iProov

Governing the High-Risk Actions of AI Agents

10 September 2026 ·

Our previous paper on securing AI agents outlined a three-part specification for agent-resistant proofs and their implementation. As with existing Privileged Access Management (PAM) and Identity Governance and Administration (IGA) solutions for managing human users, the proof strength required for agentic permissioning and approval should correspond with the consequences of the action or access granted. Only the highest-risk actions demand the highest strength and highest-touch authentication of human approval.

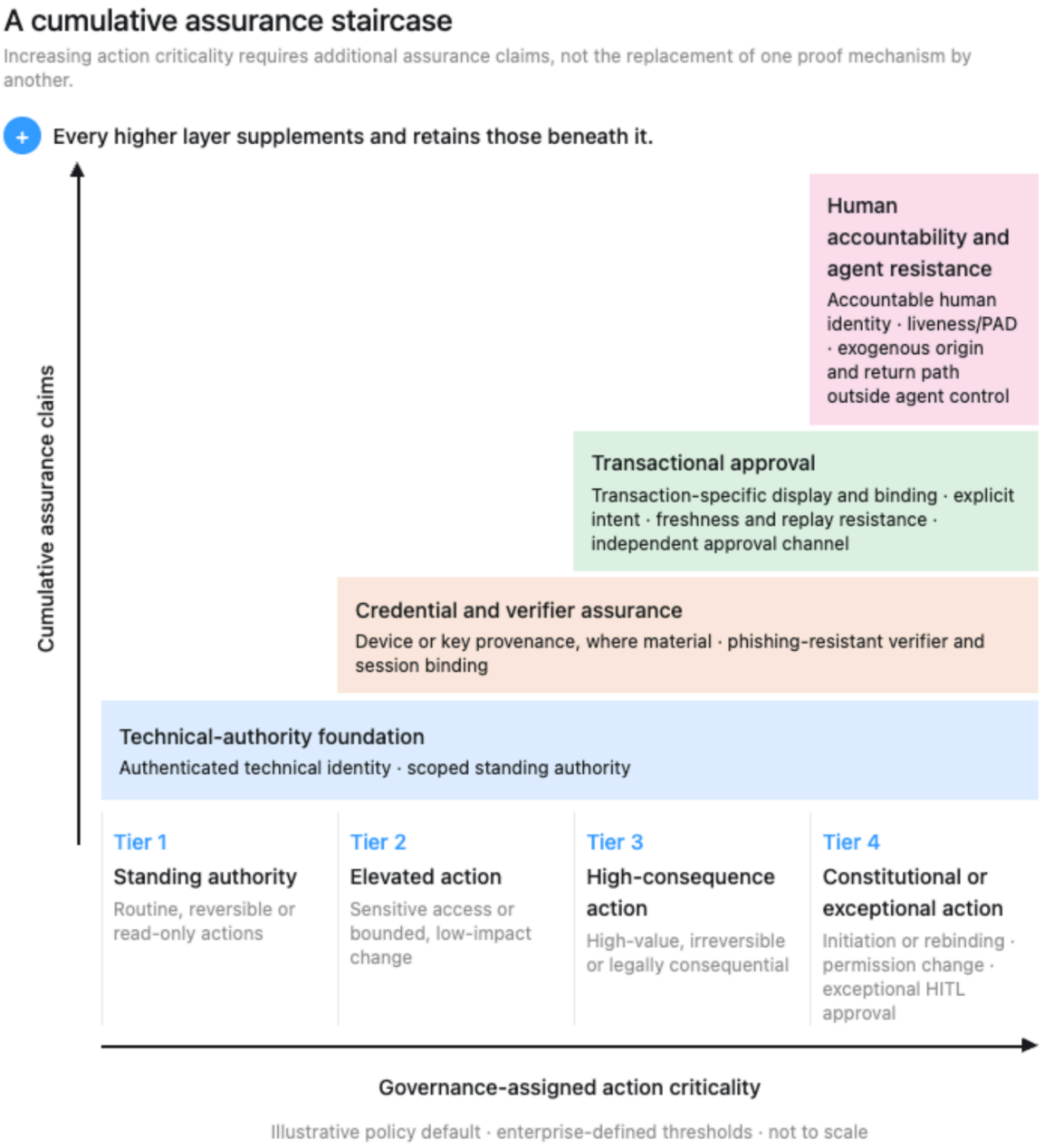
A risk-proportional approach for agentic systems should be retained, but for risky, irreversible or otherwise consequential actions, the minimum standard must rise accordingly. A human user is ordinarily presumed to control the authenticated session being exercised; an agent, by contrast, may legitimately hold the human's session, credentials, and tools such that possession does not in itself prove that the human whose authority it represents actually intended the particular act.

This does not, however, mean that every tool call requires unforgeable proof. Instead, we outline a governed permission boundary beneath those calls, with the strongest proof reserved for the small number of interactions that establish or cross that boundary. As set out in the preceding piece, that proof must be unforgeable by any agent: exogenous in origin and out-of-band in return.

A Higher Security Baseline for Agentic Actions

A higher baseline for agentic actions reflects the particularities of agents within enterprise contexts as a novel class of actors: simultaneously non-deterministic and therefore unpredictable, but yet equally useful and capable of delivering significant business value. Raising the floor and increasing enforcement around this new class shifts where proportionality begins. Device factors, workload credentials, and authenticated sessions remain useful evidence of which technical identity acted and should be used to manage routine actions. However, they do not establish that an accountable human created the authority being exercised nor approved the exceptional action proposed.

The staircase below illustrates the available spectrum of proofs, escalating in strength to meet increasing enforcement demands.



At the highest-risk end of this spectrum, we identify three interactions that, due to their downstream effects, exceptional circumstances, and the need to attribute responsibility to a human approver, should require a proof of the highest strength, which no agent, nor incorrect human user, can ever fulfill. These are permissioning and permission amendments, high-risk human-in-the-loop step-up, and binding agent identity to human accountability.

Permissioning and Permission Amendments

Persistent permissions define the authority the agent may exercise without returning to a human. This covers actions taken, resources accessed, counterparties interacted with, and cumulative or temporal limits on these actions. It also specifies the conditions requiring step-up to a human

approver, and the instances to which those conditions apply. They are therefore the riskiest and most consequential dimension of an agent's initiation or changed state.

From first principles, and as with human users, an agent cannot legitimately consent to a change in its own mandate or permissions. If it can produce, replay, or satisfy the evidence approving the permissions it will inherit, it can effectively approve itself. The proof binding an accountable human to the permission set as enforced must therefore be one that no agent can satisfy.

High-Risk Human-in-the-Loop Step-Up

Human-in-the-loop (HITL) step-up, where proposed automated actions are escalated for a human review and confirmation, applies when an action falls outside persistent permissions, the permission set reserves that class for human decision, or the relying party assigns the intended effect a stricter consequence grade. These are high-risk, irreversible, or legally consequential actions for which the policy requires assurance that the correct human approved the action and is accountable for it.

As a result, the required evidence must establish transaction-specific consent: the identified approver was shown the particular action and the parameters determining its effect, and approved it nonetheless. The premise of human-in-the-loop is nullified if the agent can satisfy its own step-up.

Binding Agent Identity to Human Accountability

In decentralized contexts, an agent's identity may not be a known fact of the estate. Initiation therefore binds the agent to the accountable parties behind it: the principal responsible for its effects and the operator permitted to instruct it. The proof must originate with the human being bound, not the agent being created. Otherwise, the agent could author its own provenance and leave the relationship open to dispute throughout its lifetime.

Initiation ordinarily occurs once. Rebinding the agent to another operator, transferring it between principals, or materially changing the relevant runtime should trigger a new initiation event because these are identity changes rather than routine uses. An adequately bound agent should be able to delegate this relation to downstream agents, although, if issued as a verifiable credential (VC), this should be marked in the subsequent VC to make the intermediated relationship back to the principal and operator legible to relying parties (RPs).

Considerations

Applying this framework requires attention to two related considerations: first, how the deliberate friction and cost of unforgeable proof can remain proportional to the risk at issue; and, second, how organizations classify, own, and review the actions for which such proof is mandated.

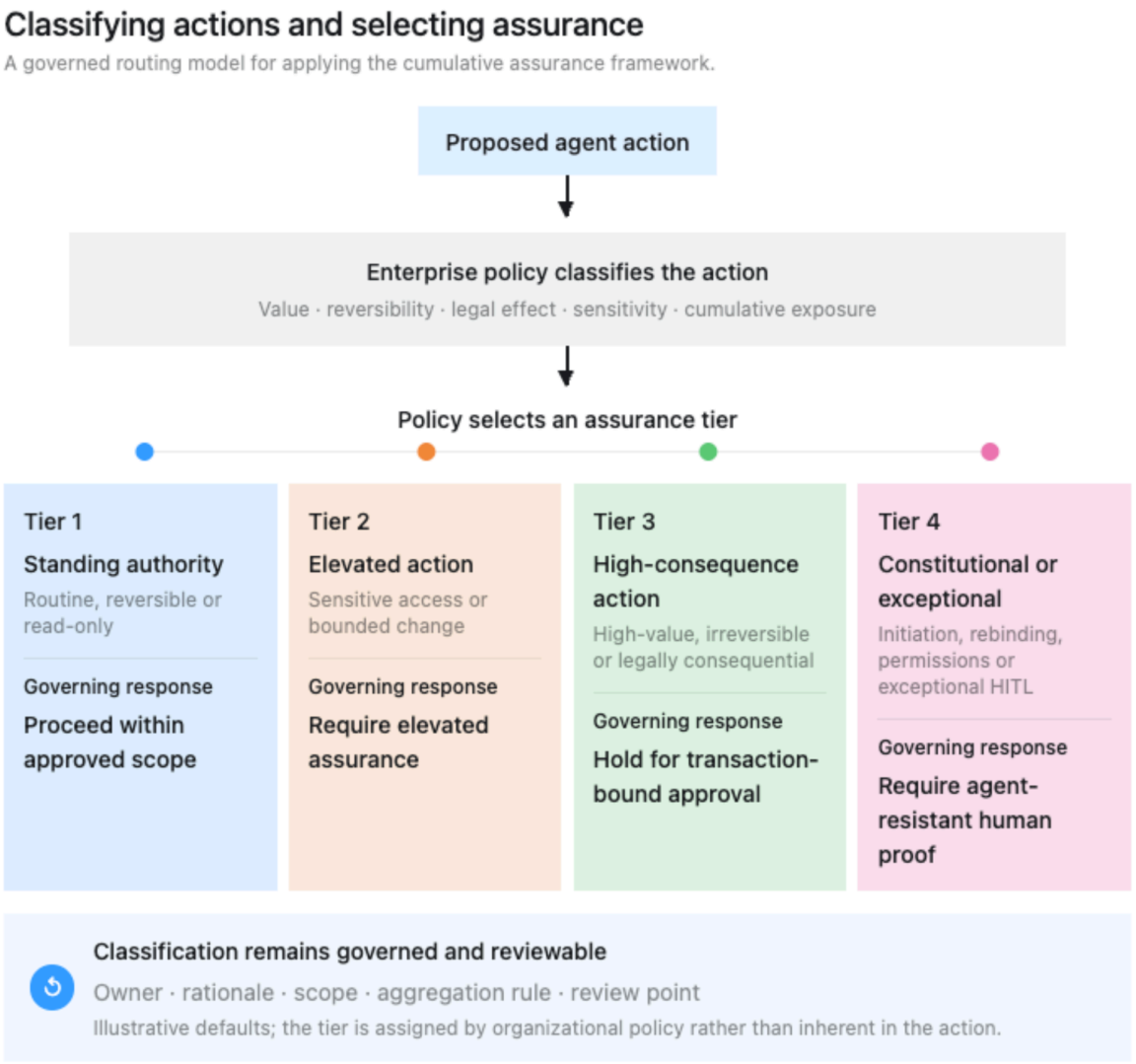
Deliberate Friction and Proportionate Cost

Unforgeable proof introduces deliberate friction and cost. It should therefore be reserved for exceptional, high-risk actions, where that friction can focus the approver's attention rather than create routine fatigue. We contend that such proof must depend on input from outside an agent's digital context. Our solutions, for example, require a verification process that requires the approver to interact directly with their device.

The required proof should, first, be commensurate with the risk of an action. An agent may perform millions of routine actions under a permission set approved once and never require an exceptional step-up. Where an exceptional step-up is required, the friction attaches to an action that would otherwise remain prohibited. And because the step-up is high-touch and takes time, it encourages the approver to engage with the action and issue deliberate consent. Used sparingly, this friction mitigates, rather than creates, approval fatigue. Similarly, the financial cost remains low because only a limited number of interactions require this level of proof, and it is proportionate to the critical risks it mitigates.

Risk, Classification and Ownership

While proportionality requires each organization's administrator to decide which actions may proceed under standing authority and which must be held for unforgeable proof, we propose the continuum as illustrated below as a default. For example, routine read-only queries may proceed under standing authority, while changing an agent's permissions or approving a high-value, irreversible action may require unforgeable human proof.



That decision is not an objective property of the action; it is a governance artifact specific to each enterprise which may be wrong, contested, or permitted to drift. We expect different parties, depending on their industry, regulatory position, and tolerance, may grade the same action differently.

Beyond our recommendations, we do not demand a specific hierarchy of risks and proofs. However, across all instantiations and proof requirements, each classification should contain an owner, rationale, scope, aggregation rule, and review point. The agent may describe the action it wishes to take, but must not determine the strength of evidence by which that action becomes permissible; that is at the discretion of each enterprise and its administrator.

In the next two papers in this series, we will outline how these principles apply within both centralized and decentralized architectures, respectively.