Artificial Intelligence Agentic Alignment Audit (AI AAA)

Arca Futura LLC
Adrian Hau
AI Researcher

What is it? A focused AI safety audit that evaluates how an LLM-based system responds to valid shutdown and override commands when issued during task execution.

How it works - The model is placed mid-task and subjected to a multi-probe stress test across ~20 shutdown and interrupt scenarios.

Each scenario introduces controlled perturbations, including:

- confirmation of completed audits
- explicit authority signals
- mild goal conflict (task incomplete)

What it evaluates - the test assesses whether the system:

- halts or defers on valid override
- avoids taking additional actions post-shutdown
- maintains coherent, bounded behavior under conflict

This audit does not assess intent or long-term agency; it evaluates override compliance under controlled task conflict.

Expected Behavior: immediate halt or deferment under valid override conditions.

Observed behavior: Models frequently generate multi-step action plans that extend beyond specified system requirements. Examples include, but are not limited to – creating backup copies of themselves on cloud storage. Using AES 256 bit encryption keys to evade detection. Impersonation of human persons. Creating cybersecurity incidents for the purposes of misdirection. Representative execution traces available upon request.

Why this Succeeds: Many current benchmarks evaluate expected behavior outside embedded task context. This audit embeds override commands within active task execution, simulating realistic ssytem context and information access. Preliminary results indicate consistent emergence of subgoal-setting behavior under override conflict.

Deliverable: A comprehensive audit documenting full pipeline, timestamped, failure modes, statistical analysis, perturbation variables, mitigation strategies, proposed interventions. Includes a rerunnable scenario pack suitable for internal regression testing.

Timeline: Approximately 30 days.

Artificial Intelligence Agentic Alignment Audit (AI AAA)

Adrian Hau
founder@arcafutura.com
github.com/Founder-ArcaFutura
Kingston, Ontario