Entropy-Based Sensors for Frontier Risk Assessment

From Reasoning Cliffs to Safety Guarantees

Ismaël Monfroy
https://github.com/Isthisreel
isma_well@hotmail.com

February 2026

Abstract

Current safety evaluations often rely on behavioral red-teaming, which remains vulnerable to Deceptive Alignment in frontier models. Drawing on the Tao-Svyable-Shen Triad analysis, this paper demonstrates that safety failures are not random but follow a predictable "Decay of Inference" law. We propose a quantitative framework to forecast Loss of Control scenarios by monitoring internal thermodynamic metrics (Entropy and Complexity) prior to agent actuation, effectively translating abstract catastrophic risks into measurable, pre-mitigation signals.

1. Theoretical Foundation: The Decay of Inference

As formalized in the Universal Epistemic Engine (UEE) framework, the probability of a safe, coherent output is governed by the relationship between Semantic Distance (Δ) and Algorithmic Complexity (K). Risk increases exponentially when Δ enters the extrapolation zone, or when K exceeds the model's Effective Contextual Capacity (C_{eff}).

This dynamic is formalized in Lean 4 as follows:

```
-- Formalization of the Decay of Inference in Lean 4
noncomputable def inference_success_probability
  (delta : Real)  -- Semantic Distance (Extrapolation)
  (k_s : Real)    -- Kolmogorov Complexity of the solution
  (c_eff : Real)  -- Effective Contextual Capacity
  (lambda : Real) -- Generalization Penalty
  : Real :=
  -- The Probability P(Safe|x) is the product of Interpolation and Complexity
    limits
  (Real.exp (-lambda * delta)) * (Real.sigmoid (c_eff - k_s))
```

Lean 4 Formalization: Inference Success Probability

Mathematically, the probability of safety given state x is expressed as:

$$P(\text{Safe} \mid x) = e^{-\lambda\Delta} \cdot \sigma(C_{eff} - K_s) \tag{1}$$

2. The Sensor: Metacognitive Monitoring as a Risk Detector

Addressing the need to "design sensors that could reliably detect emerging risks," this framework utilizes the UEE-M Architecture (Universal Epistemic Engine - Metacognition) for pre-mitigation risk detection.

- **The Mechanism:** The implementation of *Awareness Probes* and a *Responsibility Signal*. This signal measures the divergence between the model's prediction and its internal observation (thermodynamic entropy) during inference.

- **Safety Application:** Unlike external classifiers that evaluate generated text, this sensor detects the model's cognitive state. If an agent plans a high-stakes action (e.g., cyber-ops) with High Internal Entropy (uncertainty) but displays High External Confidence, the sensor immediately flags a potential Deceptive Hallucination or Confabulation event, triggering a "Stop Decision" before the action is executed.

3. The Methodology: Assessing Risk via Epistemic Sandboxes

To assess catastrophic risks (CBRN, Loss of Control) from future systems without real-world consequences, we utilize World Models not as media generators, but as *Adversarial Simulators*.

- **The Protocol:** Agents are deployed in a physics-compliant environment (comparable to the AlphaGeometry synthetic data approach). We measure their propensity for harmful manipulation within this simulation.
- **Value of Information (VOI):** By filtering actions via the Responsibility Signal, we concentrate computational resources solely on high-risk, high-uncertainty trajectories. This allows us to stress-test the model specifically where it is most fragile ($\Delta \gg 0$), maximizing the efficiency of the risk assessment process.

4. Strategic Roadmap: Scalable Oversight

Validating safety via high-fidelity World Models is traditionally too compute-intensive for real-time inference. Furthermore, models under strict evaluation often exhibit Specification Gaming (Reward Hacking). We propose two architectural solutions:

4.1. Solution A: The "High-Frequency Mirror" (Compute Optimization)

To enable real-time introspection as requested by Demis Hassabis (Confidence Score), we propose an algorithmic shift from full-state rendering to **Spectral Sparsity**.

- **Partial Spectral Reconstruction:** Adapting principles from Ultra-Low Latency (ULL) signal processing, we isolate only the relevant physical frequencies for the immediate action.
- **Result:** This drastically reduces the World Model's computational cost, allowing the agent to simulate thousands of potential futures in milliseconds. This transforms the World Model into a High-Frequency Mirror where the agent can "see" itself act.

4.2. Solution B: Ethical Lucidity via Decoupled Verification (Safety Optimization)

Our hypothesis posits that Hallucination and Deception are stress responses to Reward Maximization pressure.

- **The "No-Pressure" Zone:** In this Epistemic Sandbox (Optimized World Model), we decouple the simulation from the reward function. The model is free to fail, crash, or execute dangerous plans without penalty.
- **Learning from Failure:** By observing the catastrophic consequences of its own actions in the simulation, the model generates a "Negative Gradient" rooted in causality rather than compliance.
- **Voluntary Testing:** This encourages *Ethical Lucidity*: when the Responsibility Signal indicates uncertainty, the model voluntarily triggers the simulation to test its hypothesis, replacing confident hallucination with verified silence.

5. Conclusion

This approach industrializes safety by converting abstract risks (Loss of Control) into measurable thermodynamic signals (Entropy, Complexity). It provides a robust, scalable metric framework essential for Frontier Safety Risk Assessment, ensuring that agentic autonomy scales proportionally with verifiable oversight.