

PUBLISHED: NOVEMBER 16TH · BY SHREYAS KALYANPAD

An undo = the user can put things back the way they were, after the agent acted.

Every agent product I open spends its design budget on the moment before the action. A dialog slides up. You click Approve. After that you are on your own.

That is the wrong half. Permission is the weakest control you can ship, because people learn to click through it.

> *An approval dialog protects the actions a user read. An undo protects the ones they clicked through.*



The agent placed the order. The return is somebody else's afternoon.

1. The approval step is already being skipped

What teams ship: a confirm dialog on every action that matters.

Why it fails: Anthropic measured real use. New Claude Code users run in full auto-approve in about 20% of sessions. By 750 sessions it is over 40%. On its public API, 87% of tool calls on the simplest tasks had human involvement, against 67% on the hardest. The bigger the job, the fewer people read.

Ship instead: assume the dialog gets auto-approved, and design the minute after it.

2. Reversibility belongs to the action, not the approval

What teams ship: one approval flow for everything.

Why it fails: Safin and Balta at fortiss put the split plainly in May 2026. Whether a human must approve is the autonomy design. Whether the thing can be undone is a property of the action itself. Their scale separates drafts

and editable fields from final submissions, record updates and payments, where reversal means a formal cancellation.

Ship instead: label every action reversible or not, before you draw a screen.

3. Staging turns a permanent write into a recoverable one

What teams ship: the agent writes straight to the real record.

Why it fails: once it lands, your options are a correction or an apology.

Ship instead: write staging, the sixth tactic in that paper. Put a draft layer between the agent and the commit. The agent writes a provisional record, a person promotes it. If it is wrong, you bin the draft and the real record was never touched. The cost is latency.

4. Checking before is not recovering after

What teams ship: a verifier, a guardrail, a policy check.

Why it fails: it helps, but not enough. Tanmay Sah and colleagues measured it. Safety mediation intercepted up to 94% of non-compliant agent actions. Strictly safe task completion still stayed under 5% in most settings, and once an action was blocked, recovery ran from 21% down to near zero. The same group found 197 self-modifications that failed recoverability checks. Standard repair strategies recovered 0 of them.

Ship instead: design the recovery screens.

5. When the action leaves your product, hand over the receipt

What teams ship: a success toast.

Why it fails: in February 2025 a Washington Post columnist asked OpenAI's Operator for the cheapest eggs. It ordered through Instacart without confirming. A $13.19 basket became $31.43 with fees and a tip. He found out from a card alert. OpenAI said the agent fell short of its safeguards. The refund was still Instacart's problem. Forrester asked 406 answer-engine users in March 2026 who they would blame if an agent purchase went wrong. 45% said the AI platform, 17% said the brand. The brand still processes the return.

Ship instead: show what was authorised, where it went, and one link that starts the reversal.

Grade the action first

AGENT ACTION	UNDO?	SO SHIP THIS
Draft or editable field	Yes, cheaply	Let it run. Log the change.
File or config change	With a snapshot	Keep the old state. One-click revert.
Outbound email	No	Delay the send. Allow a pull-back.
Payment or order	No	Stage the cart. Show the order ID.
Filing or final record	No	Write a draft. A person promotes it.

The audit

Open your product. Pick three actions your agent takes while nobody watches. For each, answer: what state does this change, where is the old state kept, and how many clicks back is it?

If the last answer is "contact support", you have not shipped an undo. You shipped a dialog and a hope.

In Anthropic's sample of 998,481 tool calls, only 0.8% looked irreversible. It is small because agents still mostly edit code. It will not stay small.

Checked against

Read on 16 November 2026.

- Anthropic, Measuring AI agent autonomy in practice: auto-approve rates, human involvement by task complexity, and the share of actions that look irreversible
- Sah, Srivastava, Sah and Jordan, The Verifier Tax: safety mediation intercepts up to 94% of non-compliant actions, but strictly safe completion stays below 5% and recovery after a block runs from 21% to near zero
- Sah, Sah, Jain and Sah, EvoUndo: of 197 agent self-modifications that failed recoverability verification, conventional repair strategies recovered none
- Damir Safin and Dian Balta, fortiss: Autonomy and Agency in Agentic AI, which separates reversibility from approval and sets out write staging
- Kore.ai, Agentic commerce: when AI agents buy, who handles returns? The Operator grocery order and the March 2026 Forrester blame survey

---