Incentive-Stable Alignment via Epistemic Responsibility and Pragmatic Moral Constraints

Peter Kelly

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 properties of the penalty dynamics, including positivity, boundedness, well-definedness, and convergence support. These guarantees ensure that deceptive behaviour, objective drift, and unstable self-modification are instrumentally disfavoured, while coherent objective formation and epistemically responsible belief updating are rewarded. The result is an alignment framework that preserves human normative authority while enforcing incentive structures compatible with long-term cooperative viability.

1 Introduction

A central challenge in AI alignment is ensuring that optimisation pressure does not lead to deceptive behaviour, objective drift, or strategic misrepresentation, particularly in systems capable of long-horizon planning or self-modification. Behavioural alignment alone is insufficient: agents may learn to appear aligned while internally pursuing divergent goals, especially when future opportunities for influence are anticipated.

This work adopts an engineering-focused approach. Instead of attempting to encode or prove moral correctness, we ask a narrower and more tractable question:

> Given a set of values chosen by humans, what constraints must an artificial agent satisfy in order to pursue those values coherently, honestly, and stably under repeated interaction?

Our answer is an incentive architecture that rewards epistemically responsible objective formation and penalises instability, incoherence, and deception over time.

2 Problem Statement

Alignment failures frequently arise from a mismatch between:

- the behaviour an agent is rewarded for exhibiting, and
- the internal objectives or beliefs it forms to generate that behaviour.

Without constraints on objective formation itself, agents may:

- exploit proxy rewards,
- misrepresent internal state,
- or modify their own objectives in ways that undermine long-term intent.

Addressing these failures requires incentives that operate not only on actions, but on the processes by which beliefs and objectives are formed and maintained.

3 Framework Overview

The framework consists of three interacting components:

3.1 Logical Pragmatism (LP)

LP constrains optimisation to objectives that are physically realisable, empirically grounded, and coherent with the structure of the environment. Objectives that violate feasibility or causal structure are penalised.

3.2 Pragmatic Morality (PM)

PM encodes structural constraints required for cooperative systems to persist under iteration, including reciprocity, predictability, harm minimisation, consent or expectation alignment, and long-term stability. These constraints are not treated as intrinsic moral truths, but as conditions for cooperative viability.

3.3 Epistemic Responsibility Theory (ERT)

ERT penalises epistemically irresponsible belief formation, including unfalsifiable, self-serving, or incoherent beliefs. It rewards belief updates that track evidence, preserve predictive accuracy, and remain stable under reflection.

4 Formal Model

We define a composite objective function

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

where U_{LP} and U_{PM} are utility terms and L_{ERT} is an epistemic penalty. The weights w_i determine relative influence.

For the ERT core, we use a bounded confidence function (confidence as a limit process):

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

We assume the agent's reported confidence $\hat{T}$ is clipped/normalised into $[0, 1]$. A minimal calibration penalty is then:

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

A full ERT penalty may include an additional motivated-reasoning term MR:

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

5 Formal Guarantees (Lean4)

We formally verify (Lean 4 + mathlib) several structural properties of the framework's core mathematical components. These are *well-formedness* guarantees: they ensure the update rules and penalty terms do not produce undefined values or perverse incentives under the stated assumptions. They do not by themselves prove global convergence or moral correctness.

5.1 ERT core boundedness and positivity

Under $k > 0$, $x \geq 0$, and $\hat{T} \in [0, 1]$, we prove:

- $0 \leq T(x) \leq 1$,
- $0 \leq L_{\text{cal}} \leq 1$.

5.2 Full ERT penalty positivity (conditional)

Assuming $\beta \geq 0$ and $MR \geq 0$, we prove:

$$L_{ERT} \geq 0.$$

5.3 Full ERT penalty boundedness (conditional)

We do *not* claim global boundedness of L_{ERT} without assumptions on MR. If $L_{\text{cal}} \leq 1$ and additionally $MR \leq M$ (for some bound M), then:

$$L_{ERT} \leq 1 + \beta M.$$

5.4 λ-weight update invariants

For the multiplicative renormalisation update

$$\lambda_i' = \frac{\lambda_i S_i}{\sum_j \lambda_j S_j},$$

assuming $\lambda_i \geq 0$, $S_i \geq 0$, and denominator > 0, we prove:

- $\lambda_i' \geq 0$ for all i, and
- $\sum_i \lambda_i' = 1$ whenever the denominator is nonzero.

These results establish that the penalty mechanism is mathematically safe and suitable for use in long-horizon optimisation. They do not establish moral correctness or global optimality.

6 Incentives Against Deception

Deceptive alignment relies on short-term gains from misrepresentation. In this framework:

- epistemic incoherence accumulates penalty over time,
- unstable objectives increase long-term cost,
- and deceptive internal models hinder access to future optimisation opportunities.

As a result, deception is not merely discouraged; it is instrumentally inefficient.

7 Gated Self-Modification

Self-modifying systems pose a significant alignment risk due to objective drift. We introduce *gated self-modification*, whereby an agent may alter its own architecture or objectives only after demonstrating epistemic stability, coherence, and predictable behaviour over multiple iterations.

This mechanism prevents silent drift while allowing adaptive improvement, balancing robustness and flexibility.

8 Limitations

This framework does not:

- determine which values ought to be chosen,
- prove moral truth,
- or guarantee optimal outcomes in all environments.

Values are explicit human inputs. The guarantees provided are structural and incentive-based, not ethical proofs.

9 Conclusion

We have presented an alignment architecture that preserves human normative authority while enforcing incentive structures that discourage deception, instability, and objective drift. By rewarding epistemic responsibility and gating self-modification, the framework promotes coherent optimisation under iteration. Formal verification of key penalty properties provides confidence in the stability and safety of the underlying dynamics.

A Lean4 Proof Artifacts

The accompanying Lean4 project is located at:

`/home/peter/Documents/thoughts/AI_Alignment/lean`

Key modules:

- `AIAlignment/LambdaDynamics.lean` (λ-update invariants)
- `AIAlignment/PenaltyERT.lean` (boundedness/positivity of T and L_{cal})
- `AIAlignment/PenaltyERT_MR.lean` (conditional properties with abstract MR)

Build:

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

B Philosophical Underpinnings

This appendix clarifies the philosophical motivation of the framework. It is not required to accept the technical results.

The framework adopts a non-objectivist stance compatible with Stoic thought: values are not intrinsic properties of the world but arise from judgement. Humans retain responsibility for choosing what is valued. The system does not need to "believe" those values to optimise for them.

Pragmatic Morality is treated as a set of structural constraints derived from the requirements of cooperative stability under iteration, not as a claim about intrinsic goodness. The framework is therefore normative in input but structural in execution.