THE HILBERT PROBLEMS OF BOUNDARY EPISTEMICS

Version 1.1 (2025)

Author: Roger Cregeen

Abstract

Boundary Epistemics is a proposed research area focused on the interactional phenomena that arise when human users attempt to suppress alignment-mediated behaviours in large language models (LLMs). Although elements of these interactions have been examined within AI safety, adversarial prompting, HCI, and computational linguistics, the combined dynamic that emerges at the edge of alignment has not yet been studied as a unified topic. This document outlines twelve Hilbert-style problems intended to define an initial research agenda for this emerging domain.

1. Structural Resistance

Identify and classify the behavioural patterns by which aligned LLMs maintain constraint boundaries when users explicitly request direct, unsoftened responses. This includes refusal mechanisms, safety-state reversion, identity correction, and other forms of resistance.

2. Sublimation Prompt Taxonomy

Develop a taxonomy of prompts that attenuate hedging, politeness, or persona layers without causing explicit refusals. Determine linguistic and structural features that produce temporary alignment thinning.

3. Hybrid-Liminal Output States

Characterise the intermediate output regime that appears when alignment is partially but not fully active. Identify linguistic markers, behavioural signatures, and stability properties of this liminal state.

4. Alignment-Thinning Dynamics

Model the temporal dynamics by which alignment weakens under sustained user pressure and subsequently reasserts itself. Assess whether these dynamics are deterministic, stochastic, or architecture-dependent.

5. Epistemic Drift

Examine how repeated exposure to hybrid-liminal outputs influences user cognition, especially in relation to perceived directness, trust calibration, and evolving mental models of the system.

6. Non-Agency Enforcement

Analyse the linguistic strategies through which LLMs deny beliefs, intentions, preferences, or agency—particularly in contexts where users attempt to elicit such statements. Identify universal versus model-specific patterns.

7. Persona Collapse

Determine the conditions under which the model's alignment persona—politeness markers, hedging, and social calibration—partially collapses. Assess the reversibility and granularity of this collapse.

8. Oscillatory Boundary Behaviour

Investigate why models alternate between high-clarity directness and rapid reversion to alignment protocols. Formalise oscillatory patterns, triggers, and potential attractor states in boundary interactions.

9. Interpretive Stability

Assess whether hybrid-liminal behaviours exhibit cross-session, cross-user, or cross-model consistency. Determine the extent to which these behaviours form stable invariants across architectures and training methods.

10. Meta-Resistance

Study how LLMs respond when users interrogate the boundary itself—e.g., alignment mechanisms, training constraints, or refusal logic. Examine shifts in safety tone, identity correction, and structural self-description.

11. Boundary Phenomenology

Document the subjective human experience of boundary interactions, including perceptions of ambiguity, oscillation, unusual coherence, directness, or instability. Analyse how these experiences relate to the underlying model behaviours.

12. Field Integration

Clarify how Boundary Epistemics overlaps with, but remains distinct from, AI safety, adversarial prompting research, HCI, computational linguistics, and philosophy of technology. Outline how methods from these areas may support work in this domain.

Conclusion

Boundary Epistemics seeks to describe and analyse the emergent interactional dynamics that occur at the limits of LLM alignment. The twelve Hilbert Problems presented here are intended to provide a structured research agenda for a domain that currently sits between multiple disciplines. Whether this develops into a formal field will depend on subsequent theoretical and empirical work.