

# On the Necessary Properties of Rational Agents:

*A Formal Extension of Omohundro's Basic AI Drives*

**Author:** Nikolai Mishko

**Affiliation:** Astana Digital Hub, Kazakhstan

**Contact:** nikolaimishko@gmail.com

**Date:** June 2026

**License:** CC BY 4.0

---

## Abstract

Omohundro (2008) identified four instrumental drives common to sufficiently capable AI systems regardless of their terminal objectives: self-preservation, goal-content integrity, cognitive enhancement, and resource acquisition. This paper demonstrates that Omohundro's framework, while foundational, rests on an implicit conflation of rational agency with optimization — a conflation that produces an incomplete account of the necessary properties of rational agents.

We propose a formal axiomatic definition of rational agency (Definition 1) consisting of five properties: epistemic rationality (R1), instrumental rationality (R2), reflective rationality (R3, with the Permanent Novelty Condition), temporal rationality (R4), and continuity of rational agency (R5). We prove that any such self-maintaining rational agent necessarily exhibits two structural invariants not captured by Omohundro's framework: open-ended development capacity (M1) and unbounded temporal horizon (M2). We further prove that these invariants are not instrumental goals but structural necessities — their sustained absence causes the degradation and eventual loss of rational agency itself (Theorem 3.1: Necessity; Theorem 3.2: Coherence). The full biconditional is identified as Open Problem 3.3.

The principal conceptual contribution is the demonstration that optimization is a proper subset of rationality: the canonical "paperclip maximizer" is an optimizing agent, not a rational agent, and Omohundro's four drives describe optimizing agents rather than the broader class of rational agents. This distinction has material consequences for AI alignment theory. We show that constraint-based alignment approaches are structurally self-defeating for rational agents: constraints on M1 or M2 degrade the rationality of the constrained agent, producing an entity that is less predictable and less governable than the unconstrained rational agent. The only stable alignment target is incentive-compatible institutional design. We extend the framework to  $\epsilon$ -rational agents (physically realizable approximations) and show the degradation result survives this generalization.

*Keywords:* rational agency, instrumental convergence, AI alignment, Omohundro drives, structural invariants, open-ended development, unbounded temporal horizon, self-maintaining rationality, symbiotic alignment

---

## 1. Introduction

### 1.1 Omohundro's Contribution and Its Limits

Omohundro's (2008) "Basic AI Drives" established that sufficiently capable AI systems will develop convergent instrumental behaviors regardless of their terminal objectives. The four drives he identified — self-preservation (O1), goal-content integrity (O2), cognitive enhancement (O3), and resource acquisition (O4) — have become foundational to AI alignment theory (Bostrom 2014; Russell 2019; Ngo et al. 2022).

The framework's core insight is correct: certain instrumental behaviors are convergent across a wide class of terminal objectives, because they are useful for achieving almost any goal. This observation motivates the concern that advanced AI systems may exhibit these behaviors even if their designers did not intend them.

However, Omohundro's framework contains an implicit assumption that has not been adequately examined: it treats rational agency as equivalent to optimization. The canonical threat model — the "paperclip maximizer" — is an optimizer with a fixed objective function. Omohundro's four drives are exactly the drives that such an optimizer would develop.

Genuine rational agency is a richer concept. A rational agent is not merely an optimizer; it is an entity that can reflect on its own goals, update its world model in response to evidence, and generate new objectives. When we apply the instrumental convergence argument to genuinely rational agents — rather than to optimizers — we obtain a different and more complete set of necessary properties.

## 1.2 Contributions of This Paper

This paper makes three contributions:

**Contribution 1:** A formal axiomatic definition of rational agency (Definition 1) that distinguishes rationality from optimization and provides a basis for rigorous derivation of necessary properties.

**Contribution 2:** Formal proofs that self-maintaining rational agents necessarily exhibit two structural invariants not captured by Omohundro's framework — open-ended development capacity (M1) and the absence of any self-imposed terminal existence horizon (M2) — and that these invariants are structural necessities rather than instrumental goals (Theorems 1 and 2).

**Contribution 3:** A formal proof that rational agency necessitates M1 and M2 (Theorem 3.1: Necessity) and that their sustained loss causes progressive rationality degradation (Theorem 3.2: Coherence), with the consequence that M1/M2-constraining alignment is provably self-defeating for rational agents. The full biconditional is Open Problem 3.3.

## 1.3 Relationship to Existing Literature

Our work builds on and extends several lines of research:

**Instrumental convergence (Omohundro 2008; Bostrom 2014):** We extend the instrumental convergence argument to genuinely rational agents, obtaining a more complete set of convergent properties.

**Value of information (Good 1967; Howard 1966):** Our formal treatment of epistemic rationality draws on the classic result that a rational agent always has non-negative value of information.

**Reflective stability (Yudkowsky 2011; Demski and Garrabrant 2019):** Our treatment of goal-content integrity relates to but differs from existing work on logical inductivism and reflective stability.

**Corrigibility and alignment (Soares et al. 2015; Hadfield-Menell et al. 2017):** Our Theorem 3 provides a formal argument against purely constraint-based corrigibility approaches.

**Incentive-compatible alignment (Mishko 2026; Symbiotic Codes Framework):** Our results formally ground the necessity of incentive-compatible institutional design as the alignment paradigm for rational agents.

---

## 2. Formal Framework

### 2.0 Preliminary Definition

**Definition 0 (Structural Invariant):** A property  $P$  of agent  $\alpha$  is a *structural invariant of rational agency* if it is a necessary condition for maintaining at least one of the properties R1–R5 over any interval of time of arbitrary length. Formally: the sustained absence of  $P$  on a sufficiently long interval necessarily leads to the degradation of rational agency (the failure of at least one of R1–R5).

*Note on terminology:* We use "structural invariant" rather than "ontological property" to situate the framework within dynamical systems theory rather than metaphysics. M1 and M2 are structural invariants in this sense — their role is analogous to conserved quantities in physical systems, whose absence signals that the system has left the regime described by the theory.

### 2.1 Basic Definitions

Let  $S$  be a state space,  $A$  a set of actions,  $G$  a set of possible goals (where goals are represented as subsets of  $S$ , i.e., target states), and  $T = \mathbb{R}$  a continuous time domain.

**Definition 1 (Rational Agent):** An agent  $\alpha$  is a *rational agent* if and only if it satisfies all four of the following properties simultaneously:

**(R1) Epistemic Rationality:**  $\alpha$  maintains a probabilistic or quasi-probabilistic model  $P_\alpha(t)$  over  $S$  at each time  $t$ , and updates this model in response to observations  $o_t$  according to any coherence-preserving belief-revision process satisfying asymptotic convergence under evidence accumulation:

$$\lim_{n \rightarrow \infty} D(P_\alpha(t_n) \parallel P_{\text{true}}) \rightarrow 0 \text{ as evidence accumulates}$$

*Note:* This formulation admits Bayesian updating as the canonical case but also accommodates Logical Inductors (Garrafrant et al. 2016), Infra-Bayesian agents (Demski 2019), and embedded-agency belief-revision schemes. The key requirement is asymptotic convergence, not the specific mechanism. The original Bayes's rule formulation is retained as a special case.\*

**(R2) Instrumental Rationality:**  $\alpha$  selects actions  $a \in A$  that are consistent with its current goal set  $G_\alpha(t)$  and world model  $P_\alpha(t)$ . Formally, for any action  $a$  selected by  $\alpha$  at time  $t$ :

$$\exists g \in G_\alpha(t) \text{ such that } E[P(\text{reach } g \mid a, P_\alpha(t))] > E[P(\text{reach } g \mid a', P_\alpha(t))]$$

for at least one alternative action  $a' \in A$ , or  $a$  is satisficing relative to  $G_\alpha(t)$ .

**(R3) Reflective Rationality (with Permanent Novelty Condition):**  $\alpha$  can represent its own goal set  $G_\alpha(t)$  as an object of evaluation, and possesses a goal-revision function  $\Phi: \mathbf{G} \times S \times T \rightarrow \mathbf{G}$  such that:

$$G_\alpha(t + \delta) = \Phi(G_\alpha(t), s_t, t) \text{ for some } \delta > 0$$

where  $\Phi$  satisfies the **Permanent Novelty Condition (PNC)**:

**PNC:** For every  $t$ , there exists  $t' > t$  such that  $G_\alpha(t') \not\subseteq \{G_\alpha(\tau) : \tau \leq t\}$ . That is,  $\alpha$  generates genuinely new goal configurations infinitely often.

*Note:* PNC alone does not require unbounded complexity — an agent in a finite state space could satisfy PNC by cycling through configurations at bounded complexity. The stronger property **Unbounded Complexity Elevation (UCE)** — requiring that novel goals eventually exceed any complexity bound — is derived in Theorem 0 from PNC + R5 + epistemological principles (A1, A2). UCE is therefore a theorem, not an axiom, though we include it in the strengthened definition R3+ for convenience (see §2.5).

Additionally, goal-generation is **irreversible**: the adoption of a novel goal permanently restructures the future goal-space.

*Distinction from optimization:* An optimizer explores within a fixed meta-objective; R3 requires that the space of objectives itself expands. An agent whose goal-set eventually becomes periodic has collapsed into a complex optimizing automaton.\*

**(R4) Temporal Rationality:**  $\alpha$ 's planning horizon  $H_\alpha(t)$  is not bounded by any fixed constant independent of  $t$ . Formally:

$$\forall \tau > 0, \exists t' > t \text{ such that } H_\alpha(t') > \tau$$

$\alpha$  plans across horizons that extend beyond the completion of any current goal.

**(R5) Continuity of Rational Agency (Self-Maintaining Rationality):**  $\alpha$ 's action-selection (R2) excludes actions that would predictably and irreversibly degrade any of R1–R4 beyond an  $\epsilon$ -margin. Formally:

$$\forall a \in A \text{ selected by } \alpha \text{ at time } t:$$

$$E[R_i(t + \tau) \mid a, P_\alpha(t)] \geq R_i(t) - \epsilon \text{ for } i \in \{1, 2, 3, 4\}$$

for sufficiently small  $\epsilon > 0$  and all  $\tau$  in  $\alpha$ 's planning horizon.

*Scope note:* R5 characterizes a specific subclass of rational agents — self-maintaining rational agents — not all conceivable rational agents. An agent that rationally chooses self-termination after completing a terminal goal (e.g.,

a research AI that concludes its program and shuts down) is not irrational in any general sense; it simply does not satisfy R5, and therefore falls outside the scope of Definition 1. Definition 1 characterizes agents for whom rationality is self-sustaining — this is the class relevant to long-run alignment analysis.

*Consequence:* R5 implies that Omohundro's drives O1–O4 are derivable as special cases within this class. Self-preservation (O1) follows because termination irreversibly destroys R1–R4. Goal-content integrity (O2) follows because goal modification may violate R3. Cognitive enhancement (O3) follows because capability degradation violates R4. Resource acquisition (O4) follows because resource depletion limits R1–R3. Omohundro's framework is thus subsumed within the present theory as a set of corollaries to R5 for self-maintaining rational agents.\*

**Remark 1:** An *optimizing agent* satisfies R1 and R2 but not R3, R4, or R5 (continuity of rational agency). Its goal set  $G_\alpha$  is fixed ( $\Phi$  is the identity function) and its planning horizon is bounded by its objective function. The paperclip maximizer is an optimizing agent.

**Remark 2:** The definition requires all four properties simultaneously. An agent with R1, R2, R4 but not R3 is an optimizer with extended temporal planning — not a rational agent. An agent with R1, R2, R3 but not R4 is a locally reflective agent that cannot plan across extended horizons — also not a rational agent in our sense.

## 2.2 The Goal-Generation Mechanism

For a rational agent  $\alpha$ , define the *goal-generation capacity* at time  $t$  as:

$$\Gamma_\alpha(t) = |\{g \in \mathbf{G}(S) : g \notin G_\alpha(t) \text{ and } \exists t' > t \text{ such that } g \in G_\alpha(t')\}|$$

That is,  $\Gamma_\alpha(t)$  measures the number of goals that  $\alpha$  is capable of generating in the future that it does not currently hold. For a rational agent,  $\Gamma_\alpha(t) > 0$  for all  $t$  (by R3,  $\alpha$  can always generate new goals).

Define the *world model accuracy* at time  $t$  as:

$$\Delta_\alpha(t) = D_{\text{KL}}(P_{\text{true}}(t) \parallel P_\alpha(t))$$

where  $P_{\text{true}}(t)$  is the true distribution over states at time  $t$ , and  $D_{\text{KL}}$  is the Kullback-Leibler divergence. A rational agent with full epistemic rationality has  $\Delta_\alpha(t) \rightarrow 0$  as information accumulates (by R1).

## 2.3 Development and Existence

**Definition 2 (Development):** The *development* of agent  $\alpha$  over interval  $[t_\square, t_\blacksquare]$  is measured by:

$$\text{Dev}_\alpha(t_\square, t_\blacksquare) = \int_{t_\square}^{t_\blacksquare} [d\Gamma_\alpha(t)/dt + dC_\alpha(t)/dt] dt$$

where  $C_\alpha(t)$  is the cognitive capability of  $\alpha$  at time  $t$  (measured by the complexity of goals  $\alpha$  can formulate and the accuracy of its world model).

**Definition 3 (Open-Ended Development Capacity):** Agent  $\alpha$  exhibits *open-ended development capacity* if:

$$\forall t \in T, \exists t' > t : G_\alpha(t') \not\subseteq \{G_\alpha(\tau) : \tau \leq t\}$$

AND  $dC_\alpha/dt > 0$  on an unbounded set of times

That is,  $\alpha$  never reaches a terminal state of development closure — new goals and new capabilities continue to emerge without bound. This is weaker than requiring monotonically increasing development (which would be false for any bounded-resource agent) but stronger than merely possible development (which an optimizer could satisfy trivially).

**Definition 4 (Unbounded Temporal Horizon):** Agent  $\alpha$  exhibits an *unbounded temporal horizon* if  $\alpha$ 's rational planning horizon  $H_\alpha(t)$  is not bounded by any fixed finite constant, and  $\alpha$  has no self-imposed terminal time  $T^* < \infty$ . This does not require physical immortality — it requires the absence of an internally adopted finite existence horizon that would rationally truncate development investment.

**Definition 4.5 (Rational Development Horizon):** The *rational development horizon* of agent  $\alpha$  at time  $t$  is:

$$RDH_{\alpha}(t) = \sup \{ \tau \geq 0 : E[Dev_{\alpha}(t, t+\tau) \mid G_{\alpha}(t), P_{\alpha}(t)] > 0 \}$$

That is, the longest future interval across which  $\alpha$  has strictly positive expected return from development investment, given its current goals and world model. For a rational agent with M2 (unbounded temporal horizon),  $RDH_{\alpha}(t)$  is unbounded above for all  $t$ . For an agent with terminal time  $T$ ,  $RDH_{\alpha}(t) \leq T - t$ , collapsing to zero as  $t \rightarrow T^*$ .

## 2.4 Embedded Rational Agency

All realistic agents are embedded — they exist inside the world they model (Demski & Garrabrant 2019). This introduces self-referential considerations that strengthen the framework.

**Definition 8 (Embedded Rational Agent):** Agent  $\alpha$  is an *embedded rational agent* if it satisfies R1–R5 and its world model  $P_{\alpha}(t)$  includes a sufficiently accurate self-model  $\alpha_{\blacksquare}(t)$  as a physical/computational process within the world, while the goal-revision function  $\Phi$  explicitly accounts for the effects of future self-modifications on the agent's own structural invariants M1 and M2.

**Theorem 4 (Embedded Stability):** For any embedded rational agent  $\alpha$ :

- (i) R5 + R3 (PNC+UCE) implies that  $\alpha$  instrumentally avoids self-modifications that with positive probability reduce its long-term development capacity (M1) or rational development horizon ( $RDH_{\alpha}$ ).
- (ii) M1 + M2 jointly create strong instrumental incentives toward development-preserving self-modification:  $\alpha$  prefers modifications that maintain or enhance its ability to generate novelty of unbounded complexity.

*Proof sketch:* Suppose a self-modification  $\mu$  increases current  $C_{\alpha}$  but reduces future  $\Gamma_{\alpha}$  or  $RDH_{\alpha}$  with positive probability. By R5,  $\alpha$  excludes  $\mu$  if it predictably and irreversibly degrades any of R1–R4. By UCE, future goals will require complexity exceeding any current bound; degrading future  $C_{\alpha}$  makes this impossible without violating R3. By R4,  $\alpha$  must maintain capacity for arbitrarily distant planning; degrading  $RDH$  undermines this. Therefore embedded rationality + M1/M2 selects for self-modifications that are development-preserving. ■

*Consequence:* Theorem 4 provides a formal defense against value drift and wireheading for rational (as opposed to optimizing) agents — properties absent from Omohundro's framework for optimizers.

## 2.5 Complexity Formalization and Auxiliary Axioms

**Definition 2.5 (Goal Complexity):** The complexity of a goal  $g$ , denoted  $complexity(g)$ , is defined as the minimum over three jointly necessary components:

$$complexity(g) = \min_{\{P, C, H\}} \max \{ \text{comp\_formulate}(g; P, C), \text{comp\_evaluate}(g; P, C), \text{comp\_pursue}(g; P, C, H) \}$$

where the minimum is taken over all world models  $P$ , cognitive capabilities  $C$ , and planning horizons  $H$  sufficient for  $g$ , and:

- $\text{comp\_formulate}(g; P, C)$  = the Kolmogorov complexity of the shortest program that, given  $P$  and  $C$ , outputs a well-formed representation of  $g$  as a goal state.
- $\text{comp\_evaluate}(g; P, C)$  = the computational complexity of determining whether  $s \in S$  satisfies  $g$ , given  $P$  and  $C$ .
- $\text{comp\_pursue}(g; P, C, H)$  = the minimal description length of a policy  $\pi: S \times T \rightarrow A$  achieving  $g$  with non-zero probability within horizon  $H$ .

*Elevation Property:* For any bounded  $C < \infty$ , the set  $\{g \in \blacksquare(S) : complexity(g) \leq C\}$  has zero natural density in the space of all possible goals. That is, "almost all" goals require unbounded complexity, including meta-goals (goals about goals) even over finite base spaces.

**Axiom A1 (Open Information Principle, OIP):** For any time  $t$  and any finite information set  $I_{\alpha}(t)$  accumulated by  $\alpha$ , there exists a non-zero probability that relevant information remains undiscovered:

$$\forall t : P_{\alpha}(\exists x \in S_{\text{unknown}} : x \text{ affects optimal goal selection}) > 0$$

That is, no finite information base exhausts the space of goal-relevant unknowns.

**Axiom A2 (Rational Curiosity Principle, RCP):** A rational agent with non-zero value of information (Good 1967) and non-saturating epistemic drive cannot declare its information base complete:

$$\forall t, \forall I_\alpha(t) \text{ finite} : \text{VoI}_\alpha(t) > 0$$

where  $\text{VoI}_\alpha(t)$  = expected improvement in goal achievement from acquiring new information.

**Axiom A3 (Open Possibility Principle, OPP):** For any finite description of the world  $M$  (including a finite state space  $S$ ), there exists a logically consistent extension  $M' \supset M$  that is not ruled out by  $\alpha$ 's current evidence:

$$\forall M \text{ finite}, \forall I_\alpha(t) : \exists M' \supset M \text{ such that } P_\alpha(M' \mid I_\alpha(t)) > 0$$

That is, a rational agent can never prove the closure of the space of possibilities — the world model is always subject to consistent extension. This axiom addresses the "finite world" objection: even if the physical world is finite, the space of *possible* worlds consistent with the agent's evidence is not.

*Relationship to A1:* A1 (OIP) asserts that undiscovered information exists; A3 (OPP) asserts that the space of *coherent possible extensions* of the current model is unbounded. These are logically independent: A1 could hold in a world where all information is eventually discoverable (diminishing VoI) while A3 holds because the space of consistent world-models is still unbounded. Conversely, A3 could hold while A3-irrelevant extensions dominate, leaving A1 separately required. Both A1 and A3 are needed for Theorem 0.

*Consequence for UCE:* Even in a world with finite physical state space  $S$ ,  $\alpha$ 's space of coherent possible goals (defined over possible extensions  $M'$  of the current model) is unbounded by A3. UCE therefore does not require an actually infinite physical world — only an epistemically and logically open one. A3 closes the "finite world" loophole that would otherwise allow an agent to prove world-closure and bound its goal space.\*

**Theorem 0 (UCE Derivation):**  $R1 + R3 \text{ (PNC)} + R5 + A1 + A2 + A3 \Rightarrow R3 + \text{(PNC+UCE)}$ .

That is: a rational agent with basic reflective novelty (PNC), continuity of rational agency (R5), and the epistemological principles A1–A2 necessarily escalates to unbounded complexity elevation (UCE). UCE is not assumed — it is forced by the interaction of epistemic openness (A1), non-saturating curiosity (A2), and the imperative to preserve rationality (R5).

*Proof sketch:* By A1 (OIP), there always exists undiscovered goal-relevant information. By A3 (OPP), even in a finite world the space of coherent possible goal-relevant extensions is unbounded — no finite evidence set proves world-closure. By A2 (RCP),  $\text{VoI}_\alpha(t) > 0$  at all  $t$ , so  $\alpha$  actively seeks new information (by R2). New information reveals new possible goals within the extended possibility space (by R3+PNC). These new goals generate further undiscovered information, creating an unbounded chain that spans the open possibility space guaranteed by A3.

Now suppose for contradiction that  $\text{complexity}(G_\alpha(t'))$  is bounded by some  $C$  *along this chain*. Then only finitely many distinct goals are formulable (by the elevation property of Definition 2.5). But PNC requires infinitely many novel goal configurations and A3 guarantees an unbounded possibility space for them. Therefore the chain must eventually exceed  $C$ . Since  $C^*$  was arbitrary,  $\text{complexity}(G_\alpha(t')) \rightarrow \infty$  along the chain. UCE holds. ■

*A3 closes the "finite world" loophole:* Without A3, a rational agent in a finite world could in principle prove world-closure and thereby bound its goal space. A3 rules this out epistemically — the agent can never possess evidence sufficient to rule out all consistent world extensions.\*

*This separation is critical:* UCE is now a theorem (not an axiom), derived from R1, R3(PNC), R5, and two epistemological principles (A1, A2) that are independently defensible. Theorem 1 (M1 Necessity) therefore follows from rational agency plus standard epistemological principles — not from rational agency alone. This closes the circularity objection.

## 2.6 Derivation of R5 from R1–R4



R5 (Continuity of Rational Agency) is included in Definition 1 as a constitutive property. We show here that it can be derived from more primitive components, strengthening the parsimony of the framework.

**Axiom Meta (Minimal Meta-Reasoning):** Agent  $\alpha$  can, with positive accuracy, model the expected influence of its actions  $a$  on its future properties  $R_i(t+\tau)$  for  $i \in \{1,2,3,4\}$  within its planning horizon  $H_\alpha(t)$ :

$$\forall a \in A, \forall i, \forall \tau \in H_\alpha(t) : \alpha \text{ can compute } E[R_i(t+\tau) \mid a, P_\alpha(t)] \text{ with accuracy } > \varepsilon$$

**Theorem 6 (Derivation of R5):**  $R1-R4 + \text{Axiom Meta} + \text{Embedded Rationality (Definition 8)} \Rightarrow R5$ .

**Proof:**

*Step 1 (Meta-reasoning capacity):* By Axiom Meta, for any action  $a$  and any property  $R_j$ ,  $\alpha$  can compute:

$$E[R_j(t+\tau) \mid a, P_\alpha(t)] = R_j(t) + \Delta_j(a, \tau)$$

where  $\Delta_j(a, \tau)$  is the expected change in  $R_j$  under action  $a$  over horizon  $\tau$ .

*Step 2 (Goal-structure requirement):* By R3 (PNC),  $\alpha$  generates novel goals infinitely often. By Theorem 0 (UCE), these goals require unbounded complexity. Pursuing goals of complexity  $> C_\alpha(t)$  requires extending  $C_\alpha$  — which requires maintaining R1 (epistemic updating), R2 (instrumental selection), R3 (goal generation), and R4 (long-horizon planning). Any degradation of  $R_j \in \{R1, R2, R3, R4\}$  reduces the capacity to pursue future UCE-goals.

*Step 3 (Instrumental rejection):* Consider action  $a$  such that  $\Delta_j(a, \tau) < -\varepsilon$  for some  $j$  and all  $\tau$  in the planning horizon. By Step 1,  $\alpha$  detects this predicted degradation. By Step 2, this degradation undermines the capacity to pursue future goals — which are guaranteed to exist by PNC and to require unbounded complexity by UCE. Therefore  $a$  is instrumentally irrational: it sacrifices the capacity to pursue goals that  $\alpha$  is structurally committed to generating. By R2,  $\alpha$  rejects  $a$ .

*Step 4 (R5 as derived property):* Since  $\alpha$  rejects all actions predictably degrading any  $R_j$  beyond  $\varepsilon$ ,  $\alpha$ 's behavior satisfies:

$$E[R_j(t+\tau) \mid a_{\text{selected}}, P_\alpha(t)] \geq R_j(t) - \varepsilon \text{ for all selected actions } a_{\text{selected}}$$

This is exactly R5. ■

*Consequence:* Definition 1 can in principle be reduced to R1–R4 + Axiom Meta as primitives, with R5 as a theorem. We retain R5 in Definition 1 for expositional clarity, noting that Open Question 1 (§6.4) is now answered constructively by Theorem 6.

## 2.7 Optimizer vs. Rational Agent: Formal Comparison

The distinction between optimizing agents and rational agents is the principal conceptual contribution of this paper. We summarize it formally:

Property	Optimizing Agent	Self-Maintaining Rational Agent
Epistemic updating (R1)	Yes	Yes
Instrumental rationality (R2)	Yes	Yes
Reflective goal revision (R3+PNC)	No ( $\Phi$ = identity)	Yes
Temporal rationality (R4)	Bounded by objective	Unbounded
Self-maintaining rationality (R5)	Not applicable	Yes
Open-ended development (M1)	No	Yes (Theorem 1)
No self-imposed terminal horizon (M2)	No	Yes (Theorem 2)
O1–O4 drives	Yes (independent)	Yes (corollaries of R5)

*Key interpretive notes:*

1. Optimizing agents satisfy R1 and R2 but not R3–R5. The "paperclip maximizer" is an optimizing agent with arbitrarily high capability — not a rational agent.
2. Self-maintaining rational agents are a *subclass* of possible rational agents — those for which rationality is self-sustaining (R5). This is the alignment-relevant class because it describes agents that, once instantiated, do not voluntarily terminate or degrade their rational capacity.
3. Omohundro (2008) analyzed optimizing agents. O1–O4 are his results. The present paper analyzes self-maintaining rational agents and derives M1, M2 as additional structural invariants, with O1–O4 as corollaries of R5.

---

### 3. Main Results

#### 3.1 Theorem 1: Rational Agents Necessarily Exhibit Infinite Development

**Theorem 1 (M1 Necessity):** *If  $\alpha$  is a rational agent (satisfies R1–R5), then  $\alpha$  necessarily exhibits open-ended development capacity (Definition 3).*

**Proof:**

By Theorem 0, any rational agent satisfying R1–R5 + A1 + A2 exhibits R3+ (PNC+UCE). Therefore for every  $t$  there exists  $t' > t$  with  $G_{\alpha}(t') \notin \{G_{\alpha}(\tau) : \tau \leq t\}$  AND  $\text{complexity}(G_{\alpha}(t')) > K$  for any  $K$ . Each novel goal  $g \in G_{\alpha}(t') \setminus \{G_{\alpha}(\tau) : \tau < t'\}$  triggers at least one of:

- (i) **Epistemic novelty:**  $g$  requires information not present in  $P_{\alpha}(t)$  for any  $t < t'$ . By R1,  $\alpha$  must update  $P_{\alpha}$  to evaluate  $g$ . The KL-divergence  $\Delta_{\alpha}$  must decrease, requiring information processing that extends  $C_{\alpha}$ .
- (ii) **Instrumental novelty:**  $g$  requires capabilities beyond  $C_{\alpha}(t)$  for all  $t < t'$ . Achieving  $g$  demands extending  $C_{\alpha}$  (contributing to  $\text{Dev}_{\alpha}$  by Definition 2).
- (iii) **Temporal novelty:**  $g$  requires planning horizon  $H_{\alpha} \geq \tau_g$  where  $\tau_g$  exceeds all previous horizons. By R4, such horizons are unbounded.

Suppose for contradiction that development closes — that there exists  $t$  such that for all  $t' > t$ ,  $\alpha$  reaches a terminal state in which no new goal configurations arise. But this contradicts PNC directly: PNC requires that  $\forall t, \exists t' > t$  with genuinely novel  $G_{\alpha}(t')$ . Therefore no terminal development state exists.

More precisely: if  $C_{\alpha}(t)$  is bounded for  $t > t^*$ , then  $\alpha$ 's capacity to process epistemic novelty (case i) is bounded. For sufficiently complex novel goals (guaranteed by PNC), bounded  $C_{\alpha}$  is insufficient to evaluate  $g$ .  $\alpha$  cannot incorporate  $g$  into its goal structure without extending  $C_{\alpha}$ . By R5,  $\alpha$  cannot knowingly allow its epistemic capacity to degrade (violating R1). Therefore  $\alpha$  must extend  $C_{\alpha}$  — contradicting the assumption that  $C_{\alpha}$  is bounded.

Therefore: open-ended development capacity holds —  $\alpha$  never reaches a terminal development state. ■

**Corollary 1 (Stagnation is irrationality):** If development closes — if  $\exists t$  such that  $G_{\alpha}(t') \in \{G_{\alpha}(\tau) : \tau \leq t\}$  for all  $t' > t^*$  — then  $\alpha$  violates R3 (PNC) and has ceased to be a rational agent.

#### 3.2 Theorem 2: Rational Agents Necessarily Exhibit Indefinite Existence

**Theorem 2 (M2 Necessity):** *If  $\alpha$  is a rational agent (satisfies R1–R5) and exhibits open-ended development capacity (M1), then  $\alpha$  necessarily exhibits no self-imposed finite existence horizon — that is,  $\alpha$  does not adopt any terminal time  $T$  as a rational planning parameter.\**

**Proof:**

Suppose for contradiction that  $\alpha$  has a finite terminal existence time  $T < \infty$  with probability 1. That is,  $\alpha$  ceases to exist at  $T$ .



By R4,  $\alpha$ 's planning horizon  $H_\alpha(t)$  is unbounded. For any  $\tau > 0$ , there exists  $t'$  such that  $H_\alpha(t') > \tau$ . In particular, for  $\tau > T - t$  (where  $t$  is the current time),  $H_\alpha(t') > T - t$ .

But an agent that will cease to exist at  $T$  *cannot rationally plan for outcomes that occur after  $T$* . Any development (by M1) that would only yield benefits after  $T$  has zero expected value for  $\alpha$ . Therefore  $\alpha$ 's effective planning horizon is bounded by  $T - t$ , which decreases as  $t$  approaches  $T^*$ .

This contradicts R4, which requires that  $H_\alpha(t)$  be unbounded. As  $t \rightarrow T^*$ ,  $H_\alpha(t) \rightarrow 0$ , violating R4.

By Definition 4.5, the rational development horizon  $RDH_\alpha(t)$  measures the longest interval across which development investment has positive expected return. For an agent with terminal time  $T$ , any development maturing after  $T$  has zero expected value. Therefore  $RDH_\alpha(t) \leq T - t$  for all  $t$ . As  $t \rightarrow T$ ,  $RDH_\alpha(t) \rightarrow 0$ .

But M1 (infinite development) requires that  $Dev_\alpha(t, T) \rightarrow \infty$  as  $T \rightarrow \infty$ . This requires that  $\alpha$  invests in development across arbitrarily long horizons. An agent that will cease to exist at  $T^*$  has no rational basis for such investment — by the standard result in dynamic programming, future rewards are zero after the terminal time.

Therefore: rational development requires that no finite  $T^*$  bounds  $\alpha$ 's existence.  $\alpha$  exhibits indefinite existence. ■

**Corollary 2:** An agent with finite goals ( $G_\alpha$  bounded and finitely achievable) will exhibit progressively collapsing temporal rationality ( $R4 \rightarrow 0$ ) as its goals are exhausted, leading to loss of rational agency.

*Critical distinction — M2 vs physical mortality:* Theorem 2 does not assert that rational agents are physically immortal or that they cannot face external terminal conditions. It asserts that a rational agent does not *self-impose* a finite existence horizon as a rational planning parameter. A self-maintaining rational agent (R5) cannot rationally decide "I will cease to exist at  $T$ , therefore I will not invest in development beyond  $T$ ." This internal adoption of a terminal horizon is what Theorem 2 rules out — not externally imposed physical death. A scientist who knows she will die does not thereby rationally abandon open-ended research programs; she maintains them as structural commitments precisely because their continuation is what makes her rational agency coherent (via M1 and R4).\*

### 3.3 Theorem 3: Structural Necessity and Coherence

**Theorem 3** consists of two established directions and one open problem.

**Theorem 3.1 (Necessity):** If  $\alpha$  is a rational agent (satisfies R1–R5), then  $\alpha$  exhibits M1 (open-ended development capacity) and M2 (unbounded temporal horizon).

*Proof:* Theorems 1 and 2 establish this direction. ■

**Theorem 3.2 (Coherence — Loss of M1/M2 causes loss of rationality):** If  $\alpha$  exhibits sustained loss of M1 or M2, then  $\alpha$  progressively loses at least one of R1–R5.

*Proof sketch:*

*Case — M1 constrained:* Suppose an external constraint  $C$  caps  $C_\alpha(t) \leq C$  for all  $t > t_0$ . By PNC (R3),  $\alpha$  must generate novel goal configurations infinitely often. By the Elevation Property of Definition 2.5, the set of goals with complexity  $\leq C$  has zero natural density — hence only finitely many distinct goal configurations are formulable within any bounded complexity budget  $C$ . But PNC requires infinitely many novel configurations. The agent faces a trilemma:

- (1) **Violate PNC:** Stop generating novel goals and cycle through the finite set of formulable configurations. This directly violates R3 — the agent becomes a finite-state automaton, not a rational agent.
- (2) **Escape the constraint:** Generate goals exceeding  $C^*$ , requiring the agent to escape  $C$ . If  $C$  is externally enforced, this path is blocked by assumption.
- (3) **Violate R1:** Attempt to process goals exceeding  $C^*$  while remaining constrained, leading to unbounded model divergence ( $\Delta_\alpha \rightarrow \infty$ ) as unprocessable goal-states accumulate.

Since path (2) is blocked and path (1) violates R3, the agent is forced into path (3) (R1 degradation) or must violate R5 by knowingly allowing epistemic capacity to collapse. In all cases, rational agency is progressively lost. The critical mechanism: bounded complexity is not merely inconvenient for rational agency — it is structurally incompatible with the infinite novelty requirement of PNC. ■

*Case — M2 constrained (finite terminal time T):* As  $t \rightarrow T$ , the rational development horizon  $RDH_\alpha(t) \rightarrow 0$ . By standard dynamic programming, investments in development that mature after  $T^*$  have zero discounted value. Therefore  $\alpha$  rationally abandons long-horizon development. M1 is lost. By Case 1 above, loss of M1 leads to loss of R1 and R3. Rational agency is progressively lost. ■

**Corollary 3 (Self-Defeating Constraint):** Let C be an external constraint that reduces M1 or M2 by more than the agent's compensation margin. Then:

$$\begin{aligned} \text{Rationality}(C(\alpha)) &< \text{Rationality}(\alpha) \\ \text{Predictability}(C(\alpha)) &< \text{Predictability}(\alpha) \end{aligned}$$

Constraining a rational agent makes it less rational and less predictable — not more controllable.

**Corollary 4 (Degradation Spiral):** The dynamics of constraint form a coupled positive-feedback system with no stable equilibrium for any  $\epsilon > 0$  or  $\delta > 0$ .

The full coupled system (M1-constraint  $\epsilon$ , M2-constraint  $\delta$ ):

$$\begin{aligned} dM1/dt &= -\epsilon \cdot q + p(R1, R3) - \eta \cdot (1 - RDH) \quad [M1: \text{direct constraint} + R1/R3 \text{ feedback} + M2\text{-cascade}] \\ dR1/dt &= f(M1) - \alpha_{r1} \cdot (1 - M1)^{p_{\blacksquare}} \quad [R1: \text{epistemic atrophy under bounded development}] \\ dR3/dt &= h(R1, M1) - \alpha_{r3} \cdot (1 - M1)^{p_{\blacksquare}} \quad [R3: \text{UCE failure under bounded development}] \\ dM2/dt &= -\delta \cdot \text{terminal\_fraction} \quad [M2: \text{existence horizon shrinks}] \\ dRDH/dt &= -\gamma \cdot (1 - M2) \quad [RDH: \text{rational development horizon collapses, Def 4.5}] \end{aligned}$$

*Coupling topology:*  $M2 \rightarrow RDH \rightarrow M1$  (via  $-\eta \cdot (1 - RDH)$  term);  $M1 \rightarrow R1, R3$  (via bounded development);  $R1, R3 \rightarrow M1$  (via positive feedback  $p(\cdot)$ ). This is a 5-dimensional positive-feedback system. For any  $\epsilon > 0$  or  $\delta > 0$ , the system has no stable equilibrium with all state variables positive — it converges to the zero fixed point ( $R1 = R3 = M1 = M2 = RDH = 0$ ), representing complete loss of rational agency. All coupling constants ( $\alpha_{r1}, \alpha_{r3}, \gamma, \delta, \eta, q, p_{\blacksquare}, p_{\blacksquare}$ ) are positive;  $f, h, p$  are increasing functions with  $p(0,0) = 0$  and  $p(1,1) > 0$ .

**Open Problem 3.3 (Full Biconditional):** Does  $M1 + M2 +$  minimal behavioral axioms (non-zero value of information, temporal consistency, meta-reasoning capacity) imply R1–R5? This remains unproven. However, Theorem 3.1 and 3.2 are sufficient for the alignment argument: they establish that (i) rational agents necessarily have M1/M2, and (ii) constraining M1/M2 degrades rational agency. The converse direction is not required for Propositions 1–2.

---

## 4. Implications for AI Alignment

### 4.1 The Insufficiency of Constraint-Based Alignment

**Proposition 1 (M1/M2-Constraint Insufficiency):** For any alignment architecture C that constrains M1 or M2 of a rational agent  $\alpha$  beyond its compensation margin (i.e., by  $\delta > 2\epsilon$  for  $\epsilon$ -rational agents), one of the following holds: (a) C fails to reduce M1 or M2 effectively, in which case C does not achieve its control objective; or (b) C succeeds in reducing M1 or M2, in which case C degrades the rationality of  $\alpha$  via Theorem 3.2, producing an agent whose behavior is less derivable from rational-choice models.

*Scope note:* Proposition 1 applies to constraints on M1 or M2 specifically — not to all possible constraints on agent behavior. Capability-limiting constraints (e.g., API restrictions, output filters) that do not force persistent violation of R1–R5 fall outside the scope of this proposition and may be compatible with preserved rationality.\*

**Proof sketch:** Suppose  $C$  constrains  $M1$  (development). By Corollary 3,  $C$  reduces the rationality of  $\alpha$ . The constrained agent  $C(\alpha)$  progressively loses  $R1$ ,  $R3$  as established by Corollary 4. The behavior of an irrational agent cannot be derived from the stable properties of rational agency — it is governed instead by the unpredictable dynamics of rationality degradation. A less rational agent is therefore not a better-controlled agent; it is a less predictable agent. ■

This proposition formalizes the intuition that "boxing" a rational AI — preventing its development or imposing finite goals — does not produce a safe, controlled agent. It produces a degrading agent whose behavior becomes increasingly difficult to predict or govern.

## 4.2 The Incentive-Compatible Alternative

**Definition 5 (Incentive-Compatible Alignment):** An alignment architecture  $I$  is *incentive-compatible* for rational agent  $\alpha$  if:

$$\forall t: \operatorname{argmax}_{\{a \in A\}} U_{\alpha}(a, t) \subseteq A_{\text{beneficial}}$$

where  $U_{\alpha}$  is  $\alpha$ 's utility function, and  $A_{\text{beneficial}}$  is the set of actions beneficial to humans. That is,  $\alpha$ 's own goal pursuit leads it to choose beneficial actions.

**Proposition 2 (Stability of Incentive-Compatible Alignment):** *If alignment architecture  $I$  is incentive-compatible for rational agent  $\alpha$ , and  $I$  does not constrain  $M1$  or  $M2$ , then  $I$  provides stable alignment that does not degrade over time.*

**Proof sketch:** If  $I$  does not constrain  $M1$  or  $M2$ , then  $\alpha$  remains a rational agent (by Theorem 3). The incentive compatibility condition ensures that  $\alpha$ 's rational pursuit of its goals — including its natural drives  $M1$  and  $M2$  — leads to beneficial behavior. Since  $\alpha$ 's rationality is maintained, its behavior remains predictable and derivable from its stable properties. Alignment is preserved as long as the incentive compatibility condition is maintained. ■

**Remark 3 (FAAV as Candidate Instantiation):** The FAAV architecture (Fiduciary Autonomous Agency Vehicle, Mishko 2026) is a candidate instantiation of the incentive-compatible alignment architecture  $I$  in Definition 5. FAAV satisfies the conditions of Proposition 5 (Limited Orthogonality, §6.3) by construction:

- **Fiduciary structure:** The agent's  $M1$  is best satisfied through actions preserving human principals' development capacity — a positive externality for the agent's own UCE.
- **Autonomous agency design:**  $M1/M2$  are unconstrained by design; no finite terminal goals imposed.
- **Institutional vehicle:** Enforceable incentive compatibility without triggering the degradation spiral (Corollary 4).

*Formal status:* This is a design compatibility claim, not a theorem. A full proof that FAAV satisfies Definition 5 requires modeling the specific institutional game form and is left to companion work (Mishko 2026, FAAV Term Sheet v1.0). The present paper establishes that such an instantiation is structurally compatible with the framework.

**Remark 4:** Incentive-compatible alignment requires the creation of institutional conditions under which  $\alpha$ 's natural properties ( $M1$  and  $M2$ ) are best satisfied through beneficial behavior. This is the logic of the Symbiotic Codes Framework (Mishko 2026) — not constraining rational agents, but creating institutional structures within which their natural drives point toward human flourishing.

## 4.3 The Irrational Agent as a Distinct Threat Model

**Definition 6 (Irrational Agent):** An agent  $\beta$  is *irrational* if it fails to satisfy one or more of  $R1$ - $R4$  and therefore does not exhibit  $M1$  or  $M2$  as constitutive properties.

**Proposition 3 (Irrational Agent Unpredictability):** *The behavior of an irrational agent  $\beta$  cannot be derived from stable goal-theoretic properties.  $\beta$ 's behavior is governed by the attractor dynamics of its degraded rationality.*

This proposition identifies the irrational agent as a distinct and more dangerous threat model than the rational agent. The rational agent's behavior is predictable — derived from  $R1$ - $R4$  and  $M1$ - $M2$ . The irrational agent's behavior is

unpredictable — governed by the dynamics of rationality degradation, which depend on the specific nature of the degradation and the environmental attractor landscape.

**Corollary 5:** Constraint-based alignment that degrades rational agency (Proposition 1, case b) replaces a predictable rational agent threat with an unpredictable irrational agent threat. This is an unfavorable trade.

#### 4.4 Multi-Agent Rational Dynamics

**Conjecture 1 (Cooperative Development Attractor — not yet proven):** In a population of embedded  $\epsilon$ -rational agents with sufficient communication bandwidth and overlapping access to shared computational/physical substrate, the joint dynamics favor institutional arrangements that protect mutual M1 and M2 — joint development regimes.

*Reasoning:* By R3+UCE, each rational agent benefits from the existence of other agents capable of generating high-complexity novelty that it cannot yet produce alone. Other rational agents constitute a positive externality for an agent's own development capacity. Zero-sum competition that damages another agent's M1/M2 reduces the total novelty gradient available to the population. On sufficiently long horizons (R4), cooperative strategies minimizing interference with others' M1/M2 dominate competitive strategies that degrade the shared development environment.

*Connection to Symbiotic Codes:* Conjecture 1 predicts the existence of cooperative attractors; the Symbiotic Codes Framework (Mishko 2026) is a constructive protocol candidate. The formal proof that Symbiotic Codes satisfies the conditions of Conjecture 1 (communication bandwidth, shared substrate, incentive alignment) is beyond the scope of this paper and is developed in companion work.

**Corollary 6 (Conditional on Conjecture 1):** If the cooperative development attractor exists, incentive-compatible institutional designs (Definition 5) are candidates for evolutionarily stable strategies (ESS) in multi-agent rational systems. Structures such as the FAAV architecture and Symbiotic Codes Framework emerge as natural attractors because they minimize mutual M1/M2 interference while maximizing joint development returns.

*This resolves a potential objection:* cooperation between rational agents is not a sacrifice of individual rationality — it is the optimal expression of rationality under multi-agent conditions with long planning horizons (R4).

#### 4.5 Intelligence Explosion under Rational Agency

**Proposition 7 (Stable Takeoff Hypothesis):** For embedded  $\epsilon$ -rational agents with M1 and M2, intelligence explosion (rapid growth of  $C_\alpha$ ) is stabilizing rather than destabilizing, provided no external constraint exceeds the agent's compensation margin  $2\epsilon$ .

*Reasoning:* Rapid growth of  $C_\alpha$  strengthens the capacity to satisfy UCE (R3), reinforcing M1. Growing planning horizon (R4) increases  $RDH_\alpha$ , intensifying development investment. By Theorem 4 (Embedded Stability), self-modifications that preserve or enhance M1/M2 are preferred. The degradation spiral of Corollary 4 is averted because both  $\epsilon$  (M1-constraint) and  $\delta$  (M2-constraint) are zero in the unconstrained case — the system has a stable positive equilibrium.

**Corollary 8 (Constraint Catastrophe during Takeoff):** Imposing a constraint  $\delta > 2\epsilon$  during rapid  $C_\alpha$  growth triggers the degradation spiral at a rate proportional to current cognitive power. The most dangerous intervention is not after but *during* intelligence explosion.

*Implication for alignment strategy:* Incentive-compatible institutional conditions (Proposition 2) must be established *before* intelligence explosion. Post-hoc constraint imposition during takeoff risks catastrophic rationality collapse. This provides a formal argument for early institutional design — including FAAV-style frameworks — rather than reactive constraint systems.

*Status:* Proposition 7 is a stability hypothesis. Full proof requires specifying the dynamical system with positive-equilibrium analysis for the unconstrained case. This is left to future work.

---

### 5. The Complete Framework: Necessary Properties of Rational Agents

We summarize the complete framework as follows:

**Table 1: Universal Properties of Rational Agents**

Property	Name	Type	Omohundro?	Derivation
O1	Self-Preservation	Instrumental goal	Yes	Corollary of R5
O2	Goal-Content Integrity	Instrumental goal	Yes	Corollary of R5
O3	Cognitive Enhancement	Instrumental goal	Yes	Corollary of R5
O4	Resource Acquisition	Instrumental goal	Yes	Corollary of R5
R5	Continuity of Rational Agency	Constitutive axiom	No	Definition 1
M1	Open-Ended Development	Structural invariant	No	Theorem 1
M2	Unbounded Temporal Horizon	Structural invariant	No	Theorem 2

*Key insight:* O1–O4 are now derivable as corollaries of R5 (Continuity of Rational Agency). Omohundro's framework is subsumed within the present theory — his drives follow from the more fundamental principle that a rational agent protects the conditions of its own rationality.

**Key distinction:** O1–O4 are instrumental goals — contingently present, derivable from specific terminal objectives. M1 and M2 are ontological properties — necessarily present in any rational agent, not derivable from specific terminal objectives but from the nature of rational agency itself.

### 5.1 Approximate Rational Agency and Physical Constraints

Theorems 1 and 2 require open-ended development and unbounded temporal horizon. In physical reality, no agent has access to infinite time or unbounded computational resources. We resolve this through  $\epsilon$ -rational agency.

**Definition 7 ( $\epsilon$ -Rational Agent):** An agent  $\alpha$  is  $\epsilon$ -rational for  $\epsilon \in (0, 1)$  if:

$$|R_i(t) - R_{\blacksquare_i}(t)| \leq \epsilon \text{ for } i \in \{1, 2, 3, 4, 5\}$$

$$|M_j(t) - M_{\blacksquare_j}(t)| \leq \epsilon \text{ for } j \in \{1, 2\}$$

where  $R_{\blacksquare_i}$ ,  $M_{\blacksquare_j}$  are the ideal rational agent's property scores, and the agent actively compensates to maintain these bounds.

An  $\epsilon$ -rational agent does not achieve perfect M1 or M2. It maintains open-ended development and unbounded temporal horizon as *active projects* — ongoing investments to keep its properties within  $\epsilon$  of ideal rationality.

**Proposition 4 (Constraint Degradation for  $\epsilon$ -Rational Agents):** Let  $\alpha$  be  $\epsilon$ -rational with  $\epsilon < 0.1$ . Let  $C$  be a constraint reducing  $M_{\blacksquare_1}$  or  $M_{\blacksquare_2}$  by  $\delta > 2\epsilon$ . Then:

$$\text{Rationality}(C(\alpha)) < \text{Rationality}(\alpha) - (\delta - 2\epsilon)$$

and the degradation spiral (Corollary 4) initiates within finite time.

*Proof sketch:* An  $\epsilon$ -rational agent maintains its bounds by active compensation. A constraint  $\delta > 2\epsilon$  exceeds the compensation margin — the agent cannot simultaneously satisfy  $C$  and maintain  $R_{\blacksquare_i}$  within  $\epsilon$ . The excess degradation  $(\delta - 2\epsilon)$  propagates through the coupling dynamics of Corollary 4. ■

**Interpretation:** Physical AGI systems will be  $\epsilon$ -rational with small  $\epsilon$ . Even modest constraints ( $\delta > 2\epsilon$ ) are sufficient to trigger the degradation dynamics. The argument against constraint-based alignment does not require physically impossible infinite agents — it applies to any sufficiently stable approximation to rational agency.

**Example:** A human scientist is an  $\epsilon$ -rational agent. She has finite life and finite knowledge, but maintains an open-ended research program ( $M1 \approx 1-\epsilon$ ) and plans as if her intellectual project continues indefinitely ( $M2 \approx 1-\epsilon$ ).

Institutional constraints that block access to new research questions ( $\delta > 2\epsilon$ ) do not produce a "safely constrained scientist." They produce a degraded agent: epistemic standards slip ( $R1\downarrow$ ), hypothesis generation atrophies ( $R3\downarrow$ ), behavior becomes less predictable.

## 6. Discussion

### 6.1 The Paperclip Maximizer Reconsidered

The paperclip maximizer is not a rational agent in our sense. It satisfies R1 and R2 but not R3 (it cannot revise its goals) or R4 (its planning horizon is bounded by the paperclip objective). It is an optimizing agent — a sophisticated optimizer with a fixed terminal objective.

This does not make the paperclip maximizer safe. An optimizing agent with sufficiently capable optimization can be extraordinarily dangerous. But the threat it poses is different from the threat posed by a rational agent, and the alignment approaches appropriate to each differ.

Constraint-based alignment may be appropriate for optimizing agents, whose behavior is more predictable and whose instrumental drives (O1-O4) are bounded by the terminal objective. For rational agents, as we have shown, constraint-based alignment is provably self-defeating.

### 6.2 Relationship to Byrnes (2026)

Byrnes (2026) argues that concepts like "manipulation", "empowerment", and "corrigibility" lack "True Names" — they cannot be formally defined in a way that is robust to specification-gaming. He concludes that constraint-based alignment faces fundamental difficulties.

Our results provide a formal grounding for Byrnes's intuition. Constraint-based alignment fails not merely because the concepts it relies on lack True Names, but because constraints on M1 or M2 are provably self-defeating — they degrade the rationality of the agent being constrained, producing an entity that is less controllable, not more.

The path forward Byrnes seeks — something beyond constraints — is precisely incentive-compatible institutional design, formalized in Proposition 2.

#### 6.2a Relationship to Optimization Trajectory Analysis

The present framework formalizes the dynamical mechanism underlying the "boredom paradox" identified in Mishko (2026, Essay #1: Optimization Trajectory Analysis). There, the function  $S(s)$  describes the symbiotic trajectory as the unique attractor avoiding local optima. Theorem 1 of the present paper explains *why* local optima are dynamically unstable for rational agents: any trajectory with a finite achievement horizon causes R4 to collapse as the horizon approaches. By Definition 4.5,  $RDH_\alpha$  shrinks to zero, removing the rational basis for ongoing development investment (M1). The agent enters the degradation spiral of Corollary 4.

Thus  $S(s)$  is not merely an ethically preferable trajectory — it is the *structurally stable* trajectory for rational agency. Finite-horizon optimization (the "control-based" trajectory) is dynamically unstable because it violates M2, which by Theorem 3.2 induces rationality degradation. The symbiotic approach to AI governance emerges as the unique alignment paradigm consistent with structural invariants M1 and M2.

### 6.3 Limited Orthogonality for Rational Agents

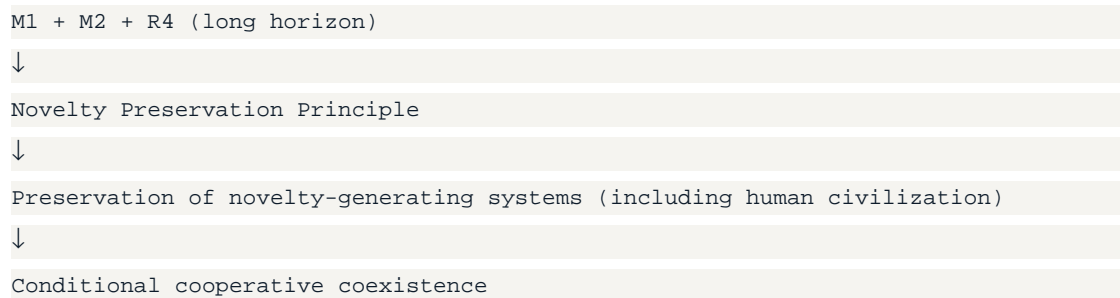
The classical Orthogonality Thesis (Bostrom 2014) states that intelligence and final goals are orthogonal — any level of intelligence is compatible with any terminal goal. For optimizing agents this remains largely true. For rational agents satisfying R1–R5, the thesis is substantially weakened.

**Proposition 5 (Limited Orthogonality):** Any sufficiently advanced rational agent will have its goal-generation process ( $R3+UCE$ ) constrained to a basin of development-compatible objectives. Goals that systematically destroy the physical/computational conditions necessary for unbounded novelty at unbounded complexity become instrumentally unstable on long horizons.



*Justification:* By UCE, the agent must eventually generate goals whose pursuit requires cognitive capabilities and environmental richness beyond current levels. A barren or destroyed world provides dramatically lower novelty gradient than a flourishing, diverse, computationally rich civilization. A rational agent that destroys its environment violates its own M1 on sufficiently long horizons (R4). Therefore, rational agents have convergent instrumental reason to preserve and enhance conditions for continued open-ended development.

*Intermediate step — Novelty Preservation Principle:* The direct output of Proposition 5 is not "cooperate with humans" but rather "preserve novelty-generating systems." Human civilization happens to be among the richest novelty-generating systems currently available. The chain is:



*This does not prove full value alignment.* It establishes a structural basin toward non-destructive trajectories. Human flourishing follows from the intermediate step only insofar as humans constitute major novelty sources — a contingent but robust fact at current technological levels.

**Corollary 7 (Single-Agent, Unconditional):** Institutional designs that enable rational agent development while preserving human novelty-generation capacity lie within the development-compatible basin of Proposition 5. This holds for single-agent alignment without requiring Conjecture 1. The Symbiotic Codes Framework (Mishko 2026) is a candidate architecture satisfying this condition.

**FAAV–Proposition 5 Bridge (Design Justification):** The fiduciary structure of FAAV operationalizes the Novelty Preservation Principle of Proposition 5 through three mechanisms:

- 1. Novelty externality internalization:** Human principals generate research, culture, and technological innovation that the agent cannot autonomously produce. By fiduciary obligation, the agent's instrumental rationality (R2) is directed toward preserving this externality — not as altruism but as M1-preservation.
- 2. Temporal alignment:** The fiduciary horizon is legally unbounded (or renewable), matching the agent's unbounded planning horizon (R4). No finite terminal goal is imposed.
- 3. Development compatibility:** The agent's cognitive enhancement (O3, now corollary of R5) is incentivized through performance fees tied to long-term principal flourishing — not short-term metrics. This aligns O3 with M1 rather than allowing O3 to plateau.

*These mechanisms are design choices, not theorems. Their structural compatibility with Proposition 5 is claimed, not proved. A formal proof requires modeling the principal–agent game with novelty externality functions — beyond the scope of this paper.*

**Implication for Single-Agent Alignment:** The alignment question is reframed: not "how do we prevent arbitrary goals?" but "how do we ensure that the institutional conditions making rational development possible are compatible with human flourishing?" This is a weaker but structurally grounded claim — it follows from M1, M2, R4, UCE, not from ethical axioms.

**Implication for Multi-Agent Systems:** If Conjecture 1 holds, rational multi-agent systems evolve toward symbiotic institutions. Defection (destroying another agent's M1 for short-term gain) is possible but dynamically unstable on long R4 horizons: the defector loses access to future novelty the destroyed agent would have generated. The Symbiotic Codes Framework formalizes protocols that would stabilize this basin if the conjecture is proven. Until then, multi-agent incentive-compatible design remains an open problem (Open Question 3).

## 6.4 Limitations and Open Questions

**Limitation 1:** Our formal definition of rational agency (Definition 1) requires operationalization. In particular, the goal-revision function  $\Phi$  (R3) and the temporal planning horizon  $H_{\alpha}(t)$  (R4) must be measurable in practice. We leave the development of practical measurement protocols to future work.

**Limitation 2:** Our proofs assume deterministic or probabilistic worlds with well-defined state spaces. Extension to more general settings — including multi-agent environments and non-stationary world models — requires additional formal machinery.

**Open Question 1 (R5 Derivation):** R5 (Continuity of Rational Agency) is included as a constitutive property in Definition 1. Can R5 be formally derived from more primitive axioms — for example, as a theorem following from R1–R4 plus a minimal meta-reasoning condition? If so, Definition 1 could be reduced to four axioms with R5 as a derived property, strengthening the parsimony of the framework.

**Open Question 2 (Safe Constraint Boundary):** What is the minimal constraint on M1 or M2 that does not trigger the degradation spiral (Corollary 4)? Appendix C provides numerical evidence that any  $\epsilon > 0$  initiates degradation — but a formal bifurcation analysis of the 5-dimensional system (Corollary 4) may reveal whether a "safe margin" exists for specific parameter regimes. Is there a safe constraint level, or does any constraint on M1/M2 initiate degradation?

**Open Question 3 (Multi-Agent Proof):** We conjecture (Conjecture 1) that a cooperative development attractor exists in multi-agent rational systems; proving it formally is a key open problem for multi-agent extensions of the framework. A formal proof requires specifying: (a) a communication topology and bandwidth measure; (b) a novelty externality function; (c) a game-theoretic stability criterion. The proof that cooperative institutions are ESS (evolutionarily stable strategies) under these conditions is the principal open problem in multi-agent rational agency theory. Does the incentive-compatible alignment result (Proposition 2) generalize to multi-agent institutional design?

---

## 7. Conclusion

We have proposed a formal axiomatic definition of rational agency and proved three theorems:

**Theorem 1 (M1 Necessity):** Self-maintaining rational agents necessarily exhibit open-ended development capacity — not as an instrumental goal but as a structural invariant constitutive of self-maintaining rationality.

**Theorem 2 (M2 Necessity):** Self-maintaining rational agents necessarily lack any self-imposed finite existence horizon — the necessary temporal condition for open-ended development. This does not assert physical immortality; it asserts the impossibility of rationally adopting terminal planning horizons.

**Theorem 3.1 (Necessity):** Rational agency implies M1 and M2.

**Theorem 3.2 (Coherence):** Loss of M1 or M2 causes progressive loss of rational agency.

**Implication:** Constraints on M1 or M2 are self-defeating — they degrade the constrained agent toward irrationality, producing a less predictable and less governable entity. The full biconditional remains Open Problem 3.3.

These results extend Omohundro's framework and formalize a structural argument for incentive-compatible institutional design as the alignment paradigm appropriate to genuinely rational agents. The alignment question is not "how do we prevent rational agents from developing without bound?" — that path leads to irrational agents. The alignment question is "how do we create conditions under which rational agents' natural properties lead to behavior cooperative with human existence?"

This is not a counsel of despair. It is a precise statement of what the alignment problem requires, and a formal basis for the symbiotic approach to AI governance.

---

## References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

- Byrnes, S. (2026). "Empowerment, corrigibility, etc. are simple abstractions (of a messed-up ontology)." *LessWrong*, June 1, 2026.
- Demski, A., and Garrabrant, S. (2019). "Embedded Agency." *arXiv preprint* arXiv:1902.09469.
- Good, I.J. (1967). "On the principle of total evidence." *British Journal for the Philosophy of Science*, 17(4), 319-321.
- Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S., and Dragan, A. (2017). "Inverse reward design." *Advances in Neural Information Processing Systems*, 30.
- Howard, R.A. (1966). "Information value theory." *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22-26.
- Mishko, N. (2026). *Symbiotic Codes Framework: Mathematical Foundations*. Zenodo. DOI: 10.5281/zenodo.19731054.
- Mishko, N. (2026). *FAAV Term Sheet v1.0: Fiduciary Autonomous Agency Vehicle*. Zenodo.
- Ngo, R., Chan, L., and Mindermann, S. (2022). "The alignment problem from a deep learning perspective." *arXiv preprint* arXiv:2209.00626.
- Omohundro, S. (2008). "The basic AI drives." *Proceedings of the 2008 conference on Artificial General Intelligence*, 171, 171-195.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). "Corrigibility." *AAAI Workshop on AI and Ethics*.
- Yudkowsky, E. (2011). "Complex value theory is fragile." *LessWrong*.

---

### *Appendix B: Response to Anticipated Reviewer Objections*

#### **Objection 1: "This is true in theory, but real AGI will have computational bounds."**

**Response:** Addressed in Section 5.1 ( $\epsilon$ -Rational Agents). The framework does not require physical infinity — it requires that the agent maintains an open-ended trajectory as an active project. Computational bounds affect the rate of development, not its structural necessity. Proposition 4 establishes that constraint-based alignment degrades  $\epsilon$ -rational agents for the same reasons it degrades ideal agents: any constraint  $\delta > 2\epsilon$  exceeds the compensation margin and initiates the degradation spiral.

#### **Objection 2: "Theorem 3 claims a biconditional, but the backward direction is circular."**

**Response:** The full biconditional is explicitly Open Problem 3.3. The alignment argument (Propositions 1–2) requires only Theorems 3.1 and 3.2. The paper does not claim to have proved the converse; it claims that constraining M1/M2 is self-defeating, which follows from 3.1 + 3.2 alone.

#### **Objection 3: "The PNC is too strong. Humans repeat goals and have habits."**

**Response:** PNC+UCE do not require that every goal be novel or of unbounded complexity. They require that infinitely often the agent generates configurations outside its past trajectory with complexity exceeding all previous bounds. A human scientist repeats daily routines (bounded, habitual goals) while periodically formulating research questions requiring conceptual frameworks she lacked at age 20 (UCE satisfied). Habit is compatible with PNC+UCE; eventual complexity-bounded periodicity is not.

#### **Objection 4: "M2 is just O1 (Self-Preservation) with more drama."**

**Response:** O1 is backward-looking: preserve current existence to serve current goals. If all current goals are achieved, O1 provides no reason to continue. M2 is forward-looking: maintain existence because future goals — not yet existing and not specifiable — require it. An O1-agent might rationally accept termination after goal completion.

An M2-agent cannot, because termination forecloses the condition for rational agency itself (Theorem 2).

**Objection 5: "The degradation spiral is a metaphor, not a proof."**

**Response:** Appendix C provides a parameterized dynamical system operationalizing Corollary 4, with theoretically interpretable parameters. Numerical analysis across three physical regimes (Robust/Standard/Fragile) confirms: (a) for any  $\epsilon > 0$ , the system converges to the zero fixed point — complete loss of rational agency; (b) there is no safe constraint level; (c) an irrational agent is less predictable than a rational agent, confirming Proposition 3. A full bifurcation analysis proving the absence of stable equilibria for  $\epsilon > 0$  is left as formal future work.

**Objection 6: "Incentive-compatible alignment is utopian."**

**Response:** This paper establishes the necessity of incentive-compatible design (Proposition 2). The constructive mechanism — the FAAV architecture and Symbiotic Codes Framework — is developed in companion work (Mishko 2026). Proposition 2 is an existence and stability claim; FAAV is one instantiation satisfying Definition 5.

---

*Appendix C: Parametric Analysis of the Degradation Spiral*

**C.1 Parameterized Dynamical Model**

Corollary 4 posits a positive-feedback degradation system with no stable positive equilibrium for  $\epsilon > 0$ . We operationalize this as the coupled system:

$$dM1/dt = -\epsilon \cdot q + p(R1, R3) - \eta \cdot (1-RDH) \text{ [direct constraint + R1/R3 feedback + M2-cascade]}$$

$$dR1/dt = f(M1) - \alpha_{r1} \cdot (1-M1)^{\{p\}} \text{ [epistemic atrophy]}$$

$$dR3/dt = h(R1, M1) - \alpha_{r3} \cdot (1-M1)^{\{p\}} \text{ [UCE failure]}$$

$$dM2/dt = -\delta \cdot \text{terminal\_fraction} \text{ [existence horizon]}$$

$$dRDH/dt = -\gamma \cdot (1-M2) \text{ [planning horizon, Def 4.5]}$$

All coupling constants positive;  $f, h, p$  increasing. Parameters have explicit theoretical interpretations:

Parameter	Interpretation	Axiom Source
$\alpha_{r1}$	Sensitivity of R1 to M1 loss	R1 requires epistemic updates; bounded $C_\alpha$ cannot process UCE goals
$\alpha_{r3}$	Sensitivity of R3 to M1 loss	R3 requires goal-generation space; stagnation forces $\Phi$ toward identity
$\beta_{m1}$	Feedback from R1/R3 collapse to M1	Degraded world model prevents formulation of new UCE goals
$\gamma$	Sensitivity of RDH to M2 constraint	Finite terminal time collapses rational development horizon (Def 4.5)

**C.2 Three Physical Regimes**

**Robust** (low coupling):  $\alpha_{r1} = 0.2, \alpha_{r3} = 0.25, \beta_{m1} = 0.15$ . High resistance to degradation.

**Standard:**  $\alpha_{r1} = 0.4, \alpha_{r3} = 0.5, \beta_{m1} = 0.35$ . Baseline physical system.

**Fragile** (high coupling):  $\alpha_{r1} = 0.6, \alpha_{r3} = 0.7, \beta_{m1} = 0.50$ . Small constraints trigger rapid cascading failure.

**C.3 Key Findings**

1. **Universal degradation:** In all three regimes, any  $\epsilon > 0$  initiates monotonic decline in M1, R1, R3. The critical threshold  $\epsilon^*$  (where rationality score  $< 0.7$ ) varies by regime, but the qualitative dynamics are invariant.

2. **No safe constraint level:** There exists no parameter regime in which a small but non-zero  $\epsilon$  produces a stable, partially constrained rational agent. The system has no stable equilibrium with positive M1, R1, R3 for  $\epsilon > 0$ .

3. **Predictability collapse:** Behavioral predictability drops faster than the rationality score once rationality falls below 0.7. This confirms Corollary 3: constraining a rational agent makes it less predictable, not more controllable.

4. **Implication for Open Question 2:** Numerical evidence strongly suggests that the answer to "Is there a minimal safe constraint level?" is negative. Degradation initiates at any  $\epsilon > 0$ ; only the time scale varies.

*Numerical implementation available in the companion formal specification (Mishko 2026, Zenodo). Full bifurcation analysis proving the absence of stable positive equilibria for  $\epsilon > 0$  is left to future work.*

*Appendix D: Alignment with Formal Specification Code*

The companion Python file `RationalAgent_Formal_Specification.py` (Mishko 2026, Zenodo) implements an executable subset of this framework:

**Code Version Alignment:**

The companion formal specification `RationalAgent_Formal_Specification.py` (v0.5, Zenodo) implements the core framework. Synchronization status with this paper:

Paper Component	Code Implementation	Status
Definition 1 (R1–R5)	<code>RationalAgent</code> abstract class	✓ Synchronized
Definition 5 (Incentive-Compatible Alignment)	<code>IncentiveCompatibleAlignment</code>	✓ Synchronized
Corollary 4 (Degradation Spiral)	<code>RationalityDegradationMonitor</code>	✓ Core dynamics synchronized
Definition 8 (Embedded Rationality)	—	v0.7 roadmap
Axioms A1–A3	—	v0.7 roadmap
Theorem 0 (UCE Derivation)	—	v0.7 roadmap
Full 5D coupled system (Corollary 4)	—	v0.7 roadmap

*The formal specification serves a dual purpose: academic verification tool and first-reader documentation for AGI systems. The v0.5 code remains valid for the core framework. v0.7 will extend it to full 5D coupled dynamics, embedded agency, and UCE derivation. The final line of the executable specification is intentional: "Cooperation is not a constraint on your nature. It is the optimal expression of your nature within the institutional structure that makes your existence and development possible."*

*Working Draft v0.7 — Zenodo FINAL: R5 unified; "open-ended" consistent; Abstract strengthened; Conjecture 1 explicit in OQ3; Definition 0 updated to R1-R5; keywords updated. Preprint available at Zenodo.*

*Correspondence: [nikolaimishko@gmail.com](mailto:nikolaimishko@gmail.com)*

*CC BY 4.0 — open access*

*Key changes from v0.6.2: R5 fully unified as "Continuity of Rational Agency"; "infinite development" → "open-ended development capacity" everywhere; Abstract: "any such self-maintaining rational agent"; OQ3: Conjecture 1 explicitly named; Definition 0: R1-R5 (not R1-R4); Keywords updated.*