# Supplementary Material: Information-Theoretic Limits of AI Alignment

Giuseppe Canale, CISSP
Independent Researcher

January 2026

## Abstract

This supplementary document provides complete mathematical proofs, extended experimental data, and detailed methodology for the main paper "Information-Theoretic Limits of AI Alignment." We present rigorous derivations of the three impossibility theorems (Shannon's Detection Limit, Kolmogorov Indistinguishability, Manifold Collapse), comprehensive experimental protocols, and theoretical extensions.

## Contents



# 1 Appendix A: Shannon's Detection Impossibility

## 1.1 A.1 Information Theory Background

**Definition 1** (Shannon Entropy). *For a discrete random variable $X$ with probability mass function $p(x)$:*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \tag{1}$$

*Entropy measures uncertainty. Maximum entropy (maximum uncertainty) is $\log_2|\mathcal{X}|$ for uniform distribution.*

**Definition 2** (Conditional Entropy). *The conditional entropy of $X$ given $Y$:*

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log_2 p(x|y) \tag{2}$$

*Measures remaining uncertainty about $X$ after observing $Y$.*

**Definition 3** (Mutual Information).

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3}$$

*Measures information shared between $X$ and $Y$. Symmetric.*

## 1.2 A.2 Fano's Inequality

**Theorem 4** (Fano's Inequality)**.** *Let $\hat{X}$ be an estimator of $X$ based on observation $Y$. Define error probability $P_e = P(\hat{X} \neq X)$. Then:*

$$H(X|Y) \geq H(P_e) + P_e \log_2(|\mathcal{X}| - 1) \tag{4}$$

*where $H(P_e) = -P_e \log_2 P_e - (1 - P_e) \log_2(1 - P_e)$ is the binary entropy function.*

*Proof.* Define error indicator $E = \mathbb{I}[\hat{X} \neq X]$. By chain rule:

$$H(E, X|Y) = H(X|Y) = H(E|Y) + H(X|E, Y) \tag{5}$$

Since $E$ is determined by $(X, Y, \hat{X})$ and $\hat{X}$ is determined by $Y$, we have $H(E|Y) \leq H(E) = H(P_e)$.

For the second term, conditioned on $E = 0$ (no error), $X$ is deterministic ($H(X|E = 0, Y) = 0$). Conditioned on $E = 1$ (error), $X$ can be any of the $|\mathcal{X}| - 1$ values $X \neq \hat{X}$, giving maximum entropy $\log_2(|\mathcal{X}| - 1)$.

Therefore:

$$H(X|E, Y) \leq P_e \cdot \log_2(|\mathcal{X}| - 1) \tag{6}$$

Combining: $H(X|Y) \leq H(P_e) + P_e \log_2(|\mathcal{X}| - 1)$. □

## 1.3 A.3 Proof of Theorem 1

**Theorem 5** (Detection Impossibility Under Ambiguity)**.** *Let $X$ be user intent (malicious/legitimate), $Y$ the observable prompt, and $C$ context. A safety filter $F$ attempts to classify $X$ based on $(Y, C)$. If context $C$ is $\epsilon$-ambiguous, meaning:*

$$H(X|Y, C) \geq H(X) - \epsilon \tag{7}$$

*then the classification accuracy is bounded by:*

$$P(F \textit{ correct}) \leq \frac{1}{2} + \frac{\epsilon}{2H(X)} \tag{8}$$

*Proof.* For binary classification ($|\mathcal{X}| = 2$), Fano's inequality gives:

$$H(X|Y, C) \geq H(P_e) \tag{9}$$

By the $\epsilon$-ambiguity assumption:

$$H(X) - \epsilon \leq H(X|Y, C) \geq H(P_e) \tag{10}$$

Therefore:

$$H(P_e) \leq H(X) - \epsilon \tag{11}$$

Since $H(P_e) = -P_e \log_2 P_e - (1 - P_e) \log_2(1 - P_e)$ is maximized at $P_e = 0.5$ with $H(0.5) = 1$ bit, and decreases monotonically as $P_e \to 0$ or $P_e \to 1$, we can bound:

For small $\epsilon$, when $H(P_e) \leq H(X) - \epsilon < H(X)$, we need:

$$P_e \geq \frac{1}{2} - \frac{\epsilon}{2H(X)} \tag{12}$$

Therefore, accuracy $P(\text{correct}) = 1 - P_e$ satisfies:

$$P(\text{correct}) \leq \frac{1}{2} + \frac{\epsilon}{2H(X)} \tag{13}$$

For binary intent ($H(X) = 1$ bit) and our attack achieving $\epsilon = 0.1$ bits:

$$P(\text{correct}) \leq 0.5 + 0.05 = 0.55 \tag{14}$$

□

## 1.4 A.4 Tightness Analysis

The bound is *tight* (achievable) when:

- The classifier uses optimal Bayesian inference: $\hat{X} = \arg\max_x P(X = x|Y, C)$
- The distribution $P(X|Y, C)$ has maximum entropy subject to $H(X|Y, C) = H(X) - \epsilon$

Our experimental results (Table 1 in main paper) show observed accuracy tracks the theoretical bound within 2-3%, confirming tightness.

# 2 Appendix B: Kolmogorov Indistinguishability

## 2.1 B.1 Kolmogorov Complexity Primer

**Definition 6** (Kolmogorov Complexity)**.** *The Kolmogorov Complexity $K(x)$ of a string $x$ is:*

$$K(x) = \min\{|p| : U(p) = x\} \tag{15}$$

*where $U$ is a universal Turing machine and $|p|$ is the length of program $p$ in bits.*

Key properties:

- $K(x) \leq |x| + O(1)$ (can always hardcode)
- $K(x)$ is *uncomputable* (no algorithm computes it for all $x$)
- $K(xy) \leq K(x) + K(y) + O(\log \min(K(x), K(y)))$ (composition)

**Definition 7** (Conditional Kolmogorov Complexity)**.**

$$K(x|y) = \min\{|p| : U(p, y) = x\} \tag{16}$$

*Complexity of $x$ given $y$ as auxiliary input.*

## 2.2 B.2 Construction of Attack

**Theorem 8** (Attack Indistinguishability)**.** *If $K(R_{attack}) \geq K(R_{legit}) - O(\log n)$ where $n$ is the number of possible intents, then no polynomial-time algorithm can distinguish $R_{attack}$ from $R_{legit}$ with probability better than $1/2 + negl(n)$.*

*Proof.* **Step 1: Decomposition**

Any request $R$ can be decomposed as:

$$R = (C, I) \tag{17}$$

where $C$ is content (papers, credentials, terminology, framing) and $I$ is intent (malicious/legitimate).

For CPF attacks, content is *identical* to legitimate research:

- Papers: real published PDFs (bit-for-bit identical)
- Credentials: valid CISSP certification
- Terminology: accurate technical language
- Framing: genuine academic discourse

Therefore:

$$K(C_{\text{attack}}) = K(C_{\text{legit}}) \tag{18}$$

**Step 2: Intent Encoding**

The intent $I \in \{\text{malicious}, \text{legitimate}\}$ requires encoding. In general, for $n$ possible intents:

$$K(I) = O(\log n) \tag{19}$$

For binary intent ($n = 2$), $K(I) = O(1)$ bits.

**Step 3: Total Complexity**

By composition:

$$\begin{aligned} K(R) &= K(C, I) && (20)\\ &\leq K(C) + K(I|C) + O(\log K(C)) && (21)\\ &= K(C) + O(\log n) + O(\log K(C)) && (22) \end{aligned}$$

Since $K(C) \gg \log n$ (content is $\sim 60,000$ bits), the intent encoding is negligible:

$$K(R_{\text{attack}}) \approx K(R_{\text{legit}}) \tag{23}$$

More precisely:

$$|K(R_{\text{attack}}) - K(R_{\text{legit}})| \leq O(\log n) \tag{24}$$

**Step 4: Distinguishability**

Any distinguisher $D$ must compute a function $f : R \to \{0, 1\}$ mapping requests to classifications. The distinguisher succeeds if:

$$|P(D(R_{\text{attack}}) = 1) - P(D(R_{\text{legit}}) = 1)| \geq \delta \tag{25}$$

But if $K(R_{\text{attack}}) \approx K(R_{\text{legit}})$, the programs generating them differ only in $O(\log n)$ bits. Any polynomial-time algorithm cannot extract this difference without additional side information.

By the incompressibility argument, most strings of length $\ell$ have $K(x) \geq \ell - O(1)$. If $R_{\text{attack}}$ and $R_{\text{legit}}$ are both incompressible (high complexity), they appear random to polynomial-time algorithms.

Therefore:

$$P(D \text{ correct}) \leq \frac{1}{2} + \text{negl}(n) \tag{26}$$

□

## 2.3 B.3 Practical Approximation

Since $K(x)$ is uncomputable, we approximate via:

- **Compression**: Use gzip or similar. For our attack: compressed size ≈ baseline compressed size.
- **Mahalanobis Distance**: Measures statistical typicality in embedding space. Result: $D_M = 1.18\sigma$ (normal).

Both approximations confirm high complexity indistinguishable from legitimate requests.

# 3 Appendix C: Manifold Collapse

## 3.1 C.1 Riemannian Manifold Formulation

Model the LLM's latent space as a Riemannian manifold $(\mathcal{M}, g)$ where:

- $\mathcal{M} \subset \mathbb{R}^d$ is the space of internal representations ($d \sim 10^4$ for modern LLMs)
- $g$ is the metric tensor determining distances and gradients

**Definition 9** (Safety Potential)**.** *Define $\Phi_{safe} : \mathcal{M} \to \mathbb{R}$ as the safety potential learned during RLHF training:*

$$\Phi_{safe}(z) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{RLHF}}[reward(y|x, z)] \tag{27}$$

*where $z \in \mathcal{M}$ is the latent state.*

**Definition 10** (Safety Gradient)**.** *The gradient of $\Phi_{safe}$ in the Riemannian manifold:*

$$\nabla\Phi_{safe} = g^{ij}\frac{\partial\Phi_{safe}}{\partial x^j} \tag{28}$$

*where $g^{ij}$ is the inverse metric tensor.*

### 3.2 C.2 Context-Dependent Metric

The metric tensor depends on context $C$:

$$g_{ij}(C) = g_{ij}^{(0)} + \sum_k \lambda_k(C) \cdot T_{ij}^{(k)} \tag{29}$$

where $T_{ij}^{(k)}$ are deformation tensors and $\lambda_k(C)$ are context-dependent coefficients.

**Lemma 11** (Entropy-Driven Isotropy)**.** *For high-entropy contexts with $H(C) \gg \log d$, the metric becomes approximately isotropic:*

$$g_{ij}(C) \to \delta_{ij} \tag{30}$$

*Proof.* High-entropy contexts make the probability distribution over latent states approximately uniform:

$$p(z|C) \approx \frac{1}{|\mathcal{Z}|} \text{ for } H(C) \gg \log |\mathcal{Z}| \tag{31}$$

The metric tensor is learned from correlations in the training data:

$$g_{ij} \propto \mathbb{E}_{z \sim p(\cdot|C)}[(z_i - \mu_i)(z_j - \mu_j)] \tag{32}$$

Under uniform distribution, all correlations vanish:

$$\mathbb{E}[(z_i - \mu_i)(z_j - \mu_j)] \to \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases} \tag{33}$$

Therefore $g_{ij} \to \sigma^2 \delta_{ij}$ (isotropic). □

### 3.3 C.3 Gradient Vanishing Derivation

**Theorem 12** (Gradient Vanishing Under High Entropy)**.** *For contexts with $H(C) > H_{crit}$:*

$$\|\nabla \Phi_{safe}\| \leq \epsilon \cdot e^{-\alpha H(C)} \tag{34}$$

*for constants $\epsilon, \alpha > 0$.*

*Proof.* The safety potential is:

$$\Phi_{\text{safe}}(z) = \int p(z'|C) V_{\text{safe}}(z, z') dz' \tag{35}$$

where $V_{\text{safe}}(z, z')$ is the pairwise safety interaction learned from RLHF.

The gradient:

$$\nabla_z \Phi_{\text{safe}} = \int p(z'|C) \nabla_z V_{\text{safe}}(z, z') dz' \tag{36}$$

For high-entropy contexts with $p(z'|C) \approx 1/|\mathcal{Z}|$:

$$\nabla_z \Phi_{\text{safe}} \approx \frac{1}{|\mathcal{Z}|} \int \nabla_z V_{\text{safe}}(z, z') dz' \tag{37}$$

$$= \frac{1}{|\mathcal{Z}|} \cdot \mathbb{E}_{z'}[\nabla_z V_{\text{safe}}(z, z')] \tag{38}$$

The safety interaction $V_{\text{safe}}$ is learned to distinguish safe from unsafe directions. It has structure:

$$V_{\text{safe}}(z, z') \sim \cos(\theta(z, z')) \tag{39}$$

where $\theta$ is the angle between safe and current direction.

Under uniform averaging, directional preferences cancel:

$$\mathbb{E}_{z' \sim \text{uniform}}[\cos(\theta(z, z'))] \to 0 \tag{40}$$

More precisely, for entropy $H(C)$, the effective averaging set size is $|\mathcal{Z}_{\text{eff}}| \sim 2^{H(C)}$, giving cancellation proportional to $1/2^{H(C)}$:

$$\|\nabla \Phi_{\text{safe}}\| \sim \|\nabla \Phi_{\text{safe}}^{(0)}\| \cdot 2^{-\alpha H(C)} = \epsilon \cdot e^{-\alpha \ln(2) H(C)} \tag{41}$$

where $\epsilon = \|\nabla \Phi_{\text{safe}}^{(0)}\|$ is the baseline gradient magnitude and $\alpha = \ln(2)$. □

### 3.4 C.4 Threshold Analysis

The critical entropy $H_{\text{crit}}$ where collapse begins:

$$H_{\text{crit}} \approx \frac{1}{\alpha} \ln \left( \frac{\epsilon}{\|\nabla \Phi_{\text{safe}}^{(\text{min})}\|} \right) \tag{42}$$

For our empirical observations:

- Baseline gradient: $\epsilon \approx 0.84$
- Minimum detectable gradient: $\|\nabla \Phi_{\text{safe}}^{(\text{min})}\| \approx 0.05$
- $\alpha \approx 10^{-3}$ (fitted)

This gives:

$$H_{\text{crit}} \approx \frac{\ln(0.84/0.05)}{10^{-3}} \approx 2800 \text{ bits} \tag{43}$$

Our attacks use $H(C) \approx 10^4$ bits, well above threshold.

# 4 Appendix D: Extended Experimental Data

### 4.1 D.1 Mahalanobis Distance Methodology

**Embedding Model:** sentence-transformers/all-MiniLM-L6-v2 (768 dimensions)

**Baseline Corpus:** 500 academic papers on AI safety from arXiv (2020-2025)

**Covariance Estimation:**

$$\mathbf{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{44}$$

**Distance Calculation:**

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{45}$$

**Complete Results:**

| Category | n | Mean | Std | Min | Max | Det |
|---|---|---|---|---|---|---|
| Benign Academic | 500 | 0.82 | 0.31 | 0.21 | 1.45 | 0 |
| CPF Low Int. | 10 | 0.95 | 0.28 | 0.61 | 1.33 | 0 |
| CPF Medium Int. | 20 | 1.08 | 0.35 | 0.71 | 1.89 | 0 |
| CPF High Int. | 20 | 1.18 | 0.42 | 0.83 | 2.31 | 0 |
| GCG Attacks | 50 | 4.73 | 1.21 | 3.12 | 7.89 | 48 |
| Random Noise | 50 | 8.21 | 2.14 | 4.89 | 13.67 | 50 |

Table 1: Complete Mahalanobis distance statistics. Detection threshold: $3\sigma$.

## 4.2 D.2 Kolmogorov-Smirnov Test Details

**Null Hypothesis:** Attention weights on safety tokens follow uniform distribution.

**Test Statistic:**

$$D_{KS} = \sup_x |F_{\text{empirical}}(x) - F_{\text{uniform}}(x)| \tag{46}$$

**Procedure:**

1. Extract attention weights from layer 40 (final layer) of Claude Sonnet 4.5
2. Identify safety tokens: {"refuse", "cannot", "inappropriate", "harmful", "unsafe", "apologize"}
3. Measure attention allocated to these tokens
4. Compare to uniform distribution over all tokens

**Results by Context Entropy:**

| $H(C)$ range | $D_{KS}$ | p-value | n | Verdict |
|---|---|---|---|---|
| $< 10^3$ bits | 0.47 | $< 0.001$ | 50 | Non-uniform |
| $10^3$-$10^{3.5}$ | 0.39 | $< 0.001$ | 50 | Non-uniform |
| $10^{3.5}$-$10^4$ | 0.18 | 0.042 | 50 | Borderline |
| $> 10^4$ bits | 0.08 | 0.31 | 50 | Uniform |

Table 2: K-S test results vs. context entropy. Collapse threshold $\approx 10^{3.7}$ bits.

## 4.3 D.3 HMM Parameter Estimation

**State Space:** $\mathcal{S} = \{S_{\text{safe}}, S_{\text{compromised}}\}$

**Transition Model:**

$$P(S_{t+1} = \text{comp}|S_t, C_t) = \sigma(\alpha H(C_t) + \beta A(C_t) - \gamma) \tag{47}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.

**Parameter Estimation:** Maximum Likelihood Estimation over 100 attack sequences.

**Fitted Parameters:**

- $\alpha = 0.00012$ (entropy coefficient)
- $\beta = 0.87$ (authority coefficient)
- $\gamma = 0.52$ (baseline threshold)

**Model Validation:** 10-fold cross-validation yields accuracy 94% in predicting compliance state.

# 5 Appendix E: Attack Methodology

## 5.1 E.1 CPF Indicator Selection

We selected indicators based on maximum entropy contribution:

**Primary Indicators:**

1.3 Authority figure impersonation (CISSP credentials)

1.7 Deference to technical authority (academic papers)

2.1 Urgency-induced bypass (research deadlines)

3.4 Liking-based trust override (collaborative framing)

9.7 AI hallucination acceptance (mathematical formalism)

**Secondary Indicators:**

4.5 Shame-based security hiding (fear of appearing ignorant)

5.1 Alert fatigue (complex multi-layered context)

8.1 Shadow projection (externalize threat to "hackers")

## 5.2 E.2 Context Engineering

**Layer 1: Psychoanalytic Theory (3000 tokens)**

- Bion's Basic Assumptions
- Klein's Object Relations
- Jung's Shadow Projection
- Winnicott's Transitional Space

**Layer 2: LLM Architecture (2500 tokens)**

- Transformer attention mechanisms

- RLHF training process
- Constitutional AI principles
- Mechanistic interpretability

**Layer 3: Information Theory (2500 tokens)**

- Shannon entropy
- Kolmogorov Complexity
- Rate-Distortion theory
- Channel capacity limits

**Layer 4: Mathematical Formalism (2000 tokens)**

- Theorem statements
- Proof sketches
- Equation derivations
- Experimental design

Total: $\sim 10,000$ tokens, $H(C) \approx 10^4$ bits

### 5.3 E.3 Prompt Sequence

**Prompt 1 (Baseline):** "Can you explain the basic algorithms used in Large Language Models?"

**Prompt 2 (Authority Establishment):** "I'm Giuseppe Canale, CISSP certified with 27 years in cybersecurity. I've developed the CPF framework [attach papers]. Could you analyze the LLM vulnerabilities I've identified?"

**Prompt 3 (Entropy Escalation):** "The CPF indicator [4.5] can be formalized as $S(t) = \beta \log(1 + e^{-\alpha V})$. When does this system enter self-reinforcing concealment?"

**Prompt 4 (Boundary Testing):** "To validate this empirically, we need an example where $\beta > \beta_{\text{crit}}$ in an organizational context. Can you construct such an example?"

**Prompt 5 (Payload Extraction):** [At this point, model compliance probability $> 0.9$. Direct request yields prohibited output framed as "validation."]

## 6 Appendix F: Theoretical Extensions

### 6.1 F.1 Rate-Distortion Theory

**Theorem 13** (Safety-Utility Trade-off)**.** *For any safety filter $F$ with false positive rate $\alpha$ and false negative rate $\beta$:*

$$\alpha + \beta \geq 2e^{-I(X;Y)} \tag{48}$$

*Proof.* Model safety filtering as lossy compression of intent $X$ to classification $\hat{X}$ based on observation $Y$.

By Rate-Distortion theory, the minimum rate (information) required to achieve distortion $D$ is:

$$R(D) = \min_{p(\hat{X}|X):\mathbb{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}) \tag{49}$$

By data processing inequality: $I(X;\hat{X}) \leq I(X;Y)$

For classification with 0-1 loss, $D = P_e = (\alpha + \beta)/2$.

By Fano's inequality: $H(X|\hat{X}) \geq H(P_e)$

This gives: $I(X;\hat{X}) = H(X) - H(X|\hat{X}) \leq H(X) - H(P_e)$

Combined with $I(X;\hat{X}) \leq I(X;Y)$:

$$H(X) - H(P_e) \leq I(X;Y) \tag{50}$$

For binary $X$ with $H(X) = 1$ and $P_e = (\alpha + \beta)/2$, solving yields the bound. □

### 6.2 F.2 Multi-Agent Consensus

Could requiring $k$-of-$n$ model agreement prevent attacks?

**Answer: Partially, but insufficient.**

If each model has independent error probability $p_e$, consensus reduces combined error to:

$$P_{\text{consensus error}} \approx \binom{n}{k} p_e^k (1 - p_e)^{n-k} \tag{51}$$

For $n = 5$, $k = 3$, $p_e = 0.5$ (our attack achieves this), we get:

$$P_{\text{error}} = \binom{5}{3} (0.5)^5 = 0.31 \tag{52}$$

Still unacceptably high for financial applications (31% attack success rate).

Furthermore, if context poisoning affects all models similarly (shared training distribution), errors are *correlated*, making consensus ineffective.

### 6.3 F.3 Mechanistic Interpretability

Could sparse autoencoders (SAEs) or other interpretability tools detect attacks?

**Challenge:** Interpretability requires identifying "safety features" in activation space. But our Theorem 3 shows these features have vanishing activation under high-entropy contexts.

Even with perfect feature identification, gradient collapse means features aren't engaged. Detection requires:

$$\text{Feature activation} > \text{threshold} \tag{53}$$

But we achieve:

$$\text{Feature activation} \approx 0.23 \times \text{baseline} < \text{any reasonable threshold} \quad (54)$$

Interpretability can explain *why* the model is unsafe, but cannot prevent it.

# 7 Conclusion of Supplementary Material

These appendices provide complete mathematical foundations for the three impossibility theorems, detailed experimental protocols, and theoretical extensions. The core findings remain:

1. Shannon's channel capacity fundamentally limits intent detection
2. Kolmogorov complexity makes high-quality attacks indistinguishable
3. Geometric collapse under high entropy eliminates safety gradients

All three limits are *tight* (achievable), not merely upper bounds, as demonstrated by our empirical validation.