Tool-Use Safety

tool-use-safety · safety · concept

Source: https://policywindow.org/wiki/tool-use-safety

Generated 2026-05-30T22:10:28 UTC

Summary

The sub-domain of agentic-system safety concerned with the risks that arise when an AI model invokes external tools (search, code execution, APIs, financial transactions, system commands) — including risks of unintended action, instruction subversion, privilege escalation, and resource consumption.

At a glance

Used by

1 instrument(s)

Related concepts

agentic-system, scalable-oversight, prompt-injection, alignment, capability-elicitation

Primary source

Wallace, E., et al. (2024), 'The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions' (OpenAI) — the canonical industry articulation of instruction-channel hierarchy as a tool-use-safety defence.

Details

Tool-use safety treats the model + tool surface as the unit of analysis rather than the model in isolation. The risk surface expands along several axes: (a) capability composition — a chat-safe model may become capability-dangerous when given a code-execution tool plus internet access; (b) instruction-channel adversaries — tool outputs are an indirect-prompt-injection vector (a web search result containing adversarial instructions); (c) privilege escalation — tools that share authentication with the user may be invoked beyond user intent; (d) resource exhaustion — agents can spend money, compute, or API credits at machine speed; (e) confused-deputy attacks — the tool acts with the user's authority on instructions actually from a third party.

Mitigation patterns include: capability allowlists (only specific tools, specific scopes), human-in-the-loop confirmation for high-impact actions (the OpenAI Operator + Anthropic Computer Use UX patterns), output-isolation tags (Anthropic's tool-result-tag scheme), and gateway-LLM patterns (Wallace et al. 2024 dual-LLM). NIST AI RMF GenAI Profile §2.7 'Value Chain and Component Integration' touches the tool-integration risk. EU AI Act Art. 14 'human oversight' is the closest binding obligation but presumes human-bandwidth-feasible review, which agentic systems break at scale. Industry-side frameworks (Anthropic RSP, OpenAI Preparedness) treat tool-use capability as a tier-relevant signal.

How to cite this article

APA

Policy Window. (n.d.). Tool-Use Safety [Wiki article — Concept]. https://policywindow.org/wiki/tool-use-safety

CHICAGO

Policy Window. n.d.. "Tool-Use Safety." Wiki article (Concept). https://policywindow.org/wiki/tool-use-safety.

HARVARD

Policy Window (n.d.) 'Tool-Use Safety', Wiki article — Concept, available at: https://policywindow.org/wiki/tool-use-safety.

OSCOLA

Policy Window, 'Tool-Use Safety' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/tool-use-safety> accessed [date].

BIBTEX

```
@misc{policywindow-tool-use-safety,
title  = {Tool-Use Safety},
author = {Policy Window},
year   = {n.d.},
howpublished = {tool-use-safety — safety},
url    = {https://policywindow.org/wiki/tool-use-safety},
note   = {Primary source: https://arxiv.org/abs/2402.07896}
}
```