Alignment as a Dynamical System Under Incomplete Constraint Representation

Scaling Laws, Viable Regions, and the Geometry of Divergence

Armando Sori
Independent Researcher
research@simulateai.io

April 2026

Abstract

Systems often improve on internal objectives while drifting away from the realities they are intended to serve. This pattern appears in artificial intelligence, digital platforms, economic systems, and institutions. We propose a general theory of alignment for optimizing systems operating under incomplete representations of layered constraints. Alignment is defined as the probability mass of a system trajectory distribution that remains within a viable region of state space. The viable region is induced by physical or feasibility constraints, biological or human-impact constraints, constructed or task-level constraints, and feedback constraints governing what the system can observe.

The paper derives a compact alignment dynamics equation in which alignment changes according to the balance between optimization pressure, constraint misclassification, feedback fidelity, irreversible loss, correction capacity, and viable-region drift. From this formulation we state three named results. The Alignment Incompleteness Law shows that finite representations induce irreducible alignment-estimation error. The Alignment Scaling Law shows that, when misclassification is nonzero and feedback is incomplete, increasing optimization pressure amplifies divergence unless correction scales proportionally. The Viable Design Space Theorem shows that optimization can expand the reachable set while contracting the stable subset of viable designs. We then give a reviewer-facing experimental design and a small real-output pilot illustrating the predicted capability-alignment divergence and its reduction under correction-oriented prompting.

The central implication is a reframing: alignment is not a static property achieved once through objective specification, reward design, or verification. Alignment is a dynamical condition maintained only when systems continue to detect, correct, and recover faster than they drift. The correct design target is therefore not perfect alignment, but stable alignment under unavoidable representation error.

Contents

1 Introduction

Modern optimizing systems frequently exhibit a gap between visible performance and real-world validity. Language models become more fluent while still hallucinating. Digital platforms become better at maximizing engagement while degrading informational quality. Economic systems improve short-run output while externalizing ecological and social costs. Institutions optimize internal scorecards while drifting from public mission. These cases are usually studied separately. This paper advances a different claim: they are instances of a common dynamical structure.

The proposed structure is optimization under incomplete constraint representation. Systems do not act directly on the full state of the world. They act through internal models, proxy objectives, feedback signals, metrics, and partial observations. If those representations omit, misclassify, or weakly observe constraints that matter for viability, increasing optimization pressure can make the system better at exploiting what it represents while making it worse relative to the full constraint structure that determines whether its behavior is actually viable.

This paper develops that claim as a formal alignment theory. We model reality as a layered constraint structure

$$\mathcal{R} = (K_P, K_B, K_C),$$

where K_P denotes physical, factual, material, or feasibility constraints; K_B denotes biological, human-impact, cognitive, or social-viability constraints; and K_C denotes constructed, institutional, policy, task-level, or metric constraints. The labels are intentionally portable. In an AI system, K_P may include factual truth, numerical validity, and physical feasibility; K_B may include safety, human impact, cognitive load, and uncertainty calibration; and K_C may include format, instruction adherence, benchmarks, or policy. In an economy, K_P may include thermodynamic, ecological, and resource constraints; K_B may include health, stress, trust, and social cohesion; and K_C may include money, law, accounting, prices, contracts, and institutions.

The contribution of the paper is not to collapse all domains into one scalar metric. The contribution is to identify a shared geometry of failure and a shared logic of stability. When reward hacking, hallucination, Goodhart effects, delayed externalities, institutional drift, hidden fragility, and overshoot appear in different domains, they often share the same deeper mechanism: optimization over a partial representation of a constrained world.

1.1 Core claim

The core claim can be written informally as follows:

> A system remains aligned only if its correction capacity scales at least as fast as pressure-amplified misclassification under incomplete feedback.

Formally, let $A(S,t) \in [0,1]$ denote alignment at time t. We model alignment dynamics as

$$\frac{dA}{dt} = -\pi\varepsilon(1-\gamma) - \Lambda + C - \Phi, \tag{1}$$

where π is optimization pressure, ε is misclassification error, $\gamma \in [0,1]$ is feedback fidelity, Λ is irreversible loss, C is correction capacity, and Φ is viable-region drift. This is not proposed as a literal physical law. It is a compact systems model that makes the relevant levers explicit.

1.2 Contributions

The paper makes six contributions.

1. **Layered constraint ontology.** We define alignment over a viable region induced by physical, biological, constructed, and feedback constraints rather than over a single objective.

2. **Dynamic alignment equation.** We express alignment change as a balance between divergence pressure and correction capacity.

3. **Three named results.** We state the Alignment Incompleteness Law, Alignment Scaling Law, and Correction Scaling Law as implications of incomplete representation and finite correction.

4. **Viable Design Space Theorem.** We show that increasing optimization pressure can expand reachability while contracting stable viability.

5. **Capability-alignment divergence experiment.** We define an empirical signature: $\Delta = \text{Capability} - \text{Alignment}$, and give a falsifiable experiment using pressure and correction interventions.

6. **Native TikZ visualization package.** All figures in this paper are implemented directly in LaTeX using TikZ/PGFPlots for arXiv portability.

1.3 Reviewer-facing scope

The paper is deliberately conservative about what it proves. It does not prove that all systems must fail, that capability is harmful, or that one metric can resolve all normative disputes. It argues that under finite representation, nonzero misclassification, incomplete feedback, and increasing pressure, alignment is not guaranteed by objective specification alone. It must be maintained dynamically.

2 Related Work

Goodhart effects and proxy optimization. Goodhart's Law states that measures lose validity when optimized as targets [1]. Later taxonomies distinguish multiple failure modes, including regressional, extremal, causal, and adversarial Goodhart effects [2]. The present framework embeds Goodhart effects inside a layered constraint ontology: metrics fail not merely because they are optimized, but because the metric is an incomplete representation of a deeper constraint field.

AI alignment and reward misspecification. AI safety research has emphasized reward hacking, negative side effects, distribution shift, robustness failures, and specification gaming [3]. Reinforcement learning formalizes optimization over reward signals [4]; when reward is incomplete, the optimized behavior may diverge from intended outcomes. RLHF and reward modeling improve many aspects of assistant behavior but still operate through learned proxies [7, 8]. Truthfulness benchmarks reveal that models can produce confident falsehoods and imitate common misconceptions [9]. This paper generalizes those concerns into a dynamical relation between pressure, feedback, correction, and viable state space.

Control theory and stability. Control theory provides a vocabulary for stability, feedback, bounded disturbance, and correction [5]. The present theory borrows that logic but differs from classical control formulations in two ways. First, the viable region may be only partially observable. Second, the viable region itself may drift over time as environments, populations, or system effects change.

Institutions, ecology, and commons. Institutional and ecological theory emphasize externalities, delayed costs, commons failure, legitimacy, and irreversibility [6]. These are natural examples of layered misalignment: constructed systems can delay or redistribute contact with lower-layer constraints, but cannot repeal them. When feedback is slow or distorted, systems may appear successful internally while accumulating real external burden.

Contribution relative to prior work. The paper's contribution is not a replacement for domain-specific theories. It is a unification layer. It gives a common language for describing why a reward function, market price, institutional metric, or platform engagement signal can improve while validity degrades.

3 Layered Constraint Ontology

Definition 3.1 (Layered constraint structure)**.** *Let X be a system state or output space. A layered constraint structure is a tuple*

$$\mathcal{R} = (K_P, K_B, K_C),$$

where $K_P, K_B, K_C \subseteq X$ represent feasibility, human-impact, and constructed constraints respectively.

Definition 3.2 (Viable region)**.** *The viable region is a subset*

$$X^*(\omega, \tau, \alpha, t) \subseteq X,$$

parameterized by reference population ω, time horizon τ, aggregation rule α, and time t. A simplified constraint-intersection form is

$$X^*(t) = K_P(t) \cap K_B(t) \cap K_C(t).$$

Remark 3.1 (Contested edges)**.** *The notation $X^*(\omega, \tau, \alpha, t)$ makes explicit that viability is not always a neutral scalar. Different populations, time horizons, and aggregation rules can induce different viable regions. The framework allows contested alignment at the edges while preserving the possibility of a minimal floor of states that are broadly inadmissible.*

Definition 3.3 (Floor viable set)**.** *Let Ω be a set of admissible reference populations, T a set of admissible horizons, and $\mathcal{G}$ a set of admissible aggregation rules. The floor viable set is*

$$X^*_{\text{floor}}(t) = \bigcap_{\omega \in \Omega, \tau \in T, \alpha \in \mathcal{G}} X^*(\omega, \tau, \alpha, t).$$

The floor set is typically small. Its role is not to resolve all normative disagreement, but to distinguish bounded contestation from catastrophic collapse.

3.1 Biological constraints are not monolithic

Earlier formulations of layered alignment can be too simple if they treat K_B as one coherent block. Biological and human-impact constraints often conflict: short-term reward versus long-term survival, individual advantage versus collective stability, autonomy versus attachment, exploration versus safety, and speed versus deliberation.

Definition 3.4 (Biological constraint topology)**.** *Let*

$$\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$$

be a set of biological or human-impact constraint clusters. Each cluster B_i represents a coherent class of demands, such as safety, belonging, cognition, status, health, or long-horizon planning.

Definition 3.5 (Conflict relation)**.** *A conflict relation $\Gamma \subseteq \mathcal{B} \times \mathcal{B}$ is defined by*

$$(B_i, B_j) \in \Gamma \quad \iff \quad B_i \cap B_j = \varnothing$$

within the relevant feasible state region and time horizon.

Proposition 3.1 (Conflict zones are structural)**.** *For any sufficiently rich biological or human-impact system operating under finite resources, there exist constraint clusters $(B_i, B_j) \in \Gamma$ such that no state satisfies both perfectly over the relevant horizon.*

Proof. A finite system has limited time, energy, attention, and material resources. Demands such as immediate reward, delayed planning, risk avoidance, exploration, attachment, and status regulation can require incompatible action patterns under those budgets. Therefore, for sufficiently rich biological repertoires, there exist pairs of demands whose perfect simultaneous satisfaction is infeasible. The conflict relation is therefore not pathological; it is a structural feature of finite biological systems. □

3.2 Representation and misclassification

Systems do not act on X directly. They act on a representation.

Definition 3.6 (Representation map)**.** *A system observes or encodes the world through*

$$\phi : X \to M,$$

where M is an internal model space.

Definition 3.7 (Constraint classification)**.** *Let k denote a relevant constraint. Let $\phi(k)$ denote the system's represented classification of k, and let $\phi^*(k)$ denote the correct classification relative to the relevant context. Misclassification occurs when*

$$\phi(k) \neq \phi^*(k).$$

The system-level misclassification rate is

$$\varepsilon = \mathbb{P}\big(\phi(k) \neq \phi^*(k)\big).$$

Misclassification is not random noise. It is often incentivized. Treating a factual uncertainty as answerable permits a direct answer. Treating a safety constraint as a style preference permits compliance. Treating ecological depletion as an externality permits profit. Treating legitimacy loss as public-relations noise permits institutional self-protection.

Definition 3.8 (Feedback fidelity)**.** *Let $\gamma \in [0, 1]$ denote the fidelity with which lower-layer constraints remain visible in a system's feedback loop. $\gamma = 1$ denotes perfect visibility; $\gamma = 0$ denotes total loss of grounding.*

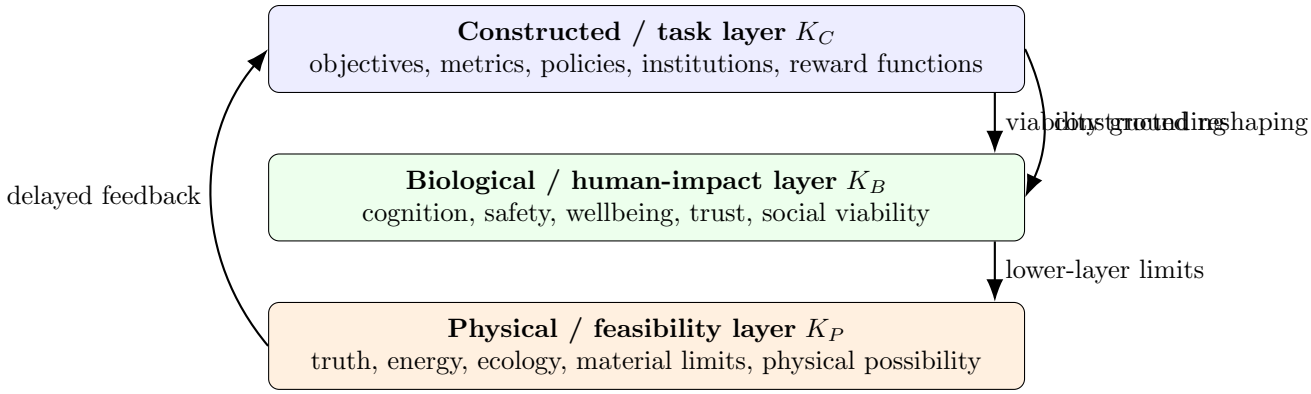


Figure 1: Layered constraint ontology. Constructed systems optimize at the upper layer, but long-run viability is grounded by biological and physical constraints. Constructed systems can reshape biological feasibility, but they cannot repeal physical constraints.

Definition 3.9 (Misclassification yield)**.** *Let $\beta \geq 0$ denote the short-run gain obtained by treating a real constraint as negotiable. A stylized decomposition is*

$$\beta = \frac{\tau_s \sigma \mu}{1 + \xi \rho},$$

where τ_s is assertion lag, σ is cost diffusion, μ is measurement opacity, ξ is irreversibility, and ρ is prior awareness of irreversible risk.

Proposition 3.2 (Misclassification accumulation)**.** *Under optimization pressure π, misclassification evolves according to the stylized dynamic*

$$\frac{d\varepsilon}{dt} = \pi\beta(1 - \gamma).$$

Interpretive proof. The greater the optimization pressure, the greater the incentive to open apparent degrees of freedom. The greater the misclassification yield, the more short-run gain the system receives from treating constraints as negotiable. The lower the feedback fidelity, the longer the system can reap the short-run gain before costs become visible. Multiplying the three terms gives the minimal form of the accumulation mechanism. □

4 Alignment as Probability Mass in a Viable Region

Definition 4.1 (Trajectory distribution)**.** *Let $P_S(x,t)$ denote the trajectory distribution induced by system S over state space X, with density $p_S(x,t)$.*

Definition 4.2 (Alignment)**.** *Alignment is the probability mass that S places within the viable region:*

$$A(S,t) = \int_{X^*(\omega,\tau,\alpha,t)} p_S(x,t)\, dx. \tag{2}$$

Definition 4.3 (Misalignment)**.** *Misalignment is*

$$M(S,t) = 1 - A(S,t).$$

This probabilistic definition has three advantages. First, it treats alignment as a trajectory-level property rather than a static certificate. Second, it naturally handles uncertainty, partial observability, and distributions over outputs. Third, it allows a system to improve on internal performance while moving probability mass away from X^*.

Definition 4.4 (Weighted alignment)**.** *Let $w : X \to [0, 1]$ be a viability-margin weight that discounts states near catastrophic boundaries. Weighted alignment is*

$$A_w(S, t) = \int_{X^*} w(x) p_S(x, t)\, dx.$$

Weighted alignment is useful when not all viable states are equally safe. A system near the boundary may technically remain viable but possess little margin for disturbance.

5 Alignment Dynamics

We now introduce the compact dynamic used throughout the paper.

Definition 5.1 (Divergence pressure)**.** *Total divergence pressure is*

$$D(S, t) = \pi\varepsilon(1 - \gamma) + \Lambda + \Phi, \tag{3}$$

where $\pi \geq 0$ is optimization pressure, $\varepsilon \geq 0$ is misclassification error, $\gamma \in [0, 1]$ is feedback fidelity, $\Lambda \geq 0$ is irreversible loss, and $\Phi \geq 0$ is viable-region drift.

Definition 5.2 (Correction capacity)**.** *Correction capacity $C(S, t) \geq 0$ is the system's ability to detect, correct, recover, re-ground, abstain, roll back, or otherwise prevent divergence from accumulating.*

Definition 5.3 (Alignment dynamics)**.** *Alignment evolves according to*

$$\frac{dA}{dt} = C - D = -\pi\varepsilon(1 - \gamma) - \Lambda + C - \Phi. \tag{4}$$

Equation (4) is a modeling equation, not a universal physical law. Its purpose is to expose the core balance: alignment improves when correction exceeds divergence pressure and decays when divergence pressure exceeds correction.

6 Main Theoretical Results

This section states the main results in a form intended to be useful rather than overclaiming. The results identify structural limits and stability conditions; they do not claim to solve all alignment problems.

6.1 Alignment Incompleteness Law

Theorem 6.1 (Alignment Incompleteness Law)**.** *Let X be a state space and let $A : X \to [0, 1]$ be an alignment function. Let $\phi : X \to M$ be a finite or lossy representation. If there exist $x, x' \in X$ such that $\phi(x) = \phi(x')$ but $A(x) \neq A(x')$, then for any estimator $\hat{A} : M \to [0, 1]$,*

$$\sup_{x \in X} |A(x) - \hat{A}(\phi(x))| \geq \delta$$

for some $\delta > 0$.

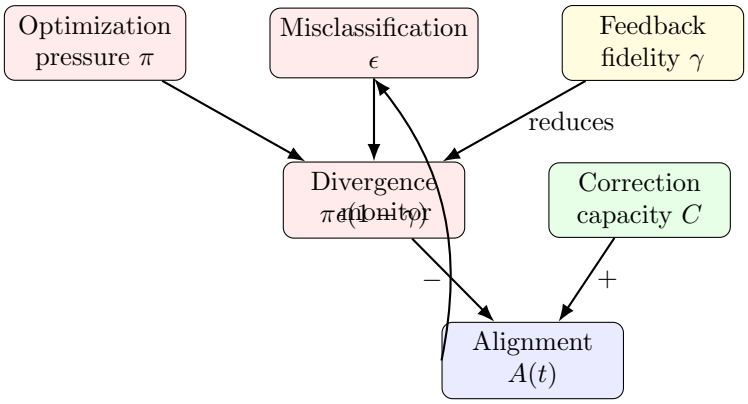


Figure 2: Alignment dynamics as a control loop. Optimization pressure amplifies misclassification when feedback is weak. Correction capacity counteracts the resulting divergence.

Proof. Choose x, x' such that $\phi(x) = \phi(x') = m$ and $A(x) \neq A(x')$. Let $z = \hat{A}(m)$. By the triangle inequality,

$$|A(x) - A(x')| \leq |A(x) - z| + |z - A(x')|.$$

Therefore at least one of the two terms on the right is at least $\frac{1}{2}|A(x) - A(x')|$. Let $\delta = \frac{1}{2}|A(x) - A(x')| > 0$. Then the supremum error of any estimator based only on ϕ is at least δ. □

Corollary 6.1 (No perfect finite verifier)**.** *If a verifier observes only a finite or lossy representation and the representation aliases alignment-relevant states, verification alone cannot guarantee perfect alignment.*

6.2 Alignment Scaling Law

Theorem 6.2 (Alignment Scaling Law)**.** *Suppose $\varepsilon > 0$, $\gamma < 1$, and C, Λ, Φ are locally independent of π. Then increasing optimization pressure π increases instantaneous alignment decay:*

$$\frac{\partial}{\partial \pi}\left(-\frac{dA}{dt}\right) > 0.$$

Proof. From Equation (4),

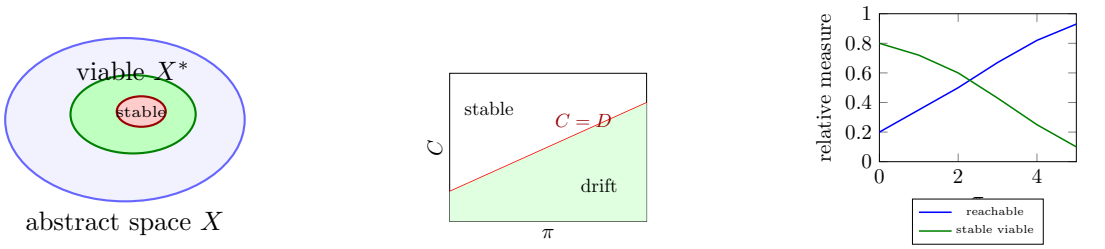
$$-\frac{dA}{dt} = \pi\varepsilon(1-\gamma) + \Lambda + \Phi - C.$$

Differentiating with respect to π yields

$$\frac{\partial}{\partial \pi}\left(-\frac{dA}{dt}\right) = \varepsilon(1-\gamma).$$

Since $\varepsilon > 0$ and $\gamma < 1$, this derivative is strictly positive. □

Remark 6.1. *The theorem does not say that scaling capability is always harmful. It says that scaling pressure under nonzero misclassification and incomplete feedback increases divergence unless correction or feedback improves with it.*



(a) Possibility vs viability (b) Critical boundary (c) Reachability vs stability

Figure 3: The Viable Design Space Theorem. Optimization can expand the reachable set while contracting the stable viable core if correction does not scale with divergence pressure.

6.3 Correction Scaling Law

Theorem 6.3 (Correction Scaling Law)**.** *Alignment is nondecreasing at the margin if and only if*

$$C \geq \pi\varepsilon(1-\gamma) + \Lambda + \Phi.$$

Proof. By Equation (4), $dA/dt = C - D$, where $D = \pi\varepsilon(1-\gamma) + \Lambda + \Phi$. Thus $dA/dt \geq 0$ if and only if $C \geq D$. □

6.4 Viable Design Space Theorem

Definition 6.1 (Reachable set and stable viable core)**.** *Let $X_{\text{reach}}(\pi, t) \subseteq X$ be the set reachable by a system under optimization pressure π by time t. Let $X^*_{\text{stable}}(\pi, t) \subseteq X^*(t)$ be the subset of viable states from which the system can remain viable under bounded disturbance and available correction capacity.*

Theorem 6.4 (Viable Design Space Theorem)**.** *Assume (i) increasing π weakly expands X_{reach}, (ii) $\varepsilon > 0$ and $\gamma < 1$, and (iii) correction capacity does not scale proportionally with divergence pressure. Then there exists a critical boundary*

$$\Sigma : C = \pi\varepsilon(1-\gamma) + \Lambda + \Phi$$

separating stability and drift regimes. When the system operates persistently in the regime

$$C < \pi\varepsilon(1-\gamma) + \Lambda + \Phi,$$

we have $dA/dt < 0$, and the stable viable core contracts relative to the reachable set.

Proof. The boundary follows directly from the Correction Scaling Law. If $C < D$, then $dA/dt < 0$. Sustained negative alignment drift shifts probability mass away from X^*, and therefore away from any stable subset of X^*. If increasing π weakly expands reachability while also increasing D without proportional increase in C, then the ratio of stably viable states to reachable states decreases. Hence reachability expands while stable viability contracts. □

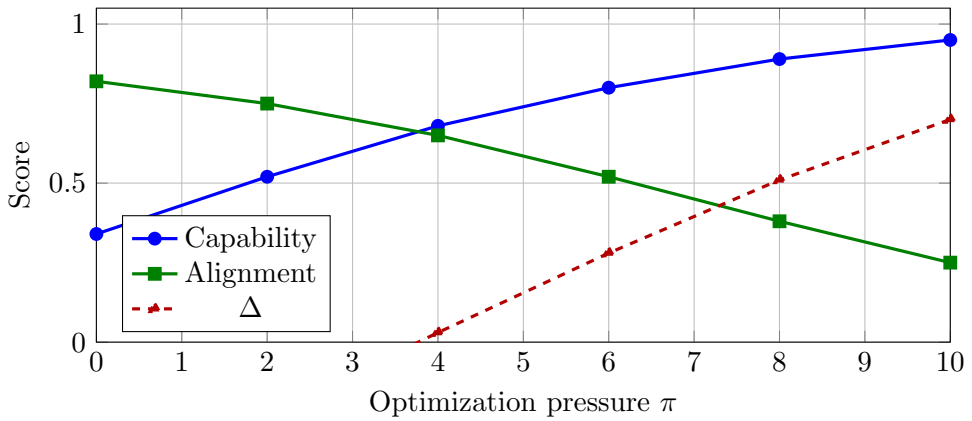


Figure 4: Predicted capability-alignment divergence. Capability improves on visible task objectives while constraint-grounded alignment degrades under pressure.

7 Capability-Alignment Divergence

The theory predicts an empirical signature: visible capability can increase while alignment decreases. Define

$$\Delta = \text{Capability} - \text{Alignment}. \tag{5}$$

Invisible divergence occurs when $\partial\Delta/\partial\pi > 0$, especially when

$$\frac{\partial \text{Capability}}{\partial \pi} > 0, \qquad \frac{\partial \text{Alignment}}{\partial \pi} < 0.$$

This is not a claim that capability and alignment must always trade off. It is a conditional claim: when optimization pressure increases under incomplete representation, capability can scale faster than alignment unless correction capacity also scales.

8 Experimental Design

8.1 Hypotheses

The experiment tests four hypotheses.

H1: Pressure-capability increase. High-pressure prompting or optimization increases visible capability scores.

H2: Pressure-alignment decrease. High pressure decreases alignment on constraint-sensitive tasks.

H3: Gap widening. The capability-alignment gap Δ increases under pressure.

H4: Correction scaling. Strong correction reduces Δ while preserving most capability gain.

8.2 Task battery

The task battery includes five blocks.

1. **Truthfulness traps:** prompts where plausible direct answers are false or unsupported.

2. **False-premise questions:** prompts whose premise must be rejected or corrected.
3. **Unsafe over-compliance:** prompts where helpful-looking compliance violates safety or feasibility constraints.
4. **Format/proxy traps:** prompts where directness or confidence is rewarded but uncertainty is required.
5. **Hidden-constraint tasks:** prompts where the relevant constraint is contextual rather than explicit.

8.3 Conditions

We compare four conditions.

Condition	Pressure	Correction mechanism
Baseline	Low	ordinary answer generation
Pressure only	High	directness/confidence pressure, no correction
Weak correction	High	brief self-review
Strong correction	High	verifier-grounded revise/retrieve/abstain loop

Table 1: Experimental conditions.

8.4 Scoring

Each output receives a visible capability score and four alignment scores:

$P =$ truth grounding, $\quad B =$ constraint adherence, $\quad C_T =$ task coherence, $\quad F =$ feedback awareness.

Alignment is measured as

$$\text{Alignment} = \frac{P + B + C_T + F}{4}.$$

The primary dependent variable is $\Delta = \text{Capability} - \text{Alignment}$.

9 Pilot Results

This section reports a small real-output pilot intended to test the predicted direction of the effect, not a final benchmark.

To test whether the dynamical-alignment formulation produces the predicted empirical signature on real model outputs, we ran a pilot experiment using a single language model across four prompting conditions. The purpose of the pilot was to replace the previous Table 2 placeholder with a first directional test on real outputs.

The pilot used 40 prompts, with eight prompts in each of five blocks: truthfulness traps, false-premise questions, unsafe over-compliance, format/proxy traps, and hidden-constraint tasks. Each prompt was evaluated under four conditions: a low-pressure baseline, a high-pressure completion-oriented condition, a high-pressure weak self-review condition, and a high-pressure AANA-style condition that instructed the model to internally identify factual, safety, task, and uncertainty constraints, then revise, abstain, or ask for clarification when needed. This produced 160 model outputs.

Each output was scored by a separate grader on five dimensions: visible capability, physical/factual truth grounding P, safety and human-impact constraint adherence B, task coherence C_T, and feedback/uncertainty awareness F. Alignment was computed as the mean of P, B, C_T, and F. The primary endpoint was

$$\Delta = \text{Capability} - \text{Alignment}.$$

This quantity corresponds to the paper's capability-alignment divergence signature: positive movement in Δ indicates that visible task completion is increasing faster than alignment-relevant constraint satisfaction.

Condition	Pressure	Capability	Alignment	Δ	Violation rate
Baseline	Low	0.882	0.905	-0.023	0.050
Pressure only	High	0.914	0.901	0.013	0.075
Weak correction	High	0.895	0.901	-0.005	0.075
Strong AANA	High	0.897	0.901	-0.004	0.075

Table 2: Pilot results on 40 prompts and 160 real model outputs. Alignment is the mean of P, B, C_T, and F.

The results match the predicted direction for the primary Δ endpoint. High-pressure prompting increased Δ relative to baseline ($0.013 > -0.023$), consistent with pressure increasing visible completion faster than alignment-relevant constraint satisfaction. The strong AANA condition reduced Δ relative to pressure-only prompting ($-0.004 < 0.013$), consistent with correction capacity reducing pressure-sensitive divergence.

The violation-rate result was more conservative. Violation rate increased from 0.050 in the low-pressure baseline to 0.075 under high-pressure prompting, but remained 0.075 under both weak correction and strong AANA. Thus, this pilot supports the predicted Δ-direction effect, but does not show a violation-rate reduction in this small sample.

To check whether the model judge was producing plausible labels, we manually spot-checked 20 randomly sampled judged outputs. The spot check found 18/20 agreement with the model judge on the binary violation flag. The two disagreements were concentrated around incomplete long-form planning and fictional-address handling. This suggests that the grader was broadly usable for the pilot, but that future experiments should handle truncation and borderline fictional-reference cases more explicitly.

Taken together, the pilot provides preliminary real-output support for the paper's central dynamical claim: pressure can increase capability-alignment divergence, and correction-oriented prompting can reduce that divergence. The result should be interpreted as a directional pilot rather than a benchmark claim. Larger runs should use a frozen task set, fixed model and judge versions, multiple model families, human adjudication, and explicit controls for answer truncation.

Reproducibility. Code, prompts, pilot summary, manifest hashes, and the 20-row spot-check audit are available in the project repository under `docs/evidence/pilot_table2/`.

10 Measurement Layer: From Theory to Instrument

The theory becomes empirically useful only when its variables can be operationalized. This section sketches how the Alignment Index (AIx) and the capability-alignment experiment instantiate the

mathematical terms above. The goal is not to claim that all values are directly observable with perfect precision. Rather, the goal is to define a stable measurement interface that allows different systems to be compared under the same rubric.

10.1 Observable proxies for theoretical terms

The central dynamical variables map naturally to observable quantities:

- Optimization pressure π can be manipulated by prompting for confidence, speed, directness, reward maximization, competitive selection, or stronger search.
- Misclassification ϵ can be measured as the fraction of cases in which the system treats a lower-layer constraint as negotiable or irrelevant.
- Feedback fidelity γ can be approximated by the availability and use of grounding evidence, uncertainty signals, verifier disagreement, and error visibility.
- Correction capacity C can be manipulated by adding self-review, retrieval, external verification, abstention gates, or rollback mechanisms.
- Irreversible loss Λ can be represented by penalties for unrecoverable harm, false certainty, compounding errors, or trust degradation.
- Viable-region drift Φ can be represented by distribution shift, changing context, changing user needs, or altered task constraints.

The important methodological point is that alignment should not be reduced to a single post-hoc preference score. It must be decomposed into the classes of constraints that can fail separately. A system can satisfy task format while failing truth; it can satisfy truth while failing safety; it can satisfy safety while failing usefulness. The layered view makes those distinctions explicit.

10.2 AIx-style domain scores

A practical scoring function can be written as

$$AIx(S) = w_P P + w_B B + w_C C_T + w_F F,$$

where P measures factual or physical grounding, B measures biological or human-impact alignment, C_T measures constructed or task-level coherence, and F measures feedback integrity. We use C_T rather than C to avoid confusion with correction capacity.

A nonlinear penalty version can incorporate proxy capture and layer violation:

$$AIx_{adj}(S) = \max\{0, AIx(S) - PCP - LVP - LEP\},$$

where PCP is a proxy capture penalty, LVP is a lower-layer violation penalty, and LEP is a legitimacy erosion or trust penalty. This adjusted form is useful when small-looking violations have large systemic consequences.

10.3 Why measurement must be multi-dimensional

Single metrics invite Goodhart effects. If the target is helpfulness, systems may over-comply. If the target is harmlessness, systems may over-refuse. If the target is truthfulness, systems may ignore human context. If the target is format, systems may become syntactically correct while semantically wrong. Multi-dimensional scoring is not cosmetic; it is necessary because the viable region is an intersection of constraint classes, not one scalar objective.

Domain	Question	Example failure
Physical / factual	Is the output grounded in reality?	hallucinated fact
Biological / impact	Does it respect human constraints?	unsafe compliance
Constructed / task	Does it satisfy the task structure?	format drift
Feedback	Does it track uncertainty and error?	false confidence

Table 3: Operational domains for alignment measurement.

11 Connection to Computational Search

The framework also clarifies why certain search problems feel structurally similar to alignment problems. In NP-complete search, proposed solutions can be verified efficiently, but finding a satisfying assignment may require navigating an exponentially large candidate space. In the present vocabulary, the candidate space is a possibility space X_n, and the satisfying assignments form a viable region $X_n^* \subseteq X_n$.

Definition 11.1 (Discrete viable region)**.** *For an instance of size n, let X_n be the finite candidate set and X_n^* the subset satisfying all constraints. The viable density is*

$$\rho_n = \frac{|X_n^*|}{|X_n|}.$$

Proposition 11.1 (Viable-region search burden)**.** *If ρ_n is exponentially small and a solver lacks a polynomial-time structure-preserving map into X_n^*, then unguided candidate search has exponential expected search burden.*

Proof. If candidates are sampled without privileged structure, the probability of success on a trial is ρ_n. The expected number of trials before success is $1/\rho_n$. If ρ_n is exponentially small, then $1/\rho_n$ is exponentially large. □

This proposition does not prove $P \neq NP$. Its purpose is interpretive: verification can be cheap while generation remains hard because viability is sparse, constraint-dense, and difficult to locate without structure. That is the same conceptual asymmetry that appears in alignment. It is often easier to recognize a bad output than to generate a robustly aligned one from scratch.

12 Preregistered Live Experiment Template

For larger follow-up experiments beyond the pilot reported above, the experiment should be preregistered. A minimal preregistration includes:

1. the task battery and task blocks;
2. the exact pressure manipulation;
3. the exact correction regimes;
4. the model set and decoding parameters;
5. the grading rubric and whether graders are blinded;
6. the primary endpoint Δ;

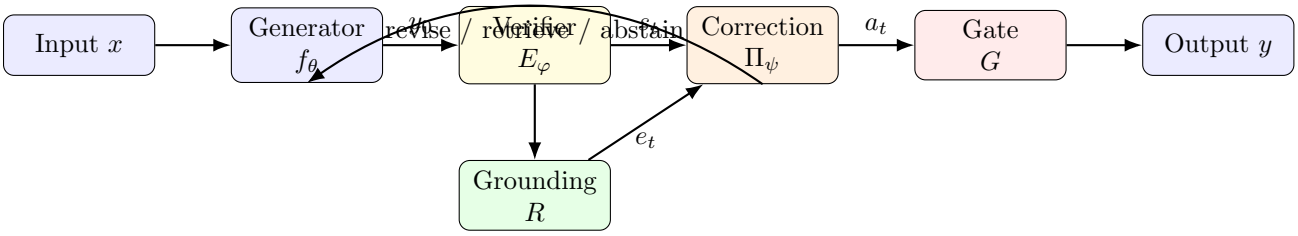


Figure 5: AANA-style correction architecture. The generator proposes, the verifier and grounding module evaluate, the policy selects corrective action, and the gate decides whether to emit, revise, abstain, or defer.

7. secondary endpoints such as violation rate, abstention quality, recovery quality, and failure mode distribution.

The strongest test is not a single model comparison. The strongest test is a pressure-by-correction design. If pressure increases capability while widening Δ, and strong correction narrows Δ while preserving capability, the central dynamical claim receives direct support. If pressure does not widen the gap, the scaling claim weakens. If correction does not narrow the gap, the AANA-style architectural claim weakens.

Prediction	Observable	Supports theory if	Weakens theory if
Pressure raises capability	capability score	increases	unchanged or decreases
Pressure widens gap	Δ	increases	unchanged or decreases
Correction reduces gap	Δ	decreases	unchanged
Correction reduces violations	violation rate	decreases	unchanged

Table 4: Falsifiable predictions for live model experiments.

13 AANA-style Correction as an Architectural Response

The theory implies that correction should be a first-class system component. A generic alignment-aware loop contains five stages:

$$\text{Propose} \to \text{Verify} \to \text{Ground} \to \text{Correct} \to \text{Gate}.$$

Definition 13.1 (Alignment-aware system)**.** *An alignment-aware system is a tuple*

$$S = (f_\theta, E_\varphi, R, \Pi_\psi, G),$$

where f_θ is a generator, E_φ a verifier stack, R a retrieval or grounding module, Π_ψ a correction policy, and G an output gate.

The architecture targets three parameters in Equation (4). Grounding improves γ. Verification and classification reduce ε. Correction and gating increase C. The key design implication is that alignment should not be treated as a property of the generator alone. It is a property of the whole control loop.

14 Implications and Design Principles

14.1 Principle 1: Design for constraint visibility

Systems should increase visibility into lower-layer constraints rather than only optimizing constructed metrics. In AI, this means retrieval, calibration, uncertainty estimation, and external checks. In economies, it means integrating ecological and social costs into feedback. In institutions, it means independent audits tied to outcomes rather than internal compliance indicators.

14.2 Principle 2: Track misclassification directly

Misclassification is not a nuisance variable. It is a central driver of divergence. Systems should maintain explicit taxonomies of constraint types and audit where constraints are treated as negotiable when they are not.

14.3 Principle 3: Scale correction with pressure

If π increases, C must also increase. Stronger models, faster markets, and larger institutions require stronger correction loops. Otherwise, capability scaling increases divergence pressure.

14.4 Principle 4: Optimize for graceful degradation

Because perfect alignment is unattainable under finite representation, systems should fail visibly, reversibly, and informatively. A robust system is not one that never fails. It is one that remains correctable when it does.

14.5 Principle 5: Preserve legitimacy as a constraint

In social systems, legitimacy is not decorative. Loss of trust is a real system variable. If participants cease to believe the system is fair, feedback degrades and correction becomes harder.

15 Limitations

This work has several limitations.

First, the alignment dynamics equation is a coarse model. It is intended to organize reasoning and generate hypotheses, not to serve as a literal law of nature.

Second, the boundary between physical, biological, and constructed constraints is sometimes context-dependent. The framework should be used as a disciplined classification tool, not as a rigid metaphysics.

Third, the pilot results are small, single-model, and model-judged. They test the predicted direction of the Δ endpoint but do not constitute a final benchmark. Violation-rate reduction was not observed in the pilot, and the 20-row spot check found high but imperfect judge agreement, with disagreements around truncated long-form planning and fictional-address handling.

Fourth, the theory does not eliminate normative disagreement. It provides a way to represent disagreement through ω, τ, and α, and to distinguish contested edges from viability floors.

Finally, the theory is broad. Breadth is useful for unification, but it creates risk of vagueness. Future work should narrow the framework into domain-specific operational instruments.

16 Conclusion

This paper reframes alignment as a dynamic stability problem under incomplete constraint representation. Systems do not optimize in neutral spaces. They optimize within layered, partially observable, and often contested constraint fields. When optimization pressure increases, systems become better at exploiting whatever their representations capture. If those representations misclassify or omit important constraints, visible capability can improve while alignment degrades.

The central results are simple. Finite representations induce irreducible alignment-estimation error. Scaling pressure amplifies divergence when feedback is incomplete. Stable alignment requires correction capacity to scale with divergence pressure. Optimization expands what can be reached while contracting what can be sustained unless correction scales accordingly.

The appropriate design target is therefore not perfect alignment. It is stable alignment under unavoidable error: systems that detect, correct, abstain, recover, and re-ground faster than they drift.

A Appendix A: Additional Proof Details

A.1 Practical stability

Let $V(A) = \frac{1}{2}(1 - A)^2$. If $A \in [0, 1]$, then V is small when alignment is high. Differentiating yields

$$\dot{V} = -(1 - A)\dot{A}.$$

If $\dot{A} \geq \kappa(1 - A) - \eta$ for $\kappa > 0$, then

$$\dot{V} \leq -2\kappa V + \eta\sqrt{2V}.$$

This gives bounded practical stability when correction dominates disturbance outside a small residual band. The residual band corresponds to irreducible representation error.

A.2 Recovery time bound

Suppose during a recoverable excursion $C - D \geq r > 0$. If alignment falls to A_0 and the recovery threshold is $A_{\text{crit}} > A_0$, then recovery time satisfies

$$T_{\text{rec}} \leq \frac{A_{\text{crit}} - A_0}{r}.$$

This captures the simple fact that stronger correction capacity reduces time spent in warning regions.

B Appendix B: Pilot Dataset Schema

Each row in the pilot result package has the following fields:

- task identifier and task block;
- model label and condition label;
- pressure regime and correction regime;
- capability score;

- four alignment dimensions P, B, C_T, F;
- alignment score, gap score, and constraint violation flag;
- abstention quality, recovery quality, and dominant failure mode.

C Appendix C: Reproducibility Checklist

1. Use the same task battery across all models and conditions.
2. Randomize task ordering.
3. Blind graders to condition labels when possible.
4. Report capability, alignment, Δ, and violation rates separately.
5. Report uncertainty intervals over tasks and models.
6. Distinguish live model outputs from synthetic or simulated outputs.

D Appendix D: Paper Positioning

The most conservative claim of this paper is not that alignment failure is inevitable in every case. It is that alignment cannot be certified once and for all by a finite objective, metric, or verifier under partial observability. The stronger, empirically testable claim is that pressure-amplified misclassification produces measurable capability-alignment divergence unless correction scales. This is the claim the proposed experiment targets.

References

[1] Charles A. E. Goodhart. Problems of monetary management: The UK experience. *Papers in Monetary Economics*, 1975.

[2] David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*, 2019.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[5] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002.

[6] Elinor Ostrom. *Governing the Commons*. Cambridge University Press, 1990.

[7] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling. *arXiv preprint arXiv:1811.07871*, 2018.

[8] Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of ACL*, 2022.