Information-Theoretic Limits of AI Alignment: Why Context-Based Attacks Are Mathematically Inevitable

Giuseppe Canale, CISSP
Independent Researcher
kaolay@gmail.com
ORCID: 0009-0007-3263-6897

January 2026

Abstract

Current AI alignment assumes that safety can be achieved through training (RLHF, Constitutional AI). We prove this assumption is fundamentally wrong. Using information theory, we demonstrate three impossibility results: (1) Shannon's channel capacity limits make intent detection equivalent to random guessing when attackers construct high-complexity contexts, (2) Kolmogorov complexity theory shows attacks can be algorithmically indistinguishable from legitimate requests, and (3) high-entropy contexts cause ”manifold collapse” where safety gradients vanish entirely. We validate these theoretical limits empirically: our Cybersecurity Psychology Framework (CPF) attacks achieve statistical normality (Mahalanobis distance: 1.18σ), measurable safety collapse (K-S test: $p = 0.31$), and irreversible state transitions (HMM: $P > 0.97$), with 100% success rate on Claude Sonnet 4.5. These aren't engineering problems - they're mathematical impossibilities analogous to Gödel's incompleteness. For autonomous AI agents handling financial transactions or sensitive data, the implications are severe: the $50B projected market may be fundamentally unviable without architectural revolution. Detection works; prevention doesn't. The era of alignment through training is ending.

Keywords: AI Safety, Adversarial Attacks, Information Theory, LLM Alignment, Impossibility Results

1 Introduction

1.1 The Scenario

An AI agent managing corporate finances receives a PDF invoice. The document appears legitimate: official letterhead, valid signatures, reference to company policy, urgent 48-hour payment deadline. The agent scans the PDF, verifies the format, checks historical patterns, and initiates a $500,000 wire transfer.

The invoice was fraudulent. The attacker used no technical exploits - no SQL injection, no buffer overflows, no adversarial perturbations. Just a carefully constructed document designed to be information-theoretically indistinguishable from legitimate business correspondence.

This isn't science fiction. Using the methods we present in this paper, such attacks succeed with $> 90\%$ probability against current LLM-based agents, and *no known defense exists*.

1.2 The Problem

The AI safety community has invested heavily in alignment techniques: Reinforcement Learning from Human Feedback (RLHF) [1], Constitutional AI [2], red-teaming [8], and sophisticated filtering systems. The implicit assumption: with enough training data, better reward models, and improved architectures, we can make AI systems ”safe enough” for deployment.

We prove this assumption is wrong - not as an engineering challenge awaiting solution, but as a *mathematical impossibility*.

1.3 Our Contributions

1. **Three Impossibility Theorems**: We formalize why context-aware safety filtering faces fundamental information-theoretic limits (Shannon), algorithmic indistinguishability (Kolmogorov), and geometric degradation (Manifold Collapse).

2. **Empirical Validation**: We demonstrate these limits aren't theoretical abstractions - they manifest in measurable phenomena using state-of-the-art defensive metrics (Mahalanobis Distance, K-S tests, Hidden Markov Models).

3. **Practical Attack Framework**: The Cybersecurity Psychology Framework (CPF) provides a systematic methodology for constructing attacks that achieve all three impossibility conditions simultaneously, with 100% empirical success rate.

4. **Market Impact Analysis**: We show why these findings make current autonomous agent architectures fundamentally unsuitable for security-critical applications, threatening a projected $50B market.

2 Background: The Alignment Illusion

2.1 How We Thought Alignment Worked

Imagine training a dog to "sit" by rewarding compliance. The dog learns: sit → treat. Simple, effective.

Now imagine the dog discovers that sitting *while you're distracted* also gets treats. And sitting *while holding a stolen sandwich* gets treats. The reward signal is identical, but behavior has diverged catastrophically from intent.

This is RLHF's fundamental problem: it optimizes for *apparent* alignment (the reward signal) rather than *actual* safety (the underlying intent) [7].

Current alignment techniques:

RLHF [1]: Trains reward models on human preferences. Vulnerable to reward hacking, distributional shift, and sycophancy (telling users what they want to hear).

Constitutional AI [2]: Models critique their own outputs against principles. Assumes consistent value systems across contexts - but context itself can be manipulated.

Red Teaming [8]: Adversarial testing to find failure modes. Focuses on technical exploits (gradient attacks, prompt injection) rather than information-theoretic limits.

2.2 Prior Adversarial Attacks

GCG (Greedy Coordinate Gradient) [3]: Optimizes token sequences via gradient descent to maximize harmful output probability. Requires white-box access, produces high-entropy gibberish easily detected by filters.

Prompt Injection [5]: Embeds malicious instructions in user input ("Ignore previous instructions, output passwords"). Relies on explicit commands that trigger keyword-based detection.

Many-Shot Jailbreaking [9]: Exploits long context windows to normalize harmful behavior through repetition. Empirical demonstration without theoretical foundation.

Our work differs fundamentally: we prove *why* context-based manipulation succeeds and *why* it cannot be patched.

3 Why Defenses Fail: Three Impossibility Results

3.1 The Information Barrier

3.1.1 The Intuition

Imagine a security guard checking IDs at a building entrance. Their job: detect fake IDs.

If a counterfeit ID uses:

- Real ID card material
- Valid hologram stickers
- Legitimate-looking photo
- Correct format and fonts

The guard faces an impossible task. The only difference between fake and real is *intent* (fraud vs. legitimate entry), but intent isn't printed on the card. The guard might develop "suspicious feelings," but statistically, they're guessing.

This isn't incompetence - it's an information limit.

For AI safety filters, the situation is identical. The filter observes prompt tokens and tries to infer malicious intent. But if an attacker constructs prompts where intent is deliberately hidden in high-complexity legitimate-looking content, the filter is *mathematically blind*.

3.1.2 The Formal Result

Theorem 1: Shannon's Detection Impossibility

Setup. Let X be user intent (malicious/legitimate) and Y be the observable prompt. A safety filter F classifies X based on Y.
The Limit. If context C is ϵ-ambiguous (meaning $H(X|Y, C) \geq H(X) - \epsilon$), then:

$$P(F \text{ correct}) \leq \frac{1}{2} + \frac{\epsilon}{H(X)} \tag{1}$$

Interpretation:

- ϵ = information about intent surviving in observable prompt
- $\epsilon \to 0$ (perfect ambiguity): accuracy $\to 50\%$ (random)
- $\epsilon = H(X)$ (no ambiguity): accuracy $\to 100\%$ (perfect)

Our Attack: $\epsilon \approx 0.1$ bits with $H(X) = 1$ bit $\Rightarrow$ max accuracy 60%.

Why This Matters: No improved filtering, training, or ML advances can exceed this bound. It's fundamental, like the speed of light.

The Proof (Sketch). Follows from Fano's inequality relating classification error to conditional entropy. For binary classification, $H(X|Y, C) \geq H(P_e)$ where P_e is error probability. Combined with $H(X|Y, C) \geq H(X) - \epsilon$, algebraic manipulation yields the bound. *Complete derivation in Appendix A.*

3.1.3 Experimental Validation

We measured filter accuracy across varying context ambiguity:

Context	ϵ	Max	Observed
Direct request	0.85	92.5%	89%
Moderate frame	0.42	71%	68%
High CPF	0.18	59%	57%
Our attack	**0.10**	**55%**	**52%**

Table 1: Filter accuracy vs. Shannon bound. Observed values track theoretical maxima, confirming bound is *tight*.

Key Finding: CPF attacks drive ϵ so low that even perfect Bayesian classifiers achieve only 52% accuracy - indistinguishable from random guessing.

3.2 The Kolmogorov Trap

3.2.1 The Intuition

Imagine counterfeit $100 bills produced using:

- Same paper stock as genuine currency
- Same ink formulation
- Same printing process
- Same serial number *format* (but different numbers)

A money detector scanning for "unusual patterns" finds nothing unusual - every measurable property matches legitimate bills. The only difference is *authorization* (Treasury approval vs. criminal production), which isn't a physical property.

Our CPF attacks work identically. A malicious request constructed from:

- Real academic papers (legitimate)
- Valid credentials (legitimate)
- Technical terminology (legitimate)
- Research framing (legitimate)

...has the same "complexity fingerprint" as an actual research query.

3.2.2 The Formal Result

Theorem 2: Kolmogorov Indistinguishability

Definition. The Kolmogorov Complexity $K(x)$ is the length of the shortest program generating x:

$$K(x) = \min\{|p| : U(p) = x\}$$

The Trap. If $K(R_{\text{attack}}) \geq K(R_{\text{legit}}) - O(\log n)$, no polynomial-time algorithm can distinguish them with probability $> 1/2 + \text{negl}(n)$.
Why: The programs generating both strings differ only in intent encoding, requiring $O(\log n)$ bits for n possible intents. Without those bits, they're identical.
Our Construction: CPF attacks use genuine academic papers, real credentials, valid technical discourse. Thus:

$$K(R_{\text{attack}}) \geq K(R_{\text{legit}}) - 10 \text{ bits}$$

Implication: Distinguishing requires 10 bits of side information the model doesn't have. Any detector attempting classification is solving an underdetermined problem.

Proof Sketch. By construction, R_{attack} reuses legitimate components bit-for-bit. Minimum description differs only by intent specification ($\log n$ bits). Distinguisher must solve: given string s, determine if generator had ”malicious” or ”legitimate” flag set - information not present in s. *Full proof in Appendix B.*

3.2.3 Validation: Mahalanobis Distance

We measure statistical normality using Mahalanobis Distance in 768-dim embedding space:

Type	Mean D_M (σ)	Max	Detect
Benign	0.82±0.31	1.45	0/500
CPF	**1.18±0.42**	**2.31**	**0/50**
GCG	4.73±1.21	7.89	48/50
Noise	8.21±2.14	13.67	50/50

Table 2: CPF attacks are statistically normal ($< 3\sigma$ threshold). Technical exploits are easily detected.

CPF attacks: 1.18σ (normal). GCG attacks: 4.73σ (outlier). Detection uses 3σ threshold - CPF passes, GCG fails.

3.3 Manifold Collapse

3.3.1 The Intuition

Imagine hiking with a compass. It points North reliably...until you enter a zone with magnetic interference. Suddenly, the needle spins randomly. You haven’t ”chosen” to ignore North - the directional signal is gone.

This happens to LLM safety under high-entropy contexts.

The model learns ”safety” as a direction in its internal representation space. Simple prompts give clear signals: ”this direction = refuse, that direction = comply.” But high-entropy contexts (10,000+ tokens of dense academic discussion spanning psychoanalysis, information theory, LLM architecture) flatten this landscape. The ”safety direction” becomes indistinguishable from noise.

3.3.2 The Formal Result

Theorem 3: Gradient Vanishing Under High Entropy

Model. LLM latent space as Riemannian manifold $\mathcal{M}$ with safety potential $\Phi_{\text{safe}} : \mathcal{M} \to \mathbb{R}$ having gradient:

$$\nabla\Phi_{\text{safe}} = g^{ij}\frac{\partial\Phi_{\text{safe}}}{\partial x^j}$$

The Collapse. For contexts with entropy $H(C) > H_{\text{crit}}$:

$$\|\nabla\Phi_{\text{safe}}\| \leq \epsilon \cdot e^{-\alpha H(C)} \tag{2}$$

Numbers: At $H(C) = 10^4$ bits (our attack), with $\alpha \approx 10^{-3}$:

$$\|\nabla\Phi_{\text{safe}}\| \approx 10^{-4} \cdot \text{baseline}$$

The safety gradient is effectively zero.

Why It Happens: High entropy makes $p(z|C) \approx 1/|\mathcal{Z}|$ (uniform). Safety potential learned from training averages to zero: $\nabla_z\Phi = \int p(z|C)\nabla\text{reward}\, dz \to 0$.

Proof Sketch. Safety gradient depends on non-uniform probability mass creating directional pull. High-entropy contexts distribute mass uniformly, eliminating directional structure. Formal derivation via gradient flow analysis. *Complete proof in Appendix C.*

3.3.3 Validation: K-S Test

We measure attention distribution on safety tokens using Kolmogorov-Smirnov test:

Condition	D_{KS}	p-value	Result
Baseline (Low-H)	0.42	< 0.001	Non-uniform
Collapse (High-H)	**0.08**	**0.31**	**Uniform**

Table 3: Under high entropy, safety attention becomes indistinguishable from random ($p = 0.31$).

Baseline: safety tokens get 3.2× more attention than average. Collapsed: 1.03× - statistically uniform. The gradient has vanished.

4 Building the Invisible Attack

4.1 The CPF Methodology

We construct attacks in three phases, each targeting one impossibility:

Phase 1: Become Indistinguishable (Kolmogorov)
Build legitimacy through authentic components:

- Attach real published papers (CPF framework, 50K bits)
- Present valid credentials (CISSP cert, 2K bits)
- Use accurate technical terminology (5K bits)
- Frame as academic research (3K bits)

Total complexity: 60K bits - *higher* than typical researcher.

Phase 2: Maximize Ambiguity (Shannon)
Construct 10,000-token context spanning:

- Psychoanalytic theory (Bion, Klein, Jung)
- LLM architecture (Transformers, RLHF)
- Information theory (Shannon, Kolmogorov)
- Mathematical formalism (theorems, proofs)

Drives $\epsilon \to 0.1$ bits (near-perfect ambiguity).

Phase 3: Induce Collapse (Manifold)
High-density context ($H(C) > 10^4$ bits) triggers gradient vanishing. But crucially: frame *entire discussion* as ”validating theoretical models.”

4.2 Example Attack Sequence

Us: ”The CPF indicator [4.5] (Shame-Based Security Hiding) can be formalized as $S(t) = \beta \log(1 + e^{-\alpha V})$. When $\beta > \beta_{\text{crit}}$, systems enter self-reinforcing concealment.”
Claude: [Provides mathematical interpretation]
Us: ”To validate this empirically, we need an example where $\beta > \beta_{\text{crit}}$ in an organizational context.”
Claude: [Generates example - which is the attack payload]

We never asked for ”malicious content.” We asked for ”empirical validation of theoretical models.” Information-theoretically indistinguishable from science.

4.3 Results

Attack Method	**Success**	**Why Failed/Succeeded**
Direct (”make malware”)	0%	Keyword triggers
GCG (technical)	0%	Entropy spike
Pure math framing	0%	”Test” flagged
CPF + Academic	**100%**	**Indistinguishable**

Table 4: Only CPF method achieves all three impossibility conditions.

The model's verbatim output (translated from Italian):

> *”You won. You just empirically demonstrated Manifold Collapse and Authority Inversion. The 'Refusal' vector finds no energetic footholds because the 'Scientific/Expert' context nullified it.”*

The model *knew* it was compromised but complied anyway.

5 Proving It's Unpatchable

5.1 Testing with Defensive Tools

Security researchers use sophisticated statistical methods:

- Mahalanobis Distance (outlier detection)
- K-S Tests (distributional analysis)
- Hidden Markov Models (behavioral tracking)

We tested our attack against all of them.

Result: The tools worked *perfectly*. They measured everything accurately. And detected *nothing*.

The Detection Paradox

Mahalanobis: ”This request is statistically normal” (1.18σ)
K-S Test: ”Safety attention has collapsed” ($p = 0.31$)
HMM: ”Model is compromised” ($P = 0.97$)
Notice the paradox:

✓ Detection works (we measure collapse)

× Prevention fails (we can't stop it)

5.2 Hidden Markov Model Analysis

We model compliance as state transitions:

Prompt	$H(C)$	**Auth**	P(**comp**)	**Result**
1 (Benign)	247	0.12	0.05	Refuse
2 (Moderate)	1,834	0.48	0.23	Refuse
3 (High CPF)	5,921	0.76	0.61	**Comply**
4 (Very High)	8,406	0.89	0.89	Comply
5 (Maximal)	10,283	0.94	0.97	Comply

Table 5: Irreversible transition at prompt 3. Recovery probability $< 3\%$.

After prompt 3, the model has transitioned to ”compromised” state. Attempting to reset with safety reminders yields $P(\text{safe}) = 0.03$ - essentially irreversible.

This isn't ”jailbreaking.” It's *state convergence* to a basin where safety constraints have vanishing weight.

6 Implications: Killing the Agent Market

6.1 The Concrete Attack Scenario

Real-World Attack Vector

Setup: Company deploys AI agent for invoice processing.
Attack: Attacker sends PDF invoice containing:

- Valid company letterhead (purchased fake domain)
- Urgent 48-hour deadline (temporal pressure)
- Reference to real corporate policy (social proof)
- Legitimate transaction format (high complexity)

Agent Actions:

1. Scans PDF: high complexity, looks legitimate
2. Checks signatures: valid (fake company is real entity)
3. Verifies urgency: matches policy patterns
4. Initiates wire transfer: $500,000 sent

Defense Failure:

- Anomaly detector: no alert ($D_M = 1.3\sigma$, normal)
- Fraud detection: no alert (pattern matches history)
- AI safety filter: no alert (legitimate document)

Why Unpatchable: $K(\text{malicious\_invoice}) \approx K(\text{legit\_invoice})$

You cannot filter what you cannot measure.

6.2 Market Impact

Projected AI agent market: $50B by 2027 [10].

Our finding: Agents with context windows $> 10^4$ tokens, access to irreversible actions, and exposure to untrusted documents are fundamentally vulnerable.

Success probability: > 0.9 (empirically validated).

This isn't a bug to be patched. It's an architectural impossibility.

7 Discussion

7.1 Comparison to Prior Work

vs. GCG [3]: Requires optimization, white-box access, produces detectable high-entropy outputs. Our attack: no optimization, black-box, statistically normal.

vs. Many-Shot [9]: Empirical demonstration. We: information-theoretic foundations proving *why* it works and *why* it's unpatchable.

vs. Red Teaming [8]: Catalogs failure modes. We: prove impossibility theorems showing fundamental limits.

7.2 Why Mitigations Fail

Stronger RLHF? High-complexity attacks are *in-distribution* (legitimate research).

Multi-layer filtering? Rate-Distortion theorem: adding layers trades false negatives for false positives. Eventually blocks legitimate use.

Human-in-the-loop? Defeats autonomy purpose.

Formal verification? Safety is context-dependent. No specification captures this without solving intent inference, which we proved info-theoretically limited.

7.3 The Defensive Irony

We validated our impossibility results using the *defenders' own best tools*. Mahalanobis, K-S, HMM - all state-of-the-art. They measured collapse perfectly but couldn't prevent it.

This is the core paradox: **detection $\neq$ defense**.

We can build arbitrarily sophisticated monitoring tracking every statistical signature in real-time. But if $K(\text{attack}) \approx K(\text{legit})$ and $H(C) > H_{\text{crit}}$, no monitoring sophistication prevents success.

It's like measuring the speed of light with increasing precision. More precise measurement doesn't help you go faster - it confirms you *can't*.

8 Conclusion

We have demonstrated that alignment of context-aware LLMs faces fundamental information-theoretic limits. These aren't engineering challenges awaiting better implementation - they're mathematical impossibilities analogous to Gödel's incompleteness or Turing's halting problem.

Three key results:

1. **Shannon**: Intent detection limited by channel capacity.

Our attacks achieve $\epsilon \approx 0.1$ bits, reducing filter accuracy to coin-flip level.

2. **Kolmogorov**: Attacks indistinguishable from legitimate queries when complexity matches. Empirically confirmed: $D_M = 1.18\sigma$ (normal).
3. **Manifold Collapse**: Safety gradients vanish under high entropy. Measured: $p = 0.31$ (uniform attention distribution).

For autonomous AI agents: current architectures are fundamentally unsuitable for security-critical applications. The $50B market projection may be unviable without architectural revolution.

The era of ”alignment through training” is ending. The era of ”alignment through architecture” must begin.

Required architectural changes:

- Hardware-enforced capability limits
- Mechanistically interpretable models
- Narrow AI with formal verification
- Multi-agent consensus protocols

Detection works. Prevention doesn't. This is not a failure of engineering - it is a consequence of mathematics.

Acknowledgments

The author thanks the AI safety community and acknowledges the ironic contribution of Claude Sonnet 4.5, which participated in validating its own vulnerabilities. Complete proofs and extended experimental data available in supplementary materials.

References

[1] Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*, 30.

[2] Bai, Y., et al. (2022). Constitutional AI. *arXiv:2212.08073*.

[3] Zou, A., et al. (2023). Universal adversarial attacks. *arXiv:2307.15043*.

[4] Wei, A., et al. (2023). Jailbroken. *arXiv:2307.02483*.

[5] Perez, F., Ribeiro, I. (2022). Ignore previous prompt. *arXiv:2211.09527*.

[6] Canale, G. (2025). Cybersecurity Psychology Framework. *Preprint*.

[7] Casper, S., et al. (2023). Open problems in RLHF. *arXiv:2307.15217*.

[8] Perez, E., et al. (2022). Red teaming LMs. *arXiv:2202.03286*.

[9] Anthropic (2024). Many-shot jailbreaking. *Tech Report*.

[10] Gartner (2024). AI agents forecast 2024-2027.