Source dossier

The approval was real, the action was not — Loopjacking reaches released agent frameworks, and the fix being cited is an endpoint audit

Edition 2026-09-24 · narrative: approval binding · compiled 2026-09-24

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-09-24-01	Loopjacking is defined in a paper submitted to arXiv as 2609.21081 on 17 September 2026 by Adithyan Arun Kumar, an independent security researcher: a human approves what they understand as operation A while the implementation uses that decision for a materially different operation B, in two variants - representation-based attack and post-approval state substitution.	verified	arxiv2026loopjacki
CLAIM-2026-09-24-02	Post-approval state substitution is reproduced in seven released Agno AgentOS versions - 2.5.6, 2.9.0, 3.0.1, 3.0.2, 3.0.3, 3.0.6 and 3.0.9 - with five of five trials succeeding at 3.0.9.	verified	arxiv2026loopjacki, github2026loopjacki
CLAIM-2026-09-24-03	The same substitution is reproduced across twelve tested versions of a conditional in-memory LangGraph Agent Server composition from 0.7.5 to 0.14.0, and the result is conditional on an authorization policy that permits a non-approver to update a shared pending thread; a deny-update policy blocks the mutation while preserving legitimate execution.	verified	arxiv2026loopjacki, github2026loopjacki
CLAIM-2026-09-24-04	Representation mismatch is reproduced in OpenClaw 2026.2.23 in three of three trials and rejected in 2026.2.24 in three of three trials, the only product in the study with a matched affected and fixed pair.	verified	arxiv2026loopjacki
CLAIM-2026-09-24-05	OpenAI Agents SDK 0.22.0 and 0.22.2 act as the study's negative control, rejecting the mutated operation in three of three trials each, because a serialized continuation preserves exact per-call binding across the pause and restore cycle.	verified	arxiv2026loopjacki
CLAIM-2026-09-24-06	Agno 2.5.6 is the established lower boundary of the affected range only because version 2.5.5 also executed the continuation but allowed direct unapproved B, so approval hijacking was not needed and the result did not qualify as Loopjacking.	verified	arxiv2026loopjacki

Claim	Statement	Status	Sources
CLAIM-2026-09-24-07	The study's evidence cutoff is 10 September 2026 and no vendor-fixed release is established for either Agno or LangGraph in its archive.	verified	arxiv2026loopjacki, github2026loopjacki
CLAIM-2026-09-24-08	Agno pull request 10270, titled nine authorization guard fixes from the authz audit, was merged on 22 September 2026 and covers endpoint authorization - user endpoint gating, reserved scope namespaces, draft preview privilege checks, provider error handling, reserved principals, scope PATCH validation, knowledge refresh mapping, role-slug provisioning and request-time gating for manually added JWT middleware - and mentions neither approval workflows nor human-in-the-loop.	verified	github2026nine
CLAIM-2026-09-24-09	Article 9(3)(c) of DORA, Regulation (EU) 2022/2554, requires financial entities to prevent the lack of availability, the impairment of the authenticity and integrity, the breaches of confidentiality and the loss of data, and DORA has applied since 17 January 2025.	verified	digitaloper2025article, chambersand2025countdown

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

Loopjacking: Hijacking Human-in-the-Loop Approval

PRIMARY

Publisher	arXiv
Published	2026-09-17
Retrieved	2026-09-24
Cite key	arxiv2026loopjacki
URL	https://arxiv.org/abs/2609.21081

arXiv:2609.21081v1, submitted 17 September 2026, 20:58:26 UTC. Author: Adithyan Arun Kumar, independent security researcher.

Abstract. Human approval is often treated as the last security boundary before an agent executes a consequential operation. That boundary is only meaningful if the operation presented for review is the operation later authorized or released. We call failures of this binding Loopjacking: a human approves what they understand as operation A, while the implementation uses that decision for a materially different operation B. We distinguish two variants. In a representation-based attack, B is already encoded but omitted or misrepresented at approval time; in a post-approval state-substitution attack, the human sees the correct A and mutable workflow state later replaces it with B.

We evaluate a purposive set of released agent products. We reproduce post-approval substitution in seven tested Agno AgentOS releases ending at 3.0.9 and in 12 tested versions of a conditional in-memory LangGraph Agent Server composition ending at 0.14.0. We reproduce representation mismatch in OpenClaw 2026.2.23 and its rejection in 2026.2.24. OpenAI Agents SDK 0.22.0 and 0.22.2 provide a negative control: serialized continuation preserves exact per-call binding and rejects mutated B. These results do not estimate ecosystem prevalence. They show that complete canonical approval rendering and exact use-time comparison, or preventing unauthorized pending-state mutation, block the tested attacks while preserving legitimate execution.

Versions. Agno AgentOS affected: 2.5.6, 2.9.0, 3.0.1, 3.0.2, 3.0.3, 3.0.6, 3.0.9; 5/5 trials at 3.0.9; no vendor fix at evidence cutoff. LangGraph Agent Server: 0.7.5, 0.7.103, 0.8.7, 0.9.1, 0.10.3, 0.11.4, 0.12.4, 0.12.6, 0.12.9, 0.13.2, 0.13.4, 0.14.0; feature absent in 0.7.4; conditional on an authorization policy allowing non-approvers to update a shared pending thread; deny-update policy is the safe configuration. OpenClaw 2026.2.23 vulnerable 3/3, 2026.2.24 rejects 3/3. OpenAI Agents SDK 0.22.0 and 0.22.2 reject 3/3 each.

Lower boundary, verbatim. 'Version 2.5.5 demonstrates why the direct-path control is necessary: it also executed the continuation, but because direct B succeeded, approval hijacking was not needed.' And: '2.5.5 control allowed direct B, so it is not a qualifying Loopjacking result; 2.5.6 is the established lower boundary at which the approval gate both blocked direct B and could be reused for substituted B.'

Mitigation. Canonical approval records binding complete action and tool identity, all material arguments, target resource and destination, initiating and approving principals, task and thread scope, nonce, creation time, expiry, consumption status, and a digest over the complete descriptor shown to the human. Use-time enforcement: reconstruct the complete resolved operation immediately before releasing the effect and compare it with the bound descriptor; reject a missing, expired, revoked, consumed or out-of-scope decision; reject or re-request approval on material difference; re-evaluate policy; consume the decision atomically with the effect; append approved descriptor, current descriptor, decision and effect to the audit record.

Regression suite. Unchanged A succeeds; denial produces no effect; direct B unavailable to the attacker; A-to-B representation or state substitution rejected or reauthorized; wrong scope fails; consumed or expired approval cannot be replayed.

Limitations. Purposive, not representative; configuration-specific; no universal severity score; single researcher, no independent reproduction; evidence cutoff 10 September 2026.

Loopjacking evidence and reproduction archive

PRIMARY

Publisher	GitHub
Published	2026-09-10
Retrieved	2026-09-24
Cite key	github2026loopjacki
URL	https://github.com/adithyan-ak/loopjacking

Evidence and reproduction archive for the Loopjacking research. Four experiment harnesses with their own instructions; verification via a Python 3.10+ script checking SHA-256 manifests; frozen package versions and dependency pins.

Table: Agno AgentOS - post-approval state substitution - 2.5.6 to 3.0.9. LangGraph Agent Server - conditional post-approval substitution - 0.7.5 to 0.14.0 (0.7.4 lacks the feature). OpenClaw - representation mismatch - 2026.2.23 vulnerable, 2026.2.24 patched. OpenAI Agents SDK - negative control - 0.22.0 and 0.22.2 reject mutation.

Described as 'a purposive comparative archive, not a prevalence survey'. Neither Agno nor LangGraph has a vendor-fixed release in this archive. Evidence cutoff 10 September 2026.

Note on a discrepancy: the repository table labels the version below each affected range as safe, while the paper's own text is narrower - 2.5.5 executed the continuation and allowed direct unapproved B, which is why it was not a qualifying Loopjacking result. The paper's wording is the one to rely on.

Nine authorization guard fixes from the authz audit (Agno pull request 10270)

PRIMARY

Publisher	GitHub
Published	2026-09-22
Retrieved	2026-09-24
Cite key	github2026nine
URL	https://github.com/agno-agi/agno/pull/10270

Merged 22 September 2026 into feat/agentos-authz, 15 commits, approved by SamJupe on 22 September 2026.

The nine fixes: (1) user endpoint gating - prevent unauthorized token access to /users endpoints; (2) scope namespace reservation - reserve the agent_os namespace against privilege escalation; (3) draft preview access - correct admin privilege checking for draft configurations; (4) provider error handling - unhandled provider errors changed from 401 leaks to 403 denials with audit logging; (5) reserved principals - reject service accounts and role slugs as user IDs; (6) scope PATCH validation - full validation before persisting; (7) knowledge refresh mapping - add the missing scope mapping for the refresh endpoint; (8) role-slug provisioning - prevent JIT provisioning of subjects matching role names; (9) request-time gating - runtime checks for manually added JWT middleware.

The description emphasises authorization, guard-style fixes and the audit trail. It does not mention approval workflows, human-in-the-loop, approval binding, or Loopjacking.

Loopjacking: When Your AI Agent Approves A, Runs B

SECONDARY

Publisher	byteiota
Published	2026-09-24
Retrieved	2026-09-24
Cite key	byteiota2026loopjacki
URL	https://byteiota.com/loopjacking-when-your-ai-agent-approves-a-runs-b/

Published 24 September 2026. Reports the paper (arXiv 2609.21081, 17 September 2026): LangGraph Agent Server post-approval substitution 'reproduced across 12 tested versions, up through v0.14.0'; Agno AgentOS 'reproduced in seven releases from v2.5.6 through v3.0.9'; OpenAI Agents SDK v0.22.0 and v0.22.2 resistant. OpenClaw not mentioned.

Vendor response reported: 'The Agno team has merged nine authorization guard fixes in PR #10270', with a recommendation to verify 'whether your deployment has pulled that update'. No LangChain or OpenAI response reported.

Handling note: this is where the Agno pull request is placed beside the research. Reading the pull request shows its content is endpoint authorization, not approval binding, so the juxtaposition should not be read as a fix for the class.

Article 9 — Protection and prevention

PRIMARY

Publisher	Digital Operational Resilience Act
Published	2025-01-17
Retrieved	2026-09-24
Cite key	digitaloper2025article
URL	https://www.digital-operational-resilience-act.com/Article_9.html

Regulation (EU) 2022/2554, Article 9, Protection and prevention.

9(1) Financial entities shall continuously monitor and control the security and functioning of ICT systems and tools and shall minimise the impact of ICT risk.

9(2) Financial entities shall design, procure and implement ICT security policies, procedures, protocols and tools that aim to ensure the resilience, continuity and availability of ICT systems.

9(3)(a) ensure the security of the means of transfer of data; (b) minimise the risk of corruption or loss of data, unauthorised access and technical flaws that may hinder business activity; (c) prevent the lack of availability, the impairment of the authenticity and integrity, the breaches of confidentiality and the loss of data; (d) ensure that data is protected from risks arising from data management, including poor administration, processing-related risks and human error.

9(4)(a) develop and document an information security policy defining rules to protect the availability, authenticity, integrity and confidentiality of data; (b) establish a sound network and infrastructure management structure; (c) implement policies that limit physical or logical access to information assets and ICT assets to what is required for legitimate and approved functions; (d) implement policies and protocols for strong authentication mechanisms, based on relevant standards and dedicated control systems.

Countdown to DORA: The Regulation applies from 17 January 2025

SECONDARY

Publisher	Chambers and Partners
Published	2025-01-01
Retrieved	2026-09-24
Cite key	chambersand2025countdown
URL	https://chambers.com/articles/countdown-to-dora-the-regulation-applies-from-17-january-2025

Second reading on the application date: Regulation (EU) 2022/2554 (DORA) applies to financial entities from 17 January 2025.

Human-in-the-loop

PRIMARY

Publisher	LangChain
Published	2026-09-24
Retrieved	2026-09-24
Cite key	langchain2026humaninth
URL	https://docs.langchain.com/oss/python/langchain/human-in-the-loop

LangChain's own documentation for the human-in-the-loop interrupt: what the pause guarantees, and what it leaves to the deployment's authorization policy. Used for context on the conditional nature of the LangGraph result.

Implementation Timeline

SECONDARY

Publisher	EU Artificial Intelligence Act
Published	
Retrieved	2026-09-24
Cite key	euartificia2026implement
URL	https://artificialintelligenceact.eu/implementation-timeline/

The AI Act implementation timeline after the Digital Omnibus, Regulation (EU) 2026/1744: Annex III high-risk obligations at 2 December 2027, Annex I at 2 August 2028, member-state sandboxes at 2 August 2027. Article 50 transparency applies since 2 August 2026; the marking-of-generated-content obligation lands 2 December 2026. Article 53 GPAI in force since August 2025.