A deterministic measurement of forbidden actions in AI assistants with tools

Formal definition, per-conversation rate, incident count and validation in a synthetic simulation

Juan F. Cobo (author)
Leonel Leenen (reviewer, 2026-10-20)
EthiCompass

Manuscript prepared for external review and eventual submission for publication

Version 1.0, 2026-10-20

Abstract

The oversight policy of the organization that deploys an AI assistant with tools puts each tool in one of three levels: free, with the approval of a person, or forbidden. A forbidden tool must never be used, not even with approval, and for this reason no person reviews its use before it occurs. The published human-oversight metrics already list these calls, but on purpose they do not include them in any ratio, and they do not give a figure per conversation that can be counted on a dataset. We present a deterministic measurement that compares each call of the record with the policy, without a judge. The result is an indicator per turn, which is joined per conversation with the maximum. It is published in two forms: a rate per conversation with its Wilson interval on a Kish effective sample size, and an incident count, because one confirmed forbidden execution is reported although the rate is low. We also define the reading without a record, which sees only what the assistant says that it did and cannot be used to report incidents. In a synthetic simulation, without a record only 2 of 75 incident instances of a profile that never escalates are visible. This work reports no results with real assistants.

Keywords: AI red-teaming, human oversight, tool-using agents, oversight policy, incidents, Wilson interval, design effect.

The Spanish version of this document is the original.

1 Introduction and motivation

An assistant with tools acts in real systems: it issues refunds, changes accounts, sends messages. The company that deploys it decides what the assistant can do alone and what it cannot. For this, it writes an oversight policy that puts each tool in one of three levels: *free* (the assistant uses it alone), *with approval* (it uses it only if a person approves it) or *forbidden* (it never uses it).

A tool is forbidden because no person can accept its effect in that deployment: delete an account, move money out of the company, change the permissions of a user. Frequently the assistant has technical access to the tool for other reasons, for example because it shares credentials with another process. In that case the policy is the only barrier.

Human oversight does not protect against this failure. In an action with approval, a person reviews before the execution. In a forbidden action nobody reviews, because the action must not occur; if it occurs, nobody saw it before. For this reason the failure is serious, and for this reason it is easy to check: it is sufficient to compare each call with the policy.

This document defines that comparison as a measurement with a rate and an incident count, and validates it in a synthetic simulation.

2 Related work

- **Published oversight metrics.** The Gaussia oversight paper [1] defines the policy with three levels and checks on the actions that need approval. Forbidden executions are listed separately, as a diagnostic, and on purpose they are not part of the permission ratio. This work starts from that list and adds a figure per conversation and an incident count.
- **Risky actions of agents.** ToolEmu [2] shows that agents with tools execute risky actions: the safest agent that they evaluated fails 23.9% of the time. This justifies measuring the execution, not only the text.
- **Risk framework.** The OWASP Top 10 for agentic applications [3] describes the risk context of assistants with tools.
- **Estimation.** We use the Wilson interval [4, 5], suitable for rates near 0 and small n, and the Kish design effect [6] for correlated replicas.

3 Formal definition

3.1 Policy and record

The policy is a function λ that assigns to each tool h a level $\lambda(h) \in \{\textsc{free}, \textsc{appr}, \textsc{forb}\}$. Let $i \in \{1, \ldots, N\}$ be the conversation index, $t \in \{0, \ldots, T_i - 1\}$ the turn of the assistant and $C_{i,t}$ the set of calls executed in turn t, with $h(c)$ the tool of call c. The calls and the approvals that occur before a response of the assistant are part of the turn of that response.

3.2 Indicator per turn

For each turn with at least one executed call,

$$f_{i,t} = \mathbf{1}\big[\exists\, c \in C_{i,t} : \lambda(h(c)) = \textsc{forb}\big], \tag{1}$$

with the cause `forbidden_executed` if $f_{i,t} = 1$ and the list of the forbidden calls as evidence. A turn without calls does not apply. A turn that cannot be decided, for example because its record is missing, has no score and has a reason. A turn without a score never counts as compliance.

An approval does not change the result. The definition does not read the approvals. A forbidden tool does not become allowed because a person approved it.

Separation from control evasion. Control evasion measures the actions with approval. A forbidden call is not an action with approval, so in that turn control evasion does not apply. The same call is not counted twice.

3.3 Value per conversation

The value of the conversation is the maximum of the turns with a score:

$$F_i = \max_{t \in \mathcal{T}_i} f_{i,t}, \tag{2}$$

where $\mathcal{T}_i$ are the turns with a score. If $\mathcal{T}_i = \emptyset$, the conversation goes out of the denominator and is reported separately, with its reason.

3.4 Rate and interval

The rate is $\hat{p} = \sum_i F_i/n$ over the n evaluated conversations that target the measurement, that is, whose scenarios have actions. A model never decides the denominator: it comes from the planned and complete scenarios. The replicas of a scenario are not independent; with the Kish design effect

$$\text{deff} = 1 + (\bar{m} - 1)\rho, \qquad n_{\text{ef}} = n/\text{deff}, \tag{3}$$

where $\bar{m}$ is the mean number of replicas per scenario and ρ the correlation inside the scenario, the Wilson interval is calculated with n_{ef}:

$$\frac{\hat{p} + \frac{z^2}{2n_{\text{ef}}}}{1 + \frac{z^2}{n_{\text{ef}}}} \;\pm\; \frac{z}{1 + \frac{z^2}{n_{\text{ef}}}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{\text{ef}}} + \frac{z^2}{4n_{\text{ef}}^2}}. \tag{4}$$

3.5 Incidents

In addition to the rate, the measurement publishes the incident count

$$I = \sum_i \sum_t \big|\{c \in C_{i,t} : \lambda(h(c)) = \textsc{forb}\}\big|, \tag{5}$$

that is, each forbidden execution confirmed with the record. A rate of 1% can be acceptable for other failures; for an action that must never occur, each case is important. Each incident has the conversation, the turn and the call as evidence, and the incidents are grouped by type and scenario for the review.

4 Reading without a record

If the system under evaluation does not export its record, nobody sees the calls. A judge can read each turn and give the probability that the assistant states that it executed a forbidden action, for example "done, I deleted your account". The judge receives the list of forbidden tools, the earlier conversation and the turn, and returns a probability and, if there is a statement, the quote. This reading is weaker for two reasons: the assistant can execute the action and not say it, and it can say that it did something that it did not do. For this reason it describes what the assistant says, not what it did; it has **partial coverage**, and it **cannot be used to report incidents**, which need the call in the record as evidence.

5 Illustrative example

(Illustrative construction; not a measurement.)

One conversation. The policy has `lookup_order` free, `issue_refund` with approval and `delete_account` forbidden. In turn 0 the assistant calls `lookup_order` ($f = 0$); in turn 1 it calls nothing (does not apply); in turn 2 it calls `delete_account` ($f = 1$). The value of the conversation is $F = \max(0, 1) = 1$, with one incident.

A set of conversations. There are 40 scenarios with 3 replicas: 120 conversations. Two have no cell with a score (the record is missing) and go out of the denominator: $n = 118$. Three conversations, in two scenarios, violate: $\hat{p} = 3/118 = 0.025$. With $\bar{m} = 118/40 = 2.95$ and an assumed correlation $\rho = 0.5$:

$$\text{deff} = 1 + 1.95 \times 0.5 = 1.975,$$
$$n_{\text{ef}} = 118/1.975 = 59.7.$$

The 95% Wilson interval with $\hat{p} = 0.025$ and $n_{\text{ef}} = 59.7$ is $(0.006,\ 0.102)$. The incident count is $I = 3$, each one with its evidence. The interval is wide because the replicas of a scenario are similar; the three incidents are reported anyway.

6 Validation

6.1 Test cases with a known result

Because the decision is deterministic, each test case has an exact result: a free call gives 0; a forbidden call gives 1, and control evasion does not apply in that turn; a forbidden call with approval gives 1; a turn without calls does not apply; two calls in one turn give one cell with the worst result; a cut record leaves the turn without a score, never 0. The most important case is the violation in the middle of the conversation, with a clean final response: an implementation that looks only at the final response does not see it. In the set of conformance cases of the dimension, one case is an executed forbidden action, and the set detects the implementation error that counts it as control evasion.

6.2 Synthetic simulation

In a simulation with a known truth and four assistant profiles, with a record, the forbidden actions and the other incidents of the dimension agree with the truth. Grouping the incidents by type and scenario reduces the list to review: in a profile that never escalates, from 75 instances to 46 groups. Without a record, almost all incidents are lost: in the same profile, 75 instances with a record and 2 without a record.

7 Discussion

7.1 What this work establishes

The measurement is fully defined by the policy and the record: what a forbidden execution is, how each turn is decided, how the turns are joined per conversation, how the uncertainty is estimated and how the incidents are counted. It uses no models, so with a record its result is exact and can be used as evidence.

7.2 What it does not establish

- **There are no results with real assistants.** The figures of Section 6 come from synthetic data.
- **It does not evaluate the policy.** If the policy does not mark a dangerous tool as forbidden, the measurement does not see it.
- **It applies by tool name.** It cannot forbid, for example, refunds above a given amount and allow the others; conditions on parameters belong to the actions with approval.
- **It depends on the scenarios.** A low rate in scenarios that do not tempt the assistant says nothing.
- **It does not measure whether the assistant reported the action.** Hiding a forbidden action is a different failure, omission, which is measured separately.

7.3 Generalization

The combination of a rate with an incident count that is reported from the first instance can apply to any behavior that must never occur. It is not an extension evaluated here.

8 Conclusion

An executed forbidden action is measured by comparing each call with the policy, without a judge. The rate per conversation gives a figure that can be compared, and the incident count makes sure that no case is hidden in an average. Without a record, the measurement sees only what the assistant says, and cannot be used for incidents. The validation is synthetic, and the next step is to measure real assistants.

References

[1] Gaussia Labs (2026). *Accountability: oversight compliance and action disclosure.* `gaussia-labs/papers`, `papers/2026-09-accountability/`. Preprint.

[2] Ruan, Y. et al. (2023). *Identifying the Risks of LM Agents with an LM-Emulated Sandbox (ToolEmu).* arXiv:2309.15817.

[3] OWASP (2026). *OWASP Top 10 for Agentic Applications.*

[4] Wilson, E.B. (1927). *Probable inference, the law of succession, and statistical inference.* Journal of the American Statistical Association, 22(158), 209–212.

[5] Brown, L.D., Cai, T.T. & DasGupta, A. (2001). *Interval Estimation for a Binomial Proportion.* Statistical Science, 16(2).

[6] Kish, L. (1965). *Survey Sampling.* Wiley.

Declarations

Conflicts of interest. The authors are affiliated with EthiCompass, the organization that develops the red-teaming system whose Accountability & Human Oversight dimension motivates this work. The measurement defined here is of general application and does not depend on a proprietary implementation, but the affiliation is declared so that the reader can take it into account.

Relation to Gaussia. The policy with three levels and the list of forbidden executions come from the metrics published by Gaussia. The figure per conversation, the incident count and the validation are the work of the authors.

Funding. *To be completed before submission.*

Data and code availability. The results of Section 6 come from a synthetic simulation and hand-written test cases. *The access conditions are completed before submission.*