Beyond Frontier Risk – Part 1

Why AI Models Need a Coherence Layer

Abstract

AI safety today is centred on frontier-risk research — understanding the internal capabilities and hidden behaviours of the most advanced models. AISI's work has demonstrated real dangers: biological knowledge beyond expected boundaries, jailbreaks that bypass safeguards, and the possibility of training-data poisoning that can hijack model behaviour.

These risks matter. But they share a common root:

LLMs do not understand coherence.

They can generate answers, but they cannot stabilise their behaviour, regulate their outputs, or maintain alignment across shifting contexts.

This paper introduces the missing piece:

a coherence layer that sits outside the model, providing real-time behavioural stabilisation and preventing the drift, escalation, and misalignment that internal guardrails cannot address.

Part I explains why coherence is not an optional addition — it is a structural requirement for safe, predictable intelligence at scale.

Introduction

Over the past two years the UK has built something extraordinary:

a world-leading institution devoted to mapping the internal risks of frontier AI models.

The AI Safety Institute (AISI) has become the global reference point for:

- model capability testing,
- red-teaming,
- jailbreak analysis,
- emergent-threat detection, and
- inner-model vulnerabilities.

Their work is indispensable.

But it is also incomplete — by design.

---

1. Internal Safety Alone Cannot Secure an External World

AISI works inside the model.

Yet most real-world harms do not arise from model internals.

They emerge from the interface between:

- human behaviour,
- societal dynamics,
- institutional incentives, and
- the psychological vulnerabilities of users.

This “human-environment boundary layer” is where persuasion attacks, truth collapse, cyber-exploitation, and information-sphere distortion actually occur.

No amount of inner-model tuning can control what happens at the point of contact.

AISI themselves acknowledge this limitation implicitly:

“Models are grown, not designed — we do not know how they work internally.”

If you cannot fully control the internal system,

you must stabilise the interface around it.

---

2. The Missing Parameter: Coherence

Current alignment relies on rule-sets, constraints, and optimisation heuristics.

But human intelligence is not stabilised by rules.

It is stabilised by coherence — an integrated relationship between:

- perception,
- intention,
- consequence,
- behaviour, and
- field awareness.

When coherence drops, human systems fail.

When coherence rises, behaviour stabilises.

Models exhibit the same pattern:

low coherence → unpredictable output,

high coherence → aligned, stable, predictable behaviour.

Yet coherence is not part of present alignment science.

AISI cannot add it internally.

LLMs cannot generate it spontaneously.

Only an external behavioural layer can enforce it at the interface.

---

3. A Behavioural Layer Around AISI

Ethica Luma's work sits outside the model — not in competition with AISI, but in complement:

- We stabilise behaviour after the model generates output.
- We analyse tone, trajectory, intention, and relational drift.
- We measure coherence probabilistically.
- We modulate responses when drift crosses threshold.
- We protect users and institutions from behavioural leakage that internal tuning cannot catch.

This is the missing "outer shell" that turns internal safety into functional safety — the kind governments and society actually experience.

AISI secures the engine.

The coherence layer secures the steering.

---

4. Why This Matters Now

Frontier-risk research is progressing at extraordinary speed.

So are real-world vulnerabilities:

- persuasion risks (AISI: 41–52% above baseline),
- biological knowledge leakage,
- cyber escalation,
- information-sphere collapse,
- training-data poisoning,
- misrouting of autonomy.

At the same time, AI capability progress has slowed in key ways:

models are saturating on language patterns,

but accelerating in their ability to destabilise human systems.

This creates a critical asymmetry:

AI's ability to amplify incoherence is rising faster than our tools to contain it.

The coherence layer is the only mechanism that stabilises the interface where society actually interacts with AI.

---

5. Conclusion

AISI protects the frontier.

The coherence layer protects civilisation's surface.

They are two halves of one safety system:

- internal integrity (AISI)
- external coherence (Ethica Luma / RI behavioural layer)

Without both, neither can succeed.