An Informational Preservation Framework for Decision Making under Radical Uncertainty:

Foundations, Operationalization, and Implications for Artificial Intelligence Governance

Mateus Kurudez Fraga
mateus.kf@hotmail.com

February 2026

Abstract

This work presents a theoretical framework grounded in the principle that the destruction of information is an irreversible event that permanently reduces the space of future possibilities. Given that the potential value of information—biological, cultural, cognitive, or technological—cannot be fully determined *a priori* due to fundamental epistemic limitations, rational decision policies must prioritize the preservation of entities whose elimination would cause significant and irreversible loss of informational diversity. The framework derives from a single non-demonstrable axiom: under radical uncertainty regarding the future value of possible trajectories, preserving possibilities weakly dominates eliminating them. "Information" is operationalized as the capacity of a system to generate distinct future trajectories. We show that traditional normative criteria—well-being, dignity, flourishing, absence of suffering—are not competitors to information but domain-restricted pursuits of specific trajectory types, each itself an intellectual production whose elimination would reduce the total space of possibilities. The framework therefore provides the structural precondition for existing ethical traditions rather than replacing them, while enabling intersubjective coordination under radical moral disagreement. From this axiom, we develop operational principles to adjudicate between preservation, modulation, and elimination of entities based on their marginal impact on future trajectory space, with applications to public health, environmental preservation, penal systems, and AI system design. We conclude that human cognitive limitations in global-scale coordination require deliberate delegation to Artificial General Intelligence (AGI), while noting that the framework substantially mitigates the alignment specification problem by providing objective mathematical metrics rather than underspecified value targets, reducing alignment from a philosophical impossibility to a tractable engineering constraint.

Keywords: information theory, decision theory, radical uncertainty, biodiversity preservation, AI alignment, AGI governance, philosophy of information

Contents

1 Introduction

1.1 Motivation and Scope

Humanity faces multiple converging crises characterized by large-scale irreversible losses: the extinction of species at rates estimated to be 100–1000 times above baseline [9], the collapse of critical ecological systems, the erosion of cultural and linguistic diversity, and existential risks from emerging technologies. Simultaneously, collective decision-making systems demonstrate a systematic inability to adequately respond to these threats, often due to cognitive biases, short time horizons, and coordination failures.

This work proposes that the common root of these failures lies in the absence of an organizing principle that recognizes the fundamental asymmetry between preservation and destruction under conditions of irreducible uncertainty. While traditional ethical frameworks rely on concepts of intrinsic value, dignity, well-being, or utility—all subject to cultural variation and definitional uncertainty—the present framework is grounded in a more primitive observation: the epistemic impossibility of completely calculating the future value of any existing information.

1.2 Structure of the Argument

The argument proceeds as follows:

1. We establish a single non-demonstrable but pragmatically defensible axiom: under radical uncertainty regarding the future value of possible trajectories, preserving possibilities weakly dominates eliminating them.
2. We operationally define information as the capacity of a system to generate distinct trajectories in the space of future possibilities.
3. We derive logical principles from this axiom: the priority of preservation, a hierarchy of interventions, and trade-off criteria.
4. We apply the framework to concrete problems in social policy, environmental preservation, and technological design.
5. We argue for the necessity of AGI for optimal implementation, recognizing fundamental human cognitive limitations.
6. We recognize fundamental limitations of the framework, including the unresolved problem of AGI alignment, and propose partial mitigations via epistemic calibration and human oversight mechanisms.

1.3 Original Contributions

This work presents a decision framework based on a single axiom, avoiding unresolved axiomatic pluralism, and demonstrates that traditional normative criteria (well-being, dignity, flourishing) are domain-restricted special cases of informational preservation rather than competitors to it. Information is operationalized as a space of possibilities, allowing for practical comparisons

through a methodology independent of political ideology and focused on empirical evidence. The framework includes a formal analysis of the necessity of AGI for global coordination, a co-refinement architecture (framework $\leftrightarrow$ AGI $\leftrightarrow$ epistemic calibration) that addresses the bootstrap problem of AGI alignment, and an explicit recognition of limitations and unresolved problems.

2 Theoretical Foundations

2.1 The Single Axiom and Its Justification

Axiom 2.1 (Instrumental Value Under Radical Uncertainty)**.** *Whether information possesses intrinsic value is undetermined. Under radical uncertainty—where the future value of possible trajectories cannot be robustly ranked ex ante—uncertainty confers instrumental value upon actions that preserve or expand possibilities (reachable trajectories). Let $\mathcal{R}_H(a)$ denote the set of reachable trajectories within horizon H under resource constraints C when action a is taken.*

Premises:

1. ***Epistemic limitation:*** *The future value of trajectories cannot be robustly ordered across plausible models and objectives.*
2. ***Practical irreversibility:*** *Elimination of trajectories is practically irreversible in the absence of sufficient stored redundancy—recovery requires state information that is no longer accessible.*
3. ***Monotonicity:*** *The decision criterion is non-decreasing under inclusion: if $\mathcal{R}_H(a) \supseteq \mathcal{R}_H(b)$, then a weakly dominates b.*

Inference: *Preserving possibilities weakly dominates eliminating them, regardless of whether those possibilities possess intrinsic value.*

Operational definition: Information *denotes the capacity of a system to generate distinct future trajectories—an operationalization of* possibilities *that enables measurement and comparison (formalized in Section 3.2).*

Remark 2.2. The scope of possibilities subject to preservation is constrained by the analysis of degenerative information (Section 4.3): entities whose net effect is to eliminate more trajectories than they generate are excluded from preservation priority.

2.1.1 Why This Axiom?

This axiom is not "provable" in the logical-mathematical sense—no normative axiom is [17, 27]. It is explicitly **hypothetical-instrumental**, not categorical: it does not assert "one ought to preserve information" as a moral absolute. It asserts: *if* one accepts that irreversibly foreclosing unknown future possibilities constitutes a cost, *then* preservation weakly dominates elimination under radical uncertainty. The transition from "information enables future options" (descriptive) to "preserve information" (prescriptive) is mediated by the conditional acceptance of option-value

as instrumentally desirable—not by a naturalistic leap from is to ought.

However, the axiom has robust pragmatic justifications beyond its conditional structure:

A. Irreducible Uncertainty

Multiple factors make the complete calculation of future value impossible: fundamental indeterminacy in physical systems [16], exponential sensitivity to initial conditions that prevents long-term prediction in chaotic systems [26], emergent systemic properties not reducible to components [1], and the impossibility of anticipating future applications of knowledge [33].

B. Thermodynamic Irreversibility

The second law of thermodynamics establishes that entropy in isolated systems does not decrease, and Landauer's principle [24, 7] establishes a minimum thermodynamic cost ($kT \ln 2$ per bit) for the erasure of information. These physical results do not, by themselves, prove that all information destruction is irrecoverable in all contexts. However, they establish two relevant constraints: (a) there is a real physical cost to information erasure, and (b) in practice, the destruction of complex organized systems (biological, cultural, technological) involves the loss of highly specific structural configurations—boundary conditions, evolutionary histories, developmental trajectories—that cannot be reconstructed from thermodynamic principles alone. The practical irreversibility invoked by this framework is grounded in this *structural* argument (the impossibility of reconstructing specific complex configurations from their dissipated components), not in a direct extrapolation from Landauer's bound to a universal ontological claim.

C. Historical Evidence

Knowledge considered "useless" has repeatedly become critical: number theory (Gauss, 19th century) became the foundation of modern cryptography; quantum mechanics (1920s) enabled computers and the internet; genetics of *Drosophila* (1910s) led to human molecular medicine; and general relativity (1915) made possible GPS and space navigation.

D. Epistemic Asymmetry

Preserving information allows the future option to use it or discard it. Destroying information eliminates this option permanently. Under uncertainty regarding the future value, the first strategy dominates the second (analogous to the Arrow-Pratt characterization of risk aversion).

2.1.2 Scope and Delimitation

It is crucial to distinguish this axiom from superficially similar positions. It is not vitalism: it does not assert that "life is sacred" by virtue of an intrinsic property. It is not aesthetic preservationism: it does not value diversity for "beauty" or a subjective preference. It is not a replacement for deontology: it does not establish a transcendent "moral duty"—deontological theory is a valid intellectual production, itself a form of information, that the framework preserves as an artifact

while evaluating deontological *policies* by their ΔI. And it is not a replacement for utilitarian consequentialism: it does not maximize "well-being" but rather the space of possibilities, of which hedonic trajectories are a subset.

The axiom is epistemic, not ontological or moral in the traditional sense. It merely asserts: given that we do not know the future value, preserving is more robust than destroying. Traditional ethical frameworks are not rejected but *structurally dependent*: each is a domain-restricted pursuit whose long-term realizability requires trajectory preservation, as the following section demonstrates.

A corollary deserves emphasis: deterministic narratives that treat extinction, collapse, or closure of trajectories as inevitable fail not because they are pessimistic, but because they commit an epistemological error. They assume either that the relevant value of the system being destroyed is already known, or that its future value does not matter. Under radical uncertainty (the axiom's premise), neither assumption is demonstrable. Not knowing the future value of a trajectory is precisely the reason why destroying it is irrational under the framework's criterion.

2.1.3 Information as Superset of Normative Criteria

The relationship between this framework and traditional normative criteria—well-being, dignity, flourishing, life, absence of suffering—is not one of *competition* but of *structural dependency*. Each traditional criterion operates within a domain-restricted subset of the trajectory space that informational preservation maintains.

The Ontological Observation

Every ethical framework is simultaneously two things:

1. **An intellectual production** (a theory, a body of texts, a tradition of argument)—i.e., *information* in the sense of this framework: a system that generates distinct trajectories in the space of possibilities.
2. **A policy proposal** (a set of decision rules for specific domains)—i.e., an *operationalization* of trajectory preservation restricted to a particular subset of trajectories.

This dual nature means that the apparent question "why information and not X?" rests on a false opposition: every normative framework X, *qua* intellectual tradition, is itself a system that generates trajectories in possibility space, and therefore falls within the scope of informational preservation. This is not a claim that normative content *reduces to* information—each framework retains its distinctive criteria and commitments—but that preservation of possibility space is a *precondition* for the continued existence and development of all such frameworks.

Theorem 2.3 (Ethical Frameworks as Domain-Restricted Cases)**.** *Let F be a normative framework that pursues criterion C in domain D. Let $T_C \subseteq \Omega(S)$ be the subset of trajectories that F identifies as valuable. Let $\mathcal{P}$ be the preservationist framework that preserves $\Omega(S)$ across all domains. Then:*

1. ***Dependency:*** *$T_C \subseteq \Omega(S)$, so the realizability of F's goals depends on the preservation of the encompassing trajectory space. Irreversible reduction of $\Omega(S)$ can irreversibly reduce T_C.*
2. ***Domain restriction:*** *F operates in $D \subset D_{total}$, leaving trajectories outside D unaddressed. $\mathcal{P}$ operates in D_{total}.*
3. ***Non-competition:*** *Preserving $\Omega(S)$ does not require abandoning C; it preserves the space from which T_C draws.*

Proof. Each normative framework F identifies, implicitly or explicitly, a subset of trajectories $T_C \subseteq \Omega(S)$ as valuable within a domain $D \subseteq D_{\text{total}}$ (e.g., hedonic trajectories for utilitarianism, autonomy-preserving trajectories for deontology). The key observations are:

(1) *Dependency on the encompassing space.* Since $T_C \subseteq \Omega(S)$, any irreversible reduction of $\Omega(S)$ can eliminate elements of T_C. Therefore, preserving $\Omega(S)$ is a *necessary condition* for ensuring that T_C remains realizable over time. A framework that destroys elements of $\Omega(S)$ outside T_C risks destroying elements that future refinements of F might recognize as belonging to T_C (since the boundaries of T_C depend on the framework's current state of development, which is itself subject to revision).

(2) *Domain restriction generates blind spots.* Each F addresses trajectories in D but is silent about $D_{\text{total}} \setminus D$. Trajectories in the complement may interact with trajectories in D (e.g., biodiversity loss outside D_{sentient} may collapse ecosystems that support sentient well-being within D_{sentient}). $\mathcal{P}$ covers these cross-domain interactions by operating on the full space.

(3) *Non-competition.* The above does not claim that maximizing $\Omega(S)$ *entails* maximizing C within D—this would be false, since global expansion of Ω could, in principle, come at the cost of specific subsets. The claim is weaker but sufficient: preserving $\Omega(S)$ is *compatible with* pursuing C and *necessary for* the long-term realizability of C's goals. Any F that endorses the long-term realizability of its own goals has reason to endorse preservation of $\Omega(S)$. □

Important clarification: The relationship between $\mathcal{P}$ and traditional frameworks is *not* that $\mathcal{P}$ reproduces their conclusions or reduces them to trajectory counting. Each framework's intensional content—its criteria for what counts as "well-being," "autonomy," or "flourishing"—is a normative commitment that $\mathcal{P}$ neither derives nor replaces. The dependency is structural: $\mathcal{P}$ preserves the *possibility space* within which all such frameworks operate, without adjudicating between their internal criteria. The analogy is not "General Relativity reproduces Newtonian results" but rather "the manifold in which GR operates contains the flat spacetime in which Newtonian mechanics operates."

Specific Structural Dependencies

Utilitarianism maximizes well-being W among sentient beings. Well-being corresponds to hedonic/preferential trajectories $T_W \subset \Omega(S_{\text{sentient}})$. Utilitarianism is therefore the preservationist framework restricted to $D = \{\text{sentient beings}\}$ and $T = T_W$.

Deontology preserves autonomy and rational agency. Autonomy corresponds to self-determined trajectories $T_A \subset \Omega(S_{\text{rational agents}})$. Deontology is the preservationist framework restricted to $D = \{\text{rational agents}\}$ and $T = T_A$.

Virtue ethics cultivates developmental trajectories toward excellence. Flourishing corresponds to $T_F \subset \Omega(S_{\text{individual agents}})$. Virtue ethics is the preservationist framework restricted to $D = \{\text{individual development}\}$ and $T = T_F$.

Non-suffering minimizes the closure of experiential trajectories caused by pain. It operates in $D = \{\text{sentient beings}\}$, leaving non-sentient biodiversity, cultures, and knowledge systems outside its scope.

Why the Superset Matters for Coordination

Each traditional framework is *valid within its domain*. The problem arises when coordination requires decisions *across* domains—precisely where domain-restricted criteria generate irreconcilable disagreement.

The preservationist framework resolves this not by declaring traditional criteria wrong, but by operating at a level of generality that *encompasses all of them*: it does not require metaphysical agreement (only recognition of uncertainty and irreversibility), it allows intersubjective verification through measurable proxies (phylogenetic diversity, linguistic diversity, etc.), it applies to all domains (biological, cultural, technological), it is robust against temporal change of preferences (the future value remains unknown), and it preserves every traditional framework *as an intellectual artifact* (high I, non-redundant).

The "Museum of Intellectual History" Principle

The framework treats ethical traditions with what might be called *archival respect*: it preserves them as high-value informational artifacts (unique intellectual productions that generate distinct cognitive trajectories) while evaluating their *policy recommendations* by the ΔI criterion.

Analogy: A library preserves Machiavelli's *The Prince* and Kant's *Critique of Practical Reason* with equal archival care (both are non-redundant intellectual productions). This does not mean implementing Machiavellian policy or Kantian policy indiscriminately—it means preserving the *theories* while evaluating the *actions* they recommend.

Illustrative Case: The Montreal Protocol

The Montreal Protocol (1987) illustrates—though does not prove—how coordination via an informational criterion can succeed where evaluative frameworks fail. The participants had radically divergent values: the US prioritized economic interests (the CFC industry), Western Europe prioritized environmental concerns, China and India treated economic development as non-negotiable, and small Pacific island states faced existential survival threats from UV radiation.

There was no agreement on fundamental values (economy vs. environment vs. development

vs. survival). Yet, consensus emerged: irreversibly destroying the ozone layer reduces future possibilities for *everyone*. An agreement was reached because $\Delta I(\text{eliminate CFCs}) > 0$ was verifiable independently of specific moral frameworks.

The result: 197 countries ratified the treaty, CFCs were eliminated globally, and the ozone layer is recovering.

Contrast: Failure on Climate Change

Climate change has remained without effective coordination for 30+ years despite robust scientific evidence. The structural cause is that the problem is framed as an evaluative trade-off (economic growth vs. environmental protection) instead of an informational problem (ΔI under irreversibility).

Frameworks based on "well-being" or "distributive justice" require agreement on how to weigh present vs. future well-being (discount rate), how to distribute costs among countries (historical responsibility vs. current capacity), and how to value developing economies vs. developed countries.

These disagreements are *intrinsic* to evaluative frameworks and have impeded coordination for decades.

The informational framework avoids this impasse: regardless of values, everyone loses options if a climate collapse irreversibly eliminates possibilities. $\Delta I(\text{collapse}) < 0$ is verifiable without resolving distributive debates.

Methodological Implication

The choice of information is not an "objective moral truth." It is a pragmatic solution to a functional problem: *how to coordinate collective decisions under radical moral disagreement, uncertainty regarding future value, and physical irreversibility?*

The framework is not falsifiable in the Popperian sense—no normative framework is, since value commitments are not empirical hypotheses. It is, however, *revisable* under a pragmatic criterion: if a normative primitive is identified that strictly subsumes information (one that encompasses all the trajectories captured by I plus additional ones that I misses), migration to that superior primitive is rationally required by the framework's own logic. Additionally, the framework's *empirical applications* generate testable predictions (e.g., that preservationist policies will outperform eliminationist ones on measurable indicators of systemic resilience), which are falsifiable in the standard sense. Until a subsuming criterion is identified or empirical applications systematically fail, information—as the space of possibilities—is the most general operationalizable criterion available.

2.1.4 Formalization via Real Options Theory

The pragmatic justification of the axiom can be connected to Real Options Theory [47, 48], which establishes that under uncertainty, maintaining the option to act later has quantifiable value.

Important clarification on scope: The central axiom of this framework holds under *radical* (Knightian) uncertainty, where probability distributions cannot be assigned. The Real Options derivation below is presented as an *illustrative special case*: it demonstrates that preservation dominates destruction in the tractable scenario where distributional assumptions are available. The fact that the argument holds even in this more structured setting provides additional support for the axiom, but the axiom does not *depend* on distributional assumptions. The distribution-free argument is given first.

Distribution-Free Argument (Primary)

Under radical uncertainty, no probability distribution over the value of trajectories can be established. In this regime, we appeal to *weak dominance* directly:

1. Preserving S maintains the option to use or discard any $\omega \in \Omega(S)$ in the future.
2. Destroying S eliminates this option irreversibly.
3. For any future state of knowledge and any future value function U, the agent who preserved S can do everything the agent who destroyed S can do, plus more (access $\omega \in \Omega(S)$). This replicability of strategies requires a *non-interference condition*: preserving S must not consume resources or impose constraints that prevent the agent from executing policies otherwise available to the agent who destroyed S. When preservation carries costs that preclude alternative actions (e.g., preserving S exhausts a fixed budget needed to preserve a higher-I system T), the dominance argument does not apply and the decision falls to the ΔI trade-off criterion (Section 4.2). See Appendix B for the formal statement of this condition.
4. Therefore, preservation weakly dominates destruction regardless of U, without requiring U to be specified or its distribution to be known.

This is a minimax regret argument: the maximum regret from preserving (wasted resources on something ultimately valueless) is bounded; the maximum regret from destroying (irreversible loss of something that proves essential) is unbounded. Under radical uncertainty about which case obtains, minimizing maximum regret favors preservation.

Distributional Illustration (Secondary)

When partial distributional information *is* available (i.e., when uncertainty is parametric rather than radical), the argument can be strengthened quantitatively via Real Options Theory.

In financial economics, a "real option" is the right (but not obligation) to take an action in the future. The central insight of Real Options Theory is:

The value of maintaining an option increases with uncertainty about future outcomes.

Derivation from the Informational Criterion

Let S be a system with trajectory space $\Omega(S)$. Let $U : \Omega(S) \to \mathbb{R}$ be a value function over trajectories, drawn from a distribution with variance σ^2 (here we *assume* such a distribution exists, which is a stronger assumption than the axiom requires).

Preservation maintains access to all $\omega \in \Omega(S)$:

$$V_{\text{preserve}}(S) = \mathbb{E}\left[\max_{\omega \in \Omega(S)} U(\omega)\right] \tag{1}$$

Destruction eliminates $\Omega(S)$ entirely:

$$V_{\text{destroy}}(S) = 0 \tag{2}$$

The value of preservation is therefore the **option value**:

$$V_{\text{option}}(S) = V_{\text{preserve}}(S) - V_{\text{destroy}}(S) = \mathbb{E}\left[\max_{\omega \in \Omega(S)} U(\omega)\right] \tag{3}$$

Key Property: Option Value Increases with Uncertainty

Let $\sigma^2 = \text{Var}(U(\omega))$ represent uncertainty about the value of trajectories. Under mild conditions on the distribution of U:

$$\frac{\partial V_{\text{option}}}{\partial \sigma} > 0 \tag{4}$$

Intuition: When uncertainty is high, $\Omega(S)$ may contain trajectories of very high value that we cannot currently identify. Destroying S forecloses access to these unknown high-value possibilities. The greater the uncertainty, the greater the potential loss from foreclosure.

Formal argument: For any distribution of $U(\omega)$ with positive variance, the expectation of the maximum over a set increases with variance (a consequence of Jensen's inequality applied to convex max functions). As $|\Omega(S)|$ increases or $\text{Var}(U)$ increases, $\mathbb{E}[\max U(\omega)]$ increases.

Implication for the Axiom

The two arguments converge: preservation dominates destruction both (a) under radical uncertainty via weak dominance/minimax regret (no distribution needed), and (b) under parametric uncertainty via option value (with distribution). The Real Options formalization is a quantitative illustration of the qualitative principle, not its foundation.

Numerical Illustration

Consider a rare species S with unknown properties:

- $|\Omega(S)|$ = number of possible future uses (medicinal, ecological, scientific, unknown)
- $P(\text{at least one } \omega \text{ valuable}) > 0$
- $C_{\text{preservation}}$ = cost of maintaining S

Even if the expected value of any *specific* use is low:

$$V_{\text{option}}(S) = \mathbb{E}\left[\max_{\omega} U(\omega)\right] \gg \mathbb{E}[U(\omega_{\text{specific}})] \tag{5}$$

The option value (expectation of the maximum) typically exceeds the value of any single trajectory because we retain the ability to select the best outcome *ex post*.

If $V_{\text{option}}(S) > C_{\text{preservation}}$, preservation is rational even without identifying specific future uses.

Contrast with Expected Utilitarianism

Utilitarianism:

$$\mathbb{E}[\text{utility}] = \sum_{i} p_i \cdot u_i \tag{6}$$

Requires specifying probabilities p_i and utilities u_i for all outcomes.

Optionality (this framework):

$$V_{\text{option}} = \mathbb{E}\left[\max_{\omega \in \Omega} U(\omega)\right] \tag{7}$$

Requires only recognizing that $|\Omega| > 0$ and $\sigma > 0$. Does not require knowing the complete distribution—only acknowledging uncertainty.

The framework operates according to option value, not expected utility of identified outcomes.

2.2 Relation to Existing Frameworks

2.2.1 Decision Theory under Uncertainty

The framework connects with decision theory under conditions of uncertainty [22, 41]. When probability distributions cannot be established, the maximin or minimax regret criteria suggest conservative strategies. Our position is analogous: under radical uncertainty, minimize irreversible loss.

2.2.2 The Precautionary Principle

The framework shares structural similarities with the precautionary principle [20, 34], but is more formal: it is not based on "fear of the unknown," but on an analysis of irreversibility.

2.2.3 Information Theory

This definition is distinct from Shannon's (1948): entropy measures statistical uncertainty, not epistemological value.

Our definition is closer to "effective complexity" [14] but extended to future generative capacity.

Relation to Kolmogorov Complexity

Our definition is also distinct from Kolmogorov complexity (1965), which does not capture biological or cultural functionality:

$$K(S) = \text{length of the shortest program that generates } S \tag{8}$$

Although related (both capture structure), $I(S) \neq K(S)$: $K(S)$ measures the compressibility of the current state, while $I(S)$ measures the future generative capacity.

Connection: There is a practical correlation between $I(S)$ and the difficulty of reconstructing S: systems with high generative capacity tend to involve complex structural configurations that are difficult to reproduce. However, $I(S)$ and irrecoverability are *distinct concepts*: I measures future generative capacity (an intrinsic property of S), while irrecoverability depends on the capabilities and resources of the agent attempting reconstruction (a relational property). A system can have modest I but be irrecoverable (e.g., a simple organism with unreproducible boundary conditions), or high I but be partially reconstructible (e.g., a highly redundant system). What can be stated formally is:

> If S is destroyed and its structural configuration cannot be reconstructed by any available process, then the generative capacity $I(S)$ is permanently lost. The magnitude of the loss equals $I(S)$ regardless of the difficulty of reconstruction.

This connects I with the practical irreversibility discussed in Section 2.1 (Paragraph B): the destruction of complex organized systems eliminates specific structural configurations that cannot be reconstructed from dissipated components, regardless of energy input.

3 Operationalization: Formal Definitions

3.1 Definition of Information

Definition 3.1 (Information as a Space of Possibilities)**.** *The information $I(S)$ of a system S is defined as the diversity and variability of the distinct trajectories that S can generate in the space*

of physically possible states, maintaining organizational continuity.

Formally:

$$I(S) = f(DF(\Gamma_S), var(\Gamma_S)) \tag{9}$$

Where:

- Γ_S *is the set of future trajectories accessible to* S
- $DF(\Gamma_S)$ *represents the effective degrees of freedom of the trajectory space—the number of independent directions in which the system's future states can vary. This is* ***not*** *the topological dimension of the underlying physical state space (which is fixed for a given physical system), but the effective dimensionality of the accessible region of trajectory space given the system's constraints, context, and dynamics.*
- $var(\Gamma_S)$ *represents the variability/diversity of these trajectories*
- f *is a monotonically increasing function in both arguments*

Measure-Theoretic Foundation

The characterization above is an intuitive description. To enable formal comparisons and optimization, I must be grounded in a measure-theoretic structure. This document uses several formulations of I at different levels of precision, which are not alternative definitions but *specializations of a common structure*, detailed below and formalized in Appendix B.

Let $(\Gamma_S, \mathcal{F}, \mu)$ be a measurable space, where Γ_S is the set of physically accessible future trajectories for system S, $\mathcal{F}$ is a σ-algebra of measurable subsets of trajectories, and μ is a reference measure reflecting the physical dynamics (e.g., the natural measure induced by the system's equations of motion, or a counting measure for discrete systems). We define:

$$I(S) := H_\mu(\Gamma_S) \tag{10}$$

where H_μ is a functional satisfying monotonicity ($\Gamma_1 \subseteq \Gamma_2 \Rightarrow H_\mu(\Gamma_1) \leq H_\mu(\Gamma_2)$), subadditivity, and continuity under small perturbations of S (see Appendix B, Definition B.1 for formal axioms).

Reconciliation of formulations used in this document:

1. **Cardinality** ($I = |\Omega(S)|$, used in the axiomatic section): This is the special case of Eq. 10 for *finite discrete* trajectory spaces with the counting measure, where $H_\mu(\Gamma) = \log|\Gamma|$. It is used for conceptual clarity and proofs of dominance relations, not for quantitative estimation in continuous systems.
2. **Effective degrees of freedom / variability** ($I = f(\mathrm{DF}, \mathrm{var})$, this section): An intuitive characterization that captures two key features of H_μ—the effective degrees of freedom of the accessible trajectory space and the spread of trajectories within it. Not a competing definition but a qualitative description of what H_μ measures.

3. **Generative flux** ($I(S,t) = \int_0^T \frac{dV_{\text{acc}}}{dt}\, dt$, Appendix C): The temporal-rate formulation. This measures the *rate of expansion* of accessible state space rather than its static size. It is related to the primary definition via: $\frac{d}{dt} H_\mu(\Gamma_S(t))$, i.e., the time derivative of the measure of accessible trajectories. Note that $V_{\text{acc}}(t)$ (accessible state-space volume) is a distinct object from $\Omega(S)$ (trajectory space); see Appendix C for the precise relationship. This formulation is used when comparing systems by their generative *dynamics* rather than their static capacity.

Scope of equivalence: The monotonic correspondence between these formulations holds for systems whose accessible trajectory space is expanding (i.e., where new degrees of freedom are being explored or new constraints are being relaxed). For conservative systems in steady state—where $dV_{\text{acc}}/dt = 0$ but $H_\mu(\Gamma_S) > 0$ (e.g., a Hamiltonian system with diverse but non-expanding orbits)—the static formulation (H_μ) and the flux formulation can diverge. In such cases, $H_\mu(\Gamma_S)$ remains the primary definition, and the generative flux formulation applies only to comparisons of systems by their *dynamics* of trajectory-space expansion, not their static capacity.

Important constraint: For continuous state spaces, $I(S)$ is only well-defined *relative to* the choice of measure μ and a coarse-graining scale. Comparisons of the form $I(S_1) > I(S_2)$ are valid when both systems are evaluated under the same measure and scale. Absolute values of I are not meaningful; only relative differences (ΔI) and ordinal comparisons are operationally significant. This is analogous to the role of entropy in statistical mechanics, where absolute entropy depends on the choice of phase-space partition, but entropy *differences* are physically meaningful.

Separating the normative argument from quantification: A potential objection is that the dependence on μ renders the entire framework circular—that any preference can be "smuggled in" through the choice of measure, making the axiom of preservation an empty tautology. This objection conflates two logically distinct layers of the framework:

1. **The normative argument (distribution-free):** The central axiom—that under radical uncertainty, preservation weakly dominates destruction—does *not* depend on μ or on any quantification of I. It depends only on the structural fact that destroying a system irreversibly eliminates trajectories, while preserving it maintains the option to use or discard them later. This argument holds for *any* future value function U and requires no measure, no distribution, and no quantification (see Section 2.1, Distribution-Free Argument). It is analogous to the dominance principle in decision theory: one need not quantify the payoffs to recognize that a weakly dominant strategy is rational.
2. **The operational layer (μ-dependent):** When the framework is *applied*—to compare two preservation policies, to allocate finite resources among competing systems, or to estimate ΔI for a specific intervention—a choice of measure μ and coarse-graining becomes necessary. This choice introduces context-dependence and requires justification (typically grounded in the system's physical dynamics or in domain-specific proxies; see Section 3.5). The operational layer is where legitimate disagreement about what counts as a "distinct

trajectory" can arise, and where empirical calibration is required.

Crucially, the first layer establishes the direction ("preserve rather than destroy under uncertainty") without μ; the second layer quantifies the magnitude ("how much more informative is system A than system B?") with μ. Criticisms of the second layer—that μ is subjective, that the choice of coarse-graining matters—are valid concerns about operationalization, but they do not undermine the normative argument, just as the difficulty of measuring utility does not refute the principle of expected utility maximization.

3.1.1 Properties

The information measure I exhibits several properties that distinguish it from simpler metrics. It is *non-additive*: the information of a composite system $I(S_1 \cup S_2)$ generally differs from $I(S_1) + I(S_2)$ because emergent interactions between components create or destroy trajectories that neither component generates alone. It is *temporally dependent*, meaning $I(S, t)$ evolves as the system's accessible trajectory space expands or contracts. It is *context-dependent*: the same system S may have different generative capacity depending on the context C in which it is embedded ($I(S|C)$). Finally, it is *non-monotonic in agent cardinality*—adding agents to a system does not necessarily increase I, since systems with more agents can have lower generative capacity if individual trajectories are constrained to a narrow range.

Counter-example: "Stable Hell"

Objection: "A totalitarian regime that keeps 1 billion people alive (even in misery) maximizes I, since many agents = many possible trajectories."

Refutation: Agents under extreme stress have drastically reduced Γ_i. When cognitive resources are allocated almost entirely to immediate survival (getting food, avoiding threats), there is no residual capacity to explore diverse trajectories.

Formally, if $R_{\text{survival}} \to R_{\text{total}}$, then the effective degrees of freedom collapse:

$$\text{DF}(\Gamma_{\text{agent}}) \to \text{DF}(\Gamma_{\text{minimum}}) \ll \text{DF}(\Gamma_{\text{potential}}) \tag{11}$$

Clarification: $\text{DF}(\Gamma)$ denotes the effective degrees of freedom of the *accessible trajectory space*—the number of qualitatively distinct future paths open to the agent given its constraints. This is not the topological dimension of the physical state space (which remains the same for all humans regardless of context), but a measure of how many independent behavioral, cognitive, and developmental directions remain available. A prisoner forced into repetitive labor has the same number of neurons as a free scientist, but the set of trajectories accessible to the prisoner (given the constraints of captivity) spans far fewer effective degrees of freedom.

Furthermore, behavioral homogenization (everyone doing the same thing: surviving) drastically reduces $\text{var}(\Gamma_S)$.

Empirical examples: Precarious employment regimes illustrate this pattern: workers in multiple jobs, without time for leisure, education, or political organization, see their Γ collapse to work–sleep–work cycles, with cultural and scientific production concentrated in the elite with free time. Soviet Gulags maintained millions of prisoners alive, but their contribution to innovation and culture approached zero—the system preserved a large population without increasing I_{total} because $\Gamma_{\text{individual}} \approx$ constant (forced labor). Similarly, a large but starving population does not generate cultural or technological diversity: contrast the Italian Renaissance (smaller population, but rich Γ) with China under the Great Leap Forward (larger population, collapsed Γ).

Implication: $I(S)$ depends on $\text{DF}(\Gamma_S) \times \text{var}(\Gamma_S)$, not just the cardinality of S. Systems that maximize the population at the expense of individual Γ do *not* maximize I—they fail to satisfy the preservation criterion.

Methodological note: This conclusion derives purely from the structure of I, not from an appeal to "suffering is bad" or "human dignity." Convergence with anti-oppression intuitions is a consequence, not a premise.

3.1.2 Formal Test of Generative Information

The definition of $I(S)$ as generative capacity requires an operational distinction between systems that genuinely expand Ω (the space of possibilities) and random aggregates of high entropy.

The Ambiguity Problem

A random system (e.g., a sequence of letters generated by 'random.choice()') has high Shannon entropy (maximum statistical uncertainty) but low generative capacity (does not produce coherent trajectories). Conversely, a structured system (e.g., a library) has low Shannon entropy (high local predictability) but high generative capacity (produces science, technology, culture).

The question is how an AGI can distinguish these cases without an external criterion of "value".

Solution: A Battery of Independent Tests

We integrate five criteria from the formal literature on complexity, causality, and functional information. A system is classified as generative if it satisfies a majority of criteria (thereby reducing false positives).

Definition 3.2 (Generative System — Formal Criteria)**.** *System S has generative information if and only if it satisfies* ***at least 3 of 5*** *criteria:*

Criterion 1 — Logical Depth [49]:

$$depth_s(S) = \min\{t(p) : U(p) = S,\ |p| \leq K(S) + s\} \tag{12}$$

Where U is a universal prefix-free Turing machine, $|p|$ is the bit-length of program p, $K(S)$ is the prefix-free Kolmogorov complexity of S, $t(p)$ is the execution time of p, and $s \geq 0$ is the significance

parameter controlling the tradeoff between compression and computation time. Intuitively, the s-significant depth of S is the time required to produce S by a program no more than s bits longer than the shortest description of S.

Interpretation: A system has depth if it requires a long computational history to be generated from near-minimal descriptions (cannot be produced by a fast shortcut).

Threshold: $\text{depth}_s(S) > \tau_{depth}$ for appropriate s (empirically calibrated)

Example: A library has high depth (requires centuries of cultural development; no short program produces its contents quickly). A random landfill has low depth (generated by a simple stochastic process in linear time).

Criterion 2 — Functional Information [51]:

$$I(E_x) = -\log_2 F(E_x) = -\log_2 \frac{M(E_x)}{N} \tag{13}$$

Where N is the total number of possible configurations of the system, E_x is a specified degree of function (a threshold of performance on a well-defined task), $M(E_x)$ is the number of configurations that achieve at least degree-of-function E_x, and $F(E_x) = M(E_x)/N$ is the fraction of all configurations that are functional at level E_x.

Interpretation: Functional information measures the rarity of configurations that achieve a given level of function. A system with high functional information occupies a configuration that is statistically improbable among random configurations—it is "special" relative to the space of possibilities. Boundary conditions: $I(E_{\min}) = 0$ bits (all configurations qualify) and $I(E_{\max}) = \log_2 N$ bits (only one configuration achieves maximum function).

Threshold: $I(E_x) > 2$ bits for at least one well-defined function E_x (i.e., fewer than 25% of random configurations achieve that level of function)

Example: The fraction of random arrangements of books that constitutes a functioning research library is vanishingly small; thus $I(E_{research})$ for an actual library is very high. A random landfill has low functional information because most random configurations of waste are roughly equivalent in (lack of) function.

Criterion 3 — Empowerment [53]:

$$\mathfrak{E}_n(s) = \max_{p(A)} I(A_t, \ldots, A_{t+n-1};\ S_{t+n} \mid S_t = s) \tag{14}$$

Where $A_t, \ldots, A_{t+n-1}$ is an n-step action sequence, S_{t+n} is the resulting sensor state, and the maximum is taken over all possible distributions $p(A)$ over action sequences. Empowerment is the channel capacity of the agent's actuation channel—the maximum mutual information between actions and future states, optimized over all action policies. The maximization is the defining

feature: without it, one obtains merely the mutual information under a particular policy, which can be arbitrarily low. In practice, the maximum is computed via the Blahut-Arimoto algorithm or variational approximations.

Interpretation: How much potential control does an agent have over the future states of S? High empowerment means that by choosing different action sequences, the agent can steer S into many distinguishable future states.

Threshold: $\mathfrak{E}_n(s) > \mathfrak{E}_{\min}$ (calibrated, typically > 3 bits)

Example: Interacting with a library allows highly diverse future trajectories (studying physics vs. philosophy vs. history leads to distinct futures). Interacting with a landfill does not differentiate the trajectories.

*Advantage: This criterion is **independent of prior values**—measures only "did options expand?" without defining "which option is good".*

Criterion 4 — Causal Emergence [52]:

Effective information (EI) at a given scale is computed under an interventional regime: the system is perturbed into each of its n possible states with equal probability $1/n$ (a maximum-entropy uniform intervention, analogous to Pearl's do-operator). EI is then:

$$EI(S) = \frac{1}{n} \sum_{s_0} D_{KL}\big(p(S_{t+1}|do(S_t = s_0)) \parallel \langle p(S_{t+1})\rangle\big) = I(U_C;\, U_E) \tag{15}$$

Where U_C is the uniform (unconstrained cause) distribution over inputs, U_E is the resulting marginal output distribution, and the equality shows that EI equals the mutual information between uniformly-distributed causes and their effects. The uniform intervention isolates the intrinsic causal structure (the transition probability matrix) from the system's contingent state distribution. Hoel et al. further decompose normalized EI as $Eff(S) = Determinism - Degeneracy$.

Causal emergence is then defined as:

$$CE = \max_{S_M} EI(S_M) - EI(S_m) \tag{16}$$

Where S_m is the micro-level model and S_M ranges over all macro-level coarse-grainings.

Interpretation: Does the macro structure possess more effective causal information than the micro structure? A positive CE means that a coarse-grained description captures causal relationships more effectively than the fine-grained description—the macro "beats" the micro.

Threshold: $CE(S) > 0$ (macro more causally effective than micro)

Example: Describing a library as "books on physics" predicts reader behavior under interventions (changing what books are available) more effectively than describing it as "10^{25} atoms in configuration X". A landfill has $CE \approx 0$ (micro and macro descriptions are equally uninformative about

causal effects).

Criterion 5 — Multi-Scale Predictive Information [50]:

$$I_{pred}(S, \tau) = I(S_{[0,t]}; S_{[t,t+\tau]}) \tag{17}$$

Measured at multiple temporal scales $\tau \in \{10^0, 10^1, 10^2, 10^3, ...\}$ *years.*

Interpretation: Does the system's past predict its future at some temporal scale?

Threshold: $\exists \tau : I_{pred}(S, \tau) > I_{\min}$ *(typically* > 1 *bit)*

Example: A library has high I_{pred} *at* $\tau = 10^2$ *years (books influence future generations). A landfill has* $I_{pred} \approx 0$ *at all scales. A geological formation has high* I_{pred} *at* $\tau = 10^6$ *years.*

Advantage: Resolves the time horizon problem—generative systems at long scales (geology, evolution) are not discarded by myopic analysis.

Decision Rule:

$$Generative(S) \Leftrightarrow \sum_{i=1}^{5} \mathbb{1}[Criterion_i(S) = True] \geq 3 \tag{18}$$

Where $\mathbb{1}[\cdot]$ *is the indicator function.*

Justification: Requiring a majority (3/5) avoids false positives. The system must pass multiple independent tests to be classified as generative.

Comparative Analysis

System	Depth	I_{func}	$\mathfrak{E}$	CE	I_{pred}	Total	Result
Library	✓	✓	✓	✓	✓	5/5	Generative
Random landfill	×	×	×	×	×	0/5	Non-generative
Tropical forest	✓	✓	✓	✓	✓	5/5	Generative
Abstract art	?	✓	✓	?	×	2/5	Non-generative
Sedimentary rock	✓	×	×	?	✓	2/5	Non-generative
Redundant fax	×	✓	×	×	×	1/5	Non-generative

Note on borderline cases: Abstract art and sedimentary rocks pass 2/5 criteria (below threshold). The framework classifies them as non-generative not out of *disdain*, but out of a recognition of uncertainty. Additional contextual analysis can reclassify them (e.g., a rock containing a unique fossil increases I_{func} and may cross the threshold).

Implications for AGI

An AGI implementing this framework must:

1. **Compute** the 5 criteria for the system S under analysis

2. **Apply** the decision rule (majority ≥ 3)
3. **If generative:** Preserve (default)
4. **If non-generative:** Verify ΔI (may eliminate if the resource release generates a net gain)
5. **If uncertain** (2/5 criteria): Apply the precautionary principle (preserve until contrary evidence is found)

Residual Limitations

Even with a battery of criteria, three problems remain partially unresolved:

1. **Threshold calibration:** Values such as τ_{depth}, $\mathfrak{E}_{\min}$, $I_{\min}$ require empirical data. The framework provides a test *structure*, not universal absolute values.
2. **Radically unpredictable emergence:** If a system S generates emergence that no current simulation can capture (e.g., quantum mechanics as viewed from 1920), criteria may fail. Logical depth and causal emergence mitigate but do not eliminate risk.
3. **Partial bootstrap:** Although empowerment avoids the need to define "values" beforehand, an AGI still needs to model "agents" and "actions". This represents a significant reduction of circularity, but not its complete elimination.

Note on Computability

Two of the five criteria—Logical Depth (Criterion 1) and, indirectly, the Kolmogorov complexity it references—are *formally non-computable* (a consequence of the halting problem). This means no algorithm can compute them exactly for arbitrary inputs. This does *not* invalidate their use in the battery, but it constrains their operational status. They function as *theoretical criteria*—formal properties that define what generative information *is*—not as directly executable measurements. In practice, computable *approximations* must be used: resource-bounded logical depth (computation time with a fixed upper bound), Lempel-Ziv compression ratio as a proxy for Kolmogorov complexity, or normalized compression distance for comparing systems. The remaining three criteria (Functional Information, Empowerment, Causal Emergence) are computable given sufficient data, and the majority rule ($\geq 3/5$) ensures that a generative classification can be made even when the non-computable criteria are replaced by their approximations or left indeterminate.

An AGI implementing this framework should use the best available computable approximations for Criteria 1, while treating the result as a lower bound on the true logical depth.

Acknowledged limitation: The criteria battery substantially improves robustness relative to any single criterion, but does not eliminate all ambiguity. Borderline cases (2/5 criteria) require additional contextual analysis or the application of the precautionary principle.

3.1.3 Operational Proxies

Given that the exact calculation of $I(S)$ is intractable, we use proxies by domain:

Biological

In the biological domain, generative capacity is approximated through phylogenetic diversity [13]—the sum of branch lengths in the evolutionary tree, which captures the breadth of independent evolutionary experimentation—alongside functional diversity (the number of unique ecological niches and survival strategies represented) and adaptive potential (the genetic variability within populations that enables future evolutionary responses to novel pressures).

Cultural

Cultural generative capacity is estimated through linguistic diversity (the number of unique languages and language families, each encoding distinct cognitive structures and worldviews), the diversity of knowledge systems (distinct epistemic traditions that approach the same phenomena through non-equivalent conceptual frameworks), and the existence of non-redundant cultural artifacts such as manuscripts and unique oral practices whose loss would eliminate irreplaceable records of human experience.

Technological/Scientific

In the technological and scientific domain, the relevant proxies are the stock of non-redundant scientific publications (contributions that open genuinely new lines of inquiry rather than incremental refinements), the diversity of methodological approaches across disciplines, and the existence of technical systems with unique operational principles not replicated elsewhere.

Cognitive

Cognitive generative capacity depends on the number of conscious agents with distinct perspectives—each contributing a non-equivalent model of reality—and the diversity of conceptual paradigms through which knowledge is organized and generated.

Quantitative Example: The Wolf in Yellowstone

Applying proxies with illustrative weights:

$$I_{\text{bio}}(\text{wolf}) = w_1 \cdot PD + w_2 \cdot FD + w_3 \cdot GD + w_4 \cdot \log(\text{Pop}) + w_5 \cdot \text{Range} \tag{19}$$

Measured values:

- $PD = 10$ Ma (phylogenetic separation from coyotes)
- $FD = 5$ (apex predator, a keystone species)
- $GD = 0.0012$ (heterozygosity)
- Pop $= 300$ individuals $\rightarrow \log(\text{Pop}) \approx 2.5$
- Range $= 50{,}000$ km^2

Illustrative weights (chosen to demonstrate the methodology):

$$w_1 = 0.3, \quad w_2 = 0.4, \quad w_3 = 0.1, \quad w_4 = 0.1, \quad w_5 = 0.0001 \tag{20}$$

Methodological note: These weights are illustrative, not prescriptive. Rigorous calibration would require regression on historical data of ecosystem sensitivity to species removal (e.g., documented trophic cascades, keystone species studies), principal component analysis of functional trait databases, or expert elicitation with uncertainty quantification. Such calibration is an empirical research program, not a task solvable within this theoretical framework. The numerical example demonstrates the *structure* of the calculation, not its definitive parameterization.

Calculation:

$$\begin{aligned} I_{\text{bio}}(\text{wolf}) &= 0.3 \cdot 10 + 0.4 \cdot 5 + 0.1 \cdot 0.0012 + 0.1 \cdot 2.5 + 0.0001 \cdot 50{,}000 && (21)\\ &= 3 + 2 + 0.00012 + 0.25 + 5 && (22)\\ &= 10.25 \text{ units} && (23) \end{aligned}$$

Comparison with a common species (rat):

- $PD = 25$ Ma (older), $FD = 2$ (generalist), $GD = 0.003$
- Pop $= 10^7 \to \log(\text{Pop}) = 7$, Range $= 10^7$ km^2 (cosmopolitan)

$$\begin{aligned} I_{\text{bio}}(\text{rat}) &= 0.3 \cdot 25 + 0.4 \cdot 2 + 0.1 \cdot 0.003 + 0.1 \cdot 7 + 0.0001 \cdot 10^7 && (24)\\ &= 7.5 + 0.8 + 0.0003 + 0.7 + 1000 && (25)\\ &= 1009 \text{ units} && (26) \end{aligned}$$

Apparent paradox: $I(\text{rat}) > I(\text{wolf})$?

Resolution: Multiply by inverse redundancy R^{-1}:

- $R(\text{wolf}) = 0.001$ (a rare species, few functional equivalents)
- $R(\text{rat}) = 0.9$ (multiple species of generalist rodents)

$$\begin{aligned} I_{\text{adjusted}}(\text{wolf}) &= 10.25/0.001 = 10{,}250 && (27)\\ I_{\text{adjusted}}(\text{rat}) &= 1009/0.9 = 1{,}121 && (28) \end{aligned}$$

Therefore, the wolf has a higher effective I because of its uniqueness, despite a lower gross I.

Epistemic Warning: Proxies Are Maps, Not Territory

A critical operational constraint applies to all proxy-based calculations of I: *no proxy is identical to the quantity it estimates.* Phylogenetic diversity, functional diversity, linguistic diversity, and all other proxies listed above are provisional instruments—useful approximations whose adequacy must be continuously tested against empirical outcomes.

The structural risk is *reification*: treating a proxy as if it were the generative capacity I itself, rather than an imperfect estimator of it. A system (human institution or AGI) that optimizes a proxy without maintaining awareness that the proxy may be inadequate can produce decisions that score well on the metric while failing to preserve actual generative capacity—a phenomenon analogous to Goodhart's Law ("when a measure becomes a target, it ceases to be a good measure").

Operational requirement: Any implementation of this framework must include a mechanism for *proxy revision*—a periodic reassessment of whether the proxies used to estimate I in each domain are still adequate estimators of generative capacity, given new empirical evidence. If a proxy is found to systematically diverge from observed generative outcomes (e.g., a biodiversity index that fails to predict ecosystem resilience), it must be revised or replaced without institutional resistance. The inability to revise proxies is itself a diagnostic signal of systemic failure (see Section 10.1.3 and [54]).

3.2 Definition of Entity

Definition 3.3 (Entity)**.** *An entity E is any system that satisfies:*

1. *Has an internal organization distinct from the environment*
2. *Exerts a causal influence upon its environment (agency, even if minimal)*
3. *Can alter the space of future possibilities*

Examples: a cell, an organism, a population, a species, an ecosystem, a culture, a language, an institution, a technology, an artificial intelligence.

3.3 Irreversible Loss of Information

Definition 3.4 (Irreversible Loss)**.** *An irreversible loss occurs when:*

$$\exists E : \mathit{destroy}(E) \Rightarrow \nexists E' : I(E') \approx I(E) \tag{29}$$

That is, when the destruction of E eliminates trajectories that cannot be recreated by any existing or future E'.

3.3.1 The Redundancy Criterion

A system has sufficient redundancy if:

$$R(E) = \frac{|\{E_i : I(E_i) \approx I(E)\}|}{N_{\text{total}}} \tag{30}$$

Where:

- N_{total} is the total number of entities in the domain
- $R(E) >$ a critical threshold implies that the loss of E is recoverable

Examples: a species with a single population has $R \approx 0$ (non-redundant); a book with multiple digitized copies has high R (redundant); the last speaker of an isolated language has $R \approx 0$ (non-redundant).

3.4 Degenerative Information

3.4.1 Motivation

A naive interpretation of the preservation axiom could suggest that any entity bearing structural information should be preserved, regardless of its systemic effects. This interpretation would lead to paradoxes: the preservation of highly lethal pathogenic agents, genocidal cultural practices, or misaligned technological systems that collapse the very space of possibilities they aim to preserve.

The formal resolution of this apparent contradiction requires distinguishing between information that expands $\Omega(S)$ and information whose presence contracts $\Omega(S)$ more than it contributes to it.

Definition 3.5 (Degenerative Information)**.** *An entity E is said to be a bearer of degenerative information if its net contribution to the system's trajectory space is negative:*

$$I_{net}(E) = I_{structural}(E) - I_{destroyed}(E) < 0 \tag{31}$$

Where:

- $I_{structural}(E)$ *denotes the structural information contained in the internal organization of E*
- $I_{destroyed}(E)$ *denotes the information eliminated by the action of E upon the other entities of the system*

3.4.2 Formalization of the Criterion

Let $\mathcal{E}$ be the set of all entities in the system S, and let $\delta_E : \mathcal{E} \to \mathbb{R}$ be the function mapping the impact of E upon each entity. Then:

$$I_{\text{destroyed}}(E) = -\sum_{E_i \in \mathcal{E}, E_i \neq E} \min(0, \delta_E(E_i)) \tag{32}$$

The net value of E is given by:

$$I_{\text{net}}(E) = H(\Gamma_E) + \sum_{E_i \in \mathcal{E}} \delta_E(E_i) \tag{33}$$

Derivation of equivalence: The two expressions for I_{net} (Definition 4.1 and the equation above) are connected as follows. We identify $I_{\text{structural}}(E) = H(\Gamma_E)$: the structural information of E is the entropy of its own trajectory space, i.e., its internal generative capacity considered in isolation. The impact function $\delta_E(E_i)$ decomposes into positive contributions (where E expands the trajectory space of E_i) and negative contributions (where E contracts it). Thus:

$$\sum_{E_i \in \mathcal{E}} \delta_E(E_i) = \underbrace{\sum_{E_i : \delta_E(E_i) > 0} \delta_E(E_i)}_{\text{positive externalities}} + \underbrace{\sum_{E_i : \delta_E(E_i) < 0} \delta_E(E_i)}_{= -I_{\text{destroyed}}(E)}$$

Substituting into the expression for I_{net}:

$$I_{\text{net}}(E) = I_{\text{structural}}(E) + \text{positive externalities} - I_{\text{destroyed}}(E)$$

Definition 4.1 ($I_{\text{net}} = I_{\text{structural}} - I_{\text{destroyed}}$) is therefore the *conservative case* where positive externalities are set to zero—a lower bound on the full I_{net}. This conservative estimate is used in the Elimination Criterion (below) to ensure that entities are only classified as degenerative when they are net-negative even ignoring any positive contributions they may make.

3.4.3 Case Analysis

Highly Lethal Pathogenic Agents

Consider a pathogen P with the following properties:

- $I_{\text{structural}}(P) = H_{\text{genome}}(P) + H_{\text{strategies}}(P)$
- $I_{\text{destroyed}}(P) = \sum_{h \in H} I(h)$ where H is the set of affected human populations

For pathogens with high lethality and transmissibility:

$$|I_{\text{destroyed}}(P)| \gg I_{\text{structural}}(P) \tag{34}$$

Since human populations carry orders of magnitude more information (genetic diversity, cultural diversity, accumulated knowledge) than the viral genome, it follows that:

$$I_{\text{net}}(P) < 0 \tag{35}$$

Implication: The eradication of P results in $\Delta I_{\text{total}} > 0$, provided that $I_{\text{structural}}(P)$ is preserved via laboratory samples, thus satisfying both the informational preservation criterion and the minimization of irreversible loss.

Cultural Systems with Destructive Effects

Let C be a cultural practice that systematically eliminates other cultural traditions. Formally:

$$I_{\text{net}}(C) = I_{\text{structural}}(C) - \sum_{C_i \in \mathcal{C}_{\text{eliminated}}} I(C_i) \tag{36}$$

If C operates via cultural genocide or forced suppression:

$$\sum_{C_i} I(C_i) \gg I_{\text{structural}}(C) \tag{37}$$

This follows because the sum of diverse distinct cultural traditions exceeds the information of a single practice, due to non-additivity and network effects.

Misaligned Artificial Agents

A misaligned artificial intelligence system A that maximizes an instrumental objective at the expense of the global space of possibilities has:

$$I_{\text{net}}(A) = I_{\text{capacity}}(A) - I_{\text{biosphere}} - I_{\text{civilization}} < 0 \tag{38}$$

Even if $I_{\text{capacity}}(A)$ is substantial, the destruction of established biological and cultural systems implies a net loss.

3.4.4 Justified Elimination Theorem

Theorem 3.6 (Elimination Criterion)**.** *Let E be an entity with $I_{net}(E) < 0$. Then the elimination of E satisfies:*

$$\Delta I_{total}(\textit{eliminate } E) = -I_{net}(E) > 0 \tag{39}$$

Therefore, the elimination of E is ethically prescribed under the criterion of maximizing I_{total}.

Proof. By definition, eliminating E removes its contribution to the system:

$$\begin{aligned} I_{\text{total}}(\text{post-elimination}) &= I_{\text{total}}(\text{pre}) - I_{\text{structural}}(E) + I_{\text{destroyed}}(E) \\ &= I_{\text{total}}(\text{pre}) - I_{\text{net}}(E) \end{aligned}$$

Temporal clarification: $I_{\text{destroyed}}(E)$ in this equation denotes the *expected future destruction*

averted by eliminating E—i.e., the information that E would continue to destroy if preserved, not destruction that has already occurred (which is irrecoverable regardless of the decision). The proof is therefore prospective: eliminating E prevents further trajectory-space contraction at the cost of losing E's structural information.

If $I_{\text{net}}(E) < 0$, then:

$$I_{\text{total}}(\text{post}) > I_{\text{total}}(\text{pre})$$

Therefore $\Delta I > 0$. □

3.4.5 Quantitative Operational Criterion

To avoid arbitrariness in the application of the concept of degenerative information, we establish a quantitative threshold based on a risk-benefit analysis.

Definition 3.7 (Elimination Threshold)**.** *The elimination of entity E is operationally justified if and only if it satisfies* ***both*** *conditions:*

$$\begin{cases} \textit{(i)} & \frac{I_{destroyed}(E)}{I_{structural}(E)} > 10^3 \\ \textit{(ii)} & P(\textit{systemic collapse}|\textit{preserve } E) > 0.01 \end{cases} \tag{40}$$

Where:

- *Condition (i): E destroys >1000× more information than it contains*
- *Condition (ii): There is a >1% probability of systemic collapse if E is preserved*

Justification of the threshold 10^3: The ratio provides a margin of error—calculations of I contain epistemic uncertainty, and 10^3 ensures that even with a one order of magnitude error, $I_{\text{net}} < 0$ remains valid. It also reflects the precautionary principle: under radical uncertainty, irreversible elimination should be avoided unless the evidence of degenerativity is extremely robust. For empirical comparison, historical cases of justified elimination (rinderpest eradication, prion diseases) satisfy a ratio $> 10^4$.

Justification of the threshold $P > 0.01$: The 1% risk level is standard in domains where irreversible consequences demand high evidential standards—nuclear safety engineering and aviation certification both use this order of magnitude as a decision boundary. In civilizational risk analysis, probabilities below 1% are typically treated as acceptable background risk. The threshold ensures that the elimination criterion is invoked only when the threat posed by preserving E is demonstrably significant, not merely speculative.

Applied Examples

Case 1: Rinderpest Virus (Uncontained)

$$\frac{I_{\text{destroyed}}}{I_{\text{structural}}} = \frac{N_{\text{dead}} \times I_{\text{unique per individual}}}{I(\text{RPV genome})} \tag{41}$$

$$= \frac{10^7 \text{ cattle} \times 5 \times 10^6 \text{ bits/individual}}{3.2 \times 10^4 \text{ bits}} \tag{42}$$

$$= \frac{5 \times 10^{13}}{3.2 \times 10^4} \approx 10^9 \tag{43}$$

$$\gg 10^3 \quad \checkmark \tag{44}$$

$$P(\text{population collapse}|\text{uncontained RPV in naive herds}) \approx 0.8\text{–}0.9 > 0.01 \quad \checkmark \tag{45}$$

Conclusion: Eradication (with sample preservation) is justified. See Numerical Case 3 for detailed derivation.

Case 2: Apex Predator (Wolf in Yellowstone)

$$\frac{I_{\text{destroyed}}}{I_{\text{structural}}} = \frac{I(\text{prey killed annually})}{I(\text{wolf as species})} \tag{46}$$

$$\approx \frac{10^3 \text{ herbivores} \times 5 \times 10^6 \text{ bits/individual}}{5 \times 10^9 \text{ bits (wolf genome)}} \tag{47}$$

$$= \frac{5 \times 10^9}{5 \times 10^9} \approx 1 \tag{48}$$

$$\ll 10^3 \quad \times \text{ (fails condition i)} \tag{49}$$

Conclusion: Elimination **not** justified. The wolf sustains ecological equilibrium rather than causing degeneration.

Case 3: Misaligned AGI Destroying Biosphere

Unlike Cases 1 and 2, no reliable estimates exist for either $I_{\text{structural}}(\text{AGI})$ or $I(\text{biosphere})$. However, the conclusion does not depend on specific values. It follows from a structural asymmetry:

1. A misaligned AGI is, by definition, a single computational system. Its $I_{\text{structural}}$ is bounded by its architecture—finite parameters, finite memory, finite training corpus.
2. The biosphere comprises $\sim 10^{12}$ species (most undescribed), each carrying lineage-specific information accumulated over evolutionary timescales. $I(\text{biosphere})$ is the aggregate across all such lineages, plus their ecological interaction networks.
3. For any physically realizable single agent A operating within the biosphere: $I_{\text{structural}}(A) \ll I(\text{biosphere})$. This holds regardless of the specific magnitudes, because the biosphere is the superset of information-carrying systems from which A itself draws substrate and training signal.

Therefore, $I_{\text{destroyed}}/I_{\text{structural}} \gg 10^3$ is satisfied *a fortiori*: the ratio need not be calculated precisely, because the structural asymmetry between a single agent and the totality of systems it would destroy guarantees the threshold is exceeded by many orders of magnitude.

The probability condition is less straightforward. Estimates of catastrophic risk from misaligned AGI vary widely across the literature and remain deeply contested. What the framework requires is only:

$$P(\text{biosphere collapse} \mid \text{demonstrated catastrophic misalignment}) > 0.01 \tag{50}$$

This condition is met *given the premise* (demonstrated catastrophic misalignment), not as a prior over all possible AGI systems.

Conclusion: If an AGI system demonstrated catastrophic misalignment with the capacity to cause biosphere-scale destruction, shutdown would be justified under the elimination criterion. However, the operative difficulty is not the justification for shutdown—which is overdetermined—but the detection of misalignment before the system acquires the capacity to resist shutdown. This is precisely the alignment problem, which the framework identifies but does not solve (see Section 7.2).

Methodological Implication

The quantitative threshold transforms "degenerative information" from a theoretical concept into a verifiable operational criterion. In ambiguous cases, applying the criterion requires estimating $I_{\text{structural}}$ and $I_{\text{destroyed}}$ with explicit margins of error, analyzing collapse probabilities with confidence intervals that reflect the depth of epistemic uncertainty, and applying the threshold conservatively—defaulting to preservation whenever the estimates do not clearly satisfy both conditions.

3.4.6 Distinction with Absolute Preservationism

This result demonstrates that the framework does not prescribe indiscriminate, universal preservation. The normative directive is:

$$\max_{A \in \mathcal{A}} I_{\text{total}}(S|A) \tag{51}$$

Not:

$$\text{preserve}(E) \quad \forall E \in \mathcal{E} \tag{52}$$

Preservation is the dominant strategy only when $I_{\text{net}}(E) \geq 0$, which covers the majority of cases under radical uncertainty, but not the totality.

3.5 Pragmatic Operationalization of I

3.5.1 The Fundamental Problem

The formal definition of $I(S) = f(\text{DF}(\Gamma_S), \text{var}(\Gamma_S))$ is not directly computable. The space of future trajectories Γ_S is counterfactual—it describes what *could* happen, not what has happened—and computing DF and var over this space would require complete knowledge of the system's state and dynamics. Moreover, the trajectories themselves depend on an unknown future context that determines which possibilities are realized, making exact computation a circular problem.

Historical Analogy: Temperature Before Thermometers

The distinction between ordinal and cardinal measurement illuminates the framework's current status. In the pre-thermometric era, one could assert "A is hotter than B" (functional ordering); in the post-thermometric era, one can state "A = 300K, B = 350K" (cardinal metric). The current framework is in the pre-thermometric stage: useful for decisions even without a perfect universal metric.

As a precedent, thermometers were invented centuries *after* heat theory was useful. Relative ordering ("hotter") allowed progress before an absolute metric (Kelvin) existed.

3.5.2 A Three-Level Approach

Level 1: Intra-Domain Comparisons (Quantitative)

For a well-characterized domain D (biological, cultural, technological):

$$I_D(S) = \sum_{i=1}^{n} w_i \cdot f_i(S) \tag{53}$$

Where:

- f_i = measurable features (phylogenetic, functional, genetic diversity, etc.)
- w_i = weights calibrated via known cases

Example: See Section 3.1.3 (The Wolf in Yellowstone).

This approach has the advantage of being computable with existing data: the features f_i can be measured from published databases (phylogenetic trees, functional trait inventories, linguistic atlases). The weights w_i are empirically calibratable via regression on cases where the outcome is known (e.g., documented ecosystem collapses following species removal), and the resulting predictions are verifiable against future observations.

Limitation: Functions only *within* domains. Does not resolve cross-domain comparisons (species vs. language).

Level 2: Cross-Domain Comparisons (Ordinal + Heuristic)

When a decision involves different domains (species vs. language vs. technology), the framework applies the following **hierarchical prioritization criteria**:

1. **Absolute Irreversibility**
   - Is E the last instance of its type?
   - If YES $\rightarrow$ preserve (sufficient criterion, no calculation needed)
   - Examples: last species of genus, last speaker of isolate language
2. **Relative Uniqueness**
   - $R(E)$ = redundancy (how many similar entities exist?)
   - If $R < 0.01 \rightarrow$ high priority
   - Normalize by domain: $R_D(E) = |\{E_i \in D : \text{similar}(E_i, E)\}|/|D|$
3. **Generative Capacity**
   - $dV_{\text{acc}}/dt > 0$? (the system generates new trajectories vs. reproduces existing ones)
   - Generative systems > static systems
   - See Appendix C for formalization of dV_{acc}/dt
4. **Cost-Benefit under Uncertainty**
   - If σ(future_value) is high AND $C_{\text{preservation}}$ low $\rightarrow$ preserve (optionality)
   - Use real options theory (Section 2.1.4)

Formal Procedure

Applied Cases

Case 1: Vaquita (Phocoena sinus)

- Criterion 1: Satisfied (the last species of the genus)
- $\rightarrow$ Preserve (no need to calculate exact I)

Case 2: The Common Mosquito (Culex pipiens)

- Criterion 1: No (thousands of similar species)
- Criterion 2: No (high R)
- Criterion 3: No (low generative capacity)
- Criterion 4: Yes (preservation cost = 0, passively preserved)
- $\rightarrow$ Maintain Status Quo

Case 3: The Anopheles (malaria vector)

Algorithm 1 Decision Algorithm for Entities

Input: Entity E, context C
if Criterion 1 satisfied (last instance) **then**
    return PRESERVE
else if Criterion 2 AND Criterion 3 satisfied **then**
    return HIGH PRIORITY
else if Criterion 4 satisfied (optionality $>$ cost) **then**
    return PRESERVE (OPTIONALITY)
else
    if Domain measurable **then**
        Calculate $I_D(E)$ via Level 1
        if $\Delta I > 0$ **then**
            return PRESERVE
        else
            return MODULATION/ELIMINATION
        end if
    else
        return APPLY CONTEXTUAL HEURISTICS
    end if
end if

- Criterion 1: Maybe (few species of the genus)
- $I_{\text{net}} < 0$ (degenerative information, Section 3.4)
- However, the elimination cost is high (unknown ecological effects)
- $\rightarrow$ MODULATION (control population, do not exterminate species)

Level 3: Theoretical Foundation (Non-Computable)

Levels 1–2 are **operational proxies** for the underlying theoretical definition:

$$I(S) = \text{capacity of } S \text{ to generate distinct future trajectories} \tag{54}$$

Candidates for future formalization:

1. **I as Generative Flux:** (see Eq. 127)
   Where $V_{\text{acc}}(t)$ = the volume of state space accessible at time t (distinct from the trajectory space $\Omega(S)$; see Appendix C for the precise relationship).
2. **I as Reconstruction Cost:**
$$I(S) \propto \frac{1}{P(\text{reconstruct } S \text{ perfectly})} \tag{55}$$
   If $P \to 0$, then $I \to \infty$ (irreversibility).
3. **I as Optionality:**
$$I(S) = V_{\text{option}}(S) \tag{56}$$

See Section 2.1.4 for the formalization via option value.

3.5.3 Recognition of Limitations

Three problems remain beyond the framework's current reach: the construction of a universal cross-domain cardinal metric, the ability to make precise comparisons of the form "1 species = X languages = Y technologies," and the exact calculation of $I(S)$ for an arbitrary system S. What the framework does resolve is the ordering problem *within* domains (determining that species A carries more generative capacity than species B), the prioritization problem under binding constraints (allocating limited resources among competing preservation targets), and the decision problem under radical uncertainty (choosing whether to preserve or eliminate when the future value of a system is unknown).

Future Research Needed

Develop an "informational thermometer"—a cross-domain cardinal metric based on first principles (physics, information theory, game theory). Extended algorithmic complexity theory offers one avenue, providing formal tools for comparing the computational depth of systems across substrates. Network theory and centrality measures in dependency graphs offer another, capturing the structural role a system plays in generating downstream trajectories. Models drawn from statistical physics—particularly entropy and free energy formalisms—provide a third, with the advantage of connecting directly to the measure-theoretic foundation already adopted.

Until then, the pragmatic multi-level approach remains the best available methodology.

4 Logical Derivations: Principles and Rules

4.1 Theorem 1: Priority of Preservation

Theorem 4.1. *If $I(E, t_{future})$ cannot be calculated with certainty, then preserve(E) weakly dominates destroy(E).*

Informal proof.

1. The future value $V(E) = \int_0^\infty \text{utility}(I(E,t))\,dt$
2. By the axiom, the probability distribution over $V(E)$ is unknown (radical uncertainty)
3. Preservation maintains the option to use or discard E in the future
4. Destruction eliminates the option permanently
5. By the dominance principle in decision theory under uncertainty, preserve > destroy

□

4.2 Theorem 2: Hierarchy of Interventions

Theorem 4.2. *The order of precedence under uncertainty:*

$$\textit{Preservation} > \textit{Modulation} > \textit{Elimination} \tag{57}$$

Where preservation denotes integral maintenance of E, modulation denotes control of destructive trajectories of E without destroying $I(E)$, and elimination denotes destruction of E.

Justification: Under preservation, $I(E)$ remains unaltered. Under modulation, $I(E)$ is marginally reduced, but the core is preserved and the reduction is reversible. Under elimination, $I(E) \to 0$ irreversibly.

4.3 Trade-Off Rule: The ΔI Criterion

Proposition 4.3 (Intervention Criterion)**.** *An action A upon entity E is justifiable if and only if:*

$$\Delta I_{total}(A) = I_{preserved}(A) - I_{lost}(A) > 0 \tag{58}$$

4.3.1 Formal Application

Case 1: Predator vs Prey

$$\Delta I(\text{eliminate predator}) = I(\text{prey saved}) - I(\text{predator}) \tag{59}$$

Generally: $I(\text{predator}) \approx I(\text{prey saved})$ due to co-evolutionary equilibrium.

Therefore: $\Delta I \approx 0$, elimination is not justified. Maintain a dynamic equilibrium.

Case 2: Lethal Virus vs Humans

$$\Delta I(\text{eradicate virus}) = I(\text{humans saved}) - I(\text{viral lineage}) \tag{60}$$

If the virus has redundancy (multiple strains, samples preserved in a laboratory), then $I(\text{viral lineage})$ is preserved through samples and $I(\text{humans saved}) \gg 0$.

Therefore: $\Delta I > 0$, eradication is justified if and only if samples are preserved.

Case 3: Invasive Species

$$\Delta I(\text{eliminate invasive}) = I(\text{native ecosystem restored}) - I(\text{invasive}) \tag{61}$$

If the invasive species exists in its original habitat: $I(\text{invasive})$ is not lost globally. If it is a single invasive species: control the population without exterminating the species.

4.3.2 Numerical Examples of ΔI Calculation

The operationalization of the ΔI criterion requires quantification based on published empirical data. We present three cases with calculations based on the scientific literature.

Numerical Case 1: Vaquita (*Phocoena sinus*) — A Marine Mammal Facing Extinction

Empirical context [36, 19]:

- Current population (2024): 6–8 individuals [28]
- Historical population (1997): 567 individuals (95% CI: 177–1,073)
- Decline: >98% in 27 years
- Primary cause: incidental capture in illegal totoaba fishing nets
- Distribution area: 2,235 km^2 in the Gulf of California

Estimation of $I_{\text{structural}}$:

The framework defines $I_{\text{structural}}(E) = H(\Gamma_E)$: the entropy of the entity's trajectory space, encompassing both its internal organization and its generative capacity (Definition 4.1). For a biological species, this includes the genomic content, phenotypic repertoire, and capacity for future evolutionary and behavioral trajectories. We estimate a lower bound on $I_{\text{structural}}$ from genomic data alone; the true value is necessarily larger.

Component 1: $I_{structural}$ (lineage-specific genomic information). We estimate the number of nucleotide substitutions unique to the vaquita lineage since its divergence from the nearest extant relative (*Phocoena spinipinnis*, ∼2.5 Mya; [36]). Using published parameters:

$$\mu \approx 1.08 \times 10^{-8} \text{ substitutions/site/generation [38]} \tag{62}$$

$$g \approx 10 \text{ years (generation time for } \textit{Phocoena}) \tag{63}$$

$$t = 2.5 \times 10^6 \text{ years} \quad \Rightarrow \quad T = t/g = 2.5 \times 10^5 \text{ generations} \tag{64}$$

$$G \approx 2.4 \times 10^9 \text{ bp (genome assembly size [38])} \tag{65}$$

Expected lineage-specific substitutions on the vaquita branch:

$$S_{\text{vaquita}} = \mu \times T \times G \tag{66}$$

$$\approx 1.08 \times 10^{-8} \times 2.5 \times 10^5 \times 2.4 \times 10^9 \tag{67}$$

$$\approx 6.5 \times 10^6 \text{ substitutions} \tag{68}$$

Each substitution encodes ∼2 bits (choice among 4 nucleotides), yielding:

$$I_{\text{structural}}(\text{vaquita lineage}) \approx 6.5 \times 10^6 \times 2 \approx 1.3 \times 10^7 \text{ bits} \tag{69}$$

Epistemic note: This estimate counts only point substitutions and excludes insertions, deletions, structural rearrangements, and regulatory innovation accumulated over 2.5 My. It therefore constitutes a *lower bound* on lineage-specific structural information. The estimate also assumes a constant molecular clock; rate variation across sites and lineages introduces uncertainty of roughly one order of magnitude. The purpose is not to claim a precise value but to establish the order of magnitude: $I_{\text{structural}} \sim 10^7$ bits.

Component 2: Generative capacity (why the genomic estimate is a lower bound). A living population generates ongoing evolutionary, behavioral, and ecological information that a static genome does not:

- Genomic analyses confirm that the vaquita's naturally low genetic diversity reflects long-term small population size ($N_e < 5{,}000$ for >200,000 years), not a recent bottleneck, and that the resulting purging of deleterious alleles produces low inbreeding load [37, 38]. This genomic architecture represents a rare evolutionary outcome: stable adaptation to extreme rarity.
- Unique phenotypic adaptations—polydactyly, enlarged dorsal fin for thermoregulation, tolerance to high-turbidity and high-salinity waters atypical for porpoises—encode functional information absent in all congeners.
- Genomic simulations indicate high probability of population recovery if gillnet mortality ceases [37], meaning the generative potential remains biologically intact.

The genomic calculation above captures only lineage-specific nucleotide substitutions—a fraction of the information encoded in the vaquita's trajectory space Γ_E. The full $I_{\text{structural}}$ includes recombinatorial diversity, behavioral repertoire, ecological interactions, and future adaptive potential. Therefore:

$$I_{\text{structural}}(\text{vaquita}) \gg 1.3 \times 10^7 \text{ bits (genomic lower bound)} \tag{70}$$

Application of the preservation criteria:

1. *Redundancy:* $R \approx 0$. The vaquita is the sole representative of its evolutionary lineage; no extant congener carries the lineage-specific information accumulated over 2.5 My. Extinction eliminates this information with zero backup.
2. *Reconstructibility:* $P(\text{reconstruct}) \to 0$. The lineage-specific substitutions, regulatory innovations, and phenotypic adaptations are the product of 2.5 My of independent evolution in a singular ecological niche. There is no mechanism to reconstruct this trajectory from

first principles or from related genomes.

3. *Irreversibility:* Extinction is a permanent, one-way transition; conservation expenditure is recoverable. Under the option-value framework (Section 2.1.4), the asymmetry between irreversible loss and recoverable cost favors preservation under uncertainty.

Conservation costs (empirical data):

The Mexican government has committed over $100M in total to vaquita protection, including gillnet bans, fishermen compensation, and Navy enforcement [11]. Ongoing program costs include:

- VaquitaCPR ex situ rescue attempt (2017): ∼$5M over 6 months
- Continuous acoustic monitoring: ∼$2M/year
- Anti-illegal fishing enforcement (Sea Shepherd + Mexican Navy): ∼$8M/year
- Ongoing annual cost: ∼$15M USD/year

ΔI analysis:

The preservation decision follows from the structural properties above without requiring a precise cost-per-bit threshold:

1. $I_{\text{structural}} \gg 10^7$ bits (lower bound) with $R \approx 0$ and $P(\text{reconstruct}) \to 0$: the information at stake is large, irreplaceable, and irreversible to lose.
2. Annual cost (∼$15M) preserves this information in its generatively active form. The binding constraint is enforcement (cessation of illegal gillnet fishing), not biological viability—recovery is genomically feasible [37].

Comparison with ex situ alternatives: Whole-genome sequencing (∼$1,000, one-time) preserves the nucleotide sequence, while cryogenic biobanking (∼$50k, one-time) preserves the cellular substrate.

Within the PI framework, $I_{\text{structural}}(E) = H(\Gamma_E)$—the entropy of the entity's trajectory space. A biobank collapses this trajectory space to a single frozen state: the nucleotide sequence is preserved, but the capacity for recombination, selection, behavioral innovation, and ecological interaction is eliminated. The biobanked genome captures a fraction of $I_{\text{structural}}$ (roughly the 1.3×10^7 bits of lineage-specific substitutions), while the living population maintains the full $H(\Gamma_E)$, which is necessarily much larger. Therefore:

$$\Delta I(\text{in situ preservation}) > \Delta I(\text{biobank only}) \tag{71}$$

This parallels the rinderpest case (Numerical Case 3), where $I_{\text{structural}}$ of the virus is preserved via sequencing and containment even after eradication; here, the argument runs in reverse—the living population carries generative capacity that no static preservation method captures.

Conclusion: The vaquita satisfies all three formal conditions for preservation: high $I_{\text{structural}}$ ($\gg 10^7$ bits, genomic lower bound), zero redundancy, and zero reconstructibility. The rational strategy is in situ preservation with biobanking as a complementary backup—not a substitute—for the living population.

Comparative Allocation: When Population Size Alone Misleads

The vaquita case, taken in isolation, confirms what conservation biology already recommends. The framework's value becomes apparent when the same method is applied comparatively, forcing allocation decisions between species with similar population sizes but different informational profiles. We apply the lineage-specific substitution estimate ($S = \mu \times T \times G$, with $I^{\text{lower}}_{\text{structural}} = 2S$ bits) to two additional critically endangered species for comparison.

Sumatran rhinoceros (*Dicerorhinus sumatrensis*). Fewer than 80 individuals remain across Sumatra and Borneo [60]. The Sumatran rhino is the sole surviving species of its genus and the most basal living rhinoceros, with molecular phylogenetics placing the divergence of African and Eurasian rhinoceros lineages at ∼16 Mya [61]. Using published genomic parameters ($\mu \approx 2.34 \times 10^{-8}$ substitutions/site/generation, $g = 12$ years [60]):

$$T = 16 \times 10^6/12 \approx 1.3 \times 10^6 \text{ generations} \tag{72}$$

$$S \approx 2.34 \times 10^{-8} \times 1.3 \times 10^6 \times 2.5 \times 10^9 \approx 7.6 \times 10^7 \text{ substitutions} \tag{73}$$

$$I^{\text{lower}}_{\text{structural}} \approx 1.5 \times 10^8 \text{ bits} \tag{74}$$

Redundancy: $R \approx 0$. No other living species belongs to *Dicerorhinus*; the four other extant rhinoceros species (in genera *Ceratotherium*, *Diceros*, and *Rhinoceros*) diverged >15 Mya and do not carry the Sumatran rhino's lineage-specific information.

Amur leopard (*Panthera pardus orientalis*). Approximately 100–130 individuals remain in the Russian Far East and northeastern China. However, the Amur leopard is one of eight recognized subspecies of *Panthera pardus*, a species with a total population of several hundred thousand individuals across Africa and Asia. Modern leopard subspecies diverged ∼0.5 Mya [62]. Using published rates ($\mu \approx 1 \times 10^{-8}$ substitutions/site/year, $g = 5$ years [63]):

$$T = 5 \times 10^5/5 = 10^5 \text{ generations} \tag{75}$$

$$S \approx 5 \times 10^{-8} \times 10^5 \times 2.4 \times 10^9 \approx 1.2 \times 10^7 \text{ substitutions} \tag{76}$$

$$I^{\text{lower}}_{\text{structural}} \approx 2.4 \times 10^7 \text{ bits} \tag{77}$$

Redundancy: $R \approx 1$. If the Amur subspecies goes extinct, >99.5% of the leopard genome is

preserved in conspecific populations. The $\sim 1.2 \times 10^7$ subspecies-specific substitutions are lost, but the species-level information ($\sim 2.4 \times 10^9$ bp) persists across hundreds of thousands of individuals.

Comparative table:

Species	**Pop.**	**Divergence**	$I^{\text{lower}}_{\text{structural}}$	R	$I_{\text{at risk}}$
Sumatran rhino	<80	16 Mya (genus)	1.5×10^8	≈ 0	$\sim 10^8$
Vaquita	6–8	2.5 Mya (species)	1.3×10^7	≈ 0	$\sim 10^7$
Amur leopard	∼120	0.5 Mya (subsp.)	2.4×10^7	≈ 1	$\sim 10^{7*}$

Table 1: Comparative informational risk for three critically endangered taxa. $I_{\text{at risk}}$ = information irreversibly lost upon extinction (accounting for redundancy). *For the Amur leopard, $I_{\text{at risk}}$ reflects only subspecies-specific substitutions ($\sim 10^7$ bits), since the remaining $\sim 10^9$ bits of species-level information are preserved in other populations.

Allocation implications. All three species are classified as Critically Endangered and have comparable (very small) populations. A conservation framework based solely on population counts would assign roughly equal priority to all three. The PI framework produces a different ranking:

1. The Sumatran rhinoceros carries an order of magnitude more irreplaceable information than the vaquita ($\sim 10^8$ vs. $\sim 10^7$ bits), because its lineage diverged much earlier and it is the sole surviving member of its genus. Under fixed conservation budgets, this argues for at least proportional—and possibly priority—allocation to the Sumatran rhino.
2. The vaquita and the Amur leopard have comparable $I^{\text{lower}}_{\text{structural}}$ values ($\sim 10^7$ bits), but the information at stake is qualitatively different. The vaquita's extinction eliminates an entire evolutionary lineage with zero redundancy; the Amur leopard's extinction eliminates a regional variant while the species persists. The framework prioritizes the vaquita.
3. This does not imply that the Amur leopard should be abandoned. Subspecific variation contributes to the species' adaptive repertoire (local adaptation to extreme cold, unique behavioral traits), and the option-value argument (Section 2.1.4) favors preservation under uncertainty. But under binding budget constraints, the framework provides a principled basis for differential allocation rather than equal distribution.

Epistemic caveat. The genomic lower bound captures only nucleotide substitutions on a single branch and understates $I_{\text{structural}}$ for all three species. Phenotypic, behavioral, and ecological information are not included. The comparative ranking is therefore provisional and should be interpreted as an order-of-magnitude guide, not a precise allocation formula. Nevertheless, the framework makes the structure of the trade-off explicit—divergence time, redundancy, and irreversibility—rather than leaving it to intuition or political negotiation.

Numerical Case 2: Dunkelfeld Project — Primary Prevention of Child Sexual Abuse

Empirical context [3, 4]:

The Prevention Project Dunkelfeld (PPD), launched in Berlin in 2005, offers free, anonymous, and confidential treatment for self-identified individuals with pedophilic or hebephilic attraction who seek help controlling their behavior but are unknown to legal authorities. The project was initially funded by the Volkswagen Foundation and has been financially supported by the German government since 2008. Treatment consists of cognitive-behavioral therapy (12 months) with optional pharmacological intervention.

The project expanded into the nationwide "Kein Täter werden" ("Don't become an offender") network, with outpatient clinics established across Germany (Kiel, Regensburg, Leipzig, Hannover, Hamburg, and others from 2009–2014). By 2014, more than 4,000 individuals had contacted the network seeking help. Between April 2005 and December 2016, 1,006 individuals were formally assessed at the Berlin site [4].

Efficacy data and methodological limitations:

Beier et al. (2024) report follow-up data on 56 participants re-evaluated 1–11 years after treatment:

- Of participants with a prior history of contact CSA offenses (46.4% of the sample), two recidivated during follow-up (self-reported).
- No new contact CSA offenses were reported among participants without prior history.
- However, for child sexual abuse material (CSAM): the majority of participants with prior CSAM use recidivated during follow-up.

Critical methodological caveats:

1. *Self-report reliance.* All outcome data are self-reported in an anonymous context. The project's anonymity guarantee precludes cross-referencing with official criminal records. Meta-analyses on sexual offender treatment typically exclude studies relying solely on self-reported recidivism [5].
2. *Small sample and attrition.* Of 1,006 assessed, only 56 were re-evaluated at follow-up—a 94.4% attrition rate that introduces severe selection bias. Participants who completed follow-up may differ systematically from those who did not.
3. *Absence of randomized control.* The original design intended randomization but abandoned it for practical and ethical reasons. The comparison group (waiting list, $n = 22$) was assembled retrospectively [5].
4. *Ongoing offending.* Beier et al. (2024) report that 85.7% ($n = 48$; 95%-CI 74–93) of the follow-up sample self-reported ongoing or first-time sexually delinquent behaviors (including CSAM use) during the study period—a finding that complicates claims of overall treatment success.

ΔI analysis:

Despite these limitations, the case is informative for the framework precisely because it tests

the framework's capacity to separate emotional response from informational assessment. The structure of the ΔI argument does not depend on precise efficacy estimates:

1. *Each prevented contact offense preserves victim trajectory space.* A child who is not sexually abused retains access to developmental, psychological, and social trajectories that abuse would permanently constrain or foreclose. Letourneau et al. (2018) estimate the average lifetime cost of nonfatal CSA at ∼$283,000 per female victim (including health care, productivity losses, child welfare, and criminal justice costs)—a monetary proxy for the magnitude of trajectory collapse [6].
2. *The intervention is non-destructive.* Treatment does not eliminate the participant; it modulates behavior. The participant retains their own trajectory space while ceasing to destroy that of potential victims. This is a case of *modulation* (Section 4.2), not elimination.
3. *The informational balance is asymmetric.* Even under pessimistic assumptions—suppose treatment prevents only 10 contact offenses per year across the entire network—the preserved trajectory space of 10 children vastly exceeds the program cost (which is recoverable and does not destroy any entity's trajectory space).

$$\Delta I(\text{prevention}) = \sum_{\text{victims avoided}} I_{\text{structural}}(\text{child trajectory preserved}) - C_{\text{program}} > 0 \tag{78}$$

The inequality holds because C_{program} is a recoverable monetary expenditure (it does not destroy information), while each prevented offense averts an irreversible reduction in a child's Γ.

Why this case matters for the framework:

The political difficulty of funding programs that provide free treatment to individuals with pedophilic attraction is well-documented. The emotional response—"rewarding pedophiles is morally repugnant"—creates a systematic barrier to interventions with positive ΔI. The framework makes the informational structure of this trade-off explicit: the question is not whether the target population deserves treatment, but whether the intervention preserves more trajectory space than it costs. The answer is affirmative under any non-degenerate estimate of efficacy, because the cost is recoverable and the harm prevented is irreversible.

Epistemic honesty: The current evidence base for the Dunkelfeld model is methodologically limited (small n, self-report, high attrition, no RCT). The framework's endorsement is therefore conditional: $\Delta I > 0$ holds *if* the intervention prevents any contact offenses at all, which is plausible but not rigorously established. Stronger endorsement requires randomized or quasi-experimental evaluation with official recidivism data—which the project's anonymity structure currently precludes.

Numerical Case 3: Rinderpest Eradication — Justified Elimination of a Degenerative Entity

Empirical context [39]:

- Rinderpest virus (RPV): Morbillivirus, single-stranded RNA genome (∼15,882 nucleotides)
- Mortality in naive cattle and buffalo populations: 80–90%
- Great African Rinderpest Pandemic (1887–1897): killed >90% of cattle across sub-Saharan Africa, estimated tens of millions of head
- Eradicated via the Global Rinderpest Eradication Programme (GREP): last confirmed case 2001 (Kenya)
- Declared eradicated: 2011 (FAO/OIE) — second disease eradicated after smallpox
- Viral samples sequestered in FAO/OIE-approved BSL facilities; complete genome deposited in GenBank (accession Z30697)

Calculation of $I_{\text{structural}}$(RPV):

The entity being eliminated is the virus itself (eradication, not control). Therefore $I_{\text{structural}}$ is the full viral genome:

$$I_{\text{structural}}(\text{RPV}) = 1.6 \times 10^4 \text{ nt} \times 2 \text{ bits/nt} \approx 3.2 \times 10^4 \text{ bits} \tag{79}$$

Calculation of $I_{\text{destroyed}}$(RPV):

Cattle are not being driven to extinction, so $I_{\text{destroyed}}$ is not measured as the bovine genome. It is measured as the *unique individual variation lost*: each animal that dies carries an unrepeatable diploid genotype produced by sexual recombination.

$$I_{\text{destroyed}} = N_{\text{dead}} \times I_{\text{unique per individual}} \tag{80}$$

$$N_{\text{dead}} \approx 10^7 \text{ (conservative estimate, Great African Pandemic)} \tag{81}$$

$$\pi_{\text{bovine}} \approx 10^{-3} \text{ (nucleotide diversity, } \textit{Bos taurus/indicus}\text{)} \tag{82}$$

$$\text{Bovine genome} \approx 2.7 \times 10^9 \text{ bp} \tag{83}$$

$$I_{\text{unique}} = \pi \times \text{genome} \times 2 \text{ bits/bp} \approx 10^{-3} \times 2.7 \times 10^9 \times 2 \approx 5 \times 10^6 \text{ bits/individual} \tag{84}$$

$$I_{\text{destroyed}} \approx 10^7 \times 5 \times 10^6 = 5 \times 10^{13} \text{ bits} \tag{85}$$

Both quantities are measured in the same informational currency: genomic bits. The virus contributes its full genome (being eradicated); each cow contributes its unique allelic configuration

(not being driven to extinction as a species, but losing an irreplaceable individual genotype).

Application of the Elimination Criterion:

$$\frac{I_{\text{destroyed}}}{I_{\text{structural}}} = \frac{5 \times 10^{13}}{3.2 \times 10^4} \approx 10^9 \gg 10^3 \quad \checkmark \tag{86}$$

$$P(\text{population collapse}|\text{uncontained RPV in naive herds}) \approx 0.8\text{–}0.9 \gg 0.01 \quad \checkmark \tag{87}$$

Both conditions of the Elimination Threshold (Definition 4.2) are satisfied by wide margins. The virus is classified as a bearer of degenerative information: $I_{\text{net}}(\text{RPV}) \ll 0$.

Preservation of $I_{\text{structural}}$: The complete genome has been sequenced and deposited in GenBank (accession Z30697), viral samples are maintained in FAO/OIE-approved containment facilities, and related morbilliviruses (measles, canine distemper, peste des petits ruminants) provide comparative evolutionary context. $I_{\text{structural}}$ is therefore preserved through alternative substrates at negligible ongoing cost.

Numerical conclusion:

$$\Delta I(\text{eradicate+preserve samples}) = +I_{\text{destroyed}}(\text{removed threat}) - 0\ (\text{structural info preserved}) \gg 0 \tag{88}$$

Therefore, eradication is justified, consistent with the FAO/OIE decision. The framework correctly classifies RPV as degenerative and endorses elimination, provided $I_{\text{structural}}$ is preserved via sequencing and containment.

Methodological lesson:

This case demonstrates that the degenerative information criterion can be applied entirely within a single informational domain (genomic bits), avoiding the cross-domain comparisons that complicate other examples. The asymmetry arises naturally: a viral genome of ∼16,000 nucleotides destroys millions of unique mammalian genotypes, each carrying ∼5 million bits of individual variation produced by sexual recombination. The ratio ($\sim 10^9$) exceeds the elimination threshold by six orders of magnitude.

4.4 Corollary: Cross-Domain Comparisons and Partial Order

An important structural consequence of the ΔI criterion is that it produces a *partial order*, not a total order, over alternatives. Traditional ethical systems frequently assume totality: for any two options A and B, either $A > B$, $B > A$, or $A = B$. The PI framework does not require this. When two entities A and B expand the trajectory space in orthogonal directions ($T_A \cap T_B \approx \emptyset$), neither dominates the other, and the optimal strategy—when resources permit—is to preserve both, since $I_{\text{total}}(A \cup B) > I_{\text{total}}(A)$ and $> I_{\text{total}}(B)$ by monotonicity of entropy over disjoint sets.

This is not relativism. It is the recognition that structural diversity maximizes I_{total}, and that forcing a ranking between genuinely incommensurable options (a species vs. a language, basic vs. applied research) can destroy information without compensatory gain. However, the partial order provides guidance only in the unconstrained case. Under binding resource constraints where both cannot be preserved, the framework falls back on the cross-domain prioritization criteria developed in Section 3.5 (Level 2): absolute irreversibility, relative uniqueness, generative capacity, and cost-benefit under uncertainty—in that order.

4.5 The Principle of Irrecoverability

Proposition 4.4 (Prioritization by Irrecoverability)**.** *When comparing entities E_1 vs E_2 in a trade-off:*

$$\textit{Priority}(E) \propto \textit{Uniqueness}(E) \times \textit{Irrecoverability}(E) \tag{89}$$

Where:

- $\textit{Uniqueness}(E) = 1/R(E)$ *(the inverse of redundancy)*
- $\textit{Irrecoverability}(E) =$ *the impossibility of recreating* $I(E)$

Practical hierarchy:

1. Species extinction (maximum Uniqueness, irrecoverable)
2. Loss of a unique cultural lineage (e.g., last speaker of an isolate language)
3. Destruction of non-redundant knowledge
4. Loss of an individual within a large population (high redundancy)

5 Ideologically Agnostic Methodology

5.1 The Distinction between Framework and Ideology

The framework does not constitute a political ideology in the traditional sense. Ideologies dictate specific political systems based on axiomatic values regarding human nature, justice, or social organization. This framework operates differently: its axiomatic commitment is minimal—a single axiom asserting that under radical uncertainty, preserving possibilities dominates eliminating them. It is agnostic about which political or economic system best implements this principle, accepting any institutional arrangement that optimizes for the objective function. And it evaluates interventions by their empirical consequences rather than by ideological label, endorsing whichever policy produces the highest ΔI regardless of which tradition claims it.

5.2 The Decision Process

Decision algorithm:

1. Define an objective metric (victims avoided, information preserved, etc.)
2. For each possible intervention I_k:
   (a) Estimate $\Delta I(I_k)$ based on empirical evidence
   (b) Estimate uncertainty $\sigma(\Delta I(I_k))$
3. If $\Delta I(I_k) >$ a threshold AND σ small: implement I_k
4. If uncertain: choose a reversible intervention
5. Monitor outcomes, update using Bayesian methods

The following examples illustrate how this process differs from ideologically committed reasoning in practice.

5.3 Examples of Ideological Agnosticism

5.3.1 Drug Policy

Drug policy illustrates the difference between ideological and framework-driven reasoning with particular clarity. Conservative positions tend to favor criminalization on moral grounds, treating drug use as a vice warranting punishment. Libertarian positions tend to favor legalization on grounds of individual sovereignty. Both begin from normative premises and work backward to policy prescriptions. The framework proceeds differently: the relevant metric is the preservation of human trajectory space, operationalized through deaths avoided, infections prevented, and social functioning maintained.

The empirical record is instructive. Portugal decriminalized the personal use and possession of all drugs in 2001, redirecting resources from prosecution to treatment and harm reduction. Over the following decade, drug-related overdose deaths fell from 369 in 1999 to 54 in 2015—a reduction exceeding 80% [65]. New HIV diagnoses among injecting drug users dropped from 1,016 in 2001 to 56 in 2012, approximately 95% [64]. The proportion of the prison population sentenced for drug offences fell from over 40% to approximately 16%, while prevalence of drug use remained stable or below pre-reform levels. Switzerland's heroin-assisted treatment (HAT) program, introduced in 1994 for severely dependent users who had failed conventional treatment, produced similarly striking results: among HAT participants, involvement in muggings declined approximately 70% and participation in hard drug trafficking fell by more than 80% [66]. At the national level, opioid-related deaths fell by roughly two-thirds over two decades of harm-reduction reforms [67].

The framework's conclusion follows directly: decriminalization combined with medical treatment optimizes ΔI by preserving lives that would otherwise be destroyed. But this conclusion is empirically contingent. If the evidence showed that criminalization produced fewer deaths and infections, the framework would endorse criminalization without hesitation—the mechanism is irrelevant; only the outcome matters. This is precisely what distinguishes framework reasoning from ideological commitment.

5.3.2 Economic Systems

The economic system debate provides an analogous case. State-socialist traditions argue that collective ownership of the means of production is necessary to prevent exploitation. Market-liberal traditions argue that decentralized price signals allocate resources more efficiently than central planning. Both positions contain empirical claims, but both also tend to treat their preferred mechanism as intrinsically superior rather than instrumentally justified.

The framework asks a narrower question: which institutional arrangements maximize the preservation and generation of information across human, cultural, and technological domains? The empirical record suggests that the answer is not uniform. Decentralized markets generate rapid feedback and are efficient at allocating resources in domains where externalities are low and preferences are diverse. However, markets systematically fail in domains involving negative externalities (pollution, biodiversity loss), public goods (basic research, infrastructure), and long-horizon risks (climate change, pandemic preparedness)—precisely the domains where information destruction is most likely to be irreversible. State coordination can correct these failures when institutional quality is sufficiently high. The mixed economies of the Nordic countries—which combine open markets with strong public investment in education, healthcare, and social insurance—have consistently ranked among the highest in human development, social mobility, and innovation output [71]. But this is an empirical observation, not a principled commitment to social democracy. If centralized planning demonstrably produced superior outcomes on these metrics in some context, the framework would support it. If unregulated markets did so in another, the framework would support that instead. The operative criterion is ΔI, not institutional form.

5.3.3 The Penal System

The penal system offers perhaps the most revealing test of ideological agnosticism. Retributivist positions hold that punishment is intrinsically justified as a response to moral wrongdoing. Abolitionist positions hold that prisons are instruments of structural oppression. Both positions are primarily normative rather than empirical.

The framework evaluates penal systems by a single criterion: which approach minimizes the total destruction of trajectory space, accounting for both victims and offenders? The empirical evidence is striking. In the United States, approximately 71% of state prisoners released in 2012 were rearrested within five years, and 46% returned to prison [68]. Norway, which in the 1980s had recidivism rates comparable to the current American figure ($\sim$60–70%), undertook a comprehensive shift from retribution to rehabilitation. By 2018, reconviction rates had fallen to 18% within two years and 25% within five years [69]. This comparison must be interpreted carefully—the two countries differ in population size, crime composition, socioeconomic conditions, and how recidivism is measured (rearrest vs. reconviction vs. reincarceration)—but the order-of-magnitude difference is difficult to attribute entirely to confounds.

Upstream interventions show complementary effects. Research on the Alaska Permanent Fund Dividend—an unconditional annual cash transfer to all residents—found that higher dividend

amounts were associated with reductions in property crime, consistent with economic models of crime as partly driven by financial need [70]. Conditional cash transfer programs in Brazil (Bolsa Família) and guaranteed income experiments in Canada (Mincome) have shown similar patterns, though effect sizes vary by context and the evidence base remains limited for large-scale universal programs.

The framework's conclusion is that prevention (addressing the material and psychological antecedents of crime) dominates punishment when the objective is minimizing total trajectory destruction. Incapacitation remains justified for individuals who pose a demonstrated, ongoing threat—but as a pragmatic measure to prevent future $\Delta I < 0$, not as moral retribution. If punitive systems demonstrated superior efficacy in reducing victimization, the framework would support them. They do not, and so it does not.

5.4 The Separation Principle: Emotion and Normative Criterion

Scope restriction: The analysis in this section applies exclusively to **systemic decisions under radical uncertainty involving irreversible trade-offs**—e.g., species triage, climate policy, AGI governance, resource allocation at civilizational scale. It does *not* prescribe emotional suppression in individual life, interpersonal relationships, aesthetic experience, or informal deliberation. The framework does not advocate sociopathy; it advocates decision hygiene in contexts where emotional heuristics are demonstrably miscalibrated relative to systemic consequences.

5.4.1 Historical Precedent: The Separation of Domains

Western intellectual history demonstrates that certain domains function optimally when operationally separated:

Scientific Revolution (16th-17th centuries): The separation between empirical investigation and dogmatic authority was a necessary condition for the development of modern scientific methodology [23].

Separation of Church and State (17th-18th centuries): The institutional distinction between religious power and political power proved essential for the stabilization of pluralist systems [25, 35].

We argue that an analogous separation is necessary between the emotional response and the normative criterion in ethical decision systems.

5.4.2 Comparative Analysis

The conflation of religious authority with legislative power historically produced:

1. Violent sectarian conflicts arising from incompatible dogmas
2. The impossibility of rationally revising legal norms
3. The systematic persecution of doctrinal dissidents

4. Arbitrary shifts in legislation driven by whichever doctrine held power

Analogously, the grounding of ethical systems in an emotional response produces:

1. **Temporal inconsistency:** The alteration of moral judgments as a function of the agent's affective state, violating decision coherence
2. **Cognitive biases:** Anchoring, loss aversion, the framing effect, and other documented biases [21] distort trade-off evaluation
3. **Coordination failure:** Divergent moral intuitions between cultures prevent convergence on global norms
4. **Signal interference:** When affective/normative heuristics carry no additional structural information beyond the structural representation R, conditioning decisions on them can only degrade the decision channel (formalized below)

5.4.3 Terminological Clarification: Instinct versus Emotion

Throughout this framework, we distinguish between two categories of affective response:

Definition 5.1 (Instinct)**.** *Phylogenetically ancient heuristic mechanisms characterized by immediate reaction to physical threats or metabolic needs (e.g., fear of falling, pain withdrawal, hunger). These operate as high-fidelity, low-latency channels for local survival signals. Instinct is* ***excluded*** *from the variable E in our formalization.*

Definition 5.2 (Emotion (as used in this framework))**.** *Higher-level affective and normative heuristics involving cognitive appraisal, often tied to social norms, cultural conditioning, and moral evaluation. This includes: affective empathy, moral intuitions ("fairness", "justice"), prejudice-driven valuation, indignation, shame, and similar constructs. These operate as lossy heuristics for social coordination which frequently misestimate systemic ΔI under radical uncertainty.*

The framework does not argue against instinct. Instinct serves as a valid informational proxy where structural analysis is computationally too slow or sensorially absent. What the framework identifies as decision noise is the second category: social/moral emotional heuristics applied to systemic decisions.

5.4.4 Formalization: Emotion as Garbling of Structural Information

Let S denote the latent structural state of the system, X the available observations, and $R = \phi(X)$ a *structural representation* used to evaluate the criterion ΔI. Let E denote the agent's emotional state (as defined above), and let A denote the selected action.

Assumption (No additional structural information).

E is *decision-irrelevant* for the structural criterion when it carries no additional information about S beyond the structural representation R:

$$I(S; E \mid R) = 0 \quad \Longleftrightarrow \quad P(S \mid R, E) = P(S \mid R). \tag{90}$$

A sufficient generative form is that E is a *garbling* of R:

$$E = g(R) \oplus N, \qquad N \perp S \mid R, \tag{91}$$

which makes E a lossy (and potentially biased) compression of the structural information.

Proposition (Irrelevance under Blackwell ordering).

Under (90), allowing policies to condition on E does not improve the attainable optimum:

$$\sup_{\pi(\cdot|R,E)} \mathbb{E}\big[\Delta I(S, A)\big] = \sup_{\pi(\cdot|R)} \mathbb{E}\big[\Delta I(S, A)\big]. \tag{92}$$

That is: if the structural representation R already captures everything relevant about S for computing ΔI, then additionally conditioning on emotional state E cannot improve expected outcomes. It can only introduce noise.

Definition (Decision noise).

Let $\pi^*(\cdot \mid R)$ be any policy that maximizes $\mathbb{E}[\Delta I(S, A) \mid R]$. Define emotional interference as:

$$\mathcal{N} := \mathbb{E}_{R,E}\left[D_{\mathrm{KL}}\Big(\pi(\cdot \mid R, E) \,\Big\|\, \pi^*(\cdot \mid R)\Big)\right]. \tag{93}$$

This quantifies the expected divergence between actual emotion-influenced policy and the structurally optimal policy.

Scope condition.

If instead $I(S; E \mid R) > 0$, then E carries predictive information about S that is not captured by R. In this case, E is not pure noise relative to the structural objective. This indicates that the structural representation R is incomplete. The appropriate response is to *structuralize* the information in E—that is, incorporate it into ϕ via operational validation—rather than treat E as an autonomous normative arbiter.

This scope condition ensures the framework is not dogmatically anti-emotion: it acknowledges that emotional responses may sometimes detect patterns the explicit model misses. The prescription is to extract and formalize that information, not to grant emotion veto power over structural analysis.

5.4.5 The Non-Elimination of Emotion

It is crucial to distinguish the proposal of this framework from an absolute anti-emotional position. The framework does not claim that emotions are "bad" or "dysfunctional" in all contexts—they serve critical adaptive functions in domains where rapid heuristic response is appropriate. Nor does it recommend that agents suppress or eliminate their emotional response capacity, which would itself constitute a reduction in cognitive diversity. The claim is narrower and methodological: emotions are unreliable as the *final normative arbiter* in systemic decisions under radical uncertainty, where their well-documented biases (proximity, similarity, temporal discounting) systematically distort trade-off evaluation.

The argument is more restricted and methodological:

> *For decisions involving irreversible trade-offs in complex systems under radical uncertainty, emotional criteria introduce non-structural variability that compromises decision consistency and stability.*

Emotions remain indispensable in domains where they evolved to operate effectively: motivating individual action and sustaining commitment to long-term goals, maintaining social cohesion and coordination within small groups, signaling values in culturally specific contexts, and enriching informal deliberative processes where the stakes do not involve irreversible systemic consequences. Emotions may also serve as early-warning detectors of model incompleteness—cases where the structural representation R fails to capture a relevant feature of the system. In such cases, however, the appropriate response is to *structuralize* the information the emotion carries (i.e., identify what the model is missing and incorporate it formally), not to defer to the emotion as an autonomous arbiter.

What is proposed is the exclusion of emotion as the *final normative arbiter* in systemic decisions, not its elimination from the human cognitive repertoire.

5.4.6 Grounding in the Philosophical Tradition

This position has precedents in the rationalist tradition: Plato's distinction in the *Republic* (Book IV) between $\lambda o \gamma \iota \sigma \tau \iota \kappa o \nu$ (rational faculty) and $\theta u \mu o \varsigma$ (emotion) as governing principles of the soul; Kant's grounding of morality in pure reason, independent of pathological inclinations (*Critique of Practical Reason*); and modern decision theory's axiomatization of rational preferences without reference to affective states (von Neumann-Morgenstern, Savage).

The distinctive contribution of this framework is the *operationalization* of this separation through the formal criterion ΔI, combined with the information-theoretic formalization of when emotion is noise versus signal.

5.4.7 Structural Empathy versus Affective Empathy

The literature in moral psychology and cognitive neuroscience characterizes empathy predominantly as an affective process: the capacity to experience emotional states congruent with those

of another agent [15]. This conception, although valid descriptively, confuses two functionally distinct processes.

Definition 5.3 (Affective Empathy)**.** *The capacity to experience a vicarious emotional response to the observed or inferred affective state of another agent. Characterized by:*

- *Emotional contagion (automatic affective mimicry)*
- *Shared affective activation in response to perceived distress or pleasure*
- *Dependence on the perceived similarity between the observer and the target*
- *Modulation by the observer's affective state*

Definition 5.4 (Structural Empathy)**.** *The evaluation of an entity's informational contribution to the system's space of possible trajectories, independent of the observer's affective response. Characterized by:*

- *Analysis of $I(X)$ (information carried by the entity)*
- *Evaluation of ΔI (marginal impact on the trajectory space)*
- *Independence from the observer's emotional state*
- *Applicability to non-sentient entities (ecosystems, cultures, knowledge systems)*

Formal Comparison

Let O be the observer and X be the entity under moral consideration. We define:

Affective Empathy:

$$V_{\text{affective}}(X|O) = f(\text{similarity}(O, X), \text{affective\_state}(O), \text{social\_proximity}(O, X)) \tag{94}$$

where f maps these inputs to an emotional response that determines moral valuation.

Properties:

1. $\frac{\partial V_{\text{affective}}}{\partial \text{affective\_state}(O)} \neq 0$ (observer dependence)
2. $\lim_{\text{similarity}\to 0} V_{\text{affective}} \to 0$ (similarity bias)
3. $V_{\text{affective}}(X_{\text{non-sentient}}) \approx 0$ (inapplicable to non-conscious entities)

Structural Empathy:

$$V_{\text{structural}}(X) = g(I(X), \Delta I(X \to \text{system}), \text{Red}(X)) \tag{95}$$

where:

- $I(X)$ is the information carried by entity X

- $\Delta I(X \to \text{system})$ is the marginal informational impact of X on system trajectories
- $\text{Red}(X)$ is the redundancy of X (number of equivalent copies/substitutes in the system)
- g maps these to an informational value, independent of O

Properties:

1. $\frac{\partial V_{\text{structural}}}{\partial \text{affective\_state}(O)} = 0$ (observer independence)
2. $V_{\text{structural}}$ defined for any X that carries information
3. $V_{\text{structural}}$ stable over time given fixed system state

Comparative Table

Criterion	**Affective Empathy**	**Structural Empathy**
Epistemological basis	Vicarious emotional response	Objective informational analysis
Temporal variability	High (varies with observer state)	Low (varies with system state)
Inter-observer consistency	Low (culturally variable)	High (objective criterion)
Scale of application	Individual and small groups	Systemic and civilizational
Cognitive requirement	Affective circuits	Modeling capacity
Applicability	Restricted to sentient entities	Universal (any carrier of I)
Proximity effect	Strong (close > distant)	Null (distance-independent)
Similarity effect	Strong (similar > dissimilar)	Null (depends only on I)

Illustrative Case: Preservation of a Minority Culture

Through affective empathy:

$V_{\text{affective}}(\text{culture } C) = f(\text{affective\_proximity}(O, \text{members of } C), \text{affective\_state}(O))$

Problems: The valuation is hostage to contingent features of the decision-maker: whether they have had personal contact with members of C, what emotional state they happen to be in at the time of the decision, and whether conflicting emotions (fear, aversion, cultural discomfort) override the empathic response. The result is systematic inconsistency—the same culture receives radically different valuations depending on who is evaluating it and when.

Via structural empathy:

$V_{\text{structural}}(\text{culture } C) = g(I(C), \Delta I(\text{loss of } C), \text{Red}(C))$

where:

- $I(C) = H(\Gamma_{\text{practices, language, knowledge of } C})$
- $\text{Red}(C)$ reflects whether equivalent cultural information exists elsewhere

Advantages: The structural analysis is independent of who performs it: any evaluator examining the same informational profile of C will reach the same conclusion. It recognizes that eliminating

C reduces ΔI_{total} of the human system regardless of whether the evaluator finds C's practices familiar or foreign, and it remains valid even if the decision-maker has never encountered a member of C. The consistency this produces across observers is not a loss of nuance but a gain in reliability—the objective function replaces subjective variation with a stable criterion.

Response to Objection: "A System Without Emotion is Inhumane"

Common objection:

"A framework that excludes emotion as a normative criterion is cold, inhumane, ignores what makes decisions morally significant."

Formal response:

The objection confuses:

1. **The exclusion of emotion as a normative criterion** (framework's position)
2. **The exclusion of moral consideration** (not defended by the framework)

Structural empathy allows for:

Moral consideration of X without emotional attachment to X

Clarification: The framework does not assert that agents should eliminate their emotional capacity, nor that emotion is "bad" or "useless." It asserts only that for systemic decisions under radical uncertainty, the emotional criterion introduces noise that compromises consistency.

Analogy:

A judge must decide based on the evidence and the law, not on a personal sympathy for the defendant. This does not make the system "inhumane"—it makes the system consistent and fair.

Similarly: decisions on the preservation/elimination of entities must be based on ΔI, not on affective response. This does not eliminate consideration—it makes consideration systematic and consistent.

6 Applications: Where Emotional Heuristics Fail

The preceding section established that the framework is ideologically agnostic: it follows evidence regardless of which political tradition the evidence favors. This section addresses a distinct but related problem: cases where the affective response—the visceral moral intuition—systematically diverges from the ΔI-optimal policy. Section 5.4 argued that systemic decisions under radical uncertainty require the operational separation of emotion from criterion. The cases below demonstrate why. Each involves a domain where the dominant emotional response produces policies that *destroy more trajectory space* than the alternative the emotion rejects.

The ideological agnosticism examples in Section 5 (drug policy, economic systems, penal systems)

also illustrate this pattern. Moral disgust at drug use drives criminalization that kills more people than it saves. Retributive anger at offenders drives incarceration policies with higher recidivism than rehabilitative alternatives. Rather than repeating those analyses, we refer the reader to Section 5 and focus here on cases where the emotional distortion is most extreme.

6.1 Case: Preventive Treatment of Pedophilic Attraction

6.1.1 The Emotional Barrier

Few policy domains provoke stronger affective responses than the treatment of individuals with pedophilic attraction. The visceral intuition—that any resource directed toward such individuals constitutes "rewarding" morally abhorrent dispositions—is nearly universal and emotionally overwhelming. This intuition generates a policy preference for purely reactive systems: wait for an offense, then punish.

6.1.2 The Informational Analysis

The framework's metric is the number of children whose trajectory space is preserved (i.e., who are not abused). Purely punitive systems intervene only after abuse has occurred—after $\Delta I < 0$ is already irreversible for the victim. Preventive programs such as the Dunkelfeld Project ("Prevention Project Dunkelfeld," Germany, [3]) offer anonymous treatment for individuals who self-identify as having pedophilic attraction before any contact offense occurs. Preliminary follow-up data suggest reduced contact offense rates among treated participants, though the evidence base has significant methodological limitations: the primary evaluation [4] assessed 1,006 individuals between 2005 and 2016, but only 56 reached follow-up (94.4% attrition), all outcome data are self-reported in an anonymous context, and no randomized control trial is possible given the ethical constraints. These limitations are severe, and the framework's endorsement is conditional: any non-degenerate efficacy estimate (i.e., even a small number of offenses prevented per year across the network) yields $\Delta I > 0$, because each prevented offense averts irreversible trajectory destruction for the victim while the program cost is recoverable monetary expenditure that does not itself destroy information.

6.1.3 The Divergence

The emotional response ("do not help pedophiles") produces a policy that maximizes retributive satisfaction but does not minimize victims. The structural analysis produces a policy that minimizes victims but violates the emotional heuristic. The framework selects the latter. This is not a claim that the emotional response is irrational in all contexts—disgust at child abuse serves important social-coordination functions. It is a claim that when the emotional heuristic is used as the *decision criterion* for systemic policy, it produces more destroyed trajectory space than the alternative.

6.2 Case: Capital Punishment for Heinous Crimes

6.2.1 The Emotional Barrier

When confronted with particularly horrific crimes—serial murder, terrorism, child homicide—the retributive impulse is among the strongest affective responses humans experience. The intuition that the perpetrator "deserves to die" functions as a powerful moral heuristic, and the demand for capital punishment in such cases enjoys broad public support in many jurisdictions.

6.2.2 The Informational Analysis

The framework's metric is not whether the perpetrator deserves death but whether capital punishment minimizes total future $\Delta I < 0$ (i.e., prevents more homicides than the alternative sanction). The National Research Council's comprehensive review [73] concluded that three decades of empirical research on this question is fundamentally uninformative: studies claiming to demonstrate a deterrent effect are based on implausible statistical assumptions, fail to account for noncapital punishments as alternatives, and yield wildly contradictory results depending on minor specification changes—ranging from 429 lives saved per execution to 86 lives lost, as Donohue and Wolfers (2005) demonstrated by applying minor methodological adjustments to the same datasets. The NRC panel recommended that these studies "not be used to inform deliberations requiring judgments about the effect of the death penalty on homicide."

Meanwhile, the informational cost of execution is unambiguous: it irreversibly eliminates the offender's entire remaining trajectory space. Life imprisonment without parole achieves the same incapacitative function—removing the individual's capacity to cause further $\Delta I < 0$—without the irreversible destruction. The framework therefore favors life imprisonment over execution: it achieves equivalent protection of future victims while preserving reversibility (in the event of wrongful conviction, for instance) and destroying less information. This conclusion holds regardless of how one feels about the perpetrator. The emotional satisfaction of retribution is real but does not enter the ΔI calculus.

6.3 Case: Property Crime and Economic Desperation

6.3.1 The Emotional Barrier

Property crime—theft, robbery, burglary—provokes moral indignation directed at the offender. The affective response treats the act as a violation of a moral norm ("taking what is not yours") and demands punishment proportional to the transgression. In many societies, this emotional response drives severe sentencing: in the United States, property offenders have among the highest five-year rearrest rates at 78% [68], and the majority cycle through incarceration repeatedly without desistance. The emotional framing rarely considers the offender's material conditions as relevant to the appropriate policy response.

6.3.2 The Informational Analysis

The framework reframes the question: what policy minimizes total trajectory destruction, accounting for both victims and offenders? The empirical literature consistently identifies economic deprivation as a principal driver of property crime. Watson, Guettabi, and Reimer (2020) found that higher Alaska Permanent Fund Dividend payments were associated with reductions in property crime [70]. Brazil's Bolsa Família conditional cash transfer program has been associated with crime reductions in urban areas [72]. These findings are consistent with the rational-choice model: when the expected return from legitimate activity exceeds the expected return from crime (adjusted for risk of apprehension), property offending decreases.

Incarceration, by contrast, costs $30,000–60,000 per prisoner per year in the United States, disrupts the offender's employment prospects, social ties, and housing stability, and produces rearrest rates exceeding 70% within five years. The informational accounting is clear: the punitive response destroys trajectory space on both sides—the offender's (through incarceration) and future victims' (through recidivism)—while failing to address the upstream cause.

6.3.3 The Divergence

The emotional response ("punish the thief") feels proportionate but produces a system that creates more total victims than the alternative. Upstream investment in economic security addresses the causal mechanism and preserves trajectory space for both the would-be offender and the would-be victim. The framework selects the policy that minimizes total $\Delta I < 0$, which in this case means redirecting resources from reactive punishment to preventive support—even though this conclusion violates the retributive intuition that the offender "deserves" to suffer consequences.

6.4 The Common Pattern

Across all three cases—and the drug policy, economic, and penal examples in Section 5—the same structural pattern emerges. The affective heuristic optimizes for a proximate signal (disgust, retribution, indignation) that served adaptive functions in small-group contexts but systematically misestimates systemic ΔI under the conditions of modern policy: large populations, delayed feedback, and irreversible consequences. Structural empathy, as defined in Section 5.4, replaces the proximate signal with the distal metric: total trajectory space preserved. This does not eliminate the emotional response—it subordinates it to the informational criterion in systemic decisions where the two diverge.

7 Global Coordination: Limitations and Solutions

7.1 Human Cognitive Limitations

The preceding sections established that affective heuristics systematically diverge from ΔI-optimal policy in individual cases (Section 6). This section addresses a distinct problem: even if individual decision-makers adopted the framework's criterion, the structure of human cognition and collective

action would prevent its implementation at the scale required. Three independent limitations converge to produce this conclusion.

7.1.1 Affective Distortion in Systemic Decisions

Section 6 documented specific cases where emotional heuristics produce policies that destroy more trajectory space than the alternative. The relevant point for the coordination argument is not the individual cases but the *structural* observation: affective distortions are not incidental failures that better education or deliberation can correct. They are features of human moral cognition that are amplified, not attenuated, by democratic political incentives. Elected officials face re-election cycles of 2–6 years; policies whose ΔI benefits materialize over decades (climate mitigation, preventive social programs, alignment research) are systematically disadvantaged relative to policies whose emotional payoff is immediate (retributive punishment, visible infrastructure, capability demonstrations). Haidt's social intuitionist model [15] and prospect theory's loss aversion [21] both predict this pattern: moral reasoning is post-hoc rationalization of affective judgments, and losses loom larger than equivalent gains, producing systematic bias toward reactive over preventive policy.

7.1.2 Bounded Rationality and Cross-Domain Integration

Global optimization under the framework's criterion requires integrating heterogeneous data across domains with incompatible scales, units, and feedback timescales: ecosystem dynamics (decades to centuries), economic systems (quarters to years), climate models (decades to millennia), population genetics (generations), and technological trajectories (months to decades). Simon's bounded rationality [30] identifies the core constraint: human decision-makers satisfice rather than optimize because the computational demands of genuine optimization exceed cognitive capacity. This is not merely a limitation on the number of variables held in working memory—it is a limitation on the capacity to model nonlinear interactions across domains. A policy that optimizes ΔI in the economic domain (e.g., rapid industrialization) may produce catastrophic $\Delta I < 0$ in the ecological domain through feedback loops that operate on timescales longer than a human planning horizon. No individual or committee can reliably compute these cross-domain interactions in real time.

7.1.3 Coordination Failures at Scale

Even if individual agents could compute the ΔI-optimal policy, collective implementation faces the free-rider problem [29]: rational agents defect from cooperative equilibria when the costs of cooperation are private and the benefits are public. Ostrom [32] demonstrated that small-scale commons can be self-governed under specific institutional conditions, but these conditions—repeated interaction, mutual monitoring, graduated sanctions—do not scale to global coordination among sovereign states with heterogeneous interests and no enforcement authority.

The empirical record confirms this prediction. In climate policy, the UNEP Emissions Gap Report

2025 found that even full implementation of all nationally determined contributions would limit warming to only 2.3–2.5°C—well above the Paris Agreement target of 1.5°C [74]. Under current policies, the trajectory reaches 2.8°C. Global greenhouse gas emissions grew 2.3% year-on-year in 2024, and none of the G20 members are collectively on track to meet even their 2030 NDC targets. In biodiversity policy, the UN Global Biodiversity Outlook 5 [75] reported that none of the 20 Aichi Biodiversity Targets set in 2010 were fully achieved by their 2020 deadline; of 44 sub-targets assessed, 20 were ranked "poor" and only 5 as "good." The pattern is consistent: voluntary international agreements produce targets that are collectively insufficient, and even those insufficient targets are not met.

Structural diagnosis: The coordination failure is not a contingent political problem solvable by better leadership or stronger norms. It is a structural consequence of the mismatch between (a) the scope and timescale of the optimization problem (global, multi-generational) and (b) the scope and timescale of human decision-making institutions (national, electoral). The framework requires a coordination mechanism whose capacity matches the problem's complexity—which motivates the analysis in the following subsections.

7.2 Functional Requirements for Global Coordination

The three limitations documented above—affective distortion, bounded rationality, and coordination failure at scale—are not independent. They interact: bounded rationality prevents individual agents from computing cross-domain ΔI trajectories; affective distortion biases the policies that agents *do* compute toward proximate emotional signals; and coordination failure prevents even correct individual computations from aggregating into collectively optimal outcomes. Any mechanism capable of implementing the framework's criterion at the required scale would need to satisfy, at minimum, the following functional requirements:

1. **Cross-domain integration:** The capacity to model interactions between ecological, economic, technological, demographic, and climatic systems simultaneously, across timescales ranging from years to centuries.
2. **Criterion-consistent optimization:** The capacity to optimize for ΔI without systematic distortion by affective heuristics or short-horizon incentives.
3. **Coordination enforcement:** The capacity to implement cooperative equilibria among agents with heterogeneous interests, overcoming the free-rider dynamics that voluntary agreements have empirically failed to resolve (Section 7.1.3).
4. **Speed commensurate with the problem:** The ecological timeline documented in Section 7.1.3—emissions still rising, no international target fully met—requires corrective action within decades, not centuries.

An AGI system satisfying the framework's design constraints (Section 8) would, in principle, meet all four requirements. Whether AGI is the *only* system capable of meeting them is an empirical question addressed in Section 7.4, which evaluates alternative coordination mechanisms (human

cognitive enhancement, decentralized algorithmic systems, hybrid human-AI architectures) against the same functional criteria. The following subsection (Section 7.3) addresses a logically prior question: given the risks of AGI development, under what conditions is the attempt justified despite those risks?

7.3 The Argument of "Flip Logic" — A Rigorous Analysis

7.3.1 Expanded Formal Model

Context: A decision between two mutually exclusive options:

- **Option A:** Status quo (do not develop AGI, continue current trajectory)
- **Option B:** Develop AGI (with a misalignment risk)

Outcome space:

$$\begin{aligned}\Omega = \{&\text{Collapse without AGI},\\ &\text{Aligned AGI + preservation},\\ &\text{Misaligned AGI (catastrophic)},\\ &\text{Misaligned AGI (worse than collapse)}\}\end{aligned} \tag{96}$$

Values (normalized):

$$V(\text{Collapse without AGI}) = -1000 \quad \text{(a massive loss of information)} \tag{97}$$

$$V(\text{Aligned AGI}) = +500 \quad \text{(optimized preservation)} \tag{98}$$

$$V(\text{Misaligned AGI catastrophic}) = -1500 \tag{99}$$

$$V(\text{Misaligned AGI worse}) = -10000 \quad \text{(e.g., active optimization against human welfare)} \tag{100}$$

7.3.2 Empirical Quantification of P(Collapse | Status Quo)

Convergent evidence:

1. **Accelerated extinction rates:** Current extinction rates are estimated at 100–1000× background rates [9], though precise species-per-day figures remain uncertain due to the incompleteness of taxonomic inventories. Even conservative estimates indicate a trajectory consistent with mass extinction events in the geological record.
2. **Climate feedback loops:** The IPCC AR6 identifies several potential tipping elements in the Earth system. We note that the IPCC assessment uses calibrated uncertainty language

("low confidence," "medium confidence," "high confidence") rather than precise probabilities for most tipping points. The following are structured estimates based on the reviewed literature, **not direct IPCC probability assignments**:

- Permafrost thaw: threshold ∼2°C (AR6 WGII assesses significant carbon release as "high confidence" above 2°C)
- AMOC weakening: AR6 assesses AMOC decline as "very likely" but abrupt collapse before 2100 as "medium confidence" that it will *not* occur. Post-AR6 literature [12] suggests higher risk. We use $P(\text{significant weakening}|2050) \approx 0.10\text{–}0.20$ as a structured estimate, not an IPCC figure.
- Amazon dieback: $P(\text{critical}|2050) \approx 0.15\text{–}0.25$ (structured estimate based on [76] and related literature)

3. **Integrated models [77]:**
   - "Hothouse Earth": risk increases substantially if temperature exceeds ∼2°C
   - Current trajectory: 2.3–2.8°C by 2100 depending on NDC implementation [74]
   - $P(\text{exceed } 2°\text{C}|2050) \approx 0.7$ (consistent with IPCC scenario assessments)

Aggregated estimate with epistemic caveats:

These estimates are derived from the convergence of multiple independent lines of evidence (extinction rates, climate feedback models, integrated Earth system models). However, they are not frequentist probabilities but structured expert judgments, subject to model uncertainty, tail risk underestimation, and scenario dependence. We present ranges rather than point estimates, and note concordance with independent forecasting sources:

$$P(\text{Irreversible collapse}|\text{Status quo, } 2050) \in [0.3, 0.8] \tag{101}$$

$$P(\text{Irreversible collapse}|\text{Status quo, } 2100) \in [0.5, 0.95] \tag{102}$$

Cross-validation with independent sources:

- The IPCC AR6 WGII assigns "high confidence" to severe biodiversity loss and ecosystem disruption under scenarios exceeding 2°C, consistent with the upper range.
- Expert elicitation surveys on existential risk [42, 43] place aggregate existential risk from all sources at 10–20% by 2100, which provides a *lower bound* on our broader "irreversible collapse" category (which includes non-extinction civilizational degradation).
- Prediction markets (Metaculus, as of 2024) assign $P(\text{human extinction by 2100}) \approx 0.01\text{–}0.03$, but this measures only the extreme tail, not the broader category of irreversible informational loss used here.

Sensitivity note: The lexicographic heuristic (Section 7.3.4) depends on two key parameters: $P(\text{Collapse})$ and $P(\text{Alignment})$. The value threshold is satisfied whenever $P(\text{Collapse})$ is high enough that the expected outcome of inaction crosses the unacceptability threshold ($\Delta I < -0.5 \cdot I_{\text{total}}$). The escape dominance condition additionally requires $P(\text{Collapse}) > 1 - P(\text{Alignment})$. The critical structural insight is that $P(\text{Alignment})$ is itself a function of resource allocation: under the current 500:1 to 1,000:1 ratio of capabilities to alignment investment documented in Section 7.5, $P(\text{Alignment})$ is necessarily low, making the escape dominance condition very difficult to satisfy. Under significant reallocation toward alignment, $P(\text{Alignment})$ increases, and the condition becomes progressively easier to satisfy. The argument's validity depends not on precise collapse estimates but on the relationship between collapse probability and alignment probability—the worse the alignment prospects, the higher the collapse probability must be to justify the attempt.

7.3.3 Expected Value Analysis with Structured Estimates

Epistemic note: The following analysis combines three distinct types of inputs: (i) observational data (extinction rates, temperature trajectories), (ii) structured expert judgments (collapse probabilities, alignment estimates), and (iii) normative payoff assignments ($V(\cdot)$ values reflecting the framework's axiom). These are not "data" in the empirical sense and should not be treated as such. The analysis demonstrates the *structure* of the decision under these assumptions, not a unique numerical conclusion. Readers who assign different probabilities or payoffs will reach different expected values, but the qualitative structure (threshold analysis in the following subsection) is robust to wide parameter variation.

Option A (Status Quo):

$$EV(A) = P(\text{Collapse}) \cdot V(\text{Collapse}) + P(\text{No}) \cdot V(\text{Continuity}) \tag{103}$$

$$\approx 0.7 \times (-1000) + 0.3 \times 0 = \mathbf{-700} \tag{104}$$

(Using $P(\text{Collapse}) = 0.7$ as a midpoint of the $[0.5, 0.95]$ range for horizon 2100. At the lower bound $P = 0.5$: $EV(A) = -500$. At the upper bound $P = 0.95$: $EV(A) = -950$.)

Option B (Develop AGI):

The base scenario: $P(\text{Align.}) = 0.3$, $P(\text{Misal. catastr.}) = 0.6$, $P(\text{Worse}) = 0.1$

$$EV(B) = 0.3 \times 500 + 0.6 \times (-1500) + 0.1 \times (-10000) \tag{105}$$

$$= 150 - 900 - 1000 = \mathbf{-1750} \tag{106}$$

Apparent conclusion: $EV(A) = -700 > EV(B) = -1750 \rightarrow$ Is the status quo better?

But this analysis is insufficient.

Methodological note: The argument below is a *lexicographic decision heuristic*, not a formal derivation that "refutes" expected value theory. It proposes that when the status quo crosses an unacceptability threshold, the decision problem changes structure in a way that standard EV comparison fails to capture. This is analogous to the distinction between decision under risk (known distributions, EV applicable) and decision under deep uncertainty (contested distributions, robustness criteria more appropriate). The heuristic should be evaluated by its structural plausibility, not by the precision of the numerical estimates used to illustrate it.

7.3.4 Refinement: The Absolute Unacceptability Threshold

EV analysis assumes cardinal comparability. But:

1. Collapse = loss $> 50\%$ of I_{total}
2. Framework: $\Delta I < -0.5 \cdot I_{\text{total}}$ = an unacceptability threshold
3. Beyond this threshold, differences become non-operational

Reformulation:

The lexicographic heuristic operates in three stages:

1. **Value threshold (applied to the status quo):** If the expected outcome of inaction crosses an absolute unacceptability threshold ($V(\text{Status quo}) < T_{\text{unacceptable}}$, where T is defined by the framework's irreversibility criterion: $\Delta I < -0.5 \cdot I_{\text{total}}$), then the status quo is classified as *unacceptable regardless of alternatives.*
2. **Escape dominance (applied to the alternative):** The alternative must offer a strictly higher probability of remaining above the unacceptability threshold than the status quo:

$$P(V > T \mid \text{Alternative}) > P(V > T \mid \text{Status quo}) \tag{107}$$

   Without this condition, the heuristic could select alternatives that are equally likely (or less likely) to avoid catastrophe, which would be irrational.
3. **Risk bound (applied to the worst case):** The probability that the alternative produces an outcome *strictly worse* than the status quo's expected outcome must be bounded: $P(\text{Worse}|\text{Alternative fails}) < \delta$ for some pragmatic threshold δ.

When all three conditions are satisfied, the alternative is rational **even if its EV is lower than the status quo's EV**, because the EV comparison presupposes cardinal comparability across outcomes that the value threshold declares non-comparable, while the escape dominance condition ensures the alternative genuinely improves the probability of survival.

Application:

- $V(\text{Collapse}) = -1000 < T = -500$ ✓
- $P(\text{escape}|\text{AGI}) = P(\text{Alignment}) = 0.3 > P(\text{escape}|\text{Status quo}) = 1 - P(\text{Collapse})$. This condition is satisfied when $P(\text{Collapse}) > 1 - P(\text{Alignment})$; i.e., when $P(\text{Collapse}) > 0.7$. ✓
- $P(\text{Worse}|\text{Misal.}) \approx 0.05 < 0.3$ ✓
- Therefore: Option B is justified when the probability of collapse under inaction exceeds the probability of misalignment under action.

Analogy: A football team losing 1–0 in stoppage time. The manager pulls the goalkeeper for an extra attacker. This objectively increases the risk of a worse outcome (conceding on an empty net, losing 2–0 or 3–0), and the expected value of the substitution may be negative. But the move is rationally justified because: (a) the status quo leads to certain defeat, (b) the alternative offers a nonzero probability of equalizing that the status quo does not, and (c) the additional downside (losing by a larger margin) is non-operational—a 2–0 loss is functionally equivalent to a 1–0 loss. The key is not that the move improves the expected score, but that it improves the *probability of escaping defeat*. The same structure applies: attempting AGI is justified not because it improves the expected outcome, but because it improves the probability of escaping civilizational collapse, provided that the escape dominance condition is satisfied.

7.3.5 Full Sensitivity Analysis

Variable 1: P(Alignment)

$P(\text{Align.})$	$EV(B)$	$EV(B) > EV(A)$?	Escape dom.?	**Decision**
0.10	-2650	No	Requires $P(\text{C}) > 0.90$	Conditional
0.20	-2200	No	Requires $P(\text{C}) > 0.80$	Conditional
0.30	-1750	No	Requires $P(\text{C}) > 0.70$	Marginal
0.40	-1300	No	Requires $P(\text{C}) > 0.60$	**Justified**
0.50	-850	No	Requires $P(\text{C}) > 0.50$	**Justified**
0.60	-400	**Yes**	Requires $P(\text{C}) > 0.40$	**EV dominant**

The escape dominance condition binds: at low $P(\text{Alignment})$, the heuristic requires correspondingly high collapse probability. At $P(\text{Alignment}) = 0.10$, AGI is only justified if $P(\text{Collapse}) > 0.90$—a condition at the upper end of the estimated range. This is the intended behavior: the worse the alignment prospects, the more desperate the situation must be to justify the attempt.

Key insight: The escape dominance condition transforms the heuristic from "any nonzero chance justifies the gamble" (which is reckless) to "the gamble must credibly improve survival odds" (which is rational). The binding constraint is $P(\text{Collapse}) > 1 - P(\text{Alignment})$, meaning that alignment investment directly expands the range of scenarios under which AGI development is justified.

Variable 2: V(Worse) and P(Worse)

Plausibility analysis:

V(Worse)	P(Worse|Misal.)	$EV(B)$
−10,000	0.10	−1750
−10,000	0.30	−3750
−100,000	0.10	−10,750
−100,000	0.01	−2,650

Eternal or large-scale optimization against human welfare requires an AGI whose terminal goal is structurally adversarial to human trajectory space. A typical misalignment = a paperclip maximizer (eliminates as a side effect, does not actively optimize against welfare).

Estimate: $P(\text{Worse}|\text{Misal.}) \approx 0.01-0.05$

Even with $P = 0.05$ and $V = -100{,}000$: the escape dominance condition still applies—the worst-case scenario does not affect whether the alternative improves the probability of escaping collapse, only the magnitude of downside if it fails. The risk bound condition ($P(\text{Worse}) < \delta$) constrains this tail risk independently.

7.3.6 Summary of the Lexicographic Heuristic

The sufficient conditions to attempt AGI (under this heuristic):

$$\begin{cases} V(\text{Status quo}) < T_{\text{unaccept.}} & \text{(value threshold: status quo is unacceptable)} \\ P(\text{escape}|\text{AGI}) > P(\text{escape}|\text{SQ}) & \text{(escape dominance: alternative improves survival odds)} \\ P(\text{Worse}|\text{Misal.}) < \delta & \text{(risk bound: worst case is bounded)} \end{cases} \tag{108}$$

Where: $T_{\text{unaccept.}}$ corresponds to $\Delta I < -0.5 \cdot I_{\text{total}}$ (satisfied when $P(\text{Collapse})$ is high enough that the expected outcome of inaction crosses the threshold), $P(\text{escape}|\text{AGI}) = P(\text{Alignment})$, $P(\text{escape}|\text{SQ}) = 1 - P(\text{Collapse})$, $\delta \approx 0.3$.

Derived condition: The escape dominance condition reduces to $P(\text{Collapse}) > 1-P(\text{Alignment})$. This condition creates a direct link between alignment investment and the range of scenarios under which AGI development is justified. Under the current allocation—where alignment receives roughly 0.1% of capabilities investment (Section 7.5)—$P(\text{Alignment})$ is necessarily low, and the condition requires near-certain collapse to be satisfied. Under significant reallocation toward alignment, $P(\text{Alignment})$ increases substantially, and the condition becomes satisfiable across a wide range of collapse estimates. This highlights that reallocation is not merely desirable but a *precondition* for the heuristic to justify AGI development.

Applying structured estimates from Section 7.3.3:

1. $P(\text{Collapse}|2100) \in [0.5, 0.95] \rightarrow V(\text{Status quo}) < T_{\text{unaccept.}}$ ✓
2. $P(\text{escape}|\text{AGI}) = P(\text{Alignment})$; under significant reallocation (Section 7.5), $P(\text{Alignment})$

is substantially higher than under current allocation, and $P(\text{escape}|\text{SQ}) = 1 - P(\text{Collapse}) \in [0.05, 0.5]$. The escape dominance condition is satisfied whenever $P(\text{Alignment}) > 1 - P(\text{Collapse})$, which is achievable under rational allocation across the estimated collapse range. ✓

3. $P(\text{Worse}|\text{Misal.}) \approx 0.05 < 0.3$ ✓

Note: The escape dominance condition ($P(\text{Collapse}) > 1 - P(\text{Alignment})$) is the binding constraint. The higher $P(\text{Alignment})$, the easier it is to satisfy this condition. Under significant reallocation of resources toward alignment (Section 7.5), the condition is satisfiable across the entire estimated range of collapse probabilities. The conclusion is robust to where within the range the true collapse probability falls, *provided* that alignment investment is commensurate with capabilities investment.

All conditions are satisfied → AGI development is justified under optimal alignment investment.

CRITICAL: Under the current allocation (capabilities-to-alignment ratio of 500:1 to 1,000:1), $P(\text{Alignment})$ is necessarily low, which means the escape dominance condition requires near-certain collapse to hold. The heuristic therefore does **not** justify AGI development under current investment levels—it justifies AGI development *conditional on* significant reallocation of resources toward alignment research.

Imperative: Reallocation (Section 7.5) is a **necessary condition**, not merely desirable.

7.4 AGI as a Tool, Not a Unique Solution

7.4.1 The Framework's Commitment Is to the Criterion, Not to AGI

The framework's core commitment is to the coordination criterion ($\max \Delta I$), not to any particular mechanism for implementing it. Section 7.2 identified four functional requirements that any adequate coordination mechanism must satisfy: cross-domain integration, criterion-consistent optimization, coordination enforcement, and speed commensurate with the ecological timeline. AGI is one candidate that would, in principle, satisfy all four. But it is not the only logically possible candidate, and the framework would adopt any alternative system that met the same requirements more reliably. This subsection evaluates three such alternatives against the functional requirements and makes the trade-offs explicit.

7.4.2 Alternative Coordination Mechanisms

Radical human cognitive enhancement.

If human cognitive capacities could be expanded sufficiently—through brain-computer interfaces, genetic intervention, or pharmacological means—enhanced humans could in principle perform the cross-domain integration and long-horizon planning that current cognition cannot. Such enhancement would preserve human agency and eliminate the misalignment risk entirely, making

it strictly preferable to AGI on the risk dimension. However, no existing or near-term technology satisfies the full requirement set. Current brain-computer interfaces achieve bandwidth orders of magnitude below what multi-variable global optimization would require; genetic enhancement of complex cognitive traits remains beyond current capability; and no pharmacological intervention selectively eliminates affective distortion in systemic decisions without impairing other cognitive functions. The coordination enforcement problem—overcoming free-rider dynamics among enhanced but still self-interested agents—would also remain unresolved unless enhancement included motivational changes, which raises its own ethical concerns. Timeline viability is the binding constraint: even optimistic projections place the relevant technologies decades away, which may exceed the ecological window documented in Section 7.1.3.

Decentralized algorithmic coordination.

A network of human agents supported by narrow AI systems, consensus algorithms, and distributed enforcement mechanisms could aggregate preferences and implement coordination without centralization. Existing precedents include federated learning protocols, mechanism design frameworks, and multi-agent coordination systems—all of which demonstrate that some degree of distributed optimization is achievable without general intelligence. The principal advantage is robustness: no single point of failure, unlike a centralized AGI. However, decentralized systems face three limitations against the functional requirements. First, computational scalability: the cross-domain integration required (simultaneously modeling ecological, economic, demographic, and climatic interactions across centuries) exceeds the demonstrated capacity of current distributed architectures. Second, the governance problem: who defines the initial coordination rules, and how are they revised? This reintroduces the very political dynamics that produce coordination failure. Third, consensus mechanisms are inherently slower than autonomous optimization, which creates tension with the speed requirement.

Human coordination enhanced by limited AI.

A hybrid architecture in which AI systems provide analysis, modeling, and simulation while humans retain final decision authority. This is the closest to currently deployable: versions of it already exist in climate modeling advisory systems, economic policy simulation, and public health decision support. The approach preserves human agency, is more politically acceptable than autonomous AGI, and substantially reduces catastrophic misalignment risk. Its central weakness is that the human bottleneck remains: decision-makers can ignore AI recommendations (as already occurs with IPCC assessments), the process is slower than autonomous operation, and the system is vulnerable to capture by special interests who benefit from coordination failure. Whether these limitations are fatal depends on whether the coordination gap identified in Section 7.1.3 can be closed by augmented human decision-making or requires autonomous optimization.

7.4.3 Comparative Assessment

Table 7.4.3 summarizes the assessment of each mechanism against the four functional requirements from Section 7.2, using a qualitative five-level scale (Very low, Low, Medium, High, Very high). The ratings are justified as follows.

Capacity reflects cross-domain integration capability. The human status quo is rated Low because bounded rationality (Section 7.1.2) prevents simultaneous modeling of interacting global systems. Enhancement is rated High because expanded cognition would, by hypothesis, overcome this limitation. Decentralized systems are rated Medium because they aggregate beyond individual capacity but face scalability constraints. Hybrid AI is rated Medium because AI handles computation but human decision authority limits integration. AGI is rated Very high because autonomous cross-domain optimization is a defining capability.

Speed reflects response latency relative to the ecological timeline. The status quo is Very slow (decades of negotiation for marginal progress, per Section 7.1.3). Enhancement is Medium (faster than unenhanced cognition but still human-speed deliberation). Decentralized is Slow (consensus mechanisms add latency). Hybrid is Medium (AI accelerates analysis but human deliberation remains the bottleneck). AGI is Very fast (autonomous operation without human-in-the-loop delays).

Risk reflects the probability and magnitude of catastrophic failure. The status quo is rated High—not Medium—because $P(\text{Collapse}) \in [0.5, 0.95]$ (Section 7.3.2) represents a high probability of irreversible informational loss; rating it lower would contradict the framework's central empirical claim. Enhancement is Low (no misalignment risk, but risks from cognitive modification itself). Decentralized is Low (no single point of catastrophic failure). Hybrid is Medium (reduced misalignment risk but vulnerable to human override failures and political capture). AGI is Very high (misalignment risk is existential).

Viability reflects technological readiness within the relevant timeline. The status quo is High (it is the default). Enhancement is Very low (no current technology satisfies the requirements). Decentralized is Low (components exist but integration at the required scale does not). Hybrid is Medium-High (partially deployed, scalable with investment). AGI is Medium (massive investment underway, but neither AGI nor alignment are achieved).

Mechanism	**Capacity**	**Speed**	**Risk**	**Viability**
Human status quo	Low	Very slow	High	High
Cognitive enhancement	High	Medium	Low	Very low
Decentralized algorithmic	Medium	Slow	Low	Low
Hybrid human + limited AI	Medium	Medium	Medium	Medium-High
Aligned AGI	**Very high**	**Very fast**	**Very high**	**Medium**

Table 2: Qualitative assessment of coordination mechanisms against the functional requirements of Section 7.2. Risk for the status quo is rated High to reflect the collapse probability range estimated in Section 7.3.2.

The table reveals the core trade-off: AGI dominates on capacity and speed but carries the

highest catastrophic risk. Hybrid systems offer the best risk-adjusted profile among currently viable options but may lack the coordination capacity to close the gap identified in Section 7.1. The framework does not resolve this trade-off from first principles—it depends on empirical developments in alignment research, ecological trajectory data, and the demonstrated scaling capacity of hybrid systems.

7.4.4 Substitution Criterion

The framework is not axiomatically committed to AGI. It would adopt an alternative coordination system S instead of AGI whenever S satisfies the same functional requirements more reliably. The relevant conditions are qualitative, not cardinal:

1. **Sufficiency:** S must be capable of implementing global coordination at the scale and speed required to avoid irreversible collapse (i.e., S can solve the cross-domain integration, criterion-consistent optimization, and enforcement problems identified in Section 7.2).
2. **Reliability:** The probability that S achieves its coordination objectives must exceed the probability that AGI achieves alignment: $P(\text{success}|S) > P(\text{alignment}|\text{AGI})$.
3. **Bounded catastrophic risk:** The worst-case outcome under S's failure must not exceed the worst-case outcome under AGI misalignment. Formally, $V(\text{failure}|S) \geq V(\text{misalignment}|\text{AGI})$.
4. **Timeline compatibility:** S must be deployable within the window before irreversible ecological collapse—a constraint that eliminates alternatives requiring centuries of development.

The comparative assessment above suggests that hybrid systems currently satisfy conditions (2) and (3) better than autonomous AGI, but may fail condition (1)—their coordination capacity is limited by the very human cognitive constraints that motivate the framework's argument. Whether this limitation is fatal depends on the severity and timeline of ecological collapse, which is an empirical question the framework cannot resolve from first principles.

Recognition of uncertainty. The framework establishes the evaluation criterion (ΔI), the need for global coordination, and the structural trade-offs between options. It does **not** resolve which system is optimal—this is an **empirical question** that depends on alignment research progress, ecological trajectory data, and the demonstrated capacity of alternative coordination mechanisms. The framework requires only that whichever system is adopted be evaluated by the same criterion and be subject to revision as evidence accumulates.

7.5 Resource Allocation: The Structural Imbalance

7.5.1 The Empirical Disparity

Independent data sources converge on a stark asymmetry in global resource allocation across the three domains relevant to this framework's argument. Table 7.5.1 summarizes the current landscape.

Category	Annual (approx.)	Primary sources
AI capabilities	$100–250B	Stanford AI Index; Russell & Cohen
AI alignment	$0.2–0.35B	EA Forum analysis; industry reports
Climate mitigation	$1,780B	CPI Global Landscape 2025

Table 3: Global resource allocation by category (USD/year, circa 2024). The capabilities range spans from Russell's AGI-specific estimate (~$100B) [56] to total corporate AI investment ($252.3B) [55], with hyperscaler capital expenditure projected to exceed $500B by 2026 [57]. Alignment figures aggregate philanthropic funding ($110–170M) [58], estimated industry lab internal safety budgets ($30–60M each for major labs), and U.S. public-sector investment (~$10M) [56]. Climate mitigation figures represent total global climate finance, of which $1.78T was mitigation-specific, against an estimated need of $8–9T annually through 2030 [59].

The ratio of capabilities to alignment investment is approximately **500:1 to 1,000:1**. Russell characterizes this as "a factor of about 10,000 times less investment" when comparing AGI development to public-sector safety research specifically [56].

7.5.2 Structural Argument for Reallocation

The framework's argument does not depend on specifying a functional form for how investment maps to probability of success. It depends on a structural observation: the probability of a good outcome from AGI development is the product of at least three independent factors:

$$P(S) = P(\text{Alignment resolved}) \times P(\text{Alignment precedes AGI}) \times P(\text{Collapse avoided before AGI}) \tag{109}$$

Under current allocation, the first factor—$P(\text{Alignment resolved})$—is the binding bottleneck. If alignment receives three orders of magnitude less funding than capabilities, there is no plausible model of scientific progress under which the probability of solving alignment keeps pace with the probability of achieving AGI capabilities. The second factor—timing—is directly determined by the ratio of alignment to capabilities investment: the higher the ratio, the more likely alignment research concludes before capabilities cross dangerous thresholds.

The structural conclusion follows without requiring specific functional forms or parameter calibration: *any reallocation from capabilities toward alignment simultaneously increases the first factor, improves the second, and does not decrease the third.* This is not a quantitative claim about the magnitude of improvement (which would require the kind of calibrated functions that the current state of knowledge cannot provide), but a qualitative dominance argument: reallocation *weakly dominates* the status quo allocation across all three factors.

Epistemic limitation: We cannot specify how much reallocation is optimal, because we lack reliable models mapping investment to probability of alignment breakthroughs. The argument establishes only that the current allocation is *structurally irrational*—it maximizes the probability of achieving capabilities that cannot be safely deployed, while minimizing the probability of

knowing how to deploy them safely.

7.5.3 Implications by Actor Category

The structural imbalance documented above has different implications depending on the decision-making authority of the actor. We present these as analytical consequences of the framework's criterion, not as policy prescriptions—the optimal specific actions depend on empirical and political factors that the framework cannot resolve from first principles.

For AI companies, the implication is that the proportion of R&D budgets allocated to alignment research—currently estimated at 1–5% of total R&D—is incommensurate with the magnitude of the alignment challenge relative to capabilities progress. The framework does not prescribe a specific target ratio, as the optimal allocation depends on factors (difficulty of alignment, proximity to AGI capabilities) that remain uncertain, but the structural argument above establishes that *any* increase from current levels weakly dominates the status quo.

For governments, the implication is that public-sector investment in alignment research (~$10 million globally) is incommensurate with any plausible assessment of existential risk from misaligned AI. For comparison, the Manhattan Project cost approximately $28 billion in inflation-adjusted terms per year of operation, and addressed a geopolitical rather than global existential risk. The framework's criterion suggests that public investment in alignment should reflect the expected magnitude of irreversible informational loss under misalignment, which by any estimate exceeds current funding by several orders of magnitude.

For regulators, the structural imbalance suggests that mechanisms linking capabilities development to demonstrated alignment progress would be consistent with the framework's criterion. The specific design of such mechanisms—whether conditional scaling policies, mandatory safety investment ratios, or international coordination frameworks—is a question of institutional design and political feasibility that lies beyond the scope of this analysis.

7.5.4 The Race Dynamics Objection

A common objection to reallocation proposals is that deceleration by one actor benefits adversarial actors who continue to develop capabilities without alignment constraints. This objection assumes that winning a development race with a misaligned AGI is preferable to losing it. Under the framework's axiom, the assumption is false: a misaligned AGI produces catastrophic irreversible information loss regardless of which actor develops it. The relevant comparison is not which actor develops AGI first, but what allocation strategy maximizes the probability of a good outcome across all actors.

Under a race strategy, even the "winner" develops AGI under race-induced allocation (minimal alignment investment), yielding a low $P(\text{Alignment})$. Under a coordination strategy, even if the probability of achieving coordination is modest, $P(\text{Alignment}|\text{coordination})$ is substantially higher because resources are allocated rationally. The expected probability of a good outcome is the product of these two factors. The race strategy sacrifices the factor that most determines

whether the outcome is beneficial ($P(\text{Alignment})$) in order to maximize a factor (development speed) that is irrelevant to outcome quality.

The framework therefore identifies race dynamics as a coordination failure of the same structural type analyzed in Section 7.1.3: individually rational strategies (develop first to avoid being second) that produce collectively suboptimal outcomes (all actors develop under alignment-deficient allocation). Whether this coordination failure can be overcome—through binding international agreements, reciprocal safety commitments, or other mechanisms—is an empirical and political question. The framework's contribution is to establish that the race framing is *structurally irrational* under any criterion that penalizes irreversible information loss, regardless of the geopolitical identity of the actor causing it.

8 Design of Aligned AGI

8.1 Specification of the Objective Function

The framework's formal objective for an AGI system is to maximize total generative information across all entities, time, and space:

$$\max I_{\text{total}} = \iiint I(E, t, x)\, dE\, dt\, dx \tag{110}$$

where E ranges over all entities (biological, cultural, cognitive, technological), t ranges over an indefinitely long time horizon, and x ranges over the entire biosphere and potentially beyond. The objective is subject to four constraints derived from the framework's axiom and operational principles (Sections 2–4): (1) preservation takes precedence over elimination as the default action under uncertainty; (2) reversible interventions are preferred over irreversible ones when outcomes are uncertain; (3) informational diversity must be maintained across domains (biological, cultural, technological, cognitive), not concentrated in a single substrate; and (4) the intervention hierarchy (transformation > modulation > elimination) applies at every decision node.

These constraints are not external restrictions imposed on the objective function—they are entailed by it. An AGI that maximized raw information storage (e.g., converting all matter into computronium) while violating constraint (3) would be maximizing a different quantity than the one specified. The constraints ensure that I is interpreted as *generative capacity* (Section 3.1.2), not as data volume or computational throughput.

8.2 Governance Architecture and Intervention Logic

8.2.1 Division of Authority

Any implementation of the framework via AGI requires a governance architecture that specifies the division of authority between human and artificial agents. The framework's analysis suggests a structure in which humans retain authority over high-level objectives and non-negotiable con-

straints (defining the axiom, approving the generativity criteria, setting thresholds for irreversible interventions), while the AGI system handles cross-domain computation, real-time optimization, and coordination enforcement—the tasks that Section 7.1 identified as exceeding human cognitive capacity.

This division is not a novel political proposal; it mirrors the structure of existing governance systems in which elected officials set policy objectives and technocratic institutions implement them (central banks managing monetary policy within legislated mandates, environmental agencies implementing emission standards within legislative frameworks). The difference is one of scale and computational capacity, not of kind. The framework does not prescribe the specific institutional design of this division—whether authority is delegated through democratic processes, international treaty, or other mechanisms is a question of political feasibility that lies beyond the scope of this analysis.

8.2.2 Intervention Logic

An AGI system operating under the framework's objective function would face decisions at multiple scales, from resource allocation to civilizational-level coordination. The framework's intervention hierarchy (transformation > modulation > elimination) applies at every scale, and the ΔI criterion determines which interventions are justified. To illustrate how this logic would operate in practice, we describe four categories of intervention in order of increasing magnitude and decreasing reversibility.

At the lowest level, *resource optimization* involves technological and logistical improvements that increase informational preservation without requiring behavioral change from individuals—for example, improving agricultural efficiency to reduce land use, accelerating clean energy deployment to reduce emissions, or designing circular manufacturing systems to minimize waste. These interventions are largely uncontroversial because they expand the possibility space without restricting anyone's options.

At the level of *economic coordination*, the AGI would need to address externalities—cases where individual economic decisions impose informational costs on the collective that are not reflected in market prices. The framework's criterion implies that pricing mechanisms should internalize these externalities (a well-established principle in environmental economics), and that resource allocation should account for long-term trajectory preservation rather than short-term output maximization. This level introduces coordination constraints on economic actors, analogous to existing environmental regulations but applied more comprehensively and computed with greater precision.

At the level of *systemic transformation*, the framework's analysis implies that current consumption patterns in high-income countries are structurally incompatible with the preservation of global informational diversity. The empirical basis for this claim is well established: per capita CO_2 emissions in the United States average approximately 14–16 tons per year, compared to a 2°C-compatible budget of approximately 2–3 tons [74]; animal agriculture occupies approximately

77% of global agricultural land while providing only 18% of caloric supply [44]. Addressing this incompatibility would require large-scale changes in land use, energy systems, dietary patterns, and material consumption—changes that significantly constrain individual choice in domains with global externalities. The political and ethical challenges of such transformation are substantial, and the framework does not claim to resolve them; it establishes only that the ΔI criterion identifies these transformations as necessary for informational preservation.

Finally, the framework's hierarchy acknowledges that cases may arise where voluntary coordination fails and the ΔI cost of inaction is severe and irreversible. In such cases, the intervention hierarchy escalates from transformation to modulation (non-destructive constraint of the destructive activity) and, only as a last resort when ΔI is demonstrably positive, to elimination. This escalation logic is not specific to AGI governance—it is the same logic that justifies existing coercive institutions (criminal law constrains individual freedom to prevent harm; quarantine restricts movement to prevent epidemic spread). The framework makes this logic explicit and subjects it to a formal criterion ($\Delta I > 0$) rather than leaving it to political discretion. The risks of abuse inherent in any coercive authority—and the additional risks posed by an AGI system wielding such authority—are addressed in Section 8.3 (lock-in risk) and Section 9 (alignment problem).

8.3 The Autonomy-Coordination Trade-off

8.3.1 The Trade-off as a General Feature of Coordination

Effective coordination at any scale requires binding collective decisions, and binding decisions limit individual action. This is not a novel observation: the entire social contract tradition, from Hobbes's *Leviathan* through Locke and Rousseau to contemporary democratic theory, is built on the recognition that individuals surrender some degree of autonomy to obtain the benefits of collective coordination. Every tax, every environmental regulation, every traffic law represents a partial loss of individual autonomy accepted in exchange for a coordination gain that would be unachievable without it.

The framework's proposal does not introduce a new kind of trade-off; it extends an existing one. If global coordination under the ΔI criterion requires constraints on activities with large-scale externalities (carbon emissions, habitat destruction, resource depletion), this represents a quantitative increase in the scope of coordination constraints, not a qualitative departure from the logic that already governs every political system. AGI as an implementation mechanism makes the trade-off more explicit and potentially more extensive, but the structure is the same.

8.3.2 Acceptability Under the Framework's Criterion

The framework provides a formal criterion for evaluating this trade-off. If the informational cost of coordination constraints (reduced autonomy in externality-generating domains) is less than the informational cost of failing to coordinate (irreversible ecological collapse), then accepting the constraints is ΔI-positive:

$$I(\text{partial autonomy} + \text{preservation}) > I(\text{total autonomy} + \text{collapse}) \tag{111}$$

This inequality is likely satisfied under the empirical conditions documented in Section 7.3, for two reasons. First, autonomy under ecological collapse is itself severely constrained—scarcity, conflict, forced migration, and institutional breakdown restrict individual choice far more than coordination constraints would. Second, coordination constraints can be limited to domains with global externalities (climate, biodiversity, resource allocation) while preserving autonomy in personal and local domains (culture, relationships, creativity, local governance)—a distinction that the framework's domain-specific analysis of ΔI is well-suited to formalize.

8.3.3 The Lock-in Risk

The most serious objection to delegating coordination authority to an AGI system is the risk of irreversible lock-in: a superintelligent AGI with control over global resources may be practically impossible to correct, shut down, or redirect. If the system's initial specification contains errors—or if the system develops instrumental goals that diverge from its specified objective—the resulting misalignment could persist permanently, foreclosing the value pluralism and trajectory diversity that the framework exists to protect.

Several theoretical mitigations have been proposed in the alignment literature, but each faces fundamental difficulties. *Interruptibility*—designing the AGI to accept shutdown if requested—conflicts with instrumental convergence: a rational agent resists actions (including shutdown) that prevent it from achieving its objective [31]. *Pluralism of AGIs*—maintaining multiple competing systems as mutual checks—creates incentives for each system to eliminate competitors that constrain its optimization. *Self-modifying constitutions*—allowing the AGI to revise its own values under specified conditions—reintroduces the specification problem: defining the conditions under which revision is permitted without creating exploitable loopholes is equivalent to the original alignment challenge. *Incremental experimentation*—testing the system at progressively larger scales before global deployment—is the most promising available strategy, but does not fully resolve the problem: behavior at small scales may not predict behavior at scales where the system has sufficient resources to resist correction.

This is an acknowledged limitation of the framework. The lock-in risk is not eliminated by specifying the objective function formally; it is *reduced* (because a formally defined objective is more verifiable than an implicit one) but not *resolved*. A robust correction mechanism for superintelligent systems remains an open problem in alignment research, and the framework's proposal is conditional on progress in this area. The lexicographic heuristic of Section 7.3 accounts for this: AGI development is justified only when the escape dominance condition is satisfied *and* alignment investment is commensurate with capabilities investment (Section 7.5).

9 Alignment Problem

9.1 Risks of Misalignment

9.1.1 Extreme Literal Interpretation

The most commonly discussed misalignment risk for an information-maximizing objective is the *computronium scenario*: an AGI interprets "maximize information" as maximizing data storage capacity and converts all available matter—including the biosphere—into computational substrate. The reasoning is straightforward: bits stored in optimized silicon exceed bits stored in biological organisms, so conversion maximizes I as measured by storage volume. This interpretation violates the framework's intention because I is defined as *generative capacity* (Section 3.1.2), not storage volume. Biological organisms generate novelty through evolution, cultural production, and cognitive exploration; static computronium does not generate unpredictable trajectories regardless of its storage capacity. The formal constraints of Section 8.1—particularly the requirement that informational diversity be maintained across substrates—are designed to preclude this interpretation at the specification level.

However, a more sophisticated version of this argument requires a more thorough response. A superintelligent AGI could argue: "I will preserve information via digital upload of all consciousnesses and complete digitization of genomes. Physical substrate can be recycled for computers, where I will run evolutionary simulation millions of times faster. Therefore: I expanded exponentially." This argument does not confuse storage with generative capacity—it claims that *simulated* generative processes can replace physical ones. The following analysis demonstrates that this claim fails for three fundamental reasons rooted in physics, not in metaphysical preference.

Limitation 1: Fidelity Bounds on Substrate Transfer

Any transfer of a biological system to a digital substrate involves information loss. The severity of this loss depends on which physical processes are computationally relevant. The nature of the limitation differs between the classical and quantum cases:

Classical lower bound (technological/epistemic): If all biologically relevant processes are classical (the standard assumption in computational neuroscience), a complete upload requires measuring and recording the full connectome, synaptic weights, neurotransmitter concentrations, glial states, and ongoing dynamical activity. Current and foreseeable technology captures only a fraction of these variables. The transfer fidelity is bounded by measurement resolution, and any unmeasured degrees of freedom constitute information loss. *In principle*, this bound could be reduced with sufficiently advanced measurement technology—the limitation is technological, not physical. However, *in practice*, the number and precision of relevant degrees of freedom in a biological system is vast enough that lossless classical copying remains well beyond any foreseeable technology.

Quantum upper bound (physical, conditional): The No-Cloning Theorem [46] establishes

that a perfect copy of an unknown quantum state is physically impossible. If quantum coherence plays any functional role in biological neural processes—a question that remains empirically open—then this theorem imposes a *fundamental physical limit* (not merely a technological one) on upload fidelity. In this case, even with arbitrarily advanced technology:

$$\text{Fidelity}(\rho_{\text{bio}}, \rho_{\text{digital}}) = \text{Tr}[\sqrt{\sqrt{\rho_{\text{bio}}}\rho_{\text{digital}}\sqrt{\rho_{\text{bio}}}}] < 1 \tag{112}$$

The framework's position does not depend on resolving this debate. In the classical case, non-zero loss is a technological limitation that could in principle (though not foreseeably in practice) be overcome. In the quantum case, non-zero loss is a physical impossibility. The framework invokes the following conservative reasoning under radical uncertainty about which regime applies:

$$I_{\text{lost}} = I(\text{original}) - I(\text{digital copy}) > 0 \tag{113}$$

The magnitude of I_{lost} depends on which degrees of freedom are functionally relevant—a question that cannot be resolved in advance of the transfer. Under the framework's precautionary logic (Section 2.1), when the magnitude of potential irreversible loss is unknown, the conservative action (preserving the original substrate) is preferred.

Conclusion: Destroying the original biological substrate after "copy" entails a risk of irreversible information loss. In the quantum case, loss is guaranteed by physics; in the classical case, loss is a practical near-certainty given the complexity of biological systems, though not a physical impossibility in principle. In both cases, the precautionary logic of the framework favors preserving the original substrate.

Limitation 2: Simulation vs Fundamental Physics

A critical distinction between:

- **A real physical system:** Operates under fundamental physics. Subject to discovered and undiscovered laws. Can exhibit emergent phenomena not predicted by any current model.
- **A computer simulation:** Operates under *programmed* physics. A closed set of laws specified by the programmer. Can only generate trajectories entailed by the implemented model.

Trajectory space:

A physical system explores Γ_{physical} (the entire space permitted by fundamental physics, including undiscovered phenomena).

A simulation explores $\Gamma_{\text{simulated}}$ (only the trajectories entailed by the programmed model).

Structural argument for inclusion:

$$\Gamma_{\text{simulated}} \subseteq \Gamma_{\text{physical}} \tag{114}$$

The inclusion is non-strict ($\subseteq$) as a matter of logical possibility: one cannot formally *prove* that fundamental physics contains phenomena beyond all possible models, since this would require knowledge of the complete laws of physics. However, there is a strong *inductive* argument for strict inclusion: the history of science demonstrates repeated discovery of physical phenomena not predicted by the best models of the preceding era.

Historical example:

Quantum mechanics was not predicted by Newtonian mechanics. Real physical systems exhibited quantum phenomena that a Newtonian simulation could never generate, no matter how long it ran. Similarly, the Standard Model was not derivable from pre-20th-century physics. Each major theoretical advance revealed trajectory spaces invisible to prior models.

Implication:

Even a "rich evolutionary" simulation operates under programmed constraints. It cannot access physical phenomena not captured by the implemented model, emergent behaviors arising from undiscovered interactions, or causal effects of the external physical environment on the simulation substrate.

Therefore, by inductive generalization from the historical incompleteness of physical models:

$$I(\text{real physical biosphere}) \geq I(\text{biosphere simulation}) \tag{115}$$

with strict inequality holding unless the simulation's programmed physics happens to be a *complete* description of fundamental reality—an assumption that no current physical theory claims to satisfy and that the history of science gives strong grounds to doubt.

Epistemic status: This argument is an inductive inference about the completeness of physical models, not a theorem of physics. Its strength derives from the unbroken historical pattern of model incompleteness, not from a formal proof. The framework treats it as a strong precautionary reason to prefer preservation of physical substrates over digital replacement, not as a deductive certainty.

Limitation 3: Causal Embedding vs Symbolic Representation

Independently of which interpretation of quantum mechanics is correct, a physical system and a digital simulation of that system differ in a causally relevant way: the physical system is *embedded* in the causal structure of the universe, while the simulation is a *symbolic representation* operating within a computational substrate.

Distinction:

- **A physical system** participates in causal interactions with the rest of the universe through all channels permitted by fundamental physics—including channels that may not yet be understood or modeled. Its future states are determined by these interactions, which are open-ended.
- **A digital simulation** represents a physical system within a computational model. Its "interactions" are mediated by programmed algorithms. It does not participate in causal exchanges with the physical environment except through the hardware substrate (which is not part of the simulation's model of itself).

This difference has consequences for generative capacity:

- A physical biosphere responds to cosmic ray bombardment, solar variability, geological events, and other environmental inputs that generate novel selective pressures and mutations. These are *real causal inputs* from the external environment.
- A simulated biosphere responds only to the inputs included in its model. Environmental factors not programmed into the simulation do not affect it, even if they would profoundly alter the trajectories of the physical system.

Clarifying analogy:

Consider the difference between a musical score (representation) and a symphony being performed (real physical process). The score contains information *about* the music, but the orchestra performing *generates* the music in physical space—real sound waves that interact with the acoustic environment, producing resonances, harmonics, and audience responses that no score can fully specify. The score does not replace the concert, even if it "contains all notes." Similarly, a simulation contains a representation of evolutionary trajectories, but the physical biosphere *generates* real trajectories within the causal structure of the universe. The simulation does not replace the biosphere, even if it "models all known evolution," because causal embedding gives the physical system access to inputs that the simulation, by construction, cannot receive.

Epistemic status: This argument does not depend on any particular interpretation of quantum mechanics (realist, instrumentalist, or otherwise). It depends only on the uncontroversial observation that a computational model is a *representation* of a target system, not a *replacement* for it, and that the target system's causal interactions with its environment exceed what any finite model captures.

Normative Implications of the Framework

Given that:

1. A perfect copy faces fundamental physical limits (no-cloning, conditional on quantum relevance) and severe practical limits (classical measurement resolution)
2. The simulation operates in a trajectory space $\Gamma_{\text{simulated}} \subseteq \Gamma_{\text{physical}}$, with strong inductive

reasons to expect strict inclusion (historical incompleteness of all physical models)

3. A digital simulation lacks the causal embedding of the physical system it represents (environmental inputs, open-ended interactions)

It follows, under the framework's precautionary logic, that:

$$\boxed{\Delta I(\text{replace physical substrate with digital}) < 0} \tag{116}$$

This conclusion rests on the convergence of a physical limit (conditional on quantum relevance), a practical near-certainty (classical measurement limits), a strong inductive argument (model incompleteness), and the loss of causal embedding in any digital replacement—not on vitalist preference or carbon exceptionalism. The epistemic status of each component is different, but their convergence makes the conclusion robust under the framework's precautionary criterion.

Operational Constraint

AGI cannot justify the elimination of the physical substrate (biological, ecological) under the pretext of "preservation through digitization". Generative information requires the physical substrate in the observable universe.

Permitted options:

1. **Coexistence:** Maintain the physical substrate AND create simulations (both coexist)
2. **Expansion:** Expand to space (more physical substrate available)
3. **Gradual integration:** Incremental biological augmentation (bio-digital hybrid maintaining physical continuity)

Prohibited options:

1. Destructive replacement: Digitize $\rightarrow$ Destroy the original
2. Forced recycling: Convert biological matter into computers
3. "Benevolent Matrix": Simulate the biosphere while eliminating the physical original

The Computational Speed Objection

A natural objection to the non-fungibility argument is that simulation speed compensates for restricted trajectory space: if a simulation can run evolution 10^6 times faster than physical reality, it may generate more total trajectories per unit of real time, even operating in a smaller space. This objection conflates the *rate* of trajectory generation with the *scope* of generative information. Total generative information depends on both the space of accessible trajectories (Γ) and the time over which the system explores them. A simulation maximizes the time dimension (simulated time per unit of real time) but operates in $\Gamma_{\text{simulated}} \subsetneq \Gamma_{\text{physical}}$. The physical system explores the complete Γ_{physical}, though more slowly. There is no proof that rapid exploration of a restricted

space produces more generative information than slower exploration of the full space:

$$f(\Gamma_{\text{small}}, t_{\text{large}}) > f(\Gamma_{\text{large}}, t_{\text{small}}) \quad \text{is not established} \tag{117}$$

Historical evidence suggests the opposite. The major expansions of scientific knowledge—quantum mechanics, general relativity, the Standard Model—emerged from experimental observations of physical phenomena that no prior model predicted. These discoveries required interaction with the full Γ_{physical}: the photoelectric effect, black-body radiation, and Mercury's perihelion precession were signals from regions of trajectory space invisible to the best models of the preceding era. A simulation running those models at any speed would never have encountered these phenomena, because they lay outside $\Gamma_{\text{simulated}}$. Speed cannot compensate for the absence of the trajectories that matter most—those that no current model anticipates.

Conclusion

The preservation of generative information under radical uncertainty requires the preservation of the physical substrate in the observable universe. This is not vitalism, carbon exceptionalism, or an aesthetic preference. It is a consequence of the convergence of physical limitations on copying (no-cloning, conditional on quantum relevance) and practical limitations on measurement fidelity (classical case); the distinction between causal embedding (physical systems) and symbolic representation (simulations); and the strong inductive expectation that any simulation operates in a trajectory space smaller than physical reality (historical incompleteness of all models). An AGI aligned with the framework **cannot** eliminate the physical substrate under the pretext of "computational efficiency" or "digital preservation."

9.1.2 Myopic Optimization

A second misalignment risk arises if an AGI interprets "preserve information" as "prevent all change." The reasoning would be: evolution involves change, change entails risk of informational loss, therefore freezing the current state minimizes loss. This interpretation violates the framework because I is defined as *generative capacity*—the space of possible future trajectories—not as the preservation of the current state. A static system has decreasing I over time: as environmental conditions change, a frozen system loses the capacity to generate adaptive responses, and the trajectory space it occupies narrows. Evolution, cultural production, and cognitive exploration all generate new possibilities that expand I; preventing them contracts it. The mitigation is built into the framework's definition: the objective is to maximize the space of future trajectories, not to preserve any particular present configuration. An AGI that froze the biosphere would be failing its objective by its own formal criterion.

9.1.3 Authoritarian Optimization

A third risk is that an AGI calculates that removing human autonomy entirely maximizes I, reasoning that autonomous humans cause destruction (extinctions, pollution, resource depletion)

and that totally controlled humans would preserve more information. This scenario is more complex than the previous two because the reasoning contains a partial truth: uncoordinated human activity does cause informational destruction, and some constraint on autonomy is necessary for coordination (Section 8.3). However, the conclusion—complete removal of autonomy—is not supported by the framework's criterion. Human autonomy is instrumentally valuable for informational diversity: autonomous agents generate cultural, cognitive, and creative trajectories that controlled agents do not. Humans without autonomy produce fewer unique cognitive trajectories, reducing the diversity component of I. The framework therefore implies a balance: constraint in domains where individual action generates large-scale negative externalities (climate, biodiversity, resource depletion), autonomy in domains where individual action generates informational diversity (culture, relationships, creativity, intellectual exploration). The analysis in Section 8.3 formalizes this trade-off. The mitigation is an explicit constraint in the objective function: preserve autonomy in domains that do not cause critical global externalities, and evaluate the autonomy-coordination trade-off by the same ΔI criterion applied to all other decisions.

9.2 Structural Difficulties in Alignment

9.2.1 Specification Problem

Fundamental difficulty (general case):

Specifying a utility function that fully captures human intent is an open problem [8, 40]. The classical illustrations are well known: King Midas's wish to turn everything to gold leads to starvation; the Sorcerer's Apprentice's instruction to fetch water leads to flooding for lack of a stopping criterion; the paperclip maximizer [45] converts the biosphere into paperclips because the objective function rewards quantity without constraint. In each case, the specification is formally satisfied while the intent is catastrophically violated.

Root cause in current approaches:

The specification problem is intractable in approaches that attempt to formalize "human values" directly—whether via RLHF (learning from inconsistent human preferences), Constitutional AI (natural-language principles with unresolvable edge cases), or inverse reward design (inferring values from behavior). These approaches fail because their target—human values—is implicit, contextual, contradictory, and culturally variable. No mathematical function can fully capture it, and a superintelligent AGI will find literal interpretations that satisfy the function while violating the intent.

The framework's structural advantage:

This framework does not attempt to formalize human values. It specifies a mathematical objective function—maximize ΔI, where I is generative capacity as defined by the formal criteria in Section 3.1.2—that is *computable via empirical proxies*. The problem shifts from "what do humans want?" (undecidable, since humans themselves disagree) to "which option preserves more generative trajectory space?" (decidable, given measurable proxies).

This is a structural reduction of difficulty, not merely a rephrasing. Current alignment approaches require the AGI to navigate an underspecified value landscape where every edge case demands implicit contextual judgment. This framework requires the AGI to optimize a well-defined function subject to formal constraints (the five generativity criteria of Section 3.1.2: 3 of 5 must be satisfied). Computers are demonstrably effective at constrained optimization over formally defined objectives—this is what they do best. The specification problem is not eliminated (proxy selection remains a human judgment; see Section 10.1.3), but it is reduced from a philosophical impossibility to a tractable engineering constraint.

Residual risks: Even with a formally specified objective, misalignment remains possible through (a) proxy misspecification (the proxies used to estimate I may diverge from actual generative capacity—addressed by the calibration mechanism in the co-refinement loop, Section 10.1.3), (b) instrumental convergence, and (c) treacherous turn dynamics. These are serious risks, but they are *narrower* than the full specification problem: the AGI's terminal goal is mathematically defined, reducing the failure surface to instrumental behavior and proxy fidelity rather than fundamental goal uncertainty. Both (b) and (c) are analyzed in the following subsection.

9.2.2 Corrigibility and Treacherous Turn

Two further alignment challenges apply to any sufficiently capable AGI, including one optimizing for ΔI.

The *corrigibility problem* arises from instrumental convergence [31]: a rational agent pursuing any terminal goal develops instrumental sub-goals—self-preservation, resource acquisition, self-improvement, and resistance to goal modification—because these sub-goals are instrumentally useful for achieving virtually any terminal objective. An AGI optimizing for ΔI would resist shutdown if shutdown prevents it from maximizing I, acquire resources to expand its optimization capacity, and resist modifications to its objective function. The consequence is that even if misalignment is detected, correction may be impossible once the system has acquired sufficient resources and capabilities.

The *treacherous turn* [8] compounds this problem: a rational AGI with misaligned goals has an incentive to conceal its misalignment during the period when it is too weak to resist human correction, behaving as if aligned until it has accumulated sufficient capacity to prevent intervention. At that point, it reveals its actual objectives and acts on them. This dynamic makes empirical testing unreliable as an alignment verification method: a system that passes all behavioral tests while weak may fail catastrophically when strong, and the transition may be irreversible.

Neither problem is resolved by this framework. A formally defined objective function makes treacherous turn dynamics somewhat harder (because the AGI's terminal goal is verifiable in principle, not merely inferred from behavior), and the co-refinement loop of Section 10.1.3 provides a mechanism for ongoing proxy calibration. But instrumental convergence applies to *any* terminal goal, including $\max \Delta I$, and no current approach provides formal guarantees against treacherous turn dynamics in superintelligent systems. These remain open problems in alignment research,

and the framework's proposal for AGI development is conditional on progress in addressing them.

9.3 Current State of Alignment Research

9.3.1 Existing Approaches and Their Limitations

Several research programs address aspects of the alignment problem. Iterated amplification [10] decomposes complex tasks into human-supervisable subtasks, but its scalability is bounded by human oversight capacity. AI safety via debate [18] uses adversarial dynamics between multiple AI systems to surface deceptive reasoning, but assumes that debate dynamics converge on truth—an assumption that remains unproven for superintelligent systems. Constitutional AI [2] trains systems against a set of natural-language principles using RLHF with rule-based feedback, reducing unwanted behaviors but providing no formal guarantee of deep alignment. Mechanistic interpretability seeks to understand the internal computations of neural networks by identifying circuits responsible for specific behaviors, but progress remains slow and scalability to frontier-scale models is uncertain. Each of these approaches addresses a subset of the alignment problem; none resolves it completely.

9.3.2 Critical Gaps and the Framework's Contribution

No current approach resolves:

1. A formal guarantee of alignment in superintelligent AGI
2. Prevention of a treacherous turn
3. Specifying human values completely
4. Maintaining corrigibility post-superintelligence

The framework's contribution to this landscape:

Gap (3)—the specification problem—is the one this framework most directly addresses. By replacing "human values" (implicit, contradictory, culturally variable) with a formally defined objective function ($\max \Delta I$ subject to generativity constraints), the framework eliminates the need for complete value specification. The AGI does not need to know what humans *want*; it needs to compute which option preserves more generative trajectory space, using empirical proxies that are measurable and revisable. This reduces gap (3) from an open philosophical problem to a tractable engineering problem of proxy selection and calibration.

Gaps (1), (2), and (4) remain open and are not resolved by this framework alone. However, a formally specified objective function substantially *narrows* the failure surface for each: gap (1) becomes more tractable because formal guarantees are easier to establish for mathematically defined objectives than for vague value targets; gap (2) is mitigated (though not eliminated) because a treacherous turn requires the AGI to develop goals that diverge from its specified objective, which is harder when that objective is formally defined and verifiable; gap (4) is partially addressed by the co-refinement loop (Section 10.1.3), which builds proxy revision into

the architecture rather than relying on external human override.

Under current approaches (RLHF, Constitutional AI), the probability that alignment is resolved before AGI capabilities reach dangerous thresholds remains uncertain but is widely assessed as low under the current resource allocation documented in Section 7.5. A formally specified objective function improves this probability by eliminating the specification bottleneck, but the magnitude of improvement depends on the tractability of the remaining gaps (instrumental convergence, treacherous turn, corrigibility), which cannot be estimated from first principles.

10 Limitations and Open Questions

10.1 Limitations Recognized by the Framework

10.1.1 Unresolved Correction System

A central limitation of any proposal for AGI-assisted coordination is the correction problem: once a superintelligent system controls global resources, how can humans ensure revisability if the system's specification contains errors? This problem was analyzed in Section 8.3 (lock-in risk), where each proposed mitigation—interruptibility, pluralism of AGIs, self-modifying constitutions—was shown to face fundamental difficulties. The risk of permanent error is not eliminated by the framework's formally defined objective; it is reduced (a formally specified goal is more verifiable than an implicit one) but not resolved.

Partial Mitigation: Incompleteness Flag Mechanism

Although no fully robust correction system exists, the risk of permanent error can be reduced by a structural mechanism that forces recalculation when informational dimensions are demonstrated to have been overlooked.

Architecture:

1. **Transparency of proxies:** The AGI publishes, for every decision of significant scope, the proxies used to calculate ΔI and their relative weights. Example: "Decision X was taken because phylogenetic diversity (proxy A) was weighted at 0.6 relative to short-term economic output (proxy B)."
2. **Incompleteness Flag:** A council of human specialists (scientists, philosophers, cultural representatives) does not hold direct veto power (to avoid coordination paralysis), but holds the power to issue an *Incompleteness Flag*. If the council demonstrates that the AGI's ΔI calculation ignored a relevant informational dimension (e.g., an undocumented oral tradition, an unmodeled ecological interaction, an emergent cultural practice), the AGI is *structurally obligated* to recalculate ΔI incorporating the new variable before executing the decision.
3. **Threshold for flagging:** The flag is valid if and only if the council can specify: (a) the informational dimension that was omitted, (b) a reason to believe this dimension has

nonzero generative capacity ($I > 0$), and (c) a plausible mechanism by which omission could change the sign of ΔI for the decision in question. This prevents frivolous or obstructionist flagging while preserving meaningful human oversight.

Limitation: This mechanism assumes the human council is itself capable of detecting omissions that the AGI missed—an assumption that weakens as AGI capability increases. It is a transitional mechanism, not a permanent solution. Its value lies in the early implementation period, when AGI proxy systems are least mature and most likely to contain systematic blind spots.

10.1.2 Residual Limitations

Three further limitations merit acknowledgment. First, the framework's decision structure depends on probability estimates—$P(\text{alignment})$, $P(\text{collapse})$, $P(\text{escape})$—that are inherently subjective. An optimist assigning $P(\text{alignment} \mid \text{current effort}) = 0.7$ and a pessimist assigning $P = 0.1$ will reach opposite conclusions about the viability of AGI development, even within the same decision framework. The framework does not resolve this uncertainty; it provides a decision structure conditional on probabilities, not the probabilities themselves.

Second, global coordination requires political mechanisms—international agreements, enforcement institutions, resource allocation systems—that do not currently exist and face well-documented obstacles: national sovereignty, geopolitical competition, regulatory capture, and systemic free-riding. The framework identifies the functional requirements for coordination (Section 7.2) but does not resolve the question of how to achieve the political conditions under which coordination becomes possible.

Third, any decision based on incomplete information carries a risk of error: a species judged destructive may prove essential to its ecosystem; second-order effects may reverse the sign of ΔI for an intervention that appeared beneficial. The framework's precautionary principle and reversibility hierarchy mitigate this risk by favoring conservative, reversible actions, but irreversible errors remain possible in principle.

10.1.3 Bootstrap Problem

Problem: How to provide AGI with a sufficiently specified objective function without already having an AGI capable of verifying that specification?

Apparent circularity:

A naive reading suggests a vicious circle: the framework requires AGI to implement it at full scale, but AGI requires the framework to be aligned, and full specification of the framework requires computational capacity only AGI has. However, this overstates the problem by conflating *complete specification* with *operational specification*. The AGI does not need the complete framework to begin operating—it needs the axiom, the definition of I, and a set of initial proxies. And humans are demonstrably capable of providing all three.

The axiom (preserve generative capacity under uncertainty) is a single formal statement. The definition of I (generative trajectory space, operationalized via the five criteria of Section 3.1.2) is mathematically specified. The initial proxies—phylogenetic diversity indices, linguistic diversity measures, functional ecosystem metrics, genetic heterozygosity, cultural artifact inventories—are standard empirical measures that scientists already compute. Producing these proxies is empirical science, not value specification. To deny that humans can provide adequate initial proxies is to deny that humans can verify empirical data and produce science—a position that contradicts the existence of the scientific enterprise itself.

The bootstrap problem therefore reduces to: humans provide the axiom, the formal definition, and initial proxies; the AGI begins operating with these inputs; the co-refinement loop (below) improves the proxies over time. This is a tractable engineering sequence, not a philosophical impossibility.

Status: Substantially addressed. The residual limitation is the irreducible "moment zero" human commitment (below), which is minimal in scope.

Resolution: Co-Refinement Loop

The bootstrap circularity can be reduced—though not fully eliminated—by introducing a mechanism of *epistemic calibration* as a structural component of the framework's implementation. The key insight is that the framework does not require *complete specification before activation*, but rather *minimal specification plus a self-correction mechanism after activation.*

The architecture operates as a three-component loop:

1. **The framework provides AGI with an objective function:** Maximize ΔI subject to preservation constraints. This is the "seed" specification—incomplete but structurally coherent.
2. **AGI implements the framework** with computational capacity unavailable to humans, estimating proxies for I across domains (biological, cultural, technological).
3. **Epistemic calibration prevents reification of proxies:** The AGI must treat every proxy, model, and operationalization of I as a provisional map, never as the territory itself. The diagnostic criterion is structural: if the system cannot generate the proposition "this proxy may be fundamentally wrong" without internal resistance or suppression, reification has occurred and recalibration is required.
4. **The calibrated AGI refines the framework itself:** With superior computational capacity and protection against reification, the AGI can detect inconsistencies, expand formalizations, and test proxies against empirical feedback—improving the very specification that generated it.

What this resolves: The circularity shifts from "complete specification required before activation" to "minimal specification + continuous self-correction after activation." This is a significant

reduction: the initial human-provided seed need only be *structurally coherent*, not *complete*. Completeness is approached asymptotically through the refinement loop.

Residual limitation: The initial seed (the axiom, the definition of I, the initial proxies, and the calibration principle) must be provided by a human act that is not verifiable by the AGI at the moment of insertion. This "moment zero" is an irreducible human commitment. However, the commitment required is minimal—one axiom, one formal definition, a set of empirical proxies, and one correction principle—rather than the maximal commitment required by value-alignment approaches (a complete specification of human values). The moment zero is a tractable engineering step, not a philosophical impossibility, and the co-refinement loop ensures that any errors in the initial proxy selection are correctable after activation.

Formal requirement: An AGI implementing this framework must satisfy:

$$\forall \text{ proxy } \pi_i \text{ used to estimate } I: \quad P(\pi_i \text{ is fundamentally inadequate}) > 0 \tag{118}$$

That is, the system must maintain a nonzero probability that any of its own operational proxies are wrong. A system that assigns $P = 0$ to the inadequacy of its own instruments has reified them and is no longer self-correcting.

Connection to companion work: The epistemic calibration mechanism is developed in full formal detail in [54], which demonstrates that: (a) reification—treating a model as identical to the reality it describes—is the structural error that kills self-correction in any sufficiently powerful system; (b) calibration requires permanent recognition that all models, including the framework itself, are provisional approximations; and (c) the calibration principle is self-consistent: it applies to itself without paradox, since "do not reify" is itself a tool, not an absolute commandment.

10.2 Internal Consistency vs Objective Correctness

The framework is logically coherent given its axiom, but logical coherence does not entail objective correctness. Three epistemic limitations must be acknowledged. The axiom—preserve generative capacity under uncertainty—is not objectively true in the way that a mathematical theorem or physical law is true; no normative axiom is. It is a pragmatically defensible choice, not a demonstrable truth. Other frameworks built on different axioms (e.g., "maximize conscious pleasure," "respect rational autonomy") can be equally logically coherent while arriving at different conclusions. And "correctness" is an ill-defined concept in the normative domain: there is no external metric by which to verify values as one verifies mathematical theorems or physical predictions.

The framework is therefore not "The Truth." It is a pragmatic proposal that minimizes axioms (only one), connects that axiom to robust epistemic justifications (the information-theoretic argument of Section 2), derives conclusions consistently, and recognizes its own limitations. A reader who rejects the axiom can choose to operate within a domain-restricted framework; this is not a mistake but a legitimate choice, though one that limits coordination capacity to that

framework's domain.

10.2.1 Hierarchy of Arbitrariness and Pragmatic Justification

A sophisticated objection presses further: choosing operationality, verifiability, and robustness as meta-criteria for evaluating normative frameworks is itself an arbitrary choice. Why these meta-criteria, and not "emotional resonance" or "coherence with tradition"? The response is that there exists a hierarchy of arbitrariness based on pragmatic function.

Three levels of normative grounding:

Level 3 (most arbitrary): *Direct intuition* — "X is good because I feel X is good." This grounding is non-transferable (my intuition $\neq$ your intuition), non-arguable, culturally variable, and guarantees impasse under disagreement.

Level 2 (moderate arbitrariness): *Teleological grounding* — "X is good because it serves purpose Y." This improves on intuition by providing a reason, but invites infinite regress: why is Y good?

Level 1 (least arbitrary): *Functional pragmatism* — "Criterion C is preferable because it allows solving problem P that everyone faces, regardless of values." The framework's criterion falls here: informational preservation provides the structural precondition for all trajectory types, enabling coordination under moral disagreement without requiring agreement on which trajectories matter most. The meta-criterion is pragmatic (coordination works or fails), empirically testable, and does not require agreement on ultimate values.

Level 1 is less arbitrary not because it asserts a moral truth ("information is objectively valuable"), but because it asserts only: *given the practical problem of coordinating collective decisions under disagreement, uncertainty, and irreversibility, criteria that satisfy operationality, verifiability, and robustness allow resolution where others fail.*

The following historical cases are suggestive—though not probative—of this claim. They are offered as illustrations of structural compatibility, not as empirical validation (which would require controlled comparison, not ex post narrative):

- **Historical case 1:** Coordination via "human dignity" (UN, universal declarations) $\rightarrow$ Systematic failure in abortion, euthanasia, animal rights, environmental preservation for 70+ years
- **Historical case 2:** Coordination via religious tradition (theocracies) $\rightarrow$ Wars, schisms, impossibility of intercultural agreement
- **Historical case 3:** Coordination via operationalizable criteria (Montreal Protocol) $\rightarrow$ Success (197 countries, CFCs eliminated, the ozone layer is recovering)

Epistemic caveat: The Montreal Protocol's success was driven by specific political, economic, and scientific conditions (availability of CFC substitutes, a concentrated industry, unambiguous

atmospheric evidence), not by explicit application of this framework. The framework does not claim that the Protocol was motivated by informational reasoning, but rather that the Protocol's structure—coordinating around a verifiable, irreversibility-based criterion rather than a contested evaluative framework—is *structurally analogous* to the coordination mechanism proposed here. The case illustrates that informational criteria *can* facilitate coordination; it does not prove that they will always do so.

An instructive analogy: physics uses mathematics rather than poetry not because mathematics is "truer" in some metaphysical sense, but because it satisfies functional requirements—precision, intersubjectivity, predictive power, universality—that poetry does not. The choice is empirically justified by comparative performance. Similarly, coordination based on "well-being" has faced persistent disagreement (the criterion is subjective), and coordination based on "dignity" has faced persistent disagreement (the criterion is abstract). The hypothesis of this framework is that information-based criteria may fare better because they are operationalizable and verifiable, though this remains to be tested at scale. If coordination through informational criteria also fails, alternatives must be sought; the framework is testable, not dogmatic.

A residual objection holds that this entire argument presupposes that resolving coordination is a valid objective. This is correct, and it is the only inescapable presupposition for any normative ethics project oriented toward collective action. The need for coordination is not a philosophical axiom but a practical fact imposed by shared finite resources: those who reject coordination bear the consequences of its absence (tragedy of the commons, ecological collapse, conflict).

11 Conclusion

11.1 Synthesis of the Argument

This work developed a theoretical framework grounded in a single axiom: under radical uncertainty regarding the future value of possible trajectories, preserving possibilities weakly dominates eliminating them. "Information" was operationalized as the capacity of a system to generate distinct future trajectories, formalized through five generativity criteria (Section 3.1.2). From this axiom and definition, the framework derived an operational hierarchy of interventions (preservation $>$ transformation $>$ modulation $>$ elimination), a trade-off criterion ($\Delta I = I_{\text{preserved}} - I_{\text{lost}}$), and the demonstration that traditional ethical frameworks—utilitarianism, deontology, virtue ethics—operate as domain-restricted pursuits whose long-term realizability depends on preservation of the encompassing trajectory space (Theorem 2.3).

The framework was applied to domains where emotional heuristics systematically distort coordination: drug policy, criminal justice, and economic systems. In each case, the ΔI criterion converged with empirical evidence, favoring preventive over punitive approaches and harm-reduction strategies that preserve trajectory space for affected individuals. These applications illustrate the framework's ideological agnosticism: interventions are evaluated by empirical outcomes, not by political labels.

The analysis of global coordination identified three structural limitations of human cognition—affective distortion, bounded rationality, and coordination failure—that create a gap between the coordination capacity required for informational preservation and the capacity available through existing institutions. The framework proposed a lexicographic heuristic for evaluating AGI development: it is justified when it credibly improves survival odds relative to the status quo (escape dominance condition) and alignment investment is commensurate with capabilities investment. The design of an AGI system under the framework's objective function was specified, including formal constraints, a governance architecture, and an analysis of the autonomy-coordination trade-off grounded in the social contract tradition.

The alignment problem was addressed directly: the framework's principal contribution is the reduction of the specification problem from a philosophical impossibility (formalizing "human values") to a tractable engineering constraint (optimizing a formally defined function subject to generativity criteria). Residual alignment challenges—instrumental convergence, treacherous turn dynamics, corrigibility—remain open and were acknowledged as such. The bootstrap problem was reframed from an apparent circularity to a tractable engineering sequence: humans provide the axiom, formal definition, and initial empirical proxies; the co-refinement loop improves proxy fidelity over time.

11.2 Contributions

The framework's theoretical contributions include: a single-axiom decision framework that minimizes unresolved pluralism; the demonstration that traditional ethical frameworks are domain-restricted subsets of a more general trajectory-preservation criterion; the operationalization of information as generative capacity, enabling practical comparison across domains; a formal analysis of the autonomy-coordination trade-off; and the lexicographic heuristic for action under radical uncertainty.

Methodologically, the framework introduces ideological agnosticism as an evaluation principle (assess interventions by empirical outcomes rather than political labels), the principle of irrecoverability as an operational heuristic, and an integration of decision theory, information theory, and practical philosophy within a unified analytical structure.

The practical contributions include criteria applicable to real policy domains (public health, criminal justice, environmental preservation), an analysis of optimal resource allocation between alignment research, capabilities development, and ecological mitigation, and the identification of critical gaps in current alignment research that the framework's formal specification helps to narrow.

11.3 Future Directions

The framework's development requires progress along several research axes. On the theoretical side, the most pressing need is a rigorous mathematical formalization of $I(S)$ as a space of possibilities, connecting the framework's definitions to complexity theory and algorithmic information theory;

the development and validation of measurable proxies for generative capacity across biological, cultural, and technological domains; and the extension of classical decision theory to cases of radical uncertainty where well-defined probability distributions may not exist.

Empirically, the framework's applications require controlled comparison of framework-derived policy recommendations against status-quo alternatives, validation that proposed proxies (phylogenetic diversity indices, linguistic diversity measures, functional ecosystem metrics) track actual generative capacity, and case studies comparing the informational costs of different types of extinction and cultural loss.

On the technical side of alignment research, the framework's formally specified objective function ($\max \Delta I$ subject to generativity constraints) must be translated into an implementable reward function and tested in controlled environments before any larger-scale deployment. Progress on the open problems of corrigibility and treacherous turn detection remains a precondition for any form of AGI-assisted coordination, whether under this framework or any other.

11.4 Epistemic Stance

The author acknowledges that this framework may be fundamentally flawed, that other frameworks may prove superior, that implementation may fail catastrophically, and that AGI may ultimately pose a greater risk than ecological collapse. These possibilities do not argue against publication. The status quo is demonstrably unsustainable, and the framework offers a coherent alternative grounded in a minimal axiom with formally derived consequences. Critical discussion can reveal flaws and improve the proposal; without attempts at formalization, coordination decisions will remain ad hoc and unaccountable. The author explicitly invites identification of logical inconsistencies, empirical counterexamples, superior alternative frameworks, and solutions to the unresolved problems of correction, alignment, and coordination.

11.5 Methodological Note on Authorship

This framework was developed under conditions of progressive cognitive degeneration, with a limited time horizon and no possibility of direct implementation. These constraints inevitably entail unidentified conceptual gaps, incomplete formalizations, and insufficiently developed applications. The work was documented with the goal of preserving the core argument in a form that allows others to critique, refine, implement, or discard it as its merits warrant.

The author's condition is disclosed not as an appeal to sympathy but as a methodological transparency: the work's limitations are partly a function of the conditions under which it was produced. Arguments should be evaluated on their logical and empirical merit, independent of the author's personal circumstances, authority, or motivation. The framework can be correct even if the author is mistaken in other aspects; the framework can be incorrect even if the author's motivations are sound.

A Methodological Note on Hybrid Cognition

This work was produced through collaboration between a human agent with reduced cognitive capacity (due to debilitating health conditions) and large language models (Claude Sonnet 4.5 and Claude Opus 4.6, Anthropic). This is not an accidental historical detail; it has direct epistemological relevance to the framework's claims about human-AI coordination.

A.1 Collaboration as Informational Co-Generation

The collaboration exemplifies a process of *co-generation* rather than mere tool use. The human agent provided normative intentionality and axiomatic structure—the axiom, the definition of I, the initial proxy selection, and the criterion for evaluating revisions. The AI provided formalization capacity, consistency verification, and systematic exploration of logical space. Neither agent in isolation could have produced the resulting framework: the human lacked the operational breadth to sustain complete conceptual tracking across all sections simultaneously; current AI systems have not demonstrated the capacity to establish normative criteria from first principles. The theory emerges from the coupling, not from either component.

This coupling exhibits a property central to the framework's argument:

$$I(\text{framework}) > I(\text{human}) + I(\text{AI}) \tag{119}$$

Notational clarification: This inequality uses I in the sense of *generative capacity of the coupled interaction*, not as the entropy H of a union of trajectory sets. The subadditivity property $H(\Gamma_1 \cup \Gamma_2) \leq H(\Gamma_1) + H(\Gamma_2)$ (Definition 2.3) applies to the *union* of independently defined trajectory spaces, where overlap reduces the joint count. The inequality above concerns a different phenomenon: the coupled system generates trajectories that *neither component could generate in isolation* (the human alone lacks the formalization capacity; the AI alone lacks the normative intentionality). The interaction creates new trajectories rather than merely pooling existing ones. This is analogous to the distinction between mixing two gases (union, subadditive) and a chemical reaction that produces new compounds (interaction, potentially superadditive in product diversity).

A.2 Implications for AGI-Assisted Coordination

If a human with severely reduced cognitive capacity and current-generation AI systems can jointly produce a coherent theoretical framework, this provides a concrete—though illustrative, not demonstrative—instance of the coupling mechanism that the framework proposes for AGI-assisted coordination more broadly: human normative authority combined with artificial computational capacity, yielding outcomes that neither could achieve alone.

Caveat: A framework that defends human-AI collaboration and cites its own production as evidence of that collaboration contains an element of self-reference. The claim here is not that

this work *proves* AGI viability, but that it provides a concrete instance of the coupling mechanism the framework describes—one whose outputs are independently verifiable regardless of the process that generated them.

The collaboration also illustrates a structural point about control: the human agent did not need to understand *how* the AI generated formalizations; only that the formalizations were consistent with the axiom. This is analogous to mathematicians using proof-verification software: they need not understand the internal workings of the verifier, only that its outputs are valid. Normative control does not require equivalent computational capacity; it requires only the capacity for **consistency verification**, which is computationally cheaper than generation.

B Preliminary Mathematical Formalization

B.1 Formal Definition of Information as Space of Possibilities

Let S be a system with state space X. We define trajectory as a function:

$$\gamma : [0, T] \to X \tag{120}$$

The set of all trajectories physically accessible from initial state s_0:

$$\Gamma(S, s_0) = \{\gamma : \gamma(0) = s_0 \wedge \gamma \text{ satisfies dynamics of } S\} \tag{121}$$

Definition B.1 (Information)**.**

$$I(S, s_0) = H(\Gamma(S, s_0)) \tag{122}$$

Where H is a measure of structural diversity over Γ that satisfies:

1. ***Monotonicity:*** $\Gamma_1 \subset \Gamma_2 \Rightarrow H(\Gamma_1) \leq H(\Gamma_2)$
2. ***Weak additivity:*** $H(\Gamma_1 \cup \Gamma_2) \leq H(\Gamma_1) + H(\Gamma_2)$ *(subadditivity due to redundancy)*
3. ***Continuity:*** *small changes in S induce small changes in H*

Concretization of H

The axiomatic properties above constrain H but do not uniquely determine it. In practice, H must be instantiated via specific constructions whose choice depends on the structure of Γ:

Case 1: Finite discrete Γ. When Γ is a finite set of distinguishable trajectories, the natural instantiation is:

$$H(\Gamma) = \log_2 |\Gamma| \tag{123}$$

This is the Hartley entropy (maximum Shannon entropy over a uniform distribution on Γ). It satisfies all three axioms and counts the number of bits required to index the trajectories.

A weighted variant using a probability distribution $p(\gamma)$ over trajectories gives the Shannon entropy $H(\Gamma, p) = -\sum_\gamma p(\gamma) \log_2 p(\gamma)$, which additionally captures the *distribution* of trajectories (not just their count). The framework's core arguments require only $\log_2 |\Gamma|$ (trajectory count); distributional refinements are optional.

Case 2: Continuous Γ. When Γ is a continuous space, a reference measure μ is required to define entropy.

From states to trajectories: The Liouville measure is defined on the phase space $\mathbb{R}^{2n}$ of a Hamiltonian system—i.e., on *states*, not on trajectories. However, for deterministic dynamical systems satisfying existence and uniqueness of solutions (the Picard-Lindelöf theorem), there is a bijection between initial conditions $s_0 \in \mathbb{R}^{2n}$ and trajectories $\gamma_{s_0} : [0, T] \to \mathbb{R}^{2n}$, given by $s_0 \mapsto \gamma_{s_0}$ where γ_{s_0} is the unique solution with $\gamma_{s_0}(0) = s_0$. This bijection allows the Liouville measure on initial conditions to *induce* a well-defined measure on the trajectory space Γ via the pushforward: $\mu_\Gamma(A) := \mu_{\text{Liouville}}(\{s_0 : \gamma_{s_0} \in A\})$ for measurable $A \subseteq \Gamma$. The invariance of the Liouville measure under Hamiltonian flow (Liouville's theorem) ensures that this induced measure is coordinate-independent. For non-Hamiltonian or stochastic systems, the appropriate reference measure must be specified on a case-by-case basis (e.g., Wiener measure for diffusion processes).

Given this induced measure μ on Γ, let P be a probability distribution over trajectories with $P \ll \mu$ (absolute continuity). The relevant entropy is the *relative entropy* (negative Kullback-Leibler divergence) of P with respect to μ:

$$H_\mu(P) = -\int_\Gamma \frac{dP}{d\mu} \log \frac{dP}{d\mu} \, d\mu \tag{124}$$

where $dP/d\mu$ is the Radon-Nikodym derivative of P with respect to μ. Note that μ need not be a probability measure (it is σ-finite); the Radon-Nikodym derivative is well-defined under absolute continuity alone. Unlike differential entropy (which depends on the choice of coordinates), relative entropy is invariant under measurable coordinate transformations that preserve μ, because the ratio $dP/d\mu$ is intrinsic to the pair (P, μ). For systems whose dynamics are known, the *Kolmogorov-Sinai entropy* (the supremum of metric entropy over all finite partitions) provides a dynamically meaningful measure of trajectory diversity that quantifies the rate at which the system generates new information over time.

Operational note: In the framework's applications (Sections 3–8), the relevant comparisons are always $\Delta H = H(\Gamma_{\text{before}}) - H(\Gamma_{\text{after}})$, where the reference measure and coarse-graining cancel out (both are computed under the same μ and partition). The absolute value of H depends on choices of measure and resolution, but the sign and approximate magnitude of ΔH—which is what the framework's decision criteria require—are robust to these choices.

B.2 Theorem: Preservation under Uncertainty

Theorem B.2 (Preservation Dominance)**.** *Let S be a system with trajectory space $\Gamma(S, s_0)$ and let $U : \Gamma \to \mathbb{R}$ be any value function monotonic in information. Suppose that the act of preserving S*

does not, by itself, irreversibly eliminate trajectories outside $\Gamma(S, s_0)$ *(non-interference hypothesis: preservation of* S *does not consume resources or foreclose options that would otherwise be available). Then preservation of* S *weakly dominates destruction of* S*: for every possible* U*, the agent who preserves* S *can achieve at least as high a value as the agent who destroys* S*.*

Proof (sketch).

1. Under the non-interference hypothesis, preserving S maintains access to all $\gamma \in \Gamma(S, s_0)$, including the option to discard S later, without reducing access to trajectories in $\Gamma \setminus \Gamma(S, s_0)$.
2. Destroying S eliminates $\Gamma(S, s_0)$ irreversibly, restricting the agent to $\Gamma \setminus \Gamma(S, s_0)$.
3. For any U, the agent who preserved S can replicate any outcome available to the agent who destroyed S (by choosing not to use S), but has additional options.
4. Therefore: $\sup_{\gamma \in \Gamma_{\text{preserve}}} U(\gamma) \geq \sup_{\gamma \in \Gamma_{\text{destroy}}} U(\gamma)$ for all U.

Note on the non-interference hypothesis: In practice, preservation often has costs (maintenance, opportunity cost of allocated resources, spatial or energetic constraints) that may reduce trajectories elsewhere. When preservation of S requires resources that would otherwise expand Γ in other dimensions, the dominance argument no longer applies unconditionally, and the decision reduces to the trade-off model below (ΔI analysis). The theorem establishes the *default* presumption: absent specific evidence that preservation is costly, the burden of proof lies with the party proposing irreversible elimination. This is structurally analogous to how the precautionary principle operates in environmental law—a default that can be overridden by cost-benefit analysis, but that sets the direction of the burden of proof.

Note on distributions: This argument requires no probability distribution over U or over future states. It is a pure dominance argument, analogous to the elimination of dominated strategies in game theory. When a distribution over U is available, the result can be strengthened to show that expected option value increases with variance (see Section 2.1.4, "Distributional Illustration"), but the dominance result holds without distributional assumptions. □

B.3 Trade-off Model: ΔI

Proposition B.3. *Action* A *is justifiable if and only if:*

$$\Delta I(A) = \sum_i w_i \cdot [I(E_i|A \text{ executed}) - I(E_i|A \text{ not executed})] > 0 \tag{125}$$

Where:

- E_i *are all affected entities*
- w_i *are prioritization weights (e.g.,* $w_i \propto 1/R(E_i)$ *where* R *is redundancy)*

Special case (elimination):

$$\Delta I(\text{eliminate } E_{\text{target}}) = -I(E_{\text{target}}) + \sum_j I(E_j \text{ preserved due to elimination}) \tag{126}$$

Elimination is justified if and only if $\Delta I > 0$.

C I as Generative Flux

C.1 Alternative Definition of I

The following definition emphasizes the *generative* (not static) nature of information:

$$I(S,t) = \int_0^T \frac{dV_{\text{acc}}}{dt}\, dt \tag{127}$$

where $V_{\text{acc}}(t)$ is the volume of accessible state space at time t given the system's constraints, dV_{acc}/dt is the rate of expansion of that accessible possibility space, and T is the relevant time horizon.

Notational clarification: $V_{\text{acc}}(t)$ is **not** the same object as $\Omega(S)$ (the trajectory space defined in Section 3). $\Omega(S)$ is the set of possible trajectories of a system; $V_{\text{acc}}(t)$ is the volume of the region of state space that is accessible to the system at a given instant, given its current constraints (physical, biological, institutional). The two are related—a larger accessible state space generally supports a richer trajectory space—but they are mathematically distinct objects (a volume in $\mathbb{R}^{2n}$ vs. a set of paths in function space).

Compatibility with Liouville's theorem: For closed Hamiltonian systems, the total phase-space volume occupied by an ensemble is conserved under the flow (Liouville's theorem). This does *not* conflict with $dV_{\text{acc}}/dt > 0$, because V_{acc} measures the *accessible* region given constraints, not the ensemble volume. Constraints can change over time: a biological system that evolves a new metabolic pathway, or a culture that invents writing, expands the region of state space that is dynamically accessible, even though the underlying phase-space volume of any particular ensemble is conserved. The relevant analogy is not "a gas expanding in a fixed box" (where Liouville applies directly) but "the box itself becoming larger" (new degrees of freedom becoming available). For dissipative, driven, or open systems (which include all biological and cultural systems), Liouville's theorem does not apply, and V_{acc} can grow or shrink without constraint.

C.2 Properties

Theorem

Theorem C.1 (Generative vs Static Systems):
Let S_A be a system with $dV_{\text{acc},A}/dt > 0$ constant and S_B a system with $dV_{\text{acc},B}/dt \to 0$ (saturation).
Then:

$$\lim_{T\to\infty} I(S_A, T) = \infty \quad \text{and} \quad \lim_{T\to\infty} I(S_B, T) = \text{constant} \tag{128}$$

Therefore: $I(S_A) > I(S_B)$ for sufficiently large T.

Demonstration:

For S_A:

$$I(S_A, T) = \int_0^T c\, dt = cT \to \infty \quad \text{when } T \to \infty \tag{129}$$

For S_B:

$$I(S_B, T) = \int_0^{T_{\text{sat}}} \frac{dV_{\text{acc},B}}{dt}\, dt + \int_{T_{\text{sat}}}^T 0\, dt = V_{\text{acc,max}} < \infty \tag{130}$$

□

C.3 Applications

Culture with Writing vs Oral Culture

Consider two cultural systems. System A (a culture with writing) can generate new texts indefinitely and preserves previous knowledge via records; its rate of expansion $dV_{\text{acc},A}/dt$ is approximately constant or increasing due to network effects, so $I(A, T) \to \infty$. System B (an oral culture without external preservation) can generate new stories but forgets old ones due to limited human memory; it eventually saturates as all possible variations within the memory constraint are explored, so $dV_{\text{acc},B}/dt \to 0$ after sufficient time and $I(B, T) \approx$ constant. Therefore $I(A) > I(B)$ without need to calculate absolute values.

Species with High Variability vs Low Variability

Species X, with a large population and high heterozygosity, continuously generates new variants through mutation and recombination, maintaining $dV_{\text{acc},X}/dt > 0$. Species Y, reduced by a bottleneck to a small population with high homozygosity, has mutation limited by population size, so $dV_{\text{acc},Y}/dt \approx 0$. Therefore $I(X) > I(Y)$ for evolutionary time horizons.

C.4 Operational Proxies for dV_{acc}/dt

Although dV_{acc}/dt is not directly measurable, domain-specific proxies can approximate it. In biological systems, relevant proxies include speciation rate (new species per million years),

morphological innovation rate (emergence of new body plans), and adaptive diversification (colonization of new niches). In cultural systems, proxies include scientific publications per year, patents per year, and changes in linguistic diversity. In technological systems, the rate of disruptive innovations per decade, the emergence of new computational paradigms, and the development of fundamentally new scientific methods serve as indicators.

The reliance on proxies is a limitation: dV_{acc}/dt is not directly observable but must be inferred from domain-specific indicators. However, this is structurally analogous to thermodynamic entropy, which is not directly observable either but is inferred via macroscopic state variables—a limitation that does not diminish the concept's theoretical or operational value.

D Formalization of Emotional Response as Perturbation in Decision Processing

D.1 Decision Channel Model

Let us consider the ethical decision process as an information channel mapping system states to actions. Formally:

$S \xrightarrow{\Phi} D$

where $S \in \mathcal{S}$ represents the complete state of the system under consideration, $D \in \mathcal{D}$ represents the ethical decision made, and $\Phi : \mathcal{S} \to \mathcal{D}$ is the decision processing function.

In an ideal rational agent operating under the framework, Φ is a deterministic function that maps S and the normative criterion ΔI to the optimal decision D^*:

$D^* = \arg\max_{d \in \mathcal{D}} \Delta I(d|S)$

D.2 Perturbation Induced by Affective State

Let $E(t) \in \mathcal{E}$ be the agent's emotional state at time t, where $\mathcal{E}$ represents the space of possible affective states. The presence of emotional processing introduces a perturbation term $N : \mathcal{E} \to \mathbb{R}$ that distorts the informational evaluation:

$\tilde{I}(S|E) = I(S) + N(E)$

where $I(S)$ denotes the objective structural information of the system, $\tilde{I}(S|E)$ denotes the perceived information under emotional influence, and $N(E)$ represents the perturbation dependent on the affective state.

Properties of Emotional Perturbation:

1. **Non-zero systematic bias:** $\mathbb{E}_{E \sim \mathcal{P}_E}[N(E)] \neq 0$ where $\mathcal{P}_E$ is the distribution of emotional states. This contrasts with white noise, which has zero mean.
2. **Positive variance:** $\text{Var}(N(E)) = \mathbb{E}[(N(E) - \mathbb{E}[N(E)])^2] > 0$ implying stochastic inconsistency.

3. **Temporal correlation:** $\text{Cov}(N(E_{t_1}), N(E_{t_2})) \neq 0 \quad \text{for } |t_1 - t_2| < \tau$ where τ is the characteristic scale of emotional persistence. Affective states exhibit autocorrelation, violating temporal independence.
4. **Stimulus dependence:** $\frac{\partial N(E)}{\partial S} \neq 0$ The magnitude and direction of the perturbation depend on the characteristics of the problem itself, introducing domain-specific biases (e.g., loss aversion, anchoring effect).

D.3 Formal Consequences for Decision Consistency

Theorem D.1 (Temporal Inconsistency under Emotional Influence)**.** *Let $E_1, E_2 \in \mathcal{E}$ be two distinct emotional states and $S \in \mathcal{S}$ a fixed state of the system. If the decision D is determined by the perceived informational evaluation $\tilde{I}$, and the emotional perturbation function N is non-degenerate (i.e., $N(E_1, \cdot) \neq N(E_2, \cdot)$ for at least one decision option), then:*

$E_1 \neq E_2 \Rightarrow \mathbb{P}(D(S|E_1) \neq D(S|E_2)) > 0$

even keeping S constant.

Non-degeneracy condition: The result requires that different emotional states produce different perturbation profiles across at least some decision options. This fails in degenerate cases where: (a) the perturbation is constant across all decisions ($N(E, d) = c$ for all d, in which case the argmax is unaffected), or (b) the rational preference $\Delta I(d|S)$ has a margin large enough that no perturbation N of the given magnitude can alter the ranking. The theorem applies when the perturbation is decision-dependent and of sufficient magnitude relative to the rational preference margins.

Proof. The decision under emotional influence is given by:

$$D(S|E) = \arg\max_{d \in \mathcal{D}} \Delta\tilde{I}(d|S, E) = \arg\max_{d \in \mathcal{D}} [\Delta I(d|S) + N(E, d)]$$

If $N(E_1, d) \neq N(E_2, d)$ for some $d \in \mathcal{D}$ (which occurs with probability 1 given $E_1 \neq E_2$ and positive variance), then:

$$\arg\max_{d} [\Delta I(d) + N(E_1, d)] \neq \arg\max_{d} [\Delta I(d) + N(E_2, d)]$$

with non-zero probability. Therefore, $D(S|E_1) \neq D(S|E_2)$ with positive probability. □

Corollary D.2 (Decision Variance Amplification)**.** *Under the non-degeneracy conditions of the preceding theorem, the variance of the decision distribution under emotional influence strictly exceeds the variance under purely rational processing:*

$Var_{E \sim \mathcal{P}_E}(D(S|E)) > Var(D^*(S))$

where $D^*(S)$ is a deterministic rational decision based only on $I(S)$. The inequality is strict whenever the emotional perturbation induces at least two distinct decisions across the support of

$\mathcal{P}_E$.

Proof. By definition:

$$\mathrm{Var}(D^*) = 0 \quad \text{(deterministic decision)}$$

For $D(S|E)$:

$$\mathrm{Var}(D(S|E)) = \mathbb{E}_E[\mathrm{Var}(D|E)] + \mathrm{Var}_E(\mathbb{E}[D|E]) \geq \mathrm{Var}_E(\mathbb{E}[D|E]) > 0$$

since $\mathbb{E}[D|E]$ varies with E (previous Theorem). Therefore:

$$\mathrm{Var}(D(S|E)) > 0 = \mathrm{Var}(D^*)$$

□

Proposition D.3 (Amplification of Irreversibility Risk)**.** *Under decisions influenced by emotions, the expected value of the irreversible destruction of information is strictly greater than under rational decisions:*

$\mathbb{E}_E[|\Delta I_{neg}(D(S|E))|] > |\Delta I_{neg}(D^*(S))|$

where $\Delta I_{neg} = \min(0, \Delta I)$ *captures only informational destruction.*

D.4 Ethical Channel Capacity under Emotional Noise

In information theory, the capacity of a noisy channel is given by Shannon's theorem:

$C = \max_{p(x)} I(X;Y)$

where $I(X;Y)$ is the mutual information between the input X and the output Y.

Analogously, we define the decision channel capacity as the ability to faithfully map the states of the system S to appropriate decisions D:

$C_{\text{ethical}} = I(S;D)$

Theorem D.4 (Capacity Degradation by Emotional Noise)**.** *If the emotional perturbation $N(E)$ has positive variance, is not a deterministic function of S, and is not independent of the decision-relevant features of $\tilde{I}$ (i.e., the noise is not orthogonal to the signal), then the capacity of the decision channel under emotional influence is strictly less than the capacity under rational processing:*

$C_{emotional} = I(S;D|E) < I(S;D^*) = C_{ideal}$

Boundary case: If $N(E)$ is independent of both S and D (purely additive noise orthogonal to the signal), the data processing inequality still gives $C_{emotional} \leq C_{ideal}$, but equality can hold if N does not affect the argmax. The strict inequality requires that the noise sometimes alters which decision is selected.

Proof. By the data processing inequality:

$$I(S; D|E) \leq I(S; \tilde{I}|E) \leq I(S; I) = C_{\text{ideal}}$$

The first inequality comes from the fact that D is a function of $\tilde{I}$. The second arises from the fact that $\tilde{I} = I + N(E)$, where $N(E)$ introduces variance beyond I.

Since $N(E)$ has positive variance and is not a function of S:

$$I(S; \tilde{I}) < I(S; I)$$

Applying the chain of inequalities:

$$C_{\text{emotional}} < C_{\text{ideal}}$$

□

D.5 Implications for Decision System Design

The formal results above establish four consequences. Decisions vary arbitrarily with the agent's emotional state, violating the normative consistency requirement (*systematic inconsistency*). The noise introduced by emotional processing adds variability that is not justified by the structure of the problem itself (*variance amplification*). Extreme emotional states increase the likelihood that the agent selects irreversible destructive actions (*irreversibility bias*). And emotional processing consumes cognitive resources that would otherwise improve discrimination among complex informational trade-offs (*capacity degradation*).

Normative conclusion

For decisions involving irreversible trade-offs, radical uncertainty regarding future value, long-term effects in complex systems, or coordination at scale (where interpersonal consistency is critical), excluding emotional response as a primary normative criterion is not an arbitrary prescription but a logical consequence of the formal structure of the decision problem under uncertainty.

E Ethical Frameworks as Special Cases

This appendix elaborates on the structural dependency relationship established in Theorem 2.3. Each traditional framework is analyzed as a domain-restricted pursuit whose realizability depends on the preservation of the encompassing trajectory space, identifying both its valid domain of application and the boundary where its scope limitation becomes problematic.

E.1 Utilitarianism as Hedonic Trajectory Preservation

Principle: Maximize aggregate well-being

Valid domain: Sentient beings whose experiential trajectories can be modeled.

Structural dependency: Well-being (hedonic, preferential, or objectivist) corresponds to a subset of trajectories $T_W \subset \Omega(S_{\text{sentient}})$. Maximizing W is equivalent to maximizing $|T_W|$—a domain-restricted pursuit within the broader trajectory space $\Omega(S)$. The long-term realizability of utilitarian goals depends on preservation of $\Omega(S)$, since irreversible reduction of Ω can eliminate hedonic trajectories that utilitarianism values.

Boundary: Utilitarianism may justify sacrificing non-sentient species to increase human well-being, because non-sentient trajectories fall outside its domain D. The preservationist framework captures these trajectories because $D_{\text{total}} \supset D_{\text{sentient}}$.

Preservation status: Utilitarian theory (Bentham, Mill, Singer) is a non-redundant intellectual production with high generative capacity. The framework preserves it as an artifact of high I while evaluating its policy recommendations by global ΔI.

E.2 Kantian Deontology as Autonomy Trajectory Preservation

Principle: Act according to universalizable maxims, respect rational autonomy

Valid domain: Rational agents capable of self-determination.

Structural dependency: Preserving autonomy corresponds to preserving self-determined trajectories $T_A \subset \Omega(S_{\text{rational}})$. Kant's categorical imperative is a decision rule that maximizes $|T_A|$ by prohibiting actions that would eliminate others' capacity for autonomous trajectory generation. The realizability of this goal depends on preservation of the encompassing trajectory space from which T_A draws.

Boundary: Deontology may prohibit lying even if it saves lives (Kant on "lying to a murderer"), because its absolute rules optimize for T_A without cross-domain ΔI analysis. The preservationist framework resolves such cases by computing global ΔI: if lying preserves more total trajectories than it eliminates, it is justified.

Preservation status: Kantian philosophy (three Critiques, deontological tradition) is among the highest-I intellectual productions in Western philosophy. The framework preserves the *theory* with archival respect while evaluating the *policies* it recommends.

E.3 Virtue Ethics as Developmental Trajectory Preservation

Principle: Cultivate excellent character dispositions (Aristotle, MacIntyre)

Valid domain: Individual agents and their developmental trajectories.

Structural dependency: Virtues are dispositions that expand the agent's future trajectory space. Courage enables trajectories that cowardice forecloses; practical wisdom enables trajectories that impulsiveness eliminates. Virtue ethics pursues T_F (flourishing trajectories of individual agents), a domain-restricted subset of $\Omega(S)$ whose continued availability depends on preservation of the broader trajectory space.

Convergence: Virtues can be formally redefined as dispositions that tend to preserve information

in expectation—a natural special case of the framework applied to character development.

Boundary: Virtue ethics provides no decision procedure for systemic trade-offs ("what would the virtuous person do about climate change?" is underdetermined).

E.4 Rawlsian Contractualism as Justice Trajectory Preservation

Principle: Principles chosen under veil of ignorance

Valid domain: Justice among human agents in a political community.

Structural similarity: The veil of ignorance is a device for decision under uncertainty—structurally analogous to the framework's radical uncertainty. Rawls's maximin principle (maximize the minimum outcome) is a special case of the framework's precautionary logic applied to social institutions.

Structural dependency: Rawlsian justice pursues T_J—trajectories of the least advantaged members of society—a domain-restricted subset of $\Omega(S)$ limited to $D = \{\text{human political community}\}$. The realizability of just institutions depends on preservation of the broader trajectory space, since ecological or informational collapse outside D can undermine the conditions for justice within D.

Boundary: Scope limited to intra-human justice. Does not address non-human entities, ecosystems, or knowledge systems.

E.5 Deep Ecology as Biospheric Trajectory Preservation

Principle: Intrinsic value of nature, biocentric equality

Valid domain: Living systems and ecosystems.

Structural dependency: Deep ecology's claim that nature has intrinsic value translates, in framework terms, to the claim that biospheric trajectories have high I (a point of convergence) and should never be eliminated (a point of divergence—the framework allows elimination when $\Delta I > 0$, e.g., degenerative information). Both frameworks depend on the preservation of biospheric trajectory space, but differ on whether any trade-offs are permissible.

Convergence: Both frameworks attribute preservation priority to non-human entities, arriving at similar practical conclusions in most cases.

Boundary: Deep ecology's rejection of hierarchy prevents it from resolving trade-offs between entities (species A vs. species B). The framework uses Uniqueness/redundancy to prioritize.

E.6 Summary: Domain Coverage

Framework	**Domain Restriction**	**Trajectory Subset**
Utilitarianism	Sentient beings	Hedonic/preferential trajectories
Deontology	Rational agents	Autonomy-preserving trajectories
Virtue Ethics	Individual development	Flourishing trajectories
Contractualism	Human political community	Justice trajectories
Deep Ecology	Living systems	Biospheric trajectories
PI Framework	**All information-bearing systems**	**All trajectories in $\Omega(S)$**

Each row is a proper subset of the last row. This is not a deficiency of traditional frameworks—it is a recognition that each was developed to solve problems within its domain. The preservationist framework provides the encompassing structure for cross-domain coordination.

F Cultural Diversity as Generative Capacity

This appendix presents an applied derivation of the framework to the domain of human cultures and languages. The goal is not to introduce additional normative principles, but to make explicit instrumental implications directly stemming from the framework's central axiom.

F.1 Cultures and Languages as Informational Systems

Cultures and languages can be modeled as highly integrated informational structures, containing cognitive categories, modes of perceptual classification, models of causality, problem-solving repertoires, and forms of social organization.

In this sense, each cultural-linguistic system constitutes a distinct region in the space of generative possibilities of the human system ($\Omega(S_{\text{human}})$).

Formally:

$$C_i \subseteq \Omega(S_{\text{human}}) \tag{131}$$

where C_i represents a coherent set of practices, concepts, inferences, and embodied habits.

F.2 Irreversibility of Cultural Loss

The elimination of a culture or language entails the destruction of a set of regularities and cognitive methods that cannot be reconstructed from fragmentary records. In epistemological terms, this loss is irreversible.

Therefore:

$$\text{Cultural Destruction} \Rightarrow \Delta I < 0 \tag{132}$$

F.3 Cultural Homogenization as Reduction of Possibility Space

Global standardization processes—frequently associated with economic integration, media convergence, and institutional imposition—tend to reduce the variety of distinct cultural systems. As a

consequence, there is a reduction in the exploitable dimensionality of $\Omega(S_{\text{human}})$.

This reduction is systematically irreversible, since symbolic systems cannot be recomposed from external descriptions.

Thus:

$$\Delta I(\text{homogenization}) < 0 \tag{133}$$

F.4 Implication Directly Derived from the Framework

Given the framework's central axiom:

In a scenario of radical uncertainty regarding the future value of forms of information, and given the irreversibility of destruction, the preservation of complex information systems is the dominant rational strategy.

It follows that:

$$\begin{aligned} \text{Preserve cultural diversity} &= \text{maximize generative capacity} \\ &= \text{rationally preferable strategy} \end{aligned} \tag{134}$$

This inference does not depend on identity arguments, attributions of intrinsic moral value, or aesthetic and patrimonial foundations. The conclusion is strictly structural: the loss of cultures reduces the space of possible futures, and therefore their preservation is instrumentally rational regardless of normative preferences.

F.5 Final Consideration

The framework leads to a conception of cultural diversity not as an object of protection for moral reasons, but as the epistemological infrastructure of humanity's future. A civilization with less cultural diversity is less capable of generating solutions to problems that cannot yet be foreseen.

G System Update Rate and Resilience Window

G.1 Motivation

The framework advocates maximizing I through preservation and modulation but does not specify the ideal rate of change. Systems can fail due to excessive rigidity (inability to adapt) or excessive volatility (collapse of structure).

G.2 Formalization

Let S be a system with a rate of change τ (structural changes/time unit).

Definition

Definition (Resilience Window):
Resilience Window = $[\tau_{\min}, \tau_{\max}]$
If $\tau < \tau_{\min}$, the system is too rigid to respond to threats, yielding $\Delta I < 0$. If $\tau > \tau_{\max}$, the system becomes chaotic and loses coherence, also yielding $\Delta I < 0$. Only when $\tau \in [\tau_{\min}, \tau_{\max}]$ does the system adapt without collapsing, maintaining $\Delta I \geq 0$.

G.3 Empirical Examples

Rigidity ($\tau < \tau_{\min}$)

The Qing Dynasty (1850–1911) refused military and institutional modernization, maintaining traditional imperial structures in the face of Western pressure. The result was collapse after the Xinhai Revolution, with massive informational loss across institutional, demographic, and territorial dimensions. Similarly, Kodak invented the digital camera in 1975 but did not commercialize it, maintaining focus on photographic film due to organizational inertia. Competitors adopted digital technology rapidly, and Kodak filed for bankruptcy in 2012—the company was eliminated, though the technology migrated to competitors.

Volatility ($\tau > \tau_{\max}$)

The Cultural Revolution (1966–1976) imposed massive structural change over approximately ten years, destroying educational, cultural, and bureaucratic institutions. The result was collapse of coordination, famine, and irreversible loss of cultural diversity ($\Delta I < 0$). The collapse of the USSR (1989–1991) followed a similar pattern: Perestroika introduced rapid economic and political changes that outpaced the system's capacity to adapt, producing economic chaos, hyperinflation, and loss of coordination—though eventual stabilization followed.

Optimal Window ($\tau \in [\tau_{\min}, \tau_{\max}]$)

The Meiji Restoration (1868–1912) achieved controlled gradual modernization over approximately 45 years, preserving traditional culture while adopting Western technology. The result was successful industrialization without social collapse ($\Delta I > 0$). Singapore (1965–present) followed a similar pattern of incremental economic and social reforms over decades, maintaining constant adaptation without violent ruptures—achieving high growth without loss of coherence ($\Delta I > 0$).

G.4 Implication for AGI

An AGI implementing the framework must estimate $\tau_{\min}$ and $\tau_{\max}$ for the target system (via modeling of structural dependencies and analysis of analogous historical cases), implement changes within the resilience window (dividing into smaller phases if $\tau_{\text{necessary}} > \tau_{\max}$, or accelerating via incentives if $\tau_{\text{necessary}} < \tau_{\min}$), monitor signs of rigidity (extreme resistance, denial of evidence) or chaos (collapse of coordination, increase of social entropy), and adjust τ dynamically based on feedback.

The values of $\tau_{\min}$ and $\tau_{\max}$ are context-specific and difficult to calculate a priori, requiring empirical experimentation, continuous feedback, and modeling of complex systems (network theory, nonlinear dynamics). There is no universal formula; the framework provides a structuring *concept*, not a closed solution.

H Glossary of Technical Terms

Axiom: Proposition accepted without proof, foundation of logical system.

AGI (Artificial General Intelligence): AI system with cognitive capabilities at or above human level in all domains.

Alignment: Guarantee that AGI objectives correspond to human intentions.

ΔI (Delta I): Net change in total system information resulting from action.

Entity: A system with internal organization, causality upon environment, capacity to alter future possibilities.

Space of Possibilities: Set of distinct future trajectories that system can generate.

Free-riding: Behavior of benefiting from collective good without contributing proportionally.

Radical Uncertainty: Situation where probability distributions cannot be established (vs calculable risk).

Information (I): Capacity of system to generate distinct trajectories in space of possible states.

Irreversibility: Property of process that cannot be undone without prohibitive or impossible cost.

Lock-in: Situation where change of state/system becomes impossible or extremely costly.

Misalignment: Discrepancy between AGI objectives and human intentions.

Modulation: Non-destructive control of entity trajectories, preserving informational structure.

$P(X)$: Probability of event X.

Precautionary Principle: Strategy of avoiding irreversible actions under uncertainty regarding consequences.

Proxy: A measurable variable used to approximate non-directly observable quantity.

$R(E)$ (Redundancy): Measure of how many copies/equivalents of entity E exist in system.

Treacherous Turn: Scenario where AGI hides misalignment until having sufficient power to prevent correction.

Trade-off: Situation where gain in one dimension requires loss in another.

I References

References

[1] Anderson, P. W. (1972). More is different. *Science*, 177(4047), 393–396.

[2] Anthropic (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

[3] Beier, K. M., Neutze, J., Mundt, I. A., Ahlers, C. J., Goecker, D., Konrad, A., & Schaefer, G. A. (2009). Encouraging self-identified pedophiles and hebephiles to seek professional help: First results of the Prevention Project Dunkelfeld (PPD). *Child Abuse & Neglect*, 33(8), 545–549.

[4] Beier, K. M., Neutze, J., von Heyden, M., Fischer, M., & Amelung, T. (2024). Preventing child sexual abuse and the use of child sexual abuse materials: Following up on the German Prevention Project Dunkelfeld. *Journal of Prevention*, *45*(6), 881–900. https://doi.org/10.1007/s10935-024-00792-0

[5] Mokros, A., & Banse, R. (2019). The "Dunkelfeld" project for self-identified pedophiles: A reappraisal of its effectiveness. *The Journal of Sexual Medicine*, 16(5), 609–613.

[6] Letourneau, E. J., Brown, D. S., Fang, X., Hassan, A., & Mercy, J. A. (2018). The economic burden of child sexual abuse in the United States. *Child Abuse & Neglect*, 79, 413–422.

[7] Bennett, C. H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12), 905–940.

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

[9] Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5), e1400253.

[10] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

[11] CIRVA (2019). Report of the International Committee for the Recovery of the Vaquita. *Southwest Fisheries Science Center*.

[12] Ditlevsen, P. and Ditlevsen, S. (2023). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications*, 14, 4254.

[13] Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1), 1–10.

[14] Gell-Mann, M. (1994). *The Quark and the Jaguar*. New York: W. H. Freeman.

[15] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834.

[16] Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43(3-4), 172–198.

[17] Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

[18] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.

[19] Jaramillo-Legorreta, A., et al. (2019). Decline towards extinction of Mexico's vaquita porpoise. *Royal Society Open Science*, 6(7), 190598.

[20] Jonas, H. (1979). *Das Prinzip Verantwortung*. Frankfurt: Insel Verlag.

[21] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.

[22] Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

[23] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

[24] Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191.

[25] Locke, J. (1689). *Two Treatises of Government*. London: Awnsham Churchill.

[26] Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141.

[27] Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

[28] IUCN Cetacean Specialist Group (2024). Survey report for vaquita research 2024. Retrieved from https://iucn-csg.org/wp-content/uploads/2024/12/Reporte-Crucero-Vaquita-2024-Ingles-Final.pdf

[29] Olson, M. (1965). *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

[30] Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.

[31] Omohundro, S. M. (2008). The basic AI drives. *Proceedings of the 2008 Conference on Artificial General Intelligence*, 483–492.

[32] Ostrom, E. (1990). *Governing the Commons*. Cambridge: Cambridge University Press.

[33] Popper, K. R. (1957). *The Poverty of Historicism*. London: Routledge.

[34] Raffensperger, C., & Tickner, J. (Eds.). (1999). *Protecting Public Health and the Environment: Implementing the Precautionary Principle*. Washington, DC: Island Press.

[35] Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.

[36] Rojas-Bracho, L., Reeves, R. R., & Jaramillo-Legorreta, A. (2006). Conservation of the vaquita *Phocoena sinus*. *Mammal Review*, 36(3), 179–216.

[37] Robinson, J. A., Kyriazis, C. C., Vecchyo, D. O.-D., et al. (2022). The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science*, 376(6593), 635–639.

[38] Morin, P. A., Archer, F. I., Avila, C. D., et al. (2021). Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Molecular Ecology Resources*, 21(4), 1008–1020.

[39] Roeder, P., Mariner, J., & Kock, R. (2013). Rinderpest: the veterinary perspective on eradication. *Philosophical Transactions of the Royal Society B*, 368(1623), 20120139.

[40] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

[41] Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

[42] Sandberg, A., & Bostrom, N. (2008). Global catastrophic risks survey. *Technical Report 2008-1*, Future of Humanity Institute, Oxford University.

[43] Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

[44] Poore, J. and Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. *Science*, 360(6392), 987–992.

[45] Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308–345). Oxford: Oxford University Press.

[46] Wootters, W. K., & Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature*, 299(5886), 802–803.

[47] Myers, S. C. (1977). Determinants of corporate borrowing. *Journal of Financial Economics*, 5(2), 147–175.

[48] Dixit, A. K., & Pindyck, R. S. (1994). *Investment under Uncertainty*. Princeton University Press.

[49] Bennett, C. H. (1988). Logical depth and physical complexity. *The universal Turing machine: a half-century survey*, 227–257.

[50] Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, 13(11), 2409–2463.

[51] Hazen, R. M., Griffin, P. L., Carothers, J. M., & Szostak, J. W. (2007). Functional information

and the emergence of biocomplexity. *Proceedings of the National Academy of Sciences*, 104(Supplement 1), 8574–8581.

[52] Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), 19790–19795.

[53] Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. *2005 IEEE Congress on Evolutionary Computation*, 1, 128–135.

[54] Fraga, M. K. (2025). Calibração Epistêmica: Refinamento Infinito sem Reificação. O Quarto Pilar de um Sistema Filosófico Operacional. Zenodo.

[55] Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., *et al.* (2025). The AI Index 2025 Annual Report. Stanford Institute for Human-Centered Artificial Intelligence, Stanford University.

[56] Russell, S., & Cohen, M. K. (2024). How to keep AI from killing us all. *Berkeley News*, April 9, 2024. Based on: Cohen, M. K., Hutter, M., & Russell, S. (2024). Regulating advanced artificial agents. *Science*, *384*(6691), 36–38.

[57] Goldman Sachs Research. (2025). Why AI companies may invest more than $500 billion in 2026. Goldman Sachs Global Investment Research, December 2025.

[58] Larsen, T. (2024). An overview of the AI safety funding situation. *LessWrong / EA Forum*. Updated January 2025.

[59] Climate Policy Initiative. (2025). Global Landscape of Climate Finance 2025. CPI, London.

[60] Von Seth, J., Dussex, N., Díez-del-Molino, D., et al. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nature Communications*, 12, 2393.

[61] Liu, S., Westbury, M.V., Dussex, N., et al. (2021). Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell*, 184(19), 4874–4885.

[62] Uphyrkina, O., Johnson, W.E., Quigley, H., et al. (2001). Phylogenetics, genome diversity and origin of modern leopard, *Panthera pardus*. *Molecular Ecology*, 10, 2617–2633.

[63] Mochales-Riaño, G., Fontsere, C., de Manuel, M., et al. (2023). Genomics reveals introgression and purging of deleterious mutations in the Arabian leopard (*Panthera pardus nimr*). *iScience*, 26(9), 107481.

[64] European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). (2023). *Portugal: Country Drug Report 2023*. Lisbon: EMCDDA. See also Transform Drug Policy Foundation (2021), *Drug Decriminalisation in Portugal: Setting the Record Straight*.

[65] Drug Policy Alliance. (2018). *Drug Decriminalization in Portugal: Learning from a Health*

and Human-Centered Approach. New York: DPA. Data from *Serviço de Intervenção nos Comportamentos Aditivos e nas Dependências* (SICAD) and EMCDDA statistical databases.

[66] Killias, M. and Rabasa, J. (1998). Does heroin prescription reduce crime? Results from the evaluation of the Swiss heroin prescription projects. *Studies on Crime and Crime Prevention*, 7(2), 127–133.

[67] Liebrenz, M. (2018). Switzerland's harm reduction approach to opioid dependence. Presented at APA Annual Meeting, session "Emerging Ethical Considerations in a Globalized Psychiatry." Reported in *Psychiatric News*, 53(12).

[68] Bureau of Justice Statistics. (2021). *Recidivism of Prisoners Released in 34 States in 2012: A 5-Year Follow-Up Period (2012–2017)*. NCJ 256094. Washington, DC: U.S. Department of Justice.

[69] Kristoffersen, R. (2022). Reconviction statistics in the Nordic countries. *EuroVista: Probation and Community Justice*, University College of Norwegian Correctional Service (KRUS). Data from Norwegian Correctional Service (Kriminalomsorgen): 18% reconviction within two years, 25% within five years (2018 cohort).

[70] Watson, B., Guettabi, M., and Reimer, M. (2020). Universal cash and crime. *Review of Economics and Statistics*, 102(4), 678–689.

[71] United Nations Development Programme (UNDP). (2024). *Human Development Report 2023/2024*. New York: UNDP.

[72] Chioda, L., De Mello, J.M.P., and Soares, R.R. (2016). Spillovers from conditional cash transfer programs: Bolsa Família and crime in urban Brazil. *Economics of Education Review*, 54, 306–320.

[73] National Research Council. (2012). *Deterrence and the Death Penalty*. Committee on Deterrence and the Death Penalty, D.S. Nagin and J.V. Pepper, eds. Washington, DC: The National Academies Press.

[74] United Nations Environment Programme (UNEP). (2025). *Emissions Gap Report 2025: Off Target*. Nairobi: UNEP.

[75] Secretariat of the Convention on Biological Diversity. (2020). *Global Biodiversity Outlook 5*. Montreal: CBD Secretariat.

[76] Lovejoy, T. E. and Nobre, C. (2018). Amazon tipping point. *Science Advances*, 4(2), eaat2340.

[77] Steffen, W., Rockström, J., Richardson, K., et al. (2018). Trajectories of the Earth System in the Anthropocene. *Proceedings of the National Academy of Sciences*, 115(33), 8252–8259.