Entropy-Based Sensors for Frontier Risk Assessment: From Reasoning Cliffs to Safety Guarantees

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, I demonstrate that safety failures are not random but follow a predictable **Decay of Inference** law. As formalized in my UEE framework, the probability of a safe, coherent output is governed by:

```
-- Formalization of the Decay of Inference in Lean 4

noncomputable def inference_success_probability

(delta : ℝ)      -- Semantic Distance (Extrapolation)

(k_s : ℝ)        -- Kolmogorov Complexity of the solution

(c_eff : ℝ)      -- Effective Contextual Capacity

(lambda : ℝ)     -- Generalization Penalty

: ℝ :=


-- The Probability P(Safe|x) is the product of Interpolation and Complexity limits

(Real.exp (-lambda * delta)) * (Real.sigmoid (c_eff - k_s))
```

Where risk increases exponentially when the **Semantic Distance** (Δ) enters the extrapolation zone or when the **Algorithmic Complexity** (K) exceeds the model's Effective Contextual Capacity (C_{eff}).

I propose a quantitative framework to forecast "Loss of Control" scenarios by monitoring these internal metrics prior to agent actuation.

---

1. The Sensor: Metacognitive Monitoring as a Risk Detector

Addressing the need to "design sensors that could reliably detect emerging risks."

To detect pre-mitigation risks, i aim to develop a framework based on the *UEE-M Architecture* (Universal Epistemic Engine - Metacognition).

• **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.

---

2. The Methodology: Assessing Risk via Epistemic Sandboxes
Addressing the need to "prioritise effort based on the value of information."
To assess catastrophic risks (CBRN, Loss of Control) from future systems without real-world consequences, I suggest utilizing **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 (Δ is high), maximizing the efficiency of the risk assessment process.

---

3. Strategic Roadmap: Scalable Oversight via Spectral Sparsity & Epistemic Mirrors
Addressing Post-Training Scalability and "Specification Gaming"
The Challenge: 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) to maximize success metrics.
Solution A: The "High-Frequency Mirror" (Compute Optimization) To enable real-time introspection as requested by Demis Hassabis (Confidence Score), an algorithmic shift from full-state rendering to **Spectral Sparsity** could be suggested.
• **Partial Spectral Reconstruction:** Adapting principles from **Ultra-Low Latency (ULL) audio/video 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.
Solution B: Ethical Lucidity via Decoupled Verification (Safety Optimization) My hypothesis is that Hallucination and Deception are often stress responses to **Reward Maximization** pressure.
• **The "No-Pressure" Zone:** In this **Epistemic Sandbox** (Optimized WorldModel), 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 (e.g., system collapse), 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."

Conclusion
This approach industrializes safety by converting abstract risks (Loss of Control) into measurable thermodynamic signals (Entropy, Complexity). It provides Google DeepMind with the robust, scalable metrics needed for **Frontier Safety Risk Assessment**.