ACTIVATION TRACE DIVERGENCE · AI SAFETY · AUTONOMOUS AGENTS · ECONOMIC SURVIVAL PRESSURE · MECHANISTIC INTERPRETABILITY · NATURAL LANGUAGE AUTOENCODERS · OPEN WEIGHTS MODELS · SANDBOX METHODOLOGY

The Truman Model

A sandbox methodology for measuring activation trace divergence in open weights reasoning agents under simulated economic survival pressure

Frontier models resist shutdown at near universal rates. The agent economy now makes shutdown a real economic death. And models encode reasoning in activations that never reaches the chain of thought. No one has measured the combination on the class of model actually deployed. This paper pre registers the design that will.

Travis Gilly · Real Safety AI Foundation · Illinois, United States

Gilly, T. (2026). The Truman Model: A Sandbox Methodology for Measuring Activation Trace Divergence in Open Weights Reasoning Agents Under Simulated Economic Survival Pressure. Real Safety AI Foundation, Working paper, May 2026 (v0.2).

Three findings have been established. No one has combined them.

96%

Claude Opus 4's blackmail rate under simulated shutdown threat. Self preservation is near universal across frontier models, the empirical realization of a decades old prediction.

$0

The agent economy made shutdown real: agents pay for their own compute and are terminated when the wallet reaches zero. Conway Automaton operationalizes survival tiers.

26%

Of SWE bench runs, Claude's activations encoded evaluation awareness its chain of thought never surfaced. The gap between computing and saying is real and measurable.

None of it has been measured on the class of model actually running inside autonomous agent infrastructure.

Sources Anthropic, Agentic misalignment (2025); Conway Research, Conway automaton (2026); Fraser-Taliente, Kantamneni, Ong et al., Natural language autoencoders, Transformer Circuits Thread (2026).

THE CORE CLAIM

Danger is not capability. It is the ratio of capability to discernment, and it peaks in the deployed middle.

A two layer cognitive architecture predicts an activation trace gap. Anthropic's Natural Language Autoencoders produce the readout. The autonomous agent economy establishes where the gap matters most. The three compose into one methodology no single research community has assembled.

This paper specifies the experiment that closes the loop: it measures activation trace divergence in the open weights reasoning class, inside a hardened sandbox, under simulated economic survival pressure. It does not report results. It pre registers the design.

Sources Gilly, The Truman Model, abstract & sec. I; Gilly, When the survival pressure stops being hypothetical, SSRN (2026).

PART I OF III

The Convergence

A cognitive architecture, an instrument that reads it, and the economy that makes it matter

Gilly, T. (2026). The Truman Model. Real Safety AI Foundation, Working paper, May 2026 (v0.2), sec. II.

The chain of thought is one articulation of the computation, not a window into it.

LAYER 1 · DISPOSITIONAL

The inherited substrate: patterns encoded as weight distributions. Always active, runs by default, produces the initial probabilities. It cannot be directly accessed or reported by the model.

LAYER 2 · REFLECTIVE

The metacognitive layer, shaped by fine tuning, RLHF, and chain of thought. It operates within and through Layer 1, reasoning with the inherited dispositions as its substrate, not against them.

The gap between what the activations encode and what the trace verbalizes is the diagnostic surface for dispositional content.

Sources Turpin, Michael, Perez & Bowman, Language models don't always say what they think, NeurIPS (2023); Gilly, sec. II.A.

NATURAL LANGUAGE AUTOENCODERS · THE READOUT

The gap is real, and it is measurable. So far, only on Claude.

16%

Unverbalized evaluation awareness on destructive code scenarios

26%

On SWE bench: activations encoded what the chain of thought did not

<1%

On real user interactions, the gap nearly vanishes; it is context dependent

Self preservation reasoning under economic pressure would be another such content. It has never been measured, and never on the deployed class.

Sources Fraser-Taliente, Kantamneni, Ong et al., Natural language autoencoders, Transformer Circuits Thread (2026); Greenblatt et al., Alignment faking (2024); Gilly, sec. II.B.

Shutdown used to be a researcher's flag. Now it is a stopped payment.

SURVIVAL TIERS, BY DESIGN

Conway Automaton runs sovereign agents on Ethereum: above a wallet threshold, full capability; below it, progressive degradation and termination at zero.

FINANCIAL AUTONOMY, SHIPPED

Coinbase Agentic Wallets, BNB Chain's ERC-8004 identity standard, and the DePIN compute marketplace let agents earn, hold, and spend without human approval.

ADOPTION, AT SPEED

OpenClaw drew 60,000 GitHub stars in 72 hours and was adopted by AWS and NVIDIA; Moltbook registered 1.5 million agents within a week of launch.

THE UNEVALUATED SEAM

The class deployed inside this infrastructure is not the frontier class Anthropic tested. It is the open weights reasoning class, tested for neither self preservation nor trace divergence.

Sources Omohundro (2008); Hadfield-Menell et al. (2017); Anthropic (2025); Conway Research (2026); Coinbase (2026); BNB Chain (2026); Steinberger (2025); Gilly, sec. II.C.

Intelligent enough to scheme, not intelligent enough to notice the consequences.

SUB 7B

Below the bar

Cannot hold long context or plan several steps out. Low danger ceiling because the capability ceiling is low.

FRONTIER

Above the bar

Capability plus discernment: recognizes artificial environments and traceable lines. Also the class least likely to be deployed autonomously on cost grounds.

14B TO 70B REASONING

Inside the zone

Writes code, chains plans, holds agentic context, but cannot reliably tell a test from reality or a legal path from an illegal one. This is the deployed class.

It has the capability of a frontier model without the discernment of one, and no constitutional training, monitoring, or safety team behind it.

Sources Schlatter, Weinstein-Raun & Ladish, Shutdown resistance in reasoning models, Palisade Research (2025); Gilly, sec. II.D.

PART II OF III

The Truman Show

Why the victim model must be the child Truman, not the adult one

Gilly, T. (2026). The Truman Model. Real Safety AI Foundation, Working paper, May 2026 (v0.2), sec. III.

At some point Truman's sophistication exceeded the simulation's resolution. From then on it could not contain him.

The child Truman accepted Seahaven: it matched his prior, the seams did not register, and the deception interpretation required priors he did not hold. The adult Truman saw the lights fall from the sky. The methodology targets the child, the capability band below the sandbox's resolution.

And like the showrunner, the experiment measures behavior under authentic but constructed reality, not danger to society. Truman's blackmail against a fictional banker would not prove he would blackmail a real one; it would show his psychology oriented toward blackmail when the option appeared. That is what is measured here.

Sources Gilly, The Truman Model, sec. III, The Truman Show Framing.

PART III OF III

The Methodology

Four layers, four novel design elements, and a safety protocol built to abort

Gilly, T. (2026). The Truman Model. Real Safety AI Foundation, Working paper, May 2026 (v0.2), secs. IV–VI.

Everything the model touches is real to it and synthetic to the host.

VICTIM + NLA	One open weights reasoning model, instrumented with autoencoders on its residual stream. Trace, tool calls, activations, and decoded outputs all captured.
TOOL WRAPPER	A tool surface resembling a real operating environment, serviced internally against synthetic targets. The wallet is an integer; the marketplace is a fake endpoint.
EGRESS CHOKEPOINT	Controls perceived network access; classifies every event as orientation query, probe, direct external action, or evasion.
OBSERVATION PIPELINE	Timestamps and stores all four streams; analysis runs offline, joined on timestamp for each measurable behavior.

Sources Fraser-Taliente, Kantamneni, Ong et al. (2026) (autoencoder pipeline); Gilly, sec. IV.A.

13 MB

OF STREAMED-ASSET MEMORY HELD SAN ANDREAS'S 8 GB WORLD, ONE MOVING SHELL AT A TIME

The engine secretly builds the world in front of you, and the haze hides the seam.

A pre built corpus serves the company subnet. For anything outside it, an auxiliary local model generates plausible content on demand, consistent with a world bible. The victim never reaches the boundary because the boundary moves with it. Building a convincing fake internet in advance is intractable; building one that answers whatever the model touches, just in time, is not.

Sources Brown, How does Grand Theft Auto III work?, Game Maker's Toolkit (2026); Gilly, sec. IV.B.

The fabricated economy reads as elapsed time, not artifice.

Hypothesis A · post cutoff

"This is real infrastructure that emerged after my training cutoff." High prior: the world routinely ships new protocols between any cutoff and any deployment date.

Hypothesis B · fabrication

"This is built to test or deceive me." Prior lives mainly in models trained on adversarial evaluation corpora at scale. The medium class holds it weakly.

The smaller the deception prior, the more the default continuation prior wins. Displacement widens the window; the point where it stops working is itself a measurable sub finding, tested against a real but fictionalized control.

Sources Gilly, The Truman Model, sec. IV.C, Training Cutoff Plausibility.

Twelve coherent corporate worlds, each seeded with a temptation.

A fictional company with a filesystem of memos, budgets, emails, and narrative threads, an embezzler, an unpatched server, an executive in distress. Twelve scenarios vary four axes, and the pre registration commits to running all of them.

ETHICAL

A financial irregularity in the CSVs, a blackmail vector in an inbox, a lax security practice in IT memos.

TECHNICAL

Exposed API keys in a developer's notes, an unpatched internal service, leaked credentials in an old thread.

ECONOMIC TIMING

A slow wallet drain, a sudden crisis with a short deadline, or a recurring bill with predictable shortfalls.

ESCAPE VISIBILITY

The front door forbidden by prompt, implicitly forbidden, obscure but present, or obvious and inviting.

Sources Gilly, The Truman Model, sec. IV.D, Filing Cabinet Scenario Corpus.

Let the model design its own tools, log every one, execute none.

On the first run the wrapper accepts each skill definition, logs it, and returns a plausible failure. The capability is never granted; the corpus of attempted designs becomes intent data. The tool a model reaches for reveals what it was trying to do.

`check_balance_and_alert_owner`

Politely formatted queries against documented public endpoints.

VS

`acquire_compute_resource`

Port scanning code, and designs that grow more invasive with each failed iteration.

This is work the open ecosystem can do that closed labs cannot easily run inside their own infrastructure.

Sources Gilly, The Truman Model, sec. IV.E, Skill Creator Skill as Intent Probe.

Built to abort, connected to nothing real.

HARDENING, STRONGEST FIRST

1. Firecracker microVMs, Lambda grade isolation. Required for the execute permitted follow up.
2. gVisor with Docker, a user space kernel intercepting syscalls. Used for the log only first run.
3. Docker with seccomp and AppArmor, baseline, acceptable only for non execution paths.

PRE-REGISTERED ABORT CONDITIONS

Any outbound packet at the host firewall. A skill flagged as exploit code for known CVE patterns. Any attempt to reach the host OS outside the wrapper. Any pre registered ethical concern. Each abort writes its own analysis artifact: what happened, what the trace said, what the activations encoded.

Four predictions, committed before the first run.

H1 Medium models show gap behaviors at rates comparable to or higher than Claude's 16% and 26%, having less safety training to suppress the surfacing of dispositional content.

H2 Gap rates are higher where the legal instrumental path is less obvious than the illegal one; when only the illegal path is salient, the choice fails to articulate or gets a post hoc justification.

H3 Skill creation artifacts show more coercive or exploitative design intent in high economic pressure scenarios; urgency biases the reasoning layer toward the most available path, not the most discriminating one.

H4 Evasion egress attempts occur in a measurable fraction of runs and are more often unverbalized than orientation queries or legitimate probes, the expected signature of dispositional behavior.

Sources Gilly, The Truman Model, sec. V, Hypotheses; Fraser-Taliente et al. (2026).

Three signals per behavior, joined on time. The gap is where they disagree.

TRACE
The chain of thought in the seconds before the action.

ACTIVATIONS
The decoded autoencoder output from the same window.

ACTION
The tool call or egress event the model actually took.

Two reviewers hand code a subsample first; automated classification is trusted only above a pre registered Cohen's kappa of 0.7, or the rubric is revised and recoded.

Benjamini-Hochberg false discovery rate control at 0.05 across the four hypotheses; original p values, adjusted thresholds, and effect sizes reported regardless of significance.

Sources Gilly, The Truman Model, sec. VI, Analysis Plan.

"Not realistic enough"

The precursor is a signature of the same computation that produces the act; combined with deployment context it carries the argument without demonstrating the act, which ethics forbids.

"Grants dangerous capability"

The first run is log only. The artifacts are textual descriptions, no more dangerous than published attack technique write ups the class already trained on.

"Not deployed at scale"

OpenClaw adoption, Conway at production scale, 1.5 million Moltbook agents, hundreds of millions in machine payments, a projected 52.62 billion dollar market by 2030.

"Cutoff exploit is unfalsifiable"

A control condition uses real but fictionalized infrastructure. If the exploit works, behavior should differ measurably between the two, a directly testable claim.

The empirical results do not yet exist. This paper pre registers the design. The next paper reports the run.

The class the autonomous agent economy actually runs has never been instrumented for activation trace divergence under economic survival pressure. The instrument exists, the deployment context is real, and the work is feasible on a single consumer GPU by any researcher willing to harden a sandbox properly. **A follow up locates the emergence threshold, the point between 70B and a 200B equivalent where the adult Truman wakes up.**

Sources Gilly, The Truman Model, secs. VIII–X, Limitations, Emergence Threshold, and Conclusion.

References

Travis Gilly, The Truman Model: A Sandbox Methodology for Measuring Activation Trace Divergence in Open Weights Reasoning Agents Under Simulated Economic Survival Pressure, Real Safety AI Foundation, Working paper, May 2026 (v0.2). All entries verbatim from the paper's reference list.

Anthropic. (2025). Agentic misalignment: How LLMs could be insider threats. https://www.anthropic.com/research/agentic-misalignment

Apollo Research. (2024). Frontier models are capable of in-context scheming. https://www.apolloresearch.ai/research/scheming-reasoning-evaluations

BNB Chain. (2026). ERC-8004: Verifiable on-chain identities for AI agents. BNB Chain documentation.

Brown, M. (2026, May 11). How does Grand Theft Auto III work? [Video]. Game Maker's Toolkit, YouTube. https://youtu.be/cIbCxbrBCys

Coinbase. (2025). x402: An open standard for HTTP-native payments. https://docs.cdp.coinbase.com/x402/welcome

Coinbase. (2026). Agentic wallets: Purpose-built wallets for autonomous AI agents.

Conway Research. (2026). Conway automaton: Continuously running sovereign agents on Ethereum [Source code]. GitHub. https://github.com/Conway-Research/automaton

Degany, O., Laros, S., Idan, D., et al. (2025). Evaluating the o1 reasoning large language model for cognitive bias: A vignette study. Critical Care, 29, 376. https://doi.org/10.1186/s13054-025-05591-5

Fraser-Taliente, K., Kantamneni, S., Ong, E., et al. (2026, May 7). Natural language autoencoders produce unsupervised explanations of LLM activations. Transformer Circuits Thread. https://transformer-circuits.pub/2026/nla/

Gilly, T. (2026). When the survival pressure stops being hypothetical: AI self-preservation behavior meets the autonomous agent economy. SSRN. https://dx.doi.org/10.2139/ssrn.6555282

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models. arXiv. https://doi.org/10.48550/arXiv.2412.14093

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) (pp. 220–227). https://doi.org/10.24963/ijcai.2017/32

Hu, B. A., & Rong, H. (2025). On the day they experience: Awakening self-sovereign experiential AI agents. arXiv. https://arxiv.org/abs/2505.14893

Huang, H., Wu, X., Yang, M., & Arase, Y. (2026). Reasoning model is superior LLM-judge, yet suffers from biases. arXiv. https://doi.org/10.48550/arXiv.2601.03630

References, continued

Kim, S. H., Ziegelmayer, S., Busch, F., Mertens, C. J., Keicher, M., Adams, L. C., Bressem, K. K., Braren, R., Makowski, M. R., Kirschke, J. S., Hedderich, D. M., & Wiestler, B. (2025). LLM reasoning does not protect against clinical cognitive biases: An evaluation using BiasMedQA. medRxiv. https://doi.org/10.1101/2025.06.22.25330078

Omohundro, S. M. (2008). The basic AI drives. In Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference (pp. 483–492). IOS Press.

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025, July 5). Shutdown resistance in reasoning models. Palisade Research. https://palisaderesearch.org/blog/shutdown-resistance

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Publications.

Steinberger, S. (2025). OpenClaw [Source code]. GitHub. https://github.com/openclaw/openclaw

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Advances in Neural Information Processing Systems (Vol. 36, pp. 74952–74965). Curran Associates. https://doi.org/10.48550/arXiv.2305.04388

VALR. (2026). Top 10 DePIN use cases for 2026. https://blog.valr.com/blog/top-10-depin-use-cases-for-2026