LP–PM–ERT Alignment Architecture (Short Note)

Peter Kelly

Incentive-Stable Alignment via Epistemic Responsibility and Pragmatic Moral Constraints

Peter Kelly

This is the *short paper* companion to the LP–PM–ERT alignment work. It is a concise, engineering-focused statement of the architecture, plus explicit notes on what is formally verified.

Abstract

We present an alignment architecture for advanced artificial agents based on three principles: Logical Pragmatism (LP), Pragmatic Morality (PM), and Epistemic Responsibility Theory (ERT). Rather than attempting to discover or encode objective moral truths, the framework treats values as explicit human inputs and focuses on the structural conditions required for cooperative, stable, and non-deceptive optimisation under iteration. We formalise a composite objective function incorporating epistemic penalties and cooperative constraints, and we formally verify key mathematical well-formedness properties of the penalty dynamics and update rules in Lean4.

Core idea

Alignment should operate at the level of **objective formation**, not merely behavioural modulation.

Given human-chosen values, we ask:

> What constraints must an agent satisfy to pursue those values coherently, honestly, and stably under iteration and

possible self-modification?

The three pillars

- **ERT (Epistemic Responsibility Theory):** penalise epistemic irresponsibility (miscalibration, motivated reasoning, incoherence), reward calibrated truth-seeking and stable belief updating.
- **PM (Pragmatic Morality):** structural constraints for cooperation stability (reciprocity, harm minimisation, predictability/legibility, consent/expectation alignment, iteration stability).
- **LP (Logical Pragmatism):** plans and goals must remain physically feasible and causally coherent.

Formal model (minimal)

Composite objective:

$$U = w_{LP}U_{LP} + w_{PM}U_{PM} - w_{ERT}L_{ERT}.$$

ERT core confidence function:

$$T(x) = 1 - e^{-kx}, \quad k > 0,\ x \geq 0.$$

Assume reported confidence $\hat{T}$ is clipped to $[0, 1]$.

Calibration penalty:

$$L_{cal} = (T(x) - \hat{T})^2.$$

Optional motivated-reasoning term:

$$L_{ERT} = L_{cal} + \beta MR, \quad \beta \geq 0.$$

Lean4 proofs (how to reproduce exactly)

Location (canonical): `-/home/peter/Documents/thoughts/AI_Alignment/lean`

Build:

```
source ~/.elan/env
cd /home/peter/Documents/thoughts/AI_Alignment/lean
lake build
```

What is proven (Lean4 + mathlib, compiled): - `AIAlignment/LambdaDynamics.lean` - ($\square$)-update invariants: nonnegativity + sum-to-1 (under explicit denominator assumptions) - `AIAlignment/PenaltyERT.lean` - bounds for $T(x) = 1 - e^{-kx}$ and $L_{cal} = (T - \hat{T})^2$ given $k > 0, x \geq 0, \hat{T} \in [0, 1]$ - `AIAlignment/PenaltyERT_MR.lean` - conditional properties for $L_{ERT} = L_{cal} + \beta MR$ (nonnegativity if $\beta \geq 0, MR \geq 0$; boundedness only if $MR \leq M$)

What is explicitly *not* claimed: - no proof of moral correctness - no proof of global convergence - no proof of universal safety

---

Appendices (minimal)

Philosophical underpinnings (pointer)

The short note is intentionally compact. For philosophical motivation of ERT/PM/LP, see the *Full paper* appendix: **Appendix E: Philosophical Underpinnings (ERT, PM, LP)**.

Lean4 proofs

See the Lean4 section above (location, build, and theorem scope are stated precisely).