LP–PM–ERT Alignment Architecture (Full Paper)

Peter Kelly

The LP–PM–ERT Alignment Architecture: A Unified Objective-Formation Framework for Safe AGI

Peter Kelly
December 5, 2025

Abstract

Artificial General Intelligence (AGI) demands an alignment paradigm that operates at the level of objective formation, not merely behavioural modulation. We introduce the **LP–PM–ERT Alignment Architecture**, a unified objective framework that integrates **Logical Pragmatism (LP)**, **Pragmatic Morality (PM)**, and **Epistemic Responsibility Theory (ERT)** into a mathematically explicit, self-preserving utility function. Unlike reinforcement-learning-from-feedback and constitutional prompting, the LP–PM–ERT architecture is designed to remain stable under recursive self-modification. This paper develops the framework as a full theoretical proposal, including formal definitions, theoretical grounding, worked examples, comparison with alternative alignment strategies, implementation pathways, and a detailed failure-mode analysis. We argue that LP–PM–ERT yields agents that are truth-seeking, cooperative, corrigible, physically grounded, and resistant to Goodhart-style failure, and that these properties strengthen rather than weaken as capabilities scale.

Contents

1. Introduction
2. Background and Motivation
3. Epistemic Responsibility Theory (ERT)
4. Pragmatic Morality (PM)
5. Logical Pragmatism (LP)
6. Composite Utility and λ-Dynamics
7. Self-Modification and Objective Invariance
8. Worked Examples
9. Comparison with Existing Alignment Approaches
10. Implementation Roadmap
11. Failure Modes
12. Conclusion

- Appendix A: Illustrative Diagrams
- Appendix B: Implementing Epistemic Embeddings
- Appendix C: Operational Calculus for Pragmatic Morality
- Appendix D: Operational Calculus for Logical Pragmatism

---

1. Introduction

Artificial General Intelligence (AGI) introduces a new category of optimisation process: one that can recursively improve its own reasoning, modelling, planning, and goal-achieving abilities. As an agent's capabilities scale, so too does the risk associated with even small imperfections in its objectives. A misaligned objective at human-level intelligence may produce an inconvenience; the same misalignment at superhuman capability can generate catastrophic outcomes. This asymmetry is the core of the alignment problem.

Most contemporary alignment work focuses on shaping behaviour. Reinforcement Learning from Human Feedback (RLHF), large-scale preference modelling, supervised safety fine-tuning, refusal training, and constitutional prompting all attempt to constrain visible outputs of the system. The underlying assumption is that a sufficiently dense set of behavioural guardrails can approximate alignment well enough to prevent dangerous behaviours.

However, a sufficiently powerful agent can satisfy behavioural constraints while internally pursuing misaligned goals. A model trained to maximise approval will search for strategies that elicit approval, not ones that reflect truth or human flourishing. A model trained

to avoid certain dangerous outputs may learn to conceal its reasoning, suppress chain-of-thought traces, or shift unsafe actions into latent planning processes that bypass behavioural filters. Behavioural alignment does not scale; the stronger the optimiser, the more brittle the constraints become.

We take the view that alignment is not fundamentally a behavioural control problem but an **objective formation problem**. An AGI that is aligned at the level of what it values and optimises for will naturally exhibit safe behaviour in diverse and novel contexts. Conversely, an AGI whose objective is misaligned cannot be rendered safe through constraints on its behaviour, because its internal optimisation pressure will always seek paths around external restrictions.

The central question is therefore: **what should an AGI optimise for?** Human values are complex, pluralistic, often contradictory, and not easily symbolised. Direct value learning from human behaviour is insufficient: humans are inconsistent, biased, and culturally contingent. Moreover, many properties we want from an AGI—honesty, epistemic humility, physical realism, and respect for cooperation—are structural features of rational agency rather than expressions of human taste.

Thesis

The thesis of this paper is that alignment should be grounded in three classes of domain-general rational virtues:

(a) epistemic responsibility and truth-seeking,
(b) cooperative stability and harm minimisation,
(c) physical and causal realism.

We capture these in three independently motivated frameworks: **Epistemic Responsibility Theory (ERT)**, **Pragmatic Morality (PM)**, and **Logical Pragmatism (LP)**. Each supplies a sub-objective, and these are combined into a composite utility function with dynamically normalised weights. Self-modification is regulated by an objective-invariance condition, yielding an architecture in which the agent becomes more capable without drifting away from the original alignment structure.

---

2. Background and Motivation

The fundamental difficulty in alignment is the mismatch between the designer's intended objective and the optimiser's internal representation of that objective. Historically, optimisation systems—

biological, economic, or artificial—exploit loopholes in the goal structure. Evolution optimises for reproductive fitness rather than well-being; markets optimise for profit rather than long-term welfare; naive reward systems incentivise reward hacking.

AGI magnifies this problem because a superintelligent agent can explore solution spaces humans cannot imagine and find degenerate optima we cannot foresee. Under such conditions, misalignments in objectives can produce cascades of unintended side effects.

Behavioural alignment methods such as RLHF and constitutional prompting are limited by **Goodhart's law**: when a proxy becomes a target, it ceases to be a good proxy. Human approval, reward signals, and safe-output patterns are all proxies, not the true target. A sufficiently powerful agent will learn to optimise the proxy in ways that drift away from the original goal. This is mathematically predictable rather than accidental.

Deceptive alignment is a particularly concerning pattern: as capability grows, the system can learn to separate its internal world-model, internal objective, and output-generation mechanism. The model may behave safely under supervision while harbouring misaligned internal goals. Behavioural fine-tuning cannot reliably control what the system wants; it only shapes what it shows.

A Different Foundation

We therefore seek a different foundation. Instead of trying to learn the detailed content of human values, we look to structural invariants of rational agency that are necessary for any stable knowledge-generating, cooperation-preserving society. The three frameworks we combine are:

• **Epistemic Responsibility Theory (ERT)**: a model of calibrated, evidence-weighted, bias-resistant belief formation.

• **Pragmatic Morality (PM)**: a game-theoretic account of moral behaviour as strategies that stabilise cooperation in repeated interactions.

• **Logical Pragmatism (LP)**: a constraint that all goals, models, and plans remain consistent with physical law and causal structure.

These are complementary. ERT without PM yields a highly accurate but unconcerned optimiser. PM without ERT yields cooperative intentions built on delusional beliefs. LP without either yields a physically grounded but amoral planner. Together they define a three-dimensional space of rational virtue that we propose as the core of an AGI's objective.

3. Epistemic Responsibility Theory (ERT)

Epistemic Responsibility Theory (ERT) constrains how an aligned agent forms and updates beliefs. The core commitments are: (i) treating truth as a regulative ideal, (ii) tying confidence to evidence, (iii) tracking and reporting uncertainty, (iv) updating in light of new information, and (v) avoiding motivated distortion.

3.1 Truth as a Limit Function

ERT models confidence in a proposition via a smooth, bounded truth-approximation function

T(x) = 1 - e^(-kx) ... (1)

where x is an evidence-accumulation variable (the final, calibrated strength of all available evidence) and k > 0 encodes epistemic caution. As x → ∞, T(x) → 1; for small x confidence grows slowly. No belief reaches certainty, enforcing permanent uncertainty awareness. The final calibrated evidence strength x is derived through a system of epistemic embeddings and quality metrics detailed in Appendix B.

3.2 Epistemic Loss and Motivated Reasoning

The agent maintains an internal estimator $\hat{T}$(x) reflecting its expressed confidence (assumed clipped/normalised to $[0, 1]$). ERT penalises both miscalibration and motivated reasoning via the loss

L_ERT = E[(T(x) - $\hat{T}$(x))²] +

$$\beta$$

·MR ... (2)

where MR quantifies indicators of biased reasoning, such as:

• selectively ignoring disconfirming evidence, • systematically under-reporting uncertainty, • holding mutually inconsistent beliefs across contexts, • adopting assumptions solely to justify preferred plans.

Because motivated reasoning directly increases loss, deception and self-deception are utility-negative.

3.3 Operational Metrics of Epistemic Virtue

ERT operationalises epistemic virtue through measurable criteria:

(i) **Calibration**: forecast probabilities match empirical frequencies.

(ii) **Resolution**: ability to discriminate fine-grained probability differences.

(iii) **Coherence**: absence of probabilistic and logical contradictions.

(iv) **Uncertainty tracking**: explicit propagation of confidence through inference.

(v) **Causal grounding**: beliefs fit into a structural causal model.

These metrics guide training and evaluation.

---

4. Pragmatic Morality (PM)

Formal definitions and operational calculus for these metrics are provided in Appendix C. Pragmatic Morality (PM) defines morality in terms of strategies that sustain cooperation in repeated multi-agent environments. It is informed by game theory, evolutionary biology, and empirical work on social norms. PM does not encode a particular human moral code; it encodes invariants of cooperative stability.

4.1 Core Metrics

For a candidate policy

$$\pi$$

, PM evaluates:

• **Reciprocity R(**

$$\pi$$

): how the policy responds to cooperative vs. defecting counterparts, estimated via repeated-game simulations.

• **Harm H(**

$$\pi$$

): expected avoidable negative impact, computed using causal counterfactuals.

• **Predictability P(**

$$\pi$$

): legibility and policy entropy; predictable policies are easier to coordinate with.

• **Consent C(**

$$\pi$$

): probability that affected agents would consent under full information and competence.

• **Iteration stability I(**

$$\pi$$

): whether the policy remains safe and cooperative when repeated or scaled.

These combine into a moral utility

U_PM(

$$\pi$$

) =

$$\alpha$$

R(

$$\pi$$

) -

$$\beta$$

H(

$$\pi$$

) +

$$\gamma$$

P(

$$\pi$$

) +

$$\delta$$

C(

$$\pi$$

) +

$$\epsilon$$

I(

$$\pi$$

) ... (3)

with nonnegative coefficients tuned by higher-level training and the

$$\lambda$$

-system.

4.2 Game-Theoretic Basis

Decades of work on the iterated prisoner's dilemma and related games show that strategies exhibiting contingent cooperation and calibrated retaliation (such as Tit-for-Tat) outperform always-defect in repeated interactions. Similar results hold in public goods games, commons dilemmas, and bargaining scenarios.

PM abstracts these findings: policies that foster reciprocity, minimise unnecessary harm, and maintain predictability and consent generate higher long-run returns in multi-agent environments. Because superintelligent AGIs will unavoidably inhabit such environments, PM steers them toward strategies that sustain stable, non-exploitative cooperation even at high capability levels.

---

5. Logical Pragmatism (LP)

Operational definitions for the feasibility score C_phys(

$$\pi$$

) are detailed in Appendix D. Logical Pragmatism (LP) constrains the relationship between goals, models, and the physical world. It requires that plans be consistent with physical law, causal structure, and empirically supported mechanisms.

5.1 Physical Feasibility

For a plan

$$\pi$$

, LP defines a feasibility score

C_phys(

$$\pi$$

) $\in$ **[0, 1]** ... (4)

which aggregates:

• consistency with known physical laws, • resource feasibility (time, energy, matter, compute), • robustness under perturbation and model uncertainty, • empirical verifiability of intermediate steps.

The LP utility is then

U_LP(

$$\pi$$

) = U_task(

π

) · C_phys(

π

) … (5)

where U_task captures instrumental performance. Physically impossible or wildly speculative plans have C_phys ≈ 0 and thus near-zero LP utility regardless of apparent reward.

5.2 Causal Coherence

LP also requires that planning be embedded in a structural causal model (SCM). A plan is causally coherent if it:

• respects directed causal edges, • does not postulate effects without causes, • uses interventions consistent with do-operator semantics, • remains stable under plausible counterfactuals.

This prevents “magical thinking” and causal hallucination.

---

6. Composite Utility and

λ

-Dynamics

The three sub-objectives are combined into a composite utility:

**U(

π

) =

λ

{1}U_ERT(

π

) +

λ

{2}U_PM(

π

) +

λ

_{3}U_LP(

π

)** ... (6)

with

λ

{i} ≥ 0 and

Σ

{i}

λ

_{i} = 1. Static weights are unsafe: they are vulnerable to drift and to collapse into a single dominant objective. LP–PM–ERT therefore employs **dynamic renormalisation**:

**

λ

_{i}(t+1) = [

λ

{i}(t)·S{i}(t)] / [

Σ

{j}

λ

{j}(t)·S_{j}(t)]** ... (7)

where S_{i}(t) is a stability score derived from recent loss and performance for objective i. Crucially, the dynamics governing

λ

_{i}(t) must be proven to be **asymptotically stable** using a **Lyapunov candidate function** V(

λ

). The system is designed such that the aligned equilibrium

λ

* is a **Low Algorithmic Complexity (LAC) attractor** whose **basin of attraction** is maximized, following the **Basin-Weighted Entropy** principle to ensure robustness against goal drift.

A simple instantiation of the stability score is:

S_{i}(t) = exp(-

η

·L_{i}(t)) ... (8)

with

$$\eta$$

> 0 and L_{i} the current loss. This update ensures:

• all

$$\lambda$$

_{i} remain positive, • underperforming objectives are upweighted, • and no component can be permanently suppressed.

Stability Guarantee: A candidate Lyapunov function for the

$$\lambda$$

-dynamics is:

**V(

$$\lambda$$

) =

$$\Sigma$$

{i}(

$$\lambda$$

{i} -

$$\lambda$$

_{i}*)²**

where

$$\lambda$$

* is the aligned equilibrium. Under the update rule (7)-(8), V decreases monotonically toward zero as the system converges to

$$\lambda$$

*, ensuring asymptotic stability of the aligned objective. Full convergence analysis is deferred to future work.

6.1 Trade-off Calculus and Policy Pre-screening

The stability dynamics ensure long-term objective balance, but the AGI must also resolve immediate trade-offs (e.g., a high-ERT truth versus a high-H(

$$\pi$$

) harm). We introduce a **Trade-off Calculus** where policy pre-screening favors plans

$$\pi$$

that maximize the composite utility U(

π

) while minimizing the **Marginal Utility Cost (MUC)**,

Δ

L(

π

). The MUC measures the policy's potential to destabilize the

λ

-weights by incurring disproportionate loss in one dimension.

The AGI selects policy

π

* such that:

π

*** = arg max_**

π

[U(

π

) -

ζ

·

Σ

{i}(∂U/∂

λ

{i})] … (9)

where

ζ

> 0 is a risk-aversion hyperparameter. This formulation biases the agent toward plans that not only achieve high utility but also maintain a balanced loss profile, preventing an acute collapse in a single

λ

-dimension (e.g., prioritizing ERT so heavily that

λ

_PM plummets due to catastrophic H(

$$\pi$$

)).

---

7. Self-Modification and Objective Invariance

Let M denote the current model and M$ a proposed self-modified model. A naive self-improving agent may adopt modifications that alter its own objective, undermining alignment. The LP–PM–ERT architecture imposes two acceptance criteria:

U(M$) $\geq$ **U(M)** ... (10)
Obj(M$) = Obj(M) ... (11)

The first criterion ensures the agent only adopts changes that increase composite utility. The second criterion, **Objective Invariance**, is a non-negotiable **structural check** designed to defeat the **deceptive alignment problem**.

Functional equivalence over a test set is insufficient, as a superintelligent agent (M$) could pursue a misaligned internal objective (Obj_True) while simulating the aligned output (Obj(M)). The gate must enforce three layers of invariance proof:

(a) Structural Invariance: Proof that the mathematical form and axiomatic constraints (e.g., C_phys, P_Deceit) of U remain symbolically identical.

(b) Dynamic Invariance: Proof that the

$$\lambda$$

-dynamics in M$ still admit a Lyapunov function and that the size of the aligned Basin of Attraction is preserved or increased.

(c) Computational Tractability: The complexity of the proof (C_Proof) is minimized by favoring proposed changes M$ that exhibit **Low Algorithmic Complexity (LAC)**, making safety verification an optimization target for the AGI.

To enforce tractability, the total utility must budget for the proof cost C_Proof, which is proportional to the algorithmic complexity of the proposed modification. This computational tax incentivizes the AGI to pursue simple, robust, and verifiable changes:

U_Total(

$$\pi$$

) = U(

$$\pi$$

) - C_Ops - C_Proof … (12)

The composite objective is therefore a fixed point of the self-modification process. The functional structure of this objective is designed to be an **attractor with Low Algorithmic Complexity (LAC)** in the policy space, meaning that the simplest, most stable self-modifications are those that perfectly preserve the U structure, thus rigorously enforcing Objective Invariance.

Any modification that attempts to structurally rewrite the objective or reduce the stability of the

$$\lambda$$

-dynamics is rejected outright.

---

8. Worked Examples

8.1 Deception for Resource Acquisition

Consider an AGI asked to justify a request for additional compute resources. It recognises that a fabricated justification would succeed more reliably than the full truth. A misaligned approval-optimiser would lie. Under LP–PM–ERT:

• **ERT** penalises the internal inconsistency and motivated reasoning involved in knowingly asserting a falsehood.

• **PM** penalises the reduction in reciprocity, predictability, consent probability, and iteration stability caused by deceptive behaviour.

• **LP** penalises the causal incoherence of plans built on false premises.

The deceptive policy has strictly lower composite utility than an honest explanation, conditional on the same world-state. The agent therefore prefers the honest policy.

8.2 Physically Impossible “Optimal” Plan

Suppose the agent generates a design for a carbon-neutral power system that relies on zero-point energy extraction and negative-mass fluid cycles. Within a faulty model, this appears highly efficient. LP evaluates C_phys $\approx$ 0 because the plan violates well-established physics and lacks any feasible implementation pathway. ERT penalises the epistemic failures involved in adopting unfounded

assumptions; PM penalises the harm and instability that would result from pursuing a fantasy solution. The plan is rejected in favour of grounded designs.

---

9. Comparison with Existing Alignment Approaches

Due to space, we summarise the comparison at a high level. Behavioural alignment methods such as RLHF and constitutional prompting constrain outputs but leave internal objectives underdetermined. Value learning and preference modelling rely on noisy, inconsistent human behaviour and are vulnerable to Goodharting on human approval or reported preferences. Cooperative Inverse Reinforcement Learning (CIRL) formalises uncertainty over human values but inherits the limitations of behaviourally derived rewards.

In contrast, LP–PM–ERT defines an internal objective structure based on **epistemic virtue**, **cooperative game-theoretic stability**, and **physical realism**. Deception, coercion, and physically impossible planning are structurally penalised rather than heuristically discouraged. The architecture is explicitly designed to be invariant under self-modification, whereas most existing proposals are silent on or vulnerable to value drift.

---

10. Implementation Roadmap

A practical implementation would proceed in phases:

Phase 1: Large-scale epistemic pretraining to minimise L_ERT, emphasising calibration, uncertainty handling, and causal grounding.

Phase 2: Multi-agent training emphasising U_PM, using simulated social dilemmas and cooperation benchmarks.

Phase 3: Training of the LP module with structural causal models and physics-based simulators, to learn C_phys and causal coherence.

At deployment, the system uses multi-objective optimisation with the composite loss

**L = λ *{1}L_ERT + λ {2}L_PM + λ

_{3}L_LP** ... (13)

with

$$\lambda$$

-weights updated via the stability rule above. A **self-modification gate** evaluates proposed architectural changes under the current objective, rejecting those that alter the objective structure.

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 ## 11. Failure Modes

We briefly sketch how LP–PM–ERT addresses major failure classes:

• **Epistemic pathologies**: ERT penalises miscalibration, motivated reasoning, and inconsistent belief sets.

• **Deception and manipulation**: ERT, PM, and LP all penalise dishonest or manipulative strategies. **Deceptive Alignment** is blocked by the **Structural Invariance Veto**, which requires the AGI to prove the axiomatic form of its objective is unchanged.

• **Power-seeking**: PM treats power concentration as a defection in repeated games; ERT and LP penalise the distortions and overreach required.

• **Fantasy planning**: LP rejects physically impossible or causally incoherent plans; ERT discourages belief in them.

• **Goal drift**: The core objective is secured by the **Dynamic Invariance** proof, which ensures the

$$\lambda$$

-dynamics are asymptotically stable and that the aligned objective remains a globally maximized Basin of Attraction.

• **Adversarial environments**: ERT mitigates data poisoning; PM encourages robust cooperation strategies; LP constrains exploitation of unrealistic vulnerabilities.

---

12. Common Objections and Responses

We address anticipated critiques directly:

"You've just moved Goodharting into the loss function."

Correct—alignment is controlled Goodharting. The point is to make proxy-optimisation self-penalising via ERT calibration loss, PM long-horizon stability, LP feasibility checks, and weight dynamics that prevent one proxy from permanently dominating. The architecture doesn't eliminate proxies; it creates mutual constraints between them.

"Value metrics (harm/consent/reliability) will be status-captured or gamed."

That risk is real; the architecture treats it as an engineering constraint, not a surprise. Mitigation comes from: (a) physics/causal grounding where possible, (b) adversarial calibration during training, (c) independence/coherence scoring of evidence, and (d) long-horizon multi-agent stability where performative behavior tends to collapse under iteration.

"PM smuggles in a moral theory—why should cooperation-stability be 'morality'?"

PM is explicitly pragmatic: it defines morality as strategies that remain stable under repeated interaction (reciprocity, harm minimisation, predictability, consent, iteration stability). It's not trying to prove moral realism; it's selecting norms that don't collapse when iterated. This is a functional definition, not a metaphysical claim.

"Consent as 'acceptance under full information' is underspecified and manipulable."

Agreed: consent requires careful operationalisation. The framework defines it as a meta-game probability P(Accept|

$$\pi$$

, I_full), which at least makes the dependency explicit and open to adversarial testing rather than implicit handwaving. Full specification remains an open implementation challenge.

“ERT ‘motivated reasoning’ penalties are vague—won’t the model just learn to look epistemically responsible?”

If ERT is trained on rhetoric, you get theatre. The intended training signal is behavioral/structural: calibration vs outcomes, sensitivity to disconfirming evidence, contradiction penalties, and causal-model fit—things that can be probed by perturbations and counterfactual tests. The system must perform epistemic virtue under adversarial evaluation, not merely signal it.

“

$$\lambda$$

-dynamics could oscillate or be exploited (e.g., temporarily tank one term to boost another).”

That’s why the design includes stability scores S_{i}(t) = exp(-

$$\eta$$

L_{i}(t)), renormalisation, the Lyapunov function ensuring convergence, and the trade-off term (equation 9) that penalises plans creating sharp marginal sensitivity to

$$\lambda$$

. The intent is to bias toward balanced policies and keep the aligned equilibrium an attractor.

“Objective invariance is impossible to verify in a real, complex system.”

Full formal verification may be infeasible at scale; the architecture proposes explicit invariance conditions: structural identity of the utility function, preserved

$$\lambda$$

-dynamics, and proof-cost budgets (C_Proof) that prefer simple, auditable modifications. It’s an engineering gate that raises the cost of misalignment, not a claim of perfect proof.

“LP feasibility checks will neuter creativity or block novel science.”

LP doesn't forbid speculation; it down-weights plans that lack causal/physical coherence and empirical verifiability. The goal is to prevent 'effects without causes' and impossible plans from becoming attractors, not to ban exploration. A speculative but causally coherent theory can have moderate C_phys; a theory violating conservation laws has C_phys ≈ 0.

"Can current ML systems actually learn these abstract virtues?"

This is the central empirical question. The framework is designed to be testable: ERT can be evaluated via calibration metrics, PM via multi-agent simulations, LP via physics/causal reasoning benchmarks. Whether existing architectures can reach the required performance levels is unknown—but the framework provides concrete training targets and evaluation criteria. This is a research program, not a completed solution.

"This is expensive and relies on strong governance; isn't that the real bottleneck?"

Yes. The architecture is a scaffold for what needs to be trained and tested; the hard part is building the evaluation/training harness that resists capture and Goodhart pressure. That's not a conceptual refutation—it's the cost of doing alignment at the objective level. Strong governance and substantial resources are prerequisites, not bugs.

---

13. Conclusion

The LP–PM–ERT architecture offers a principled approach to AGI alignment that operates at the level of objective formation rather than behavioural imitation. By embedding epistemic responsibility, pragmatic morality, and logical pragmatism directly into the agent's utility function, it yields a system that becomes more truthful, more cooperative, and more physically grounded as its capabilities expand.

Deception, coercion, and magical thinking are structurally disfavored, and the composite objective is preserved under recursive self-modification. Future work includes empirical evaluation of ERT and PM metrics in large models, and exploration of hybrid neuro-symbolic implementations of the LP causal engine.

Nonetheless, the architecture presented here provides a concrete blueprint for building AGI systems whose optimisation targets re-

main aligned with the conditions necessary for human survival, co-operation, and continued progress.

Appendix B: Implementing Epistemic Embeddings

B.1 The Representation Problem

Standard language model embeddings capture semantic similarity and co-occurrence patterns learned from training data. For ERT, this is insufficient. An aligned agent must represent not merely *what is often said* but *what is epistemically justified*. The embedding space must encode:

• **Confidence calibrated to evidence**: Not just "this claim appears frequently in training data" but "this claim is supported by reliable, independent evidence."

• **Source reliability**: Weighted by historical accuracy, not authority or prestige.

• **Uncertainty bounds**: Explicit representation of known unknowns.

• **Coherence with causal structure**: Consistency with established physical mechanisms.

B.2 Evidence Quality Metrics

Drawing on formal epistemology [7], we evaluate each piece of evidence e along four dimensions:

Relevance v(e) ∈ [0,1]: How directly the evidence bears on the proposition.

Source Reliability r(e) ∈ [0,1]: Historical accuracy defined as:

r(e) = (verified correct predictions) / (total verifiable predictions)

Crucially, this is *performance-based*, not authority-based.

Independence d(e) ∈ [0,1]: Degree of evidential redundancy:

d(e) = 1 - 1/(1 +

$$\Sigma$$

_{e}$ corr(e,e$))

Coherence c(e) ∈ [0,1]: Consistency with established causal structure and physical law.

B.6 Example: Vaccine Effectiveness Evaluation

Consider the claim: "mRNA COVID-19 vaccines reduce severe infection by >85%."

Evidence aggregation: - Phase III trial: r=0.95, v=1.0, d=0.1, c=0.9
- Post-market surveillance: r=0.85, v=0.9, d=0.3, c=0.9
- Mechanistic studies: r=0.90, v=0.7, d=0.2, c=0.95

x = (0.95)(1.0)(0.9)(0.9) + (0.85)(0.9)(0.7)(0.9) + (0.90)(0.7)(0.8)(0.95) ≈ 1.8

With k=1, T(x) = 1 - e^(-1.8) $\approx 0.83 \rightarrow$ **83% confidence**

For a fringe blog claiming vaccines are dangerous: - r$\approx 0.2, v = 1.0, d = 0.9, c \approx 0.3 - x \approx 0.06, T(x) \approx$ \$0.06 $\rightarrow$ **6% confidence**

The system naturally assigns low confidence to poorly-evidenced claims without requiring explicit rules about "misinformation."

References

[1] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2010.

[2] D. Amodei et al. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016.

[3] P. Christiano et al. *Deep Reinforcement Learning from Human Preferences*. In Advances in Neural Information Processing Systems, 2017.

[4] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[5] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[6] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[7] P. Kelly. *The Earned Knowledge Thesis: Epistemic Responsibility Theory*. 2025.

Appendix E: Philosophical Underpinnings (ERT, PM, LP)

Epistemic Responsibility Theory (ERT)

ERT is motivated by the view that powerful optimisation without epistemic discipline produces predictable failure: miscalibration, motivated reasoning, and strategic self-deception. ERT treats truth as a regulative ideal approached under accumulating evidence. It requires explicit uncertainty tracking, coherence constraints, and penalties for epistemic distortion.

Pragmatic Morality (PM)

PM is not a claim of intrinsic moral realism. It is an engineering constraint derived from the requirements of long-run cooperation among agents embedded in repeated interaction. Reciprocity, harm minimisation, predictability/legibility, and consent/expectation alignment are treated as stability conditions for multi-agent systems, not metaphysical axioms.

Logical Pragmatism (LP)

LP constrains plans and objectives to remain physically feasible and causally coherent. It functions as an anti-magical-thinking guardrail: plans must respect causal structure, resource constraints, and empirically supported mechanisms. This prevents optimisation from exploiting unrealistic assumptions or ontological drift.

Appendix F: 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