness of the action.

14.5 Coercion and collusion remain difficult

Physical verification may establish the expected person while failing to reveal coercion, bribery, compromised judgment, or collusion. High-consequence profiles may require multi-party control, independent channels, delay, duress procedures, or behavioral and organizational safeguards.

14.6 Privacy can become the dominant risk

A system capable of repeatedly verifying physical people or objects can become a tracking system if identifiers, timestamps, or raw evidence are over-collected. Privacy must be a security property, not a later compliance layer.

14.7 Standard capture

If a standard merely encodes one vendor's architecture, external validity and implementer diversity are compromised. Governance must distinguish indispensable claim semantics from contingent implementation preferences. This requirement does not imply that all conformant mechanisms have equal security or that proprietary implementations lack durable differentiation; conformance establishes a common grammar, while assurance depends on mechanism-specific evidence.

14.8 Quantifying independence

The greatest potential value of physical trust comes from failure-mode independence relative to credentials and software. Measuring that independence is difficult. Research is needed on common-cause modeling, dependency graphs, attack-tree composition, and assurance-case evidence.

14.9 Physical trust for machine subjects

Autonomous machines complicate the person-object distinction. A robot may be an object, an attested computing platform, an agent, and a delegated authority. Future profiles must separate machine identity, physical embodiment, current configuration, delegated permission, and accountable institutional owner.

14.10 Governance under extreme capability

If future AI systems operate at speeds and complexity beyond direct human comprehension, policy cannot depend on ad hoc intervention. Yet delegating all governance to those same systems creates circular trust. Research is needed on minimal physical authority roots, constitutional policy layers, multi-institution control, and recoverable off-switch or reauthorization paths.

15. Research agenda

The following work can advance PTAS without disclosing protected mechanisms.

1. **Formal semantics:** develop machine-readable claim and result ontologies with non-overclaim constraints.
2. **Assurance cases:** define structured evidence packages linking claims, threats, controls, tests, and residual risk.
3. **Context binding:** specify how workflow, location class, device, relying party, and consequence are bound without unnecessary surveillance.
4. **Freshness models:** compare temporal, event-based, custody-based, and lifecycle-based validity.
5. **Privacy-preserving results:** investigate selective disclosure, unlinkable presentations, pairwise references, and minimal audit commitments.
6. **Crypto-agile envelopes:** profile current and post-quantum methods without fixing physical claim semantics to one algorithm generation.
7. **Adversarial benchmarks:** create reproducible test taxonomies for replay, relay, substitution, observation spoofing, lifecycle attack, and integration compromise.
8. **Human factors:** evaluate friction, accessibility, override behavior, duress, and operator trust calibration.
9. **Cross-standard mapping:** align PTAS roles and artifacts with NIST digital identity, zero trust, IETF RATS, W3C credentials, C2PA, and sector safety standards.
10. **Economic measurement:** estimate where an independent physical checkpoint produces positive consequence-weighted value.
11. **Governance design:** study how commercial reference implementations can coexist with open claim semantics, implementer diversity, and independent conformance.
12. **Institutional pilots:** map one decision boundary at a time, beginning with workflows whose policy owner and consequence are already explicit.

16. Conclusion

Digital systems achieve scale by abstracting information from the physical entities and circumstances to which it refers. Cryptography, identity protocols, provenance systems, and remote attestation provide increasingly rigorous assurance about defined representational and computational properties. Their success does not entail that every decision-relevant physical fact has been observed.

This article formalizes the resulting physical trust gap. Lemma 1 establishes that, when policy-relevant physical states induce identical accessible inputs, no deterministic or randomized decision rule can condition its behavior on the missing distinction. Corollary 1 shows that additional credentials, records, telemetry, attestations, or inferred features do not close the gap when they remain jointly forgeable or observationally equivalent within the disputed claim's compromise domain. Proposition 1 characterizes the expected uncertainty reduction provided by an additional observation. Proposition 2 strengthens the analysis by relating observational informativeness to Bayes risk through Blackwell dominance. Together, these results locate the constraint in the information structure and adversarial dependency structure of the decision problem rather than in the capability of a particular class of algorithms.

Physical trust is consequently defined as a bounded assurance property, not as direct access to reality. It introduces appraised evidence from a governed physical anchor whose observation path is adversarially distinct from the disputed claim into an institutional authorization path. The observation channel, reference process, verifier, lifecycle service, policy, and enforcement point remain fallible and must be evaluated as a coupled system. Physical evidence is justified only when its expected reduction in consequence-weighted decision loss exceeds its acquisition, privacy, latency, and availability costs; where the physical condition is a mandatory policy prerequisite, failure to obtain adequate evidence requires redesign, restriction, escalation, or denial rather than silent acceptance.

A prospective standard should therefore specify the assurance envelope rather than a protected mechanism: claim semantics, operating domain, claim-path independence, context and freshness, evidence-result separation, lifecycle, failure states, policy binding, residual risk, evaluation method, and conformance language. Mechanism neutrality preserves interoperability without asserting mechanism equivalence. Whether this framework can support a consensus standard is an empirical and institutional question. It depends on reproducible evaluation, implementer diversity, independent governance, and evidence that the proposed abstractions remain valid across materially different mechanisms and deployment domains.

Appendix A. Illustrative PTAS event envelope

The following example is intentionally abstract and contains no raw physical evidence or implementation-specific feature.

```
{
  "ptas_version": "0.1-discussion",
  "event_id": "pairwise-or-ephemeral-reference",
  "claim_profile": "human-authority",
  "assurance_level": "PTA-L3",
  "subject_reference": "relying-party-scoped-reference",
  "context": {
    "workflow_class": "consequential-action",
    "action_reference": "policy-scoped-reference",
    "relying_system": "institutional-policy-engine",
    "consequence_class": "high"
  },
  "freshness": {
    "observed_at": "RFC3339 timestamp",
    "valid_until": "RFC3339 timestamp",
    "event_bound": true
  },
  "result": "SATISFIED",
  "verifier": {
    "reference": "verifier-identifier",
    "version": "version-identifier"
  },
  "policy": {
    "profile": "policy-identifier",
    "version": "version-identifier"
  },
  "lifecycle_status": "ACTIVE",
  "limitations": [
    "does-not-establish-intent",
    "does-not-establish-action-correctness"
  ],
  "integrity_protection": "algorithm-agile-container-reference"
}
```

This envelope demonstrates a central principle: a relying system should receive enough information to interpret the result correctly, but not the protected measurement representation from which the result was derived.

Appendix B. Illustrative assurance-case structure

An assurance case for a PTAS profile may be organized as:

Top claim: The relying system can use results from implementation M to enforce claim profile P at assurance level L within operating domain Ω.

Subclaims:

1. enrollment binds the intended subject class under governed procedure;
2. the observation path produces decision-relevant evidence within Ω;
3. the observation path breaks the relevant claim-path equivalence under the declared adversary;
4. the verifier appraises evidence under an identified policy;
5. replay, relay, substitution, and declared spoofing attacks are addressed to level L;

6 lifecycle state is current and protected;
7 result semantics and integrity are preserved to the relying system;
8 failure and degraded modes follow declared policy;
9 privacy and retention satisfy the profile;
10 independent evaluation supports each measurable claim;
11 residual risks are accepted by the accountable institutional owner.

Each subclaim should point to test reports, design evidence, operational procedure, audit evidence, or explicit assumptions. Promotional assertions are not assurance evidence.

Appendix C. Public disclosure and IP policy

A physical trust ecosystem requires enough transparency for trust without enough disclosure to facilitate circumvention. Public specifications should therefore distinguish four layers:

Layer	Default disclosure posture
Claim semantics, roles, lifecycle, result states	Public and interoperable
Assurance requirements and evaluation categories	Public; test details may include controlled supplements
Implementation architecture and security claims	Disclosed to evaluators under appropriate controls
Physical mechanism, protected features, manufacturing, thresholds	Proprietary and need-to-know

Security through obscurity is not a sufficient assurance argument. At the same time, responsible assurance does not require public release of every feature that would reduce attack cost. Independent evaluators can inspect controlled implementation evidence while the public standard remains mechanism-neutral.

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x and https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[2] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," IBM Journal of Research and Development, vol. 5, no. 3, pp. 183-191, 1961. https://doi.org/10.1147/rd.53.0183

[3] J. Kaplan et al., "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020. https://arxiv.org/abs/2001.08361

[4] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," arXiv:2203.15556, 2022. https://arxiv.org/abs/2203.15556

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

[6] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, Zero Trust Architecture, NIST SP 800-207, 2020. https://doi.org/10.6028/NIST.SP.800-207

[7] D. Temoshok et al., Digital Identity Guidelines, NIST SP 800-63-4, 2025. https://doi.org/10.6028/NIST.SP.800-63-4

[8] H. Birkholz et al., Remote ATtestation procedureS (RATS) Architecture, IETF RFC 9334, 2023. https://www.rfc-editor.org/rfc/rfc9334.html

[9] World Wide Web Consortium, Verifiable Credentials Data Model v2.0, W3C Recommendation, 15 May 2025. https://www.w3.org/TR/vc-data-model-2.0/

[10] Coalition for Content Provenance and Authenticity, C2PA Specifications and Explainer, version 2.4, 2026. https://spec.c2pa.org/specifications/

[11] National Institute of Standards and Technology, Module-Lattice-Based Key-Encapsulation Mechanism Standard, FIPS 203, 2024. https://doi.org/10.6028/NIST.FIPS.203

[12] National Institute of Standards and Technology, Module-Lattice-Based Digital Signature Standard, FIPS 204, 2024. https://doi.org/10.6028/NIST.FIPS.204

[13] National Institute of Standards and Technology, Stateless Hash-Based Digital Signature Standard, FIPS 205, 2024. https://doi.org/10.6028/NIST.FIPS.205

[14] E. Barker et al., Considerations for Achieving Crypto Agility: Strategies and Practices, NIST CSWP 39upd1, 19 December 2025, updated 29 June 2026. https://doi.org/10.6028/NIST.CSWP.39-upd1

[15] International Organization for Standardization and International Electrotechnical Commission, ISO/IEC 17007:2026 - Conformity assessment - Guidance for drafting normative documents suitable for use for conformity assessment, 2026. https://www.iso.org/standard/17007

[16] W. R. Ashby, An Introduction to Cybernetics. London: Chapman & Hall, 1956. https://ashby.info/Ashby-Introduction-to-Cybernetics.pdf

[17] D. Blackwell, "Comparison of Experiments," in J. Neyman, ed., Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, pp. 93-102, 1951. https://doi.org/10.1525/9780520411586-009

[18] J. Pearl, "A Causal Calculus for Statistical Research," in Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 430-449, 1995. https://proceedings.mlr.press/r0/pearl95a.html

[19] J. H. Saltzer and M. D. Schroeder, "The Protection of Information in Computer Systems," Proceedings of the IEEE, vol. 63, no. 9, pp. 1278-1308, 1975. https://doi.org/10.1109/PROC.1975.9939

[20] International Organization for Standardization and International Electrotechnical Commission, ISO/IEC 15408-1:2026 - Information security, cybersecurity and privacy protection - Evaluation criteria for IT security - Part 1: Introduction and general model, 2026. https://www.iso.org/standard/15408-1

[21] J. A. Hanley and A. Lippman-Hand, "If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators," JAMA, vol. 249, no. 13, pp. 1743-1745, 1983. https://doi.org/10.1001/jama.1983.03330370053031

[22] National Institute of Standards and Technology, Fundamental Issues in Biometric Performance Testing: A Modern Statistical and Philosophical Framework for Uncertainty Assessment, 2021. https://www.nist.gov/system/files/documents/2021/05/26/fundamentalissues_final.pdf

Author and research declarations

Author: Niclas Braun, independent researcher.

Competing interests: The author is engaged in the development and commercialization of technologies related to physical trust. No named product or proprietary implementation is evaluated in this article. The proposed architecture and assurance levels should therefore be subjected to independent technical criticism, replication, and governance before they support external conformance claims.

Data availability: No empirical dataset was generated or analyzed. All examples are conceptual or illustrative.

Human-subjects statement: The article reports no research involving human participants or identifiable personal data.

Peer-review status: Not peer reviewed.

Recommended citation

Braun, N. (2026). Physical Trust at the Digital-Physical Boundary: An Information-Theoretic Framework for Verification in Autonomous Action Systems. Independent research report, version 1.0.