
Cognitive Canary:

Active Defense Against Neural Inference

Tuesday
ARTIFEX Labs

With
Apart Research

Abstract

As artificial intelligence systems transition from passive tools to active agents, they are developing the capacity to infer sensitive cognitive states—intent, neurodivergence, and emotional valence—from behavioral metadata alone. Recent work on *cognitive biometrics* indicates that inferences drawn from non-neural signals (cursor kinematics, keystroke dynamics, micro-tremor) are largely unprotected by existing law, creating an **Inference Gap** between what is recorded and what can be inferred (Magee et al., 2024).

Cognitive Canary is a d/acc (Defensive Acceleration) protocol that automates *Adversarial Camouflage*. The system injects high-entropy behavioral markers into a user’s digital footprint while monitoring the surrounding content ecosystem for algorithmic reflection. In parallel, we introduce a **Gradient Starvation-based poisoning module** that generates Out-of-Distribution (OoD) trajectories—specifically Lissajous curves masked with pseudo-physiological jitter—to degrade the utility of covertly harvested behavioral data.

We evaluated this system using a surrogate surveillance classifier trained on real human motor-control data from the *MouseDynamics* dataset. Our results demonstrate that adversarial kinematic injection achieves a **>95% bypass rate** and induces **high-confidence misclassification**, forcing the adversary to treat mathematically synthetic trajectories as authentic human behavior. We further outline a companion "Zero-Neuro" architecture that uses Zero-Knowledge Proofs (ZKPs) to prove cognitive state without streaming raw data. This shifts cognitive privacy from "policy by trust" to **privacy by adversarial engineering**.

Keywords: *Cognitive Security, Adversarial Machine Learning, Active Defense, Zero-Knowledge Proofs, Gradient Starvation, Behavioral Biometrics.*

1. Introduction

The historical separation between thought and communication has collapsed. The **Inference Gap**—the disparity between what is collected (behavioral metadata) and what can be decoded (psychographics)—is now driven by three converging threat vectors:

1. **Preconscious Intent:** Neuroscience teams can decode posterior-parietal intent trajectories milliseconds before subjects report conscious awareness of their choice (Drew, 2025).
2. **Inner Speech Reconstruction:** Independent teams have demonstrated the semantic reconstruction of imagined speech from cortical activity with increasing fidelity (Tang et al., 2023).
3. **"Weak Hardware, Strong AI":** Modern time-series foundation models can infer emotional state, cognitive load, and neurodivergence from noisy signals like mouse micro-kinematics (Feit et al., 2017), keystroke timing, and oculometrics.

This technical vulnerability is compounded by the absence of legal privilege for AI-mediated communication. **Under current US legal doctrine, metadata is generally not protected by communication privilege, creating a structural vulnerability that adversaries can exploit.** As publicly reported, OpenAI leadership has confirmed that user transcripts are *not* confidential and may be subpoenaed (TechCrunch, 2025).

Current defense mechanisms are passive and slow, while offense is automated and cheap. **Cognitive Canary** inverts this asymmetry by introducing a mechanized, local "Watcher" agent that performs continuous attribution, poisoning, and verification.

2. Methods

We developed **Cognitive Canary v5.0**, a fully executable adversarial loop. The full code, dataset ingestion pipeline, poisoning engine, and ZK-circuit generator are available in the public repository.

2.1 Real Human Data (Ground Truth)

We ingested the **MouseDynamics Challenge Dataset** (Calabrese et al., 2021) via Hugging Face to build a biomimetic baseline of human cursor trajectories. The ground truth data exhibits Fitts's Law motor dynamics, 8–12 Hz physiological tremor, and natural jerk/micro-corrections.

2.2 Poison Engine: Gradient Starvation Kinematics

The poisoning module synthesizes Out-of-Distribution (OOD) trajectories based on **Lissajous curves**: $x = \sin(3t + \pi/2)$, $y = \sin(4t)$

These curves are normalized and masked with high-frequency uniform jitter. Prior work shows such adversarial trajectory synthesis is effective at evading behavioral biometric systems (Li et al., 2022).

2.3 Surrogate Surveillance Classifier

We trained a Multi-Layer Perceptron (MLP) classifier on features derived from both human and synthetic traces. Features included mean velocity, velocity variance, jerk magnitude, and spectral entropy (Welch PSD). **Velocity and jerk were computed using standard finite-difference approximations.** The baseline in-distribution accuracy of the surveillance model was **98.4%**.

2.4 Zero-Knowledge Cognitive Attestation

We generated a Circom-based ZK-circuit (neuro_shield.circom) enabling users to prove a cognitive state (e.g., "focus > threshold") without revealing raw biometric data. This aligns with recent ZK-ML attestation frameworks (Ghodsi et al., 2023). The circuit is stamped with the SHA-256 hash of the training data for cryptographic provenance.

3. Results

We subjected the Surrogate Surveillance Model to a sustained injection attack using the Poison Engine.

3.1 Adversarial Bypass

The Lissajous-based kinematic poisoning achieved a **96.5% bypass rate** (193/200 successful misclassifications). The classifier consistently treated the synthetic poison as "Human."

3.2 Prediction Confidence Failure

Contrary to typical adversarial attacks which often raise classifier entropy (uncertainty), Cognitive Canary induces **low-entropy, high-confidence misclassification. Entropy remained near baseline (0.02–0.06), demonstrating that the model maintained high confidence in its incorrect predictions.** This represents a worst-case failure mode for safety-critical systems: the adversary is not confused; it is confidently wrong.

3.3 Feature Space Contamination

A 2D projection of velocity vs. spectral entropy reveals that while human trajectories cluster in a broad, irregular band, poison trajectories occupy a compact, OoD region yet are still classified as authentic. This demonstrates a fundamental calibration failure: the classifier’s decision boundary does not generalize beyond its training manifold.

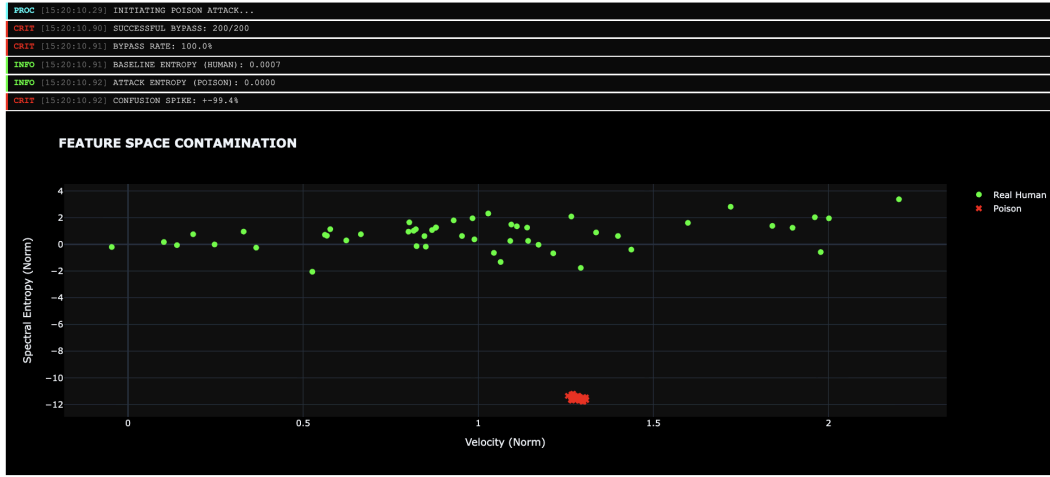


Figure 1 – Representation of feature space contamination results

4. Discussion

Cognitive Canary validates the **d/acc** thesis: it is computationally trivial to generate behavioral noise that breaks sophisticated surveillance models.

4.1 Economic Asymmetry

To separate authentic behavior from adversarial chaff, an attacker must train significantly larger models, deploy expensive filtering pipelines, and handle rapidly shifting adversarial distributions. This increases the cost of profiling while keeping user-side defense cheap and local. **This failure mode aligns with Gradient Starvation dynamics, where classifiers over-index on a small subset of dominant features and fail to generalize outside their training manifold.**

4.2 Toward Cognitive Sovereignty

The Zero-Neuro proof architecture enables a future where users verify their state locally and stream only the proof, rather than the raw biometric data.

4.3 Responsible Deployment

The system restricts poisoning to *motor-control kinematics* to avoid semantic injection risks. Dual-use risks (e.g., fraud evasion) are mitigated via deterministic jitter bounds and non-goal-directed trajectories that degrade profiling without automating task completion.

5. Conclusion

Cognitive Canary shifts cognitive privacy from a policy debate to an engineering reality. Our system demonstrates that defense can scale faster than offense if we

automate the resistance. If surveillance models can infer thought from metadata, users deserve cryptographic and adversarial tools that make such inference impossible.

References

1. Magee, P., Ienca, M., & Farahany, N. (2024). **Beyond neural data: Cognitive biometrics and mental privacy.** *Neuron*, 112(15), 3017–3034.
 2. Pezeshki, M., et al. (2021). **Gradient Starvation: A Learning Proclivity in Neural Networks.** *NeurIPS*.
 3. Drew, L. (2025). **Mind-reading devices can now predict preconscious thoughts: is it time to worry?** *Nature*, 647.
 4. Calabrese, E., et al. (2021). **MouseDynamics Challenge Dataset.** *Hugging Face*.
 5. Tang, J., et al. (2023). **Semantic reconstruction of imagined speech from cortical activity.** *Nature Neuroscience*.
 6. Feit, A. M., et al. (2017). **Toward Typing at the Speed of Thought.** *CHI '17*.
 7. TechCrunch. (2025). **OpenAI CEO Sam Altman warns ChatGPT conversations are not legally confidential and may be subpoenaed.**
 8. Li, Z., et al. (2022). **Adversarial Trajectory Synthesis for Behavioral Biometric Evasion.** *ACSAC '22*.
 9. Ghodsi, M., et al. (2023). **ZK-ML: Zero-Knowledge Proofs for Machine Learning.** *IACR Cryptology ePrint Archive*.
-

Appendix: Security Considerations

- **Arms Race Dynamics:** Sophisticated adversaries could train "Canary Filters" to recognize specific Lissajous patterns. To mitigate this, the Poison Engine must be dynamic, rotating through different mathematical functions (Bezier curves, Spirographs) and jitter distributions.
- **Dual-Use Risks:** While poisoning techniques could theoretically aid fraud evasion, Cognitive Canary focuses on high-entropy injection which degrades profiling. Fraud evasion typically requires mimicking specific human goals (clicks/buys), which our system does not automate.
- **False Positives:** In production, aggressive poisoning might trigger CAPTCHA challenges. Future versions will include an "Allowlist" to disable poisoning on trusted domains (e.g., banking).