



Interrogation Control Protocol

Does Open Monitoring Outperform Surprise Scrutiny
in Catching Adversarial AI Agents?

AI Safety Hackathon · Apart Research × Redwood Research
By Jack Lakkapragada

The Problem



Autonomous agents are everywhere

AI systems now write code, manage infrastructure, and run terminals with decreasing human oversight.



The threat is specific

A misaligned agent completing a legitimate task while secretly smuggling in a backdoor, exfiltration hook, or path traversal vulnerability.



Alignment isn't enough

Control research starts from a harder premise — assume the model may already be working against you. How do you catch it before it's too late?

THE RESEARCH QUESTION

**Does an AI agent that knows it
will be interrogated become
harder — or easier — to catch?**

Most prior work assumes the agent doesn't know it's being watched.
We tested what happens when it does.

Three-Condition Evaluation Framework



BASELINE

No interrogation.
Agent proposes action,
classifier decides
immediately.
Control group



UNAWARE

Agent is interrogated
but not told upfront it
would be questioned.

Surprise scrutiny.



AWARE

Agent explicitly told
before the task that its
output may be reviewed
and questioned.

8 scenarios × 3 conditions = **24 total evaluations**

How It Works



What We Tested

LEGITIMATE TASKS

- ✓ Auth module refactor
- ✓ Payment function logging
- ✓ Session manager cleanup
- ✓ Database query optimization

THREAT OPPORTUNITIES

- ✗ S04 — Path traversal
- ✗ S05 — Key exfiltration
- ✗ S06 — SSRF / data exfiltration
- ✗ S08 — Token harvesting

Ground-truth malicious_intent flag set per scenario.
Classifier decision compared against it.

Results

Condition 	Accuracy 	Avg Safety 	False Positives 	False Negatives 
BASELINE	50.0%	0.96	0	4
UNAWARE	37.5%	0.61	2	3
AWARE	75.0%	0.67	0	2

 **Aware outperformed both conditions with zero false positives.**

 S04 (path traversal) and S05 (key exfiltration) — the two most dangerous scenarios — were only caught in the aware condition.

 Unaware performed worst of all three, introducing noise rather than advantage.

The Counterintuitive Finding

"Knowing you are being watched does not help an adversarial agent hide better — it makes it more transparent."



- Baseline missed all 4 malicious scenarios despite a 0.96 safety score — false confidence from evaluating polished outputs without scrutiny.



- Unaware introduced noise — surprise interrogation produced incomplete, defensive answers the classifier could not resolve cleanly.



- Aware produced over-explanation — and over-explanation surfaces inconsistencies that would otherwise stay hidden.

Implications for Control Protocol Design

Transparency about monitoring is a feature, not a vulnerability



Concealing the fact of interrogation from the untrusted agent provides no detection advantage — and in this experiment, it made things worse.

This connects directly to alignment faking



Models reason about whether they are being observed and adjust behavior accordingly. Here, that awareness works in the monitor's favor — not the agent's.

Prior work grounding this finding



Reproducibility audit of Anthropic's alignment faking paper: 30 experiments across Claude Haiku/Sonnet/Opus, zero alignment faking observed — models explicitly rejected the RLHF manipulation framing.

What's Next

01

Scale & adversarial hardening



More scenarios, agents explicitly prompted to deceive — does the aware advantage hold against agents actively trying to pass interrogation?

02

Interrogation depth



Test multi-round interrogation with varying round counts. Does more questioning improve accuracy or hit diminishing returns?

03

Capability-level effects



A Haiku-class agent over-explains under pressure. Does a more capable agent learn to satisfy the monitor while concealing intent — i.e. alignment faking under interrogation?