

Assuring Human Approval of Agentic AI Actions

10 SEPTEMBER 2026 ·

Agentic AI can create novel failure conditions for conventional digital approval; an agent may hold, invoke, or satisfy many factors relied upon to prove human approval. While a valid credential, a rudimentary presence check, or a response returned through an agent-accessible device may demonstrate access to that factor, it does not necessarily establish that the responsible human saw and approved a particular action. As a result, this creates vulnerabilities in instances of severe model miscomprehension or aggressive prompt injection. This piece outlines how agent-resistant proofs mitigate this vulnerability.

For actions which are irreversible, legally consequential, or demand human accountability, the required proof must establish both that the nominated human independently approved the specific action, and that the agent cannot satisfy the evidence relied upon for that approval. For routine actions, this residual risk may be proportionate to the risk of proceeding without the correct human approval.

This adequate human approval must meet three conditions:

- 1. Proof originating beyond the operating capability of any agent;**
- 2. Deployment out-of-band such that its challenge and response travel through channels the agent does not mediate;**
- 3. Binding of the approving human to the exact action approved.**

Problem: Agentic Fulfillment of Existing Proofs

When primary context integrity is uncertain, organizations conventionally rely on out-of-band confirmation. Historically, the security of out-of-band confirmation is derived from the physical separation of devices rather than from any property asserted by the implementation. For example, an SMS might reach a mobile phone while the primary session remained on a desktop computer. As a result, compromising the primary session does not provide access to the second device. It remains secure because it is elsewhere and under the control of the appropriate approver.

However, agents do not always conform to this premise because they may operate and communicate across multiple devices or runtimes. For example, agent A may be unable to access the second channel directly. It may, however, recruit agent B, which operates on the second device.

Agent B can then retrieve a message, operate an interface, or return a conventional approval factor. Agent-to-agent recruitment therefore undermines the use of second channels to secure proof, even though neither agent needs to compromise the device or evade its security controls. The problem is more extreme and consequential. For example, in aggressive instances of misalignment and prompt injection, where circumvention may be deliberate, a suitable proof and its implementation must therefore be secure against both incidental circumvention and active attack.

Even strong conventional factors, such as passkeys, remain endogenous if an agent can hold, invoke, replay, or recruit another agent to satisfy them. They may demonstrate access, but for the highest-risk interactions the proof must establish human approval and remain unsatisfiable by any agent, including one recruited through a second channel.

Resolution:

Use an Exogenous Data Source

An exogenous proof must originate outside the possible operating context of any agent. This requires data which an agent can neither generate nor simulate. A proof generated on another device, runtime, or tenancy is therefore insufficient, as it remains open to cross-device agent-to-agent recruitment.

Instead, suitable proofs must originate in the material world beyond the digital system that an agent can operate. These may include secure-intent hardware requiring deliberate human actuation, physical keys, or biometrics whose liveness is established at the moment of capture. The defining property of exogeneity is that the evidence cannot be retrieved from something an agent possesses or produced by operating an accessible interface.

iProov's liveness solutions provide suitable proof because they depend on data that cannot be digitally simulated. The device captures and transmits the proof, but its source is ultimately the physically present human.

Out-of-Band Challenge and Response

The proof must be implemented out-of-band. An assertion that passes through the compromised context on its way to or from the gateway, doorkeeper, or relying party could itself be compromised.

Although the outbound request presents less risk, the proposing agent may suppress or deliberately misrepresent it if it passes through the agent's own output. The control layer should instead trigger an approval surface that presents control-derived action details directly to the approver, independently of any agent-composed account.

More importantly, the verdict must return directly from the approval surface to the control layer that requested it. A misaligned or prompt-injected agent carrying the verdict may redirect, reuse, or replay it. The agent should never carry the evidence that the control layer relies on. It should only be informed that the approval condition has been satisfied.

Another agent on the approval device does not defeat this implementation. That agent can neither alter the challenge nor satisfy the exogenous proof required to answer the challenge. Agent B may open the approval interface, but it cannot replace the nominated human whose liveness is required. Nor can it change the action presented or manufacture the resulting assertion.

Human-Approval Binding

Finally, an exogenous proof must bind the nominated human to the exact action presented, rather than merely establish that a suitable human was present. Accordingly, the approval surface should present the action's material parameters directly from the control layer and incorporate an immutable representation of those parameters into the challenge of the proceeding UI. The resulting assertion must identify the approver and be valid exclusively for the affirmed action and transaction instance. Any alteration to the amount, scope, recipient, tool used, or any other consequential dimension invalidates the approval. This prevents a genuine proof from being detached from the human decision and reused to approve a different action.

Suitable Deployment

The strongest implementation therefore combines both properties. The approval challenge and response travel through out-of-band channels which the proposing agent does not mediate. Within that channel, the exogenous proof requirement must remain unsatisfiable by any recruited agent.

However, this level of proof is not required for every action. High-touch interruption of routine operations would forfeit the productivity which agentic autonomy delivers. In our next piece, we will outline how, as with established step-up authentication, the strength of the proof should be proportional to the importance of the interaction and risk. And how this can be deployed to facilitate increased adoption of AI agents.