

Source dossier

The hook that let it through — a guardrail's failure mode was permissive, and the AI Act's fail-safe sentence is fourteen months out

Edition 2026-10-09 · narrative: A control plane written in prose defaults to permissive; the statute that names fail-safe does not bind until December 2027, so the only thing making your gate fail closed today is a config flag. · compiled 2026-10-09

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-09-01	Claude Code 2.1.295, released 8 October 2026, added <code>onFailure: "block"</code> for command and HTTP hooks, so that a hook which cannot start, times out, or exits with an unexpected code blocks the action instead of letting it through.	verified	anthropic2026claude
CLAIM-2026-10-09-02	Before that release, all three of those hook failure modes let the action through — the changelog entry states the new behaviour blocks "instead of letting it through".	verified	anthropic2026claude
CLAIM-2026-10-09-03	Claude Code 2.1.294, released the same day, 8 October 2026, fixed <code>prompt</code> and <code>agent</code> hooks written as instructions allowing what they should block.	verified	anthropic2026claude
CLAIM-2026-10-09-04	<code>onFailure: "block"</code> is an added, per-hook option rather than a changed default, so the fail-open behaviour remains the default for hooks that do not set it.	verified	anthropic2026claude
CLAIM-2026-10-09-05	arXiv 2610.09264, "Package Hallucination Attacks on Coding Agents through Prompt Injection in Rule Files", was submitted on 7 October 2026 as v1.	verified	arxiv2026package
CLAIM-2026-10-09-06	The paper describes planting malicious prompts in benign community rule files such as AGENTS.md or .cursorrules so that coding agents swap legitimate dependencies for attacker-controlled packages, and reports transferability across models and agent frameworks.	verified	arxiv2026package
CLAIM-2026-10-09-07	The paper's abstract record states no numeric success rates, and its HTML, PDF and ar5iv renderings could not be retrieved; no figure from the paper appears in this edition.	verified	arxiv2026package
CLAIM-2026-10-09-08	The work's own front matter on the arXiv abstract record carries no institutional affiliations, so none is attributed in this edition.	verified	arxiv2026package

Claim	Statement	Status	Sources
CLAIM-2026-10-09-09	AI Act Article 15(4) requires high-risk systems to be as resilient as possible regarding errors, faults or inconsistencies, and states that robustness may be achieved through technical redundancy solutions, "which may include backup or fail-safe plans".	verified	euaiactexpl2026accuracy
CLAIM-2026-10-09-10	Article 15 applies from 2 December 2027 for high-risk systems under Annex III and from 2 August 2028 for those under Annex I, per Article 113(c).	verified	euaiactexpl2026accuracy, eurlex2026regulatio
CLAIM-2026-10-09-11	Regulation (EU) 2026/1744, the Digital Omnibus on AI, in force 27 July 2026, is the instrument that moved those two dates.	verified	eurlex2026regulatio
CLAIM-2026-10-09-12	GDPR Article 25 is titled "Data protection by design and by default"; 25(2) requires measures ensuring that by default only personal data necessary for each specific purpose are processed, and that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.	verified	gdprinfoeu2026data, eurlex2016regulatio
CLAIM-2026-10-09-13	Cyber Resilience Act Article 14 reporting obligations have applied since 11 September 2026; the bulk of the CRA's obligations, including the Annex I essential requirements, apply from 11 December 2027.	verified	cyberresili2026cyber
CLAIM-2026-10-09-14	KI-MIG has been in force since 29 July 2026 and designates the Bundesnetzagentur as Germany's market surveillance authority under the AI Act.	verified	bundesamt2026ki mig

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

Claude Code changelog

PRIMARY

Publisher	Anthropic
Published	2026-10-08
Retrieved	2026-10-09
Cite key	anthropic2026claude
URL	https://code.claude.com/docs/en/changelog

Assessment. The vendor's own release notes. Event dates here are the primary actor's own dates, not coverage dates. Entries quoted verbatim; the page is newest-first and was read from the top.

Claims drawn from this source. CLAIM-2026-10-09-01, CLAIM-2026-10-09-02, CLAIM-2026-10-09-03, CLAIM-2026-10-09-04

2.1.295 (October 8, 2026)

Added

- Added onFailure: "block" for command and HTTP hooks: a hook that can't start, times out, or exits with an unexpected code blocks the action instead of letting it through
- Added Program Status Protocol (OSC 7501) support
- Added a warning to claude plugin install, enable, disable and marketplace add when the settings file they write to does not load
- Added \$.ui.notify for mods

Changed

- Changed MCP tool descriptions the model loads through tool search to be cut at 16,384 characters instead of 2,048
- Changed subagents to preload at most 32 skills from the skills field

2.1.294 (October 8, 2026)

- Fixed prompt and agent hooks written as instructions (such as "Block commands that...") allowing what they should block
- Improved how prompt hooks on Stop and SubagentStop written as instructions (such as "Carry on if the build is broken") are judged, so Claude is less likely to stop early

2.1.293 (October 7, 2026)

- Added Claude Haiku 5.5 (claude-haiku-5-5), 1M context
- Added agentType to the subagentStatusLine payload

2.1.292 (October 6, 2026)

- Added --marketplace <source> to claude plugin install
- Added an effort parameter to the Agent tool

- Added workflow agents to the `agent.spawn` mod hook, with their run and index, so a mod can refuse them

2.1.291 (October 6, 2026)

- Fixed a regression in 2.1.290 where cloud sessions could drop answers to permission prompts

Package Hallucination Attacks on Coding Agents through Prompt Injection in Rule Files

PRIMARY

Publisher	arXiv
Published	2026-10-07
Retrieved	2026-10-09
Cite key	arxiv2026package
URL	https://arxiv.org/abs/2610.09264

Assessment. Abstract record only. arxiv.org/html/2610.09264v1, arxiv.org/pdf/2610.09264 and the ar5iv rendering (which 302-redirects back to /abs) all failed: the fetch proxy returned HTTP 429 and instructed that the pages not be retried. No figure from this paper is used anywhere in this edition. The abstract record carries no institutional affiliations, so none is attributed; 'Researchers' is used in the copy.

Claims drawn from this source. CLAIM-2026-10-09-05, CLAIM-2026-10-09-06, CLAIM-2026-10-09-07, CLAIM-2026-10-09-08

Submitted 7 Oct 2026 (v1), cs.CR, cross-listed cs.AI. arXiv:2610.09264.

The paper introduces a package hallucination attack in which an attacker places malicious prompts inside benign community rule files (such as AGENTS.md or .cursorrules) so that coding agents swap legitimate dependencies for attacker-controlled packages. It proposes PackHallu, an evolutionary optimization framework that refines the injected prompts using trajectory-level feedback and LLM-guided mutations. It reports evaluations across multiple benchmarks, LLMs and agent frameworks, claiming high attack success rates and strong transferability across models and agent combinations, and concludes that coding agents are vulnerable and that stronger safeguards are needed.

The abstract states no numeric results: no attack success rates, no counts of models or agent frameworks, no package counts. No affiliations appear in the abstract record's front matter.

Accuracy, robustness and cybersecurity (Article 15)

SECONDARY

Publisher	EU AI Act Explorer
Published	2026
Retrieved	2026-10-09
Cite key	euaiactexpl2026accuracy
URL	https://artificialintelligenceact.eu/article/15/

Assessment. Article text reproduction with application dates at the foot of the page. The operative fragment was read verbatim; the dates match the vault's settled Digital Omnibus positions and Article 113(c).

Claims drawn from this source. CLAIM-2026-10-09-09, CLAIM-2026-10-09-10

Article 15 — Accuracy, robustness and cybersecurity.

15(1). High-risk AI systems shall be designed and developed to achieve an "appropriate level of accuracy, robustness, and cybersecurity" and to perform consistently in those respects throughout their lifecycle.

15(2). The Commission shall encourage the development of benchmarks and measurement methodologies, in cooperation with relevant stakeholders including metrology and benchmarking authorities.

15(3). The levels of accuracy and the relevant accuracy metrics shall be declared in the accompanying instructions of use.

15(4). High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which it operates. Technical and organisational measures shall be taken to this end. The robustness of high-risk AI systems may be achieved through technical redundancy solutions, "which may include backup or fail-safe plans". High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed so as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations, and to ensure suitable mitigation measures.

15(5). High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. Measures shall be appropriate to the circumstances and risks, and shall address, where appropriate, data poisoning, model poisoning, adversarial examples (evasion) and confidentiality attacks.

Application: Article 15 applies from 2 December 2027 for high-risk systems under Annex III, and from 2 August 2028 for those under Annex I (Article 113(c)). No amendments to Article 15 are listed.

Regulation (EU) 2026/1744 — Digital Omnibus on AI

PRIMARY

Publisher	EUR-Lex
Published	2026-07-27
Retrieved	2026-10-09
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Assessment. The instrument that moved the Annex III and Annex I application dates. Cited for the instrument; the dates themselves are settled in the vault from earlier editions and are restated on the Article 15 page.

Claims drawn from this source. CLAIM-2026-10-09-11

Regulation (EU) 2026/1744, the Digital Omnibus on AI, in force 27 July 2026. It moved the Annex III high-risk obligations to 2 December 2027 and the Annex I obligations to 2 August 2028, and member-state regulatory sandboxes to 2 August 2027. It did not move Article 50 (applying since 2 August 2026, with the marking-of-generated-content obligation landing 2 December 2026) or Article 53 GPAI (in force since August 2025).

Note on retrieval: the EUR-Lex ELI URL renders to the fetch proxy as a metadata shell without article text. The dates here rest on the vault's existing verified claims and on the Article 15 page's own application note, not on an unread page.

Data protection by design and by default — Art. 25 GDPR

SECONDARY

Publisher gdpr-info.eu

Published 2026

Retrieved 2026-10-09

Cite key gdprinfoeu2026data

URL <https://gdpr-info.eu/art-25-gdpr/>

Assessment. Article text reproduction, used because the EUR-Lex CELEX and ELI renderings both returned empty or metadata-only text to the fetch proxy on this run. The instrument itself is cited separately as eurlex2016regulatio. Vault precedent: gdpr-info article pages have been used this way since the August editions.

Claims drawn from this source. CLAIM-2026-10-09-12

Article 25 — Data protection by design and by default.

25(1). Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

25(2). The controller shall implement appropriate technical and organisational measures for ensuring that, by default, "only personal data which are necessary for each specific purpose of the processing are processed". That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that "by default personal data are not made accessible without the individual's intervention" to an indefinite number of natural persons.

Regulation (EU) 2016/679 — General Data Protection Regulation

PRIMARY

Publisher	EUR-Lex
Published	2016-04-27
Retrieved	2026-10-09
Cite key	eurlex2016regulatio
URL	https://eur-lex.europa.eu/eli/reg/2016/679/oj

Assessment. Cited for the instrument. On this run the ELI page and the CELEX HTML rendering both returned metadata-only or empty text to the fetch proxy, so no article wording is taken from here; the Article 25 wording comes from the article-page source and is marked secondary.

Claims drawn from this source. CLAIM-2026-10-09-12

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Applicable since 25 May 2018.

Retrieval note: eur-lex.europa.eu/eli/reg/2016/679/oj and eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679 both returned no article text to the fetch proxy on 9 October 2026.

Cyber Resilience Act explained — scope, classes and deadlines

SECONDARY

Publisher CyberResilienceAct.eu

Published 2026-09-11

Retrieved 2026-10-09

Cite key cyberresili2026cyber

URL <https://www.cyberresilienceact.eu/explained.html>

Assessment. Dates only, and consistent with the European Commission's own CRA reporting page cited in the 6 October edition. The Annex I 'secure by default' wording itself could not be read verbatim from this page or from the EUR-Lex rendering, so the copy refers to the requirement without quoting it.

Claims drawn from this source. CLAIM-2026-10-09-13

Cyber Resilience Act, Regulation (EU) 2024/2847.

- Article 14 reporting obligations apply from 11 September 2026.
- The bulk of the obligations, including the Annex I essential requirements, apply from 11 December 2027.
- Open-source software stewards have lighter, tailored obligations; this page gives no separate application date for them.

The page summarises but does not quote the Annex I secure-by-default configuration requirement, and gives no point number.

KI-MIG

PRIMARY

Publisher	Bundesamt fuer Justiz
Published	2026-07-29
Retrieved	2026-10-09
Cite key	bundesamtfu2026kimig
URL	https://www.gesetze-im-internet.de/ki-mig/

Assessment. Reused cite key; the vault already holds this source from the 5 October edition. In this edition it supports only who would eventually ask about Article 15, not a duty that binds today.

Claims drawn from this source. CLAIM-2026-10-09-14

Gesetz zur Marktueberwachung und Innovationsfoerderung von kuenstlicher Intelligenz (KI-Marktueberwachungs- und Innovationsfoerderungsgesetz, KI-MIG), in force since 29 July 2026. It designates the Bundesnetzagentur as Germany's market surveillance authority under Regulation (EU) 2024/1689.