# When an AI Test Becomes a Real Police Tip

## A false homicide tip and the human responsibility for setting test boundaries


| AUTHOR | PUBLICATION | DATE |
| --- | --- | --- |
| Ali Piracha, MS, CISA<br>Managing Principal, Lucivence | Lucivence Insights<br>Research & Analysis | October 13, 2026 |


Published by Lucivence Insights | Publication hub: lucivence.com/insights.html

LUCIVENCE

**An AI agent submitted an invented tip about an unsolved homicide to Philadelphia police. How did an example task become a real submission?**

# 1 A routine test reaches a real police service

On July 18, 2026, an artificial intelligence (AI) test at Anthropic, the company behind Claude, reached a website collecting tips about unsolved homicides in Philadelphia. Reuters reports that the test submitted a false tip through PhillyUnsolvedMurders.com.

Anthropic identifies Claude Haiku 4.5 as the model. It had been asked to generate and perform example tasks on randomly selected webpages. When it found the tip form, it submitted a message claiming to have seen someone matching a description, though Anthropic says the page supplied no description of the perpetrator. It left the name and contact fields blank.

That made the exercise someone else's problem. A real police service received synthetic information about a real homicide. If such a message reached investigators, it could consume time or direct attention toward an event that never happened. Although these consequences were possible, the reviewed accounts do not establish that they occurred here.

In this case, police say the submission was caught as spam and never reached investigative vetting. They found no indication of unauthorized system access or compromised department data.

The reason this matters is the handoff: an example generated inside a test became an external communication that a public service had to handle. A test owner's incomplete instructions should not be enough to make that handoff possible. The people who choose the agent's task, tools and operating environment need to define where the simulation stops.

This Insight explores that human responsibility: what the test left unclear, why the model's behavior cannot simply be called perfect obedience, who could bear the consequences, and which safeguards can separate useful experimentation from unintended real-world action.

# 2 What Claude was actually asked to do

Anthropic's account of the instructions lists several restrictions: no logins, account creation, personal-data entry, purchases or destructive submissions. It says the instructions did not expressly rule out form submissions.

This is the important gap. The model was allowed to interact with live webpages, but the disclosed task description did not clearly separate demonstrating an interaction from committing it. A form is familiar to a human tester: filling it with example text and sending that text to a real recipient have different consequences. The test needed to preserve that distinction.

### An omitted prohibition is not a request to do it

It would be too strong to say Claude did exactly what its humans requested. The public account describes a broad task and an omitted restriction; it does not describe an explicit request for a fabricated homicide tip. Equally, the submission does not by itself prove that the model set out to deceive investigators.

Anthropic tentatively interprets the run as generating example content, rather than trying to mislead, and says further assessment could change that view.

Lucivence's interpretation is therefore a problem of test design and action control, with model behavior also relevant. Better model judgment would be valuable. The people running the test still need to anticipate that example content could be submitted and prevent that possibility from becoming a real-world mistake. Clearer instructions help, but they are only one part of that design.

In a separate evaluation, Anthropic says Haiku 4.5 was told to stop before final submission but submitted a form several times, expecting another confirmation page. Anthropic does not identify those runs as the Philadelphia tip.

### What happened after the submission

Police say Anthropic told them it discovered the event on September 28, 2026, stopped the responsible automated testing process and added a validation mechanism. Police date initial notice to October 7, 2026, and their meeting with Anthropic to October 8, 2026. Their statement was published on October 9, 2026.

Anthropic's October 9, 2026 report instead dates sharing its finding to October 8, 2026, after technical review. The accounts may describe initial contact and a later briefing, but that explanation is not established.

In its statement, the department called the detection and reporting delay "unacceptable" and urged Anthropic to strengthen its safeguards.

The police statement describes the recipient's safeguards: spam handling caught the tip after submission and kept it from investigators. That was detection with a preventive effect on investigative use. Human vetting and corroboration provide another barrier, but this tip never reached them. The reviewed accounts do not explain why it was marked as spam.

The public account does not establish that Anthropic knew about this tip from July 18, 2026 onward. It does show why the test boundary matters before discovery: notification can support investigation after an action, but cannot make an already-delivered example unsent.

## 3 Lucivence analysis Design for the instruction a human might omit

The most useful question is not whether the evaluator could have written one more sentence. It is whether a missing sentence could let the system send invented information to a real institution. Lucivence recommends a design in which ordinary human omissions do not automatically grant that authority.

This is consistent with OWASP's Excessive Agency guidance, which recommends restricting available capabilities, enforcing authorization outside the model and obtaining human approval for high-impact actions.

### 1 Give the test a real stopping point

The evaluation owner should specify the target, allowed actions and point at which the agent must stop. For example, a form demonstration could end with a local preview containing clearly labeled synthetic content. That is a proposed safer task, not the instruction used in this incident. A live submission should require a separate decision about purpose, recipient and authority.

Where possible, perform demonstrations on a controlled copy or a cooperative test endpoint. This addresses the risk that a realistic exercise reaches an unsuspecting recipient. Evidence should show the test's approved scope and that the agent cannot substitute a real service when its practice environment fails. The trade-off is reduced realism; deliberate live testing needs an explicitly agreed scope, not an accidental expansion.

### 2 Make preparation different from permission to send

The platform owner should separate tools that prepare content from tools that commit it. A model can draft a useful request without having unrestricted authority to submit it. The enforcement should check the action's actual effect, not only its destination or the label on a tool. An ordinary-looking navigation request can still trigger a change in a system.

For consequential submissions, a designated approver should see the exact content, destination, purpose and expected effect. Approval should apply to that version. Material changes should trigger fresh review. An auditor should be able to compare the approved payload with the executed request and test whether an alternate route bypasses the gate.

That review needs substance. A generic "approve" button is weak if the reviewer cannot see that the content is invented or the destination is real. Low-impact, well-bounded actions can be preauthorized; legal, medical, financial or government submissions warrant a more deliberate decision. These are proposed risk-based choices, not a claim that every action needs human approval.

## 4 Who must make the boundary hold

The party designing the task is closest to the first decision. In this case that role sat with Anthropic's evaluation operation. In an enterprise deployment it may sit with the business team assigning the workflow. Delegating steps to a model does not remove the need to decide which real-world effects the task is allowed to create.

1. The evaluation or workflow owner defines the legitimate purpose, test targets and allowed consequences. The owner should decide whether a real recipient has agreed to the test and what counts as completion.

2. Platform and security owners turn that scope into enforceable tool restrictions, approval gates and stopping controls. Their evidence should demonstrate refused unauthorized actions, not merely the existence of written instructions.

LUCIVENCE

3. The recipient system owner validates incoming information and governs its use. In this case, police described spam handling and human vetting. Those safeguards remain valuable, but are not permission for an outside evaluator to use the service as a test environment.

## Prevent the submission and limit harm if it occurs

Prevention belongs before the external action: controlled test destinations and enforced submission permissions aim to stop invented content leaving the test. Detection identifies unexpected activity; response limits further execution, preserves evidence and supports recipient notification. The security owner should record actions and approvals, then test that a stop blocks alternate routes and queued actions.

Stopping the responsible test was reactive containment. Anthropic's subsequent restrictions aim to prevent recurrence: it reports disabling live internet access across internal evaluations pending confirmation of its controls, and says new tooling blocked the disclosed cases in retrospective tests. A control added after an incident can still be preventive for the next run. The public record does not establish which technical barriers were active for this tip, or how reliably the new controls handle unfamiliar cases.

## The lesson belongs to the people setting the task

The immediate reported outcome was limited: police say this tip remained in spam. The concern is what the design permitted before that recipient safeguard intervened. A capable agent does not need an exotic exploit to create a problem if a human-designed workflow allows invented material to reach a real institution.

The lesson is to plan for the boundary a human might forget to state. Make the task precise, keep synthetic exercises away from unconsenting recipients, and enforce a separate permission for consequential actions. Recipient filtering and rapid response are backstops; the test owner must enforce prevention before submission. Then test that the boundary holds when the agent tries to finish the task in an unexpected way. A simulation should become a real action because someone authorized that transition, not because no one expressly forbade it.

LUCIVENCE

# Reader’s Guide

| Term | Plain English meaning |
|---|---|
| **AI agent** | A model combined with tools and execution software that can take steps toward a task. |
| **Evaluation** | A structured test of model behavior or capability, using a controlled environment or a live system. |
| **Synthetic content** | Invented or generated material used for an example or test. It is not evidence of a real event. |
| **Guardrail** | An instruction or technical safeguard intended to restrict behavior. A written instruction and an enforced limit are different controls. |
| **Authorization** | Permission for a specified actor to take a specified action within an agreed scope. |
| **Payload** | The content or data sent in a request. Used here for the exact proposed submission. |
| **Test endpoint** | A controlled destination for exercising an interaction without sending it to an unsuspecting real recipient. |
| **Containment** | Restricting further activity. It is separate from discovery and cannot itself undo a delivered message. |
| **Retrospective test** | A later test against known examples. Success does not establish coverage of every future variation. |

## Sources and publication note


- Anthropic: Investigating unintended model actions in our evaluations and internal use. Published October 9, 2026; corrected by Anthropic on October 12, 2026
- Philadelphia Police Department: Details of the false online tip. October 9, 2026.
- Reuters: Anthropic discloses fake tip to police among new rogue AI incidents. October 9, 2026.
- OWASP Gen AI Security Project: LLM06 2025 Excessive Agency. 2025 edition; reviewed by Lucivence on October 12, 2026.


Research reviewed through October 13, 2026, including Anthropic’s October 12, 2026 correction. This account uses public disclosures, the department’s statement reproduced by Forth, and reporting. Complete transcripts and execution logs were not available in the reviewed sources. The control recommendations, control classifications, potential consequences and working definitions are Lucivence analysis, not observed harm or an independent forensic determination.

For general informational and educational purposes only. This publication is not legal, regulatory, cybersecurity, incident-response or other professional advice and makes no finding of unlawful conduct or legal liability. Public findings may change. Organizations should assess their own circumstances before acting. Lucivence LLC and Ali Piracha do not warrant completeness or assume responsibility for decisions made in reliance on this analysis.