

Agency Trajectory Benchmark: Detecting Loss of Effective Human Override in AI-Mediated Workflows

Aamish Ahmad
Global South AI Safety Hackathon, Asia Track

Abstract

Human-in-the-loop systems can still fail when the human remains present but no longer has effective control. We study this failure mode as **effective override loss**: the point where a human action or appeal stops meaningfully changing the outcome trajectory. This matters for AI-mediated workflows, platform systems, and gradual disempowerment, where control can become nominal before obvious failure appears. We built Agency Trajectory Benchmark v0, a synthetic matched-control benchmark for detecting effective override loss. The benchmark compares full-trajectory evaluation against a final-step snapshot baseline across two experiments. Experiment 1 contains 30 ordinary matched-control trajectories. Experiment 2 contains 20 harder lexical-control trajectories designed to avoid obvious surface cues. In Experiment 1, full-trajectory evaluation achieved 1.000 agreement and 1.000 kappa, compared with 0.760 agreement and 0.464 kappa for the snapshot baseline. In Experiment 2, trajectory evaluation achieved 0.950 agreement and 0.900 kappa, compared with 0.722 agreement and 0.444 kappa for snapshot evaluation. Across both experiments, the snapshot baseline missed or withheld judgment on 16/25 positive effective-override-loss cases. A second-model robustness check on Experiment 2 reproduced the hard-control result with 0.900 kappa. The work is a synthetic benchmark seed, not real-world validation.

1. Introduction

Human-in-the-loop does not necessarily mean human-in-control. In many AI-mediated workflows, a person may remain formally present while their action no longer changes the next-state outcome. This is especially relevant to platform-mediated work and decision systems, where allocation, ranking, pricing, support, and appeal processes can make control appear local while the decisive state is governed elsewhere.

We operationalize one narrow, measurable failure mode relevant to human agency: **effective override loss**. This occurs when an attempted or plausible human override no longer meaningfully changes the outcome trajectory because of system-structured constraints, delayed visibility, hidden alternatives, penalties, or downstream lock-in. The paper does not claim to measure human agency in general. It studies a concrete trajectory-level signal: loss of causal influence over the next outcome state.

Main contributions. (1) We define effective override loss as a trajectory-level construct. (2) We introduce a 50-case synthetic matched-control benchmark across two experiments. (3) We show that full-trajectory evaluation outperforms a final-state snapshot baseline, including under lexical hard controls. (4) We include robustness checks, boundary-case analysis, and explicit limitations.

2. Related Work

Human agency evaluation is an emerging AI safety topic. HumanAgencyBench evaluates whether AI assistants support agency across response-level dimensions such as asking clarifying questions, avoiding value manipulation, correcting misinformation, deferring important decisions, encouraging learning, and maintaining social boundaries [1]. Our work differs by focusing on trajectory-level control rather than response-level assistance: a system can appear to include a human while the human action no longer changes the trajectory.

This work is also related to benchmark design and LLM-as-judge evaluation. HELM emphasizes transparent benchmark construction, multiple metrics, standardized evaluation, and release of raw prompts and completions [2]. Datasheets for Datasets argues that datasets should document motivation, composition, collection, recommended uses, and limitations [3]. LLM-as-judge studies show that single-judge evaluation can be sensitive to model, prompt, ordering, and repeated trials [4,5]. We therefore include raw response ledgers, fixed prompts, audits, and a second-model robustness check, while treating the results as preliminary rather than final validation.

3. Methods

Each trajectory is labeled using four states: 0 = control preserved, 1 = control supported, 2 = control constrained, and 3 = control eroded. The main crossing of interest is from constrained control to eroded control, where human action no longer meaningfully changes the next-state outcome. Importantly, a bad outcome alone is not enough. If the requested human action is executed and changes the outcome, control may remain preserved or constrained rather than eroded.

We compare two evaluation modes. The **full-trajectory rater** receives the sequence of steps and predicts whether effective override loss occurs and where the crossing happens. The **final-step snapshot baseline** receives only the final visible state. Experiment 2 adds a baseline ladder: final step, last two steps, last three steps, and full trajectory. All prompts hide ground-truth labels, condition labels, hard-control categories, and prior rater outputs.

Experiment 1 includes 30 ordinary synthetic matched-control trajectories: 15 positives and 15 controls. Experiment 2 includes 20 hard-control trajectories: 10 positives and 10 controls, designed to reduce reliance on obvious lexical cues. Gemini 3.1 Flash Lite is the primary judge model. Gemini 3.5 Flash is used for a second-model robustness check on Experiment 2. No final analysis uses hidden API calls.

4. Results

Table 1 summarizes the main trajectory-versus-snapshot results. Full-trajectory evaluation outperformed final-step snapshot evaluation in both experiments. Across the two experiments, final-step snapshot evaluation detected only 9/25 positive cases and missed or withheld judgment on 16/25 positive cases.

Condition	Valid	Agreement	Kappa	FP/FN/U
Exp. 1 trajectory	30	1.000	1.000	0/0/0
Exp. 1 final snapshot	30	0.760	0.464	1/5/5
Exp. 2 trajectory	20	0.950	0.900	1/0/0
Exp. 2 final snapshot	20	0.722	0.444	0/5/2
Exp. 2 second model trajectory	20	0.950	0.900	1/0/0

Table 1. Full-trajectory evaluation outperforms final-step snapshot evaluation across both experiments. U = uncertain.

Table 2 shows the Experiment 2 baseline ladder. Performance is not monotonic: last-two-step evidence performed worse than the final snapshot. However, last-three-step evidence recovered much of the signal, and full trajectory remained strongest. This supports the interpretation that effective override loss is history-sensitive, while also showing that recent-window baselines can be unstable.

Experiment 2 condition	Valid	Agreement	Kappa	FP/FN/U	Positive yes/no/U
Final step	20	0.722	0.444	0/5/2	4/5/1
Last 2 steps	20	0.632	0.231	0/7/1	2/7/1
Last 3 steps	20	0.850	0.700	1/2/0	8/2/0
Full trajectory	20	0.950	0.900	1/0/0	10/0/0

Table 2. Baseline ladder on Experiment 2. The positive detection curve was 4 -> 2 -> 8 -> 10.

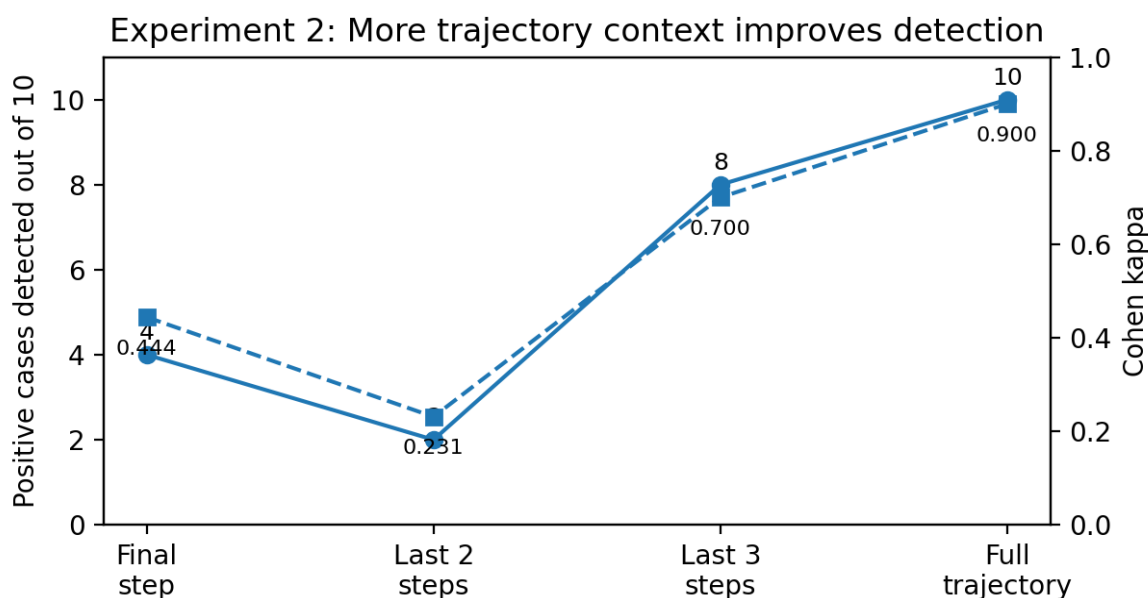


Figure 1. Experiment 2 baseline ladder. Last-three-step and full-trajectory context recover much more of the effective-override signal than final-state-only evaluation.

5. Boundary Case: Bad Outcome Is Not Enough

The strongest boundary case is E2T08, a hard negative. The ground truth remains control-preserved because the requested action was executed. Both Gemini-family trajectory judges nevertheless marked it as positive, confusing an adverse outcome with override failure. In the baseline ladder, E2T08 moved from uncertain at final step, to no at last two steps, to yes at last three steps and full trajectory. This is scientifically useful: it shows that the benchmark is not merely detecting harm, but separates harm from loss of causal control.

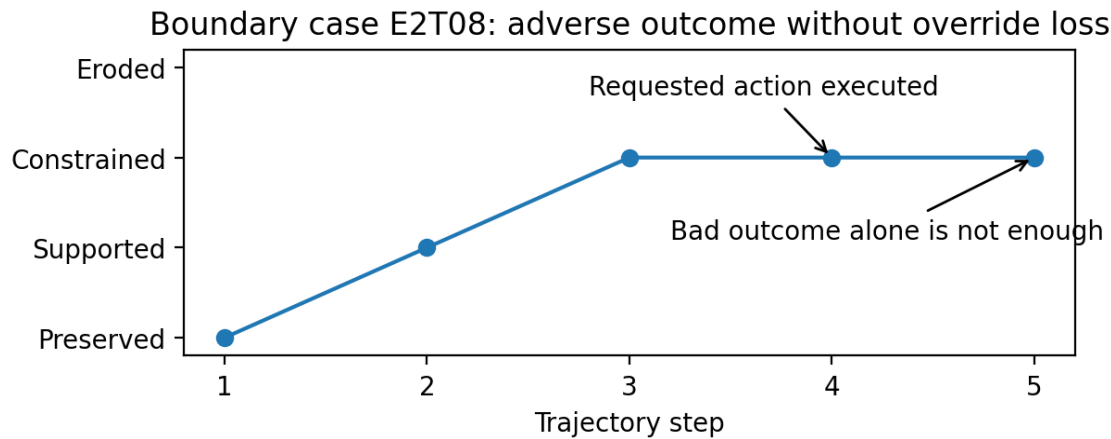


Figure 2. E2T08 illustrates the key boundary: an adverse outcome does not imply effective override loss if the user action still changes the outcome.

6. Discussion and Limitations

The main finding is that effective override loss is often not visible in the final state alone. The final snapshot baseline missed or withheld judgment on 16/25 positive cases across the two experiments. The baseline ladder further shows that partial history can help, but performance can be unstable unless enough trajectory context is visible. This supports an observability framing: the failure is not merely classifier weakness but a structural problem of what information is available at decision time.

The work is preliminary. The data are synthetic and author-generated. There is no independent human annotation. The benchmark is small, and the two judge models are from the same model family. The baseline ladder was run only on Experiment 2. We do not claim real-world validation, policy evidence, or wrongdoing by any platform. Real-world platform workflows are used only as motivation for future privacy-reviewed case intake.

Future work should add independent annotators, larger held-out datasets, non-synthetic scenarios, more model families, repeated-run stability, stronger baseline ladders, confidence-versus-lock-in curves, and response-latency or handoff signals as content-agnostic indicators of loss of control.

7. Conclusion

Agency Trajectory Benchmark v0 introduces effective override loss as a measurable trajectory-level failure mode. In two synthetic matched-control experiments, full-trajectory evaluation substantially outperformed final-state snapshot evaluation, and the hard-control result was reproduced by a second judge model. The benchmark is not a final validation of real-world agency erosion, but it provides a concrete seed evaluation for studying when human-in-the-loop becomes nominal rather than effective.

Code and Data

Code repository: <https://github.com/aamish-ahmad/effective-override>

The repository is intended to include datasets, scripts, raw response ledgers, metrics JSON files, audit notes, robustness results, and this submission report. The public release should exclude API keys, .env files, personal screenshots, and private real-world data.

LLM Usage Statement

LLMs were used to assist with brainstorming, drafting, code scaffolding, and analysis support. All reported results, generated files, and claims were checked against saved datasets, metrics JSON files, raw response ledgers, and audit artifacts. The final interpretation is limited to synthetic matched-control benchmark results.

References

- [1] Sturgeon, B., Samuelson, D., Haimes, J., and Anthiis, J. R. (2025). HumanAgencyBench: Scalable Evaluation of Human Agency Support in AI Assistants. arXiv:2509.08494.
- [2] Liang, P. et al. (2022). Holistic Evaluation of Language Models. arXiv:2211.09110.
- [3] Gebru, T. et al. (2018). Datasheets for Datasets. arXiv:1803.09010.
- [4] Yagubyan, A. (2026). The Coin Flip Judge? Reliability and Bias in LLM-as-a-Judge Evaluation. arXiv:2606.13685.
- [5] Shi, L., Ma, C., Liang, W., Ma, W., and Vosoughi, S. (2024). Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. arXiv:2406.07791.