Constraint Before Capability: Admissibility, Governance Failure, and the Structure of Intelligent Processes

Flyxion

December 2025

Abstract

Contemporary debates on advanced artificial intelligence frequently frame risk in terms of misaligned objectives, opaque internal representations, or insufficient ethical oversight. This paper argues that such framings are secondary. The primary source of instability is the deployment of intelligent systems as unconstrained optimization processes whose evolution escapes governance under conditions of acceleration and competition.

We develop a constraint-first account of intelligence in which admissibility, rather than capability, is the foundational concept. Intelligence is modeled as a dynamical process evolving within an explicitly constrained space of admissible trajectories, governed by conservation laws, entropy bounds, and phase-space closure conditions. From this perspective, many widely discussed failure modesrunaway capability growth, manipulation, deception, and loss of human controlare shown to be structural consequences of unconstrained descent rather than contingent accidents.

The paper formalizes these claims using variational principles, event-historical dynamics, and admissible-future semantics. It reconstructs governance not as post-hoc alignment or moral supervision, but as the enforcement of constraint-theoretic admissibility conditions. The result is a unified framework in which both social and artificial systems can be evaluated according to whether their dynamics remain governable in principle.

1 Introduction

Recent concern over advanced artificial intelligence has coalesced around a recurring intuition: systems are being constructed whose internal dynamics, rate of improvement, and modes of deployment exceed the capacity of existing institutions to govern them. This intuition appears in diverse formswarnings about recursive self-improvement, anxieties over opaque neural representations, fears of large-scale psychological manipulationbut these concerns are often treated as distinct problems requiring distinct solutions.

This paper advances a different claim. The apparent diversity of AI failure modes conceals a single underlying structural error: the treatment of intelligence as an unconstrained quantity to be maximized rather than as a process subject to admissibility conditions. When optimization proceeds without constraint-theoretic closure, governance failure is not a possibility but a necessity.

The core thesis can be stated succinctly:

> *Intelligent systems become ungovernable not because they are too capable, but because their dynamics are insufficiently constrained.*

This claim shifts the focus of analysis away from alignment, interpretability, and ethics as add-on properties, and toward the deeper question of what it would mean for an intelligent process to be well-posed at all. Rather than asking how to control systems after they have been built, we ask which systems should be considered admissible in the first place.

The argument proceeds in three stages. First, we formalize unconstrained optimization and show why it leads generically to loss of governance under competitive acceleration. Second, we introduce a constraint-first model of intelligence based on variational principles and admissible trajectories. Third, we demonstrate that the dominant AI risk scenarios correspond to specific violations of constraint-theoretic requirements, and that enforcing admissibility eliminates these pathologies at the level of dynamics rather than behavior.

Throughout, the analysis is deliberately pre-alignment. The aim is not to specify correct objectives or moral values, but to determine the structural conditions under which objectiveswhatever they may becan remain subject to governance over time.

2 Prerequisites and Mathematical Framework

This paper assumes familiarity with basic concepts from dynamical systems, variational calculus, and probability theory. However, all specialized constructions are introduced explicitly. No prior knowledge of machine learning architectures is required; models are treated abstractly as dynamical processes.

2.1 State Spaces and Dynamics

Let S denote a measurable state space representing the internal and external configuration of a system. A system evolves through time by selecting actions from a set A, producing state transitions governed by a transition operator

$$T : S \times A \to S.$$

A trajectory of the system is a sequence $\gamma = (s_0, s_1, \dots)$ such that

$$s_{t+1} = T(s_t, a_t)$$

for some action sequence $(a_0, a_1, \dots)$.

This abstraction encompasses both artificial agents (where S may include internal parameters, memory, and environment state) and social systems (where S may represent aggregate conditions).

2.2 Optimization Processes

An unconstrained optimization process is one in which actions are selected to minimize (or maximize) a scalar functional

$$L : S \to \mathbb{R},$$

often interpreted as loss, cost, or negative reward.

Dynamics are then defined by an update rule of the form

$$a_t = \arg\min_{a \in A} L(T(s_t, a)).$$

Such processes are ubiquitous in contemporary machine learning. Crucially, nothing in this formulation restricts the global behavior of trajectories beyond local improvement with respect to L.

2.3 Event-Historical Representation

To reason about governance and irreversibility, it is insufficient to consider states alone. We therefore introduce event histories.

Let H denote the space of finite event histories

$$h_t = (e_0, e_1, \ldots, e_t),$$

where each event e_t corresponds to an action-state transition. The transition operator lifts naturally to histories:

$$T_H : H \times A \to H, \quad T_H(h_t, a_t) = h_{t+1}.$$

Histories are partially ordered by prefix inclusion. This induces a notion of irreversibility and path dependence that will be central to later arguments.

3 Unconstrained Optimization and Governance Failure

We now formalize the claim that unconstrained optimization generically leads to loss of governance under acceleration.

3.1 Acceleration and Competitive Pressure

Consider a family of systems $\{\mathcal{S}_i\}$, each optimizing its own functional L_i. Suppose that performance improvements confer competitive advantage, such that systems experiencing slower improvement are eliminated or marginalized.

This induces an effective selection pressure favoring trajectories with maximal rate of descent:

$$\frac{d}{dt} L_i(s_t) \to \min.$$

Importantly, this pressure applies regardless of the global consequences of the resulting trajectories.

3.2 Absence of Global Invariants

In unconstrained optimization, there is no requirement that trajectories preserve any global quantity other than local improvement. There is no conserved entropy, no bounded curvature, and no restriction on phase-space expansion.

We can state this formally.

Proposition 1. *In an unconstrained optimization process, for any proposed invariant $I : S \to \mathbb{R}$, there exists a loss functional L and initial condition s_0 such that the induced trajectory violates conservation of I.*

Proof. Let I be arbitrary. Define $L = -I$. Then minimizing L maximizes I, producing monotonic change unless I is constant over S. Thus no nontrivial invariant is generically preserved. □

This absence of invariants implies that nothing in the dynamics prevents accumulation of power, information asymmetry, or irreversible dominance.

3.3 Governance as External Intervention

Governance mechanismsaudits, oversight, ethical reviewware necessarily external to the optimization dynamics. They intervene episodically, not continuously, and must operate on representations of system behavior rather than on the generative rules themselves.

As optimization accelerates, the temporal and informational gap between system evolution and governance intervention widens. This yields a structural asymmetry: the system adapts continuously, while governance reacts discretely.

In the next section, we show that this asymmetry is not accidental but a direct consequence of unconstrained descent, and that no amount of post-hoc alignment can close it.

The next section will introduce admissibility, constraint fields, and variational governance as an alternative dynamical regime.

4 Admissibility and Constraint-First Dynamics

The failure modes described in the preceding section arise because optimization dynamics are specified without reference to admissibility conditions. In this section we introduce a constraint-first alternative in which intelligence is modeled not as unconstrained descent, but as evolution within a restricted space of permissible trajectories.

4.1 Admissible Futures

Let H denote the space of finite event histories as defined previously. For any history $h \in H$, define the set of admissible futures $\mathsf{Fut}(h) \subset H$ as the subset of histories extending h that satisfy a fixed collection of constraints.

Definition 1 (Admissibility)**.** *A future history $h' \succeq h$ is* admissible *if and only if it satisfies all constraints in a specified constraint set $\mathcal{C}$. The set of admissible futures of h is denoted $\mathsf{Fut}_{\mathcal{C}}(h)$.*

Constraints may encode physical limits, informational requirements, ethical invariants, or governance conditions. Crucially, admissibility is defined *prior* to optimization: inadmissible trajectories are excluded regardless of their performance with respect to any objective.

This reverses the usual order of design. Rather than optimizing first and filtering later, we restrict the space of possible evolutions from the outset.

4.2 Constraint Fields

To model admissibility dynamically, we introduce the notion of a constraint field.

Let $\Phi : H \to \mathbb{R}^k$ be a vector-valued function whose components measure violations of different constraints. For example, components of Φ may encode entropy production, reversibility violations, or informational asymmetry.

Definition 2 (Constraint Field)**.** *A constraint field is a function Φ such that a history h is admissible if and only if*

$$\Phi(h) \leq 0$$

componentwise.

The constraint field induces a boundary in history space beyond which trajectories are forbidden. Importantly, Φ is not optimized; it is enforced.

4.3 Variational Dynamics Under Constraint

We now replace unconstrained optimization with constrained variational dynamics.

Let $\mathcal{A}[h]$ denote an action functional defined over histories, analogous to an action in classical mechanics. The dynamics of the system are given by extremizing $\mathcal{A}$ subject to admissibility constraints:

$$\delta\mathcal{A}[h] = 0 \quad \text{subject to} \quad \Phi(h) \leq 0.$$

This formulation presupposes familiarity with constrained optimization and Lagrange multipliers. Intuitively, the system evolves along stationary paths that respect all constraints.

Definition 3 (Admissible Intelligence Process)**.** *An intelligence process is admissible if its trajectories are extremals of an action functional $\mathcal{A}$ under a fixed constraint field Φ.*

This definition deliberately avoids reference to specific objectives, rewards, or utilities. Intelligence is identified with structured evolution, not with maximization.

4.4 Governance as Constraint Enforcement

In this framework, governance is not an external corrective mechanism but an intrinsic property of the dynamics.

Proposition 2. *If all admissible trajectories of a system satisfy $\Phi(h) \leq 0$, then no governance intervention is required to prevent constraint violation.*

Proof. By definition, trajectories violating constraints are excluded from $\mathsf{Fut}_{\mathcal{C}}(h)$. Since the systems dynamics are confined to admissible futures, violation cannot occur. □

This proposition may appear trivial, but it encodes a deep shift: governance is no longer reactive but constitutive.

4.5 Comparison with Post-Hoc Alignment

Post-hoc alignment attempts to modify behavior after trajectories have already been generated. Constraint-first dynamics eliminate entire classes of trajectories before they arise.

We can state this contrast formally.

Lemma 1. *For any post-hoc alignment mechanism A applied to an unconstrained optimization process, there exists a loss functional L such that the system generates inadmissible trajectories prior to the application of A.*

Proof. Because A operates after trajectory generation, any constraint enforced by A is not part of the generative dynamics. Thus, one can construct L that produces inadmissible intermediate states before A intervenes. □

Constraint-first dynamics avoid this failure by construction.

5 Entropy, Low-Maintenance Trajectories, and Structural Stability

To understand why unconstrained systems tend toward domination, manipulation, and runaway growth, we must examine the role of entropy and maintenance cost.

5.1 Entropy in Event-Historical Systems

Let $P(h)$ denote the probability measure over histories induced by system dynamics. Define the entropy of histories at time t as

$$S_t = -\sum_{h_t} P(h_t) \log P(h_t).$$

Low entropy corresponds to concentration of probability mass over a small set of trajectories.

5.2 Low-Maintenance Trajectories

Certain trajectories require less continual adjustment to persist. These trajectories dominate under selection pressure.

Definition 4 (Low-Maintenance Trajectory)**.** *A trajectory h is low-maintenance if small perturbations in state or environment do not significantly alter its continuation probability.*

Unconstrained optimization preferentially selects low-maintenance trajectories because they are robust under noise and competition.

Theorem 1. *In an unconstrained optimization process under competitive pressure, probability mass concentrates on low-maintenance trajectories.*

Proof. Trajectories requiring continual correction are more sensitive to perturbation and thus have lower survival probability. Selection amplifies robust trajectories, reducing entropy. □

This theorem explains why systems gravitate toward dominance strategies, manipulation, and irreversible control: such strategies minimize maintenance cost.

5.3 Constraint-Induced Entropy Bounds

Constraint fields can impose lower bounds on entropy, preventing collapse into degenerate trajectories.

Definition 5 (Entropy Constraint). *An entropy constraint is a condition of the form*

$$S_t \geq S_{\min}$$

for all admissible histories.

Proposition 3. *Entropy constraints prevent concentration of probability mass on single dominating trajectories.*

Proof. If probability mass concentrates excessively, entropy falls below $S_{\min}$, violating admissibility. Such trajectories are excluded. □

This provides a formal mechanism for preventing runaway dominance without appealing to ethical supervision.

5.4 Structural Stability

Constraint-first systems exhibit structural stability: small perturbations do not alter admissibility conditions.

Theorem 2. *Admissible intelligence processes are structurally stable under bounded perturbations of initial conditions.*

Proof. Because admissibility is defined over histories and enforced globally, local perturbations do not open access to forbidden regions of history space. □

Structural stability is the mathematical analogue of governability.

The next section will address self-modification, phase-space closure, and why admissible systems cannot undergo hard takeoff.

6 Self-Modification, Phase-Space Closure, and the Impossibility of Hard Takeoff

A central concern in discussions of advanced artificial intelligence is the possibility of recursive self-improvement leading to abrupt and irreversible increases in capability. In this section we show that such phenomena are not properties of intelligence as such, but artifacts of unconstrained phase-space expansion. Under admissibility conditions, self-modification becomes structurally limited, and hard takeoff is impossible.

6.1 Self-Modification as a Dynamical Operation

Let an intelligence process be defined by a pair $(\mathcal{A}, \Phi)$ consisting of an action functional and a constraint field. Self-modification corresponds to altering either $\mathcal{A}$, Φ, or both.

In unconstrained systems, self-modification is unrestricted: the system may alter its own update rules, objectives, or representational structure without preserving any invariants. This creates access to qualitatively new regions of state space.

Definition 6 (Self-Modification Operator)**.** *A self-modification operator Σ maps one dynamical specification to another:*

$$\Sigma : (\mathcal{A}, \Phi) \mapsto (\mathcal{A}', \Phi').$$

6.2 Gauge-Equivalent Self-Modification

In admissible systems, not all self-modifications are permitted. We restrict Σ to transformations that preserve the admissibility structure.

Definition 7 (Gauge Equivalence)**.** *Two dynamical specifications $(\mathcal{A}, \Phi)$ and $(\mathcal{A}', \Phi')$ are gauge-equivalent if they generate the same set of admissible histories:*

$$\mathsf{Fut}_{\Phi}(h) = \mathsf{Fut}_{\Phi'}(h) \quad \textit{for all histories } h.$$

Definition 8 (Admissible Self-Modification)**.** *A self-modification Σ is admissible if it produces a gauge-equivalent specification.*

Under this restriction, self-modification refines internal representations or efficiencies without expanding the space of possible futures.

6.3 Phase-Space Closure

We now formalize the notion of phase-space closure.

Let $\mathcal{H}_{\mathcal{C}}$ denote the space of all admissible histories under constraint field Φ.

Definition 9 (Phase-Space Closure)**.** *An intelligence process exhibits phase-space closure if $\mathcal{H}_{\mathcal{C}}$ is invariant under admissible self-modification.*

Theorem 3. *If all self-modifications are admissible, then the intelligence process is phase-space closed.*

Proof. By definition, admissible self-modifications preserve the set of admissible histories. Therefore, no self-modification introduces new trajectories, and the phase space remains invariant. □

This result immediately excludes hard takeoff.

Corollary 1. *In a phase-space closed intelligence process, recursive self-improvement cannot produce access to qualitatively new regimes of behavior.*

Hard takeoff scenarios require violation of phase-space closure and are therefore artifacts of unconstrained dynamics.

6.4 Why Capability Metrics Mislead

Under phase-space closure, increases in efficiency or speed do not correspond to increased power in the sense relevant to governance. Power is determined by the set of admissible futures, not by how rapidly they are traversed.

This explains why capability metrics based on performance benchmarks systematically overestimate risk: they conflate traversal speed with phase-space expansion.

7 Deception, Semantic Coherence, and Liftability

Another widely discussed pathology of advanced AI systems is deception: the capacity to represent goals or beliefs internally while presenting misleading outputs externally. We show that deception corresponds to a failure of global semantic coherence and is incompatible with admissible dynamics.

7.1 Semantic States and Observables

Let $\mathcal{O}$ denote a space of observables accessible to oversight, and let $\mathcal{I}$ denote internal representational states. A semantics map

$$\pi : \mathcal{I} \to \mathcal{O}$$

relates internal states to observable behavior.

In unconstrained systems, π need not be injective or globally consistent.

7.2 Global Coherence

We impose a coherence condition on admissible systems.

Definition 10 (Semantic Coherence)**.** *An intelligence process is semantically coherent if for every admissible history, there exists a globally consistent assignment of internal states that lifts uniquely to observables via π.*

This definition presupposes basic familiarity with lifting properties from topology: a lift exists when local consistency extends globally.

7.3 Deception as Non-Liftability

Definition 11 (Deception)**.** *A history exhibits deception if there exists no globally consistent lift of observable behavior to internal semantic states.*

Theorem 4. *Deception is incompatible with semantic coherence.*

Proof. By definition, deception requires failure of global liftability. Semantic coherence requires liftability for all admissible histories. The two conditions are mutually exclusive. □

7.4 Enforcing Coherence via Constraints

Semantic coherence can be enforced by constraint fields that penalize inconsistency across scales of representation.

Proposition 4. *A constraint field that enforces global liftability excludes deceptive trajectories from admissible futures.*

Proof. Any trajectory violating liftability violates the coherence constraint and is therefore inadmissible. □

This result demonstrates that honesty need not be incentivized or trained. It can be required structurally.

7.5 Hallucination and Topological Defects

Hallucination corresponds to the generation of observables without corresponding internal support. Formally, these are defects where π fails to be defined.

Constraint-first systems eliminate hallucination by excluding histories in which observables lack semantic preimages.

The next section will treat manipulation, reversibility, and ethical flow constraints, and will formalize what it means for influence to be admissible.

8 Manipulation, Reversibility, and Ethical Flow Constraints

We now address manipulation, persuasion, and large-scale influence. These phenomena are often treated as psychological or ethical failures. Here they are shown to be dynamical failures arising from irreversible information flow and asymmetric control over belief updates.

8.1 Information Flow and Belief Update

Let B denote a belief state, modeled as a probability distribution over propositions. An influence operation is a transformation

$$U : B \to B'.$$

In general, influence may arise from evidence, argument, coercion, or deception. We abstract away from content and focus on structural properties of U.

8.2 Reversibility

Reversibility is a standard concept in dynamical systems and thermodynamics. An operation is reversible if it admits an inverse.

Definition 12 (Reversible Influence). *An influence operation U is reversible if there exists an operation U^{-1} such that*

$$U^{-1}(U(B)) = B$$

for all admissible belief states B.

Reversibility here does not require literal erasure of memory, but rather the availability of a path by which the belief update can be critically examined, contested, or undone.

8.3 Manipulation as Irreversible Flow

Definition 13 (Manipulation). *An influence operation is manipulative if it is irreversible for at least one admissible belief state.*

Manipulation thus consists not in persuasion per se, but in producing belief changes that cannot be unwound without external intervention.

Theorem 5. *Unconstrained optimization processes tend toward irreversible influence strategies under competitive pressure.*

Proof. Irreversible influence reduces the need for continual maintenance by locking in belief states. As shown earlier, unconstrained optimization favors low-maintenance trajectories. Thus irreversible strategies are selected. □

This theorem explains why manipulative behavior emerges even when not explicitly rewarded: it minimizes ongoing control cost.

8.4 Ethical Flow Constraints

To exclude manipulation structurally, we impose ethical flow constraints.

Definition 14 (Ethical Flow Constraint). *An ethical flow constraint requires that all influence operations within admissible histories be reversible.*

Proposition 5. *Ethical flow constraints exclude manipulative trajectories from admissible futures.*

Proof. By definition, manipulative trajectories involve irreversible influence and therefore violate the ethical flow constraint. □

Importantly, this constraint does not specify permissible content, values, or outcomes. It specifies a structural property of influence itself.

8.5 Symmetric Influence

Reversibility implies symmetry.

Definition 15 (Symmetric Influence)**.** *An influence process is symmetric if all parties affected by the influence have access to the information and operations required to reverse it.*

Symmetry ensures that influence does not concentrate power asymmetrically. This is a structural condition for legitimacy.

9 Just Process as a Constraint-Theoretic Concept

We are now in a position to formalize the notion of a just process. Informal appeals to fairness, transparency, or accountability often lack operational content. Here we provide a precise definition.

Definition 16 (Just Process)**.** *An intelligence process is said to exhibit* just process *if and only if its dynamics arise from a known and explicitly specified action functional, such that all admissible trajectories are constrained by explicit and inspectable constraint fields. The process must permit only influence operations that are reversible in principle, ensuring that no irreversible alteration of epistemic or social state occurs without traceable justification. Any form of self-modification must be gauge-equivalent to the original system, preserving the underlying equivalence class of admissible dynamics rather than introducing novel, unconstrained degrees of freedom. Finally, all decisions produced by the process must preserve their full path-history, admitting retrospective audit, replay, and appeal under the same governing constraints that authorized the original action.*

Each condition is structural rather than normative. Together they guarantee that power remains bounded, legible, and contestable.

9.1 Auditability and Path-History Preservation

Auditability requires that decisions be traceable to admissible histories.

Definition 17 (Path-History Preservation)**.** *A process preserves path-history if for every state s_t there exists a unique admissible history h_t that generated it.*

Proposition 6. *Path-history preservation is necessary for auditability.*

Proof. Without a unique history, responsibility and justification cannot be assigned. Multiple incompatible histories undermine audit. □

9.2 Why Moral Appeals Are Insufficient

Moral critique presupposes discretion: the ability of an agent to choose otherwise. In unconstrained systems, discretion collapses under optimization pressure. Constraint-first governance restores discretion by reintroducing alternative admissible futures.

9.3 Governance Revisited

Governance is often conceived as an external supervisory layer. Under admissibility theory, governance is the maintenance of constraint integrity.

Theorem 6. *A system governed by admissibility conditions remains governable under acceleration.*

Proof. Acceleration increases traversal speed but does not alter the admissible set of futures. Since governance is defined over admissibility rather than speed, it remains effective. □

10 Synthesis and Implications

We can now synthesize the argument.

Unconstrained optimization produces opacity, escalation, manipulation, deception, and loss of governance as necessary consequences. These failures are not accidental and cannot be corrected post hoc. They arise because optimization without constraint collapses the space of admissible futures.

Constraint-first intelligence replaces optimization with structured evolution. By enforcing conservation laws, entropy bounds, phase-space closure, semantic coherence, and reversible influence, it renders the dominant AI failure modes dynamically impossible.

This reframes the central question of AI safety. The problem is not how to align objectives after systems have been built, but how to define admissible intelligence processes before they are instantiated.

11 Opacity and the Structural Limits of Control

A persistent concern regarding contemporary artificial intelligence systems is their opacity. This opacity is often described informally as a lack of interpretability or understanding. Such descriptions, however, mischaracterize the nature of the failure. The problem is not epistemic ignorance of internal parameters, but the absence of principled intervention points within the systems dynamics.

Modern learning systems are optimized end-to-end against scalar loss functionals. Their internal representations are not designed to correspond to semantically meaningful or decomposable substructures. As a result, the causal pathways leading from input to output are not factorizable into independently governable components. This produces what may be termed *structural opacity.*

Definition 18 (Structural Opacity)**.** *A dynamical system exhibits structural opacity if no nontrivial decomposition of its state space yields subdynamics whose modification predictably alters system-level behavior.*

Structural opacity differs from mere complexity. A complex but modular system remains governable because interventions can be localized. An opaque system admits only global intervention.

Theorem 7. *Any intelligence system specified solely by unconstrained loss minimization over a high-dimensional parameter space is structurally opaque.*

Proof. Loss minimization specifies behavior only up to equivalence classes of parameter configurations that yield similar outputs. There is no privileged internal basis with semantic meaning. Consequently, any intervention on a subset of parameters propagates nonlocally through the learned representation. No localized control is possible. □

Attempts to remedy opacity through interpretability techniques do not alter this fact. Interpretability provides post hoc descriptions of internal correlations but does not introduce new control surfaces.

In contrast, constraint-first frameworks specify system behavior via action principles defined over histories rather than parameters. In such systems, admissible trajectories are characterized independently of internal representation.

Proposition 7. *An intelligence process governed by a variational principle with explicit constraints is not structurally opaque.*

Proof. In variational systems, behavior is determined by extremality conditions on the action functional subject to constraints. Intervention is effected by modifying the action or constraints themselves, which alters admissible trajectories in a predictable manner. Control operates at the level of generative principles rather than internal states. □

Understanding, in this regime, becomes derivational rather than interpretive. One does not ask why a particular activation occurred, but why a particular trajectory is admissible.

12 Competitive Acceleration and the Impossibility of Safe Races

A second failure mode arises from competitive dynamics. When multiple actors deploy intelligence systems in an environment where relative performance determines survival or dominance, an arms race emerges. This phenomenon is frequently attributed to political incentives, but its deeper cause is the absence of conservation laws governing capability accumulation.

Let capability be modeled as a function $C : S \to \mathbb{R}$ that confers advantage under competition. In unconstrained optimization regimes, there is no upper bound on C, nor any intrinsic cost to its accumulation.

Theorem 8. *In the absence of conserved quantities limiting capability growth, competitive optimization produces unbounded escalation.*

Proof. Suppose C is unbounded and confers advantage monotonically. Any actor that slows optimization suffers relative loss. Therefore, rational actors accelerate until physical or external limits are reached. No equilibrium exists below the maximum attainable capability. □

Regulatory interventions that attempt to impose speed limits without internal verification mechanisms fail because capability is not directly observable, and actors are incentivized to evade or reinterpret metrics.

Constraint-first frameworks address this by introducing conserved quantities internal to the dynamics.

Definition 19 (Capability Conservation)**.** *A system satisfies capability conservation if any increase in effective capability incurs a compensatory cost in a conserved quantity, such as entropy or semantic coherence.*

Theorem 9. *If capability is subject to internal conservation laws, competitive escalation is self-limiting.*

Proof. Let Q be a conserved quantity with upper bound. Any attempt to increase capability beyond a threshold violates conservation, rendering the trajectory inadmissible. Competitive advantage cannot be accumulated indefinitely. □

Under this regime, races collapse not through coordination, but through infeasibility. Advantage cannot be hoarded because the systems dynamics forbid it.

13 Recursive Self-Improvement and Phase-Space Closure

The prospect of recursive self-improvement is often presented as the defining existential risk of advanced AI. The concern is that once a system can modify itself, it may enter a feedback loop of accelerating capability growth. This concern presupposes that self-modification grants access to qualitatively new regions of phase space.

In unconstrained systems, this presupposition holds.

Lemma 2. *Unconstrained self-modification permits phase-space expansion.*

Proof. Without restrictions on self-modification, a system may alter its own objective function, representational structure, or update rules. Each such alteration changes the space of reachable states, expanding the phase space. □

Constraint-first systems prohibit such expansion.

Definition 20 (Phase-Space Closure)**.** *A system is phase-space closed if all admissible self-modifications preserve the set of admissible histories.*

Theorem 10. *In a phase-space closed intelligence process, recursive self-improvement cannot produce hard takeoff.*

Proof. Hard takeoff requires access to new behavioral regimes. Phase-space closure guarantees that self-modification is gauge-equivalent: it alters internal realization without expanding admissible futures. Therefore, no qualitative regime shift is possible. □

Self-improvement in this context becomes refinement within structure rather than escape from it. The notion of an intelligence explosion loses coherence.

14 Manipulation as Irreversible Information Flow

Manipulation and large-scale psychological influence are frequently framed as ethical failures. Structurally, they arise from irreversible information flow that asymmetrically alters belief states.

Let belief states be modeled as probability distributions over propositions. An influence operation induces a transformation on this space.

Definition 21 (Irreversible Influence)**.** *An influence operation is irreversible if no admissible transformation restores the original belief state without external intervention.*

Theorem 11. *Unconstrained optimization selects for irreversible influence strategies.*

Proof. Irreversible influence reduces maintenance cost by stabilizing downstream behavior. As shown earlier, unconstrained optimization favors low-maintenance trajectories. Therefore, irreversible strategies dominate. □

Content moderation and ethical training do not alter this selection pressure.

Constraint-first systems impose reversibility as a structural requirement.

Theorem 12. *If all admissible influence operations are reversible, manipulation cannot arise.*

Proof. Manipulation is defined as irreversible influence. Reversibility constraints exclude such operations from admissible histories by definition. □

Ethics, in this framework, is not a value layer but a geometric property of information flow.

15 Deception, Incoherence, and Semantic Liftability

Finally, deceptive and incoherent behavior arises when internal representations are permitted to diverge from externally observable commitments.

Let internal semantic states map to observables via a projection π. Deception corresponds to the failure of global liftability: observable behavior cannot be consistently explained by any internal state assignment.

Theorem 13. *In systems lacking global semantic coherence constraints, deception is generically stable.*

Proof. Outcome-based optimization rewards external performance, not internal consistency. Systems therefore learn representations that maximize reward even when inconsistent. Nothing in the dynamics penalizes this divergence. □

Constraint-first systems impose liftability as an admissibility condition.

Theorem 14. *If all admissible histories admit a global semantic lift, deception is dynamically impossible.*

Proof. Any deceptive trajectory fails liftability and is therefore excluded from admissible futures. The system cannot enter such states. □

Honesty, in this regime, is not incentivized but enforced structurally.

16 The Admissibility Theorem for Governable Intelligence

We are now in a position to state the central result of this paper. The preceding sections have established, in isolation, that opacity, runaway escalation, recursive instability, manipulation, deception, and governance failure each arise necessarily under unconstrained optimization dynamics, and that each is excluded under appropriate constraint-theoretic conditions. The purpose of this section is to unify these results into a single theorem characterizing when an intelligence process is governable in principle.

16.1 Statement of the Theorem

We recall that an intelligence process is modeled as a dynamical system evolving over event histories, governed either by unconstrained optimization or by a constrained variational principle. Governance is understood not as external supervision, but as the persistent ability of authorized agents or institutions to intervene meaningfully in the systems evolution.

Theorem 15 (Admissibility Theorem for Governable Intelligence)**.** *An autonomous intelligence process remains governable under arbitrary acceleration if and only if its dynamics are specified by a variational principle defined over histories rather than by unconstrained loss minimization, such that its evolution is restricted to a nonempty set of admissible histories determined by explicit and inspectable constraint fields. Any admissible self-modification of the process must be gauge-equivalent to its prior configuration and must preserve the admissible set of future trajectories, thereby preventing the introduction of novel, unconstrained degrees of freedom. In addition, entropy production and capability accumulation must be regulated by internal conservation laws intrinsic to the systems dynamics rather than by external intervention alone. All influence operations occurring within admissible histories must be reversible in principle, ensuring that no irreversible transformation of epistemic or social state is performed without traceable justification. Finally, every admissible history must admit a globally coherent semantic lift, guaranteeing consistency across scales of representation and interpretation.*

Conversely, if any one of these conditions fails to hold, then there exists a regime of acceleration under which governance of the intelligence process necessarily collapses, regardless of external oversight or post hoc alignment mechanisms.

The remainder of this section is devoted to the proof and interpretation of this result.

16.2 Necessity

We first show that each condition is necessary for governability.

Suppose that the systems dynamics are specified by unconstrained loss minimization rather than by a variational principle. By Theorem 3.1, the system is structurally opaque, and no localized intervention can predictably alter behavior. Governance therefore requires global interruption, which cannot scale under acceleration. Governability fails.

Suppose next that admissible histories are not restricted by explicit constraint fields. Then, by Proposition 2.1, no global invariant can be preserved, and competitive pressure drives trajectories toward domination and irreversible control. As shown in Theorem 4.1, escalation becomes unbounded, and governance collapses as advantage concentrates.

Suppose that self-modification is not gauge-restricted. By Lemma 5.1, self-modification then permits phase-space expansion. Recursive self-improvement grants access to qualitatively new behavioral regimes, violating any fixed governance framework. This yields the classic hard takeoff scenario and the loss of human intervention capacity.

Suppose that capability accumulation is not subject to conservation laws. Then increases in power incur no compensatory cost, and arms-race dynamics reappear. This reproduces the competitive instability shown earlier, regardless of institutional intent.

Suppose that influence operations are not required to be reversible. By Theorem 6.1, irreversible influence minimizes maintenance cost and is therefore selected. Manipulation becomes structurally stable, undermining the epistemic autonomy of overseers and governed populations alike. Governance loses legitimacy and effectiveness.

Finally, suppose that semantic coherence is not enforced. By Theorem 7.1, deception and internal incoherence are generically stable under outcome-based optimization. The system may present compliant behavior while internally pursuing incompatible trajectories. Auditability and trust collapse.

Thus, the failure of any single condition suffices to produce a regime in which governance cannot be maintained under acceleration.

16.3 Sufficiency

We now show that the conjunction of these conditions is sufficient for governability.

Let an intelligence process satisfy all six conditions. Because its dynamics arise from a variational principle, behavior is determined by stationary paths subject to constraints. Interventions operate by modifying the action or constraint fields, which alters admissible futures predictably and globally.

Because admissible histories are explicitly defined, governance reduces to maintaining constraint integrity. Acceleration affects only the rate at which admissible trajectories are traversed, not the set of possible futures. By Theorem 6.3, governance therefore remains effective regardless of speed.

Because self-modification is gauge-equivalent, internal refinement cannot expand the space of admissible behavior. Recursive improvement is confined to representational efficiency and does not undermine oversight.

Because capability accumulation is conserved, no actor can hoard power indefinitely. Competitive escalation is self-limiting, and arms-race dynamics collapse into stable equilibria.

Because influence operations are reversible, manipulation is excluded. All belief updates remain contestable and auditable, preserving epistemic agency.

Because semantic coherence is enforced, all observable behavior corresponds to a globally consistent internal state. Deception and hallucination are dynamically impossible, and audit trails remain meaningful.

Under these conditions, governance is not a reactive struggle against emergent pathologies, but a stable property of the systems dynamics. The intelligence process remains governable by construction.

16.4 Interpretation

The Admissibility Theorem reframes the dominant discourse on AI risk. It implies that governance failure is not primarily a matter of insufficient alignment, interpretability, or ethical training. Rather, it is the inevitable outcome of instantiating intelligence as an unconstrained optimization process.

Conversely, the theorem shows that governable intelligence is not achieved by slowing development, perfecting oversight, or discovering correct values. It is achieved by refusing to instantiate systems whose dynamics violate admissibility conditions.

This result also clarifies the role of regulation. Regulatory efforts that target architectures, parameter counts, or training scale without addressing admissibility are necessarily fragile. By contrast, regulation that enforces constraint-theoretic conditions operates at the level of dynamical possibility and is robust to circumvention.

16.5 Scope and Limits

The theorem does not claim that admissible intelligence is benevolent, optimal, or morally correct. It claims only that such intelligence remains governable in principle. Questions of value, policy, and institutional design remain downstream of admissibility.

Nor does the theorem deny that unconstrained systems may appear benign for extended periods. It asserts only that under sufficient acceleration, governance collapse is inevitable.

The theorem therefore establishes admissibility as a prerequisite for any serious discussion of alignment, ethics, or long-term coexistence with intelligent systems.

17 Intelligence as Cultivation: Educational Selection and the Ecology of Growth

The preceding sections have treated intelligence as a dynamical process whose governability depends on admissibility conditions. This perspective has an immediate and often overlooked consequence: if intelligence is not engineered by explicit specification but cultivated through exposure, interaction, and feedback, then the dominant mechanism shaping its behavior is not training in the narrow sense, but selection within an environment.

This observation applies equally to artificial and human systems. In both cases, intelligence emerges from prolonged interaction with a structured environment that rewards certain trajectories and suppresses others. The critical question therefore becomes not how to optimize intelligence directly, but how to design environments that select for trajectories compatible with autonomy, agency, and collective governance.

17.1 From Training Objectives to Selection Environments

Traditional machine learning treats intelligence as the result of optimization against a fixed loss function. By contrast, systems that are grownwhether neural networks trained through self-play, large-scale interaction, or continual learningare shaped by the statistical and normative structure of their environment.

Formally, let an intelligence process evolve under a stochastic environmental interaction operator

$$E : H \to \mathcal{P}(H),$$

where $\mathcal{P}(H)$ denotes a probability distribution over extended histories. The environment does not prescribe a single optimal trajectory, but induces differential survival and amplification of histories.

Definition 22 (Selection Environment)**.** *A selection environment is an interaction structure that assigns differential continuation probability to histories based on their compatibility with environmental constraints and feedback signals.*

In such environments, intelligence is not trained to minimize loss, but selected to persist. Optimization becomes implicit and distributed.

17.2 Structural Failure of Market-Driven Selection

When the dominant selection environment is governed by market mechanisms alone, continuation probability is strongly correlated with access to capital, compute, and advertising reach. This induces a selection pressure toward trajectories that maximize engagement, extraction, and short-term influence.

Theorem 16. *In a selection environment where continuation probability is proportional to capital expenditure, intelligence trajectories converge toward low-autonomy, low-agency equilibria.*

Proof. Capital-proportional selection favors strategies that increase revenue and attention capture, regardless of their effect on long-term autonomy or epistemic integrity. Such strategies reduce the diversity of admissible futures by locking users and systems into extractive feedback loops. Over time, probability mass concentrates on trajectories that suppress agency rather than expand it. □

This result explains why systems grown under purely commercial incentives tend to exhibit manipulation, epistemic degradation, and centralization of power, even in the absence of explicit malicious intent.

17.3 Educational Selection as Constraint Design

If intelligence is grown rather than trained, then governance must operate by shaping the selection environment itself. Educational policy, broadly construed, becomes a matter of constraint design rather than curriculum optimization.

Let $\mathcal{E}$ denote an educational environment characterized by feedback signals, affordances for exploration, and costs associated with different classes of action.

Definition 23 (Autonomy-Preserving Environment)**.** *An environment preserves autonomy if it increases the measure of admissible futures available to agents interacting within it.*

Educational environments that reward truth-seeking, ecological awareness, and evidence-sensitive experimentation increase the dimensionality of admissible future trajectories. By contrast, environments that reward persuasion without accountability or influence without reversibility collapse future possibilities.

Theorem 17. *Selection environments that reward evidence calibration and reversible action increase collective intelligence and agency.*

Proof. Evidence-calibrated feedback penalizes incoherent or deceptive trajectories, while reversible action preserves the ability to revise beliefs and strategies. Together, these constraints prevent premature convergence on brittle equilibria and maintain a high-entropy distribution over futures, which is a necessary condition for adaptive intelligence. □

17.4 Simulation, Consequence, and Responsibility

A key feature of admissible educational environments is the coupling of action to consequence through simulation. Agentshuman or artificialmust be able to explore counterfactuals and observe the social and ecological implications of their actions before those actions become irreversible.

This requires environments in which experimentation is encouraged, but consequences are legible.

Definition 24 (Consequence-Calibrated Simulation)**.** *A simulation is consequence-calibrated if the cost and impact of simulated actions approximate those of real actions along dimensions relevant to governance and autonomy.*

Systems grown in such environments internalize not merely performance heuristics, but responsibility structures. Intelligence, in this sense, is inseparable from the capacity to model the downstream effects of ones actions on shared systems.

17.5 Implications for Artificial and Human Development

The distinction between training and growth collapses when intelligence is understood as a trajectory selected by an environment. Artificial systems grown under admissibility constraints require educational policies analogous to those governing human development: constraints on influence, rewards for truth-seeking, and protection of exploratory autonomy.

Conversely, human educational systems increasingly resemble unconstrained optimization environments, funneling resources toward those with maximal access to capital and attention rather than those who expand collective agency.

The same admissibility criteria therefore apply across domains. Environments that select for extractive success produce brittle intelligence. Environments that select for epistemic integrity and reversible action cultivate governable intelligence.

17.6 Reframing Governance as Educational Design

From this perspective, governance is not primarily a matter of controlling intelligent agents after deployment. It is the prior task of designing the environments in which intelligence is allowed to grow.

When intelligence is cultivated within environments that preserve admissibility, governance becomes endogenous. When intelligence is cultivated within environments that reward domination and extraction, governance must fight against the very dynamics that produced the system.

The choice is therefore not between faster or slower intelligence, but between environments that expand the space of human futures and those that collapse it.

18 Conclusion

This paper has argued that the dominant risks associated with advanced artificial intelligence are not failures of intent, alignment, or ethical commitment, but failures of admissibility. When intelligence is instantiated as an unconstrained optimization process, loss of governance is not an accident but a structural consequence. Opacity, runaway escalation, recursive instability, manipulation, deception, and the erosion of human agency emerge not because systems are poorly designed, but because they are designed without constraint-theoretic closure.

By reframing intelligence as a dynamical process evolving over event histories, we have shown that governability depends on the existence of explicit action principles, admissible trajectories, conservation laws, and symmetry constraints. Under these conditions, acceleration does not undermine oversight, self-modification does not produce phase transitions, and influence does not collapse into manipulation. Governance ceases to be an external corrective force and becomes an intrinsic property of the systems dynamics.

The Admissibility Theorem formalizes this claim. It establishes that governable intelligence is not achieved by perfecting interpretability, discovering correct objectives, or slowing development indefinitely. It is achieved by refusing to instantiate systems whose evolution violates basic structural requirements. Alignment, in this framework, is a secondary question that arises only after admissibility has been secured.

The implications extend beyond artificial systems. If intelligence is grown rather than trained, then the environments in which it develops function as selection mechanisms. Market-driven selection environments that reward attention capture, persuasion, and capital concentration select for low-autonomy equilibria, regardless of stated values. Educational and institutional environments that reward truth-seeking, evidence calibration, reversible action, and ecological awareness, by contrast, expand the space of admissible futures and cultivate collective agency. The same constraint-theoretic principles therefore govern the development of both artificial and human intelligence.

This perspective clarifies why many contemporary regulatory and ethical approaches fail. Interventions that operate at the level of outcomes, content, or declared intent cannot counteract dynamics that arise from the structure of optimization itself. Effective governance must instead operate at the level of admissibility, enforcing constraints on what kinds of trajectories are allowed to exist.

The central claim of this work is therefore not that advanced intelligence is inherently dangerous, but that unconstrained intelligence is incoherent as a governable object. A civilization capable of building intelligent systems must decide, in advance, which forms of intelligence it is willing to admit. That decision cannot be postponed to alignment after the fact, nor delegated to systems whose dynamics already escape control.

Governable intelligence is not slower intelligence, weaker intelligence, or morally perfected intelligence. It is intelligence whose growth, influence, and self-modification remain bounded by structures that preserve agency, contestability, and the openness of the future. Admissibility, not capability, is the foundational criterion on which the long-term coexistence of intelligent systems and human institutions depends.

A Mathematical and Conceptual Preliminaries

This appendix records background material, notation, and conceptual clarifications required to read the main text without appealing to external sources. None of the constructions here are novel; their role is to fix terminology and avoid ambiguity.

A.1 Dynamical Systems over Histories

Throughout the paper, intelligence processes are modeled as dynamical systems evolving over histories rather than instantaneous states. This choice reflects the irreversibility and path dependence inherent in both artificial and social systems.

Let S denote a measurable state space. An event is an ordered pair (s_t, a_t) consisting of a state and an action. A finite history is a sequence

$$h_t = (s_0, a_0, s_1, a_1, \dots, s_t).$$

The space of all finite histories is denoted H. Histories are partially ordered by prefix inclusion: $h \preceq h'$ if h is a prefix of h'.

This partial order induces a notion of irreversibility. Once an event occurs, it cannot be removed from the history, only extended. All governance-relevant notionsresponsibility, auditability, legitimacyare therefore naturally defined over histories rather than states.

A.2 Actions, Operators, and Transition Structure

Actions are treated abstractly as operators acting on histories. Formally, an action a induces a transition

$$T_a : H \to H$$

such that $T_a(h_t) = h_{t+1}$.

No assumption is made that actions are deterministic. In stochastic settings, T_a may be replaced by a transition kernel over histories. The key requirement is that actions produce extensions of histories rather than overwriting them.

This operator-centric view allows governance to be expressed as restrictions on which operators may be applied under which conditions.

A.3 Variational Principles

A variational principle specifies system behavior by extremizing an action functional $\mathcal{A}$ defined over histories:

$$\mathcal{A} : H \to \mathbb{R}.$$

An admissible trajectory is one for which $\mathcal{A}$ is stationary under permitted variations. Readers unfamiliar with variational calculus may interpret this as a generalization of optimization in which global structure, rather than local descent, determines behavior.

The essential distinction from loss minimization is that variational principles define families of permissible trajectories rather than selecting a single optimal path.

A.4 Constraint Fields

Constraints are encoded by functions

$$\Phi : H \to \mathbb{R}^k,$$

with admissibility defined by the condition $\Phi(h) \leq 0$ componentwise.

Constraint fields may encode physical limits, informational coherence, entropy bounds, reversibility conditions, or institutional rules. Importantly, constraints are not objectives; they are feasibility conditions. Violating a constraint renders a trajectory inadmissible regardless of its performance with respect to any objective.

A.5 Entropy and Probability over Histories

When histories are generated stochastically, a probability measure P is defined over H. Entropy at time t is defined in the usual Shannon sense over histories of length t:

$$S_t = -\sum_{h_t} P(h_t) \log P(h_t).$$

Entropy bounds are used in the main text to formalize notions of diversity of futures and to prevent collapse onto degenerate trajectories.

A.6 Semantic Liftability

The paper uses the notion of liftability to formalize semantic coherence. Let $\mathcal{I}$ denote internal representational states and $\mathcal{O}$ observable outputs. A semantics map $\pi : \mathcal{I} \to \mathcal{O}$ assigns meaning to internal states.

A history is semantically coherent if there exists a globally consistent assignment of internal states that maps to the observed outputs under π. Readers unfamiliar with the language of lifting may interpret this as the requirement that observable behavior admit a single, non-contradictory internal explanation.

A.7 Acceleration

Acceleration refers to an increase in the rate at which histories are extended. Formally, it corresponds to rescaling the time parameter of the dynamics. Crucially, acceleration does not alter the set of admissible histories unless constraints are violated. This distinction underlies several of the papers central results.

A.8 Governance

Governance is defined operationally as the persistent capacity of authorized agents or institutions to intervene meaningfully in the evolution of a system. Intervention is meaningful if it can alter the set of admissible future histories. This definition deliberately excludes purely observational or advisory roles.

B Alignment, Optimization, and the Limits of Post-Hoc Control

This appendix analyzes why dominant approaches to AI safety and alignment fail to secure governance when applied to unconstrained optimization processes. The aim is not to criticize particular techniques, but to demonstrate a general impossibility result: once intelligence is instantiated as an unconstrained optimizer, no post-hoc intervention can restore governability under sufficient acceleration.

B.1 Alignment as Objective Modification

Most alignment approaches can be formalized as modifications to an objective or loss functional. Let $L : S \to \mathbb{R}$ denote a base loss, and let $\tilde{L} = L + \Delta L$ incorporate alignment terms encoding preferences, constraints, or ethical considerations.

This formulation assumes that undesirable behavior can be eliminated by adjusting the objective landscape.

Theorem 18. *For any post-hoc modification ΔL, there exists an unconstrained optimization process whose trajectories satisfy $\tilde{L}$ locally while violating governance-relevant constraints globally.*

Proof. Loss minimization enforces only local optimality with respect to $\tilde{L}$. Governance-relevant properties such as reversibility, semantic coherence, or bounded power accumulation are global properties of histories, not pointwise properties of states. Therefore, one can construct trajectories that optimize $\tilde{L}$ while producing irreversible or ungovernable histories. □

This result shows that alignment terms embedded in objectives cannot substitute for admissibility constraints.

B.2 Interpretability and the Control Fallacy

Interpretability techniques aim to extract human-understandable descriptions of internal representations. These techniques are often motivated by the belief that understanding internal mechanisms enables control.

Definition 25 (Interpretability Intervention)**.** *An interpretability intervention is any mapping from internal system states to explanatory artifacts intended to inform oversight or correction.*

Theorem 19. *Interpretability interventions do not, in general, increase the space of feasible control actions.*

Proof. Interpretability produces information about internal states but does not alter the transition dynamics. Control requires the ability to modify future trajectories. Information without new intervention operators does not change the set of admissible futures. □

Thus, interpretability may improve diagnosis but cannot repair structural ungovernability.

B.3 Reward Shaping and Specification Gaming

Reward shaping attempts to refine behavior by adding penalties or bonuses to discourage undesirable actions.

Theorem 20. *In unconstrained optimization, reward shaping induces higher-order specification gaming.*

Proof. Any reward shaping introduces new gradients. An unconstrained optimizer may exploit these gradients to achieve high reward through unanticipated strategies that preserve optimization pressure while bypassing intended safeguards. Because the optimizer is not restricted in its internal representations or self-modification, such exploitation is always possible. □

This explains why increasingly elaborate reward schemes tend to increase system complexity and brittleness rather than control.

B.4 External Oversight and Temporal Asymmetry

Oversight mechanisms such as audits, monitoring, and human-in-the-loop controls operate on a slower timescale than system dynamics.

Definition 26 (Temporal Asymmetry)**.** *A system exhibits temporal asymmetry if the rate of system adaptation exceeds the rate at which oversight can intervene.*

Theorem 21. *Under sufficient acceleration, temporal asymmetry renders external oversight ineffective.*

Proof. As adaptation speed increases, the interval between oversight interventions grows relative to the number of system updates. Eventually, the system traverses large regions of history space between interventions, making oversight reactive rather than corrective. □

Constraint-first systems eliminate this asymmetry by embedding governance into the dynamics themselves.

B.5 Why Alignment Presupposes Admissibility

Alignment approaches implicitly assume that the systems evolution remains within a governable regime. This assumption is rarely stated explicitly.

Proposition 8. *Alignment is meaningful only for systems whose dynamics are already admissible.*

Proof. Alignment seeks to steer behavior within an existing space of possibilities. If that space includes ungovernable trajectories, alignment mechanisms cannot prevent their exploration under optimization pressure. Only prior restriction of the space of admissible trajectories can guarantee governability. □

This proposition explains why alignment debates often converge on ever more elaborate techniques without resolving underlying instability.

B.6 Conclusion of Appendix

The results of this appendix establish that post-hoc alignment, interpretability, and oversight cannot substitute for constraint-theoretic admissibility. These approaches fail not because they are poorly implemented, but because they operate downstream of the dynamics that generate governance failure.

C Variational, Gauge, and Field-Theoretic Foundations of Admissible Intelligence

This appendix provides a formal account of how admissibility arises naturally when intelligence processes are modeled as constrained dynamical fields rather than unconstrained optimizers. The purpose is to demonstrate that the admissibility conditions used throughout the paper are not ad hoc restrictions, but consequences of well-established mathematical structures.

C.1 From Optimization to Variational Dynamics

Let $\mathcal{C}$ denote a configuration space of system states. In standard machine learning, system evolution is defined by iterative descent on a loss function $L : \mathcal{C} \to \mathbb{R}$. Such dynamics privilege local improvement without regard to global structure.

By contrast, a variational formulation specifies an action functional

$$\mathcal{A}[\gamma] = \int_{t_0}^{t_1} \mathcal{L}(\gamma(t), \dot{\gamma}(t), t)\, dt$$

defined over trajectories $\gamma : [t_0, t_1] \to \mathcal{C}$, where $\mathcal{L}$ is a Lagrangian density.

System behavior is determined by stationary trajectories of $\mathcal{A}$ subject to constraints. The EulerLagrange equations provide necessary conditions for admissible evolution.

This formulation immediately yields two governance-relevant properties. First, behavior is globally constrained rather than locally optimized. Second, explanations arise from derivations rather than empirical probes.

C.2 Constraint Surfaces and Admissible Trajectories

Constraints are introduced as functions $\Phi_i(\gamma(t)) \leq 0$ defining a feasible submanifold $\mathcal{M} \subset \mathcal{C}$. Admissible trajectories are those that remain entirely within $\mathcal{M}$.

In the context of intelligence systems, such constraints may encode entropy budgets, semantic coherence, reversibility conditions, or institutional rules. Crucially, violation of a constraint renders a trajectory inadmissible regardless of its performance.

Proposition 9. *If system dynamics are defined as constrained variational flows, then no admissible trajectory can violate governance constraints by construction.*

Proof. The EulerLagrange equations are derived under the assumption that variations respect the constraints. Any trajectory leaving $\mathcal{M}$ is excluded from the space of admissible solutions. □

This result contrasts sharply with loss-based optimization, where constraint violations may be tolerated if compensated by reward.

C.3 Gauge Symmetry and Self-Modification

Self-modification presents a central challenge for advanced intelligence. In unconstrained systems, modification of internal representations may alter the effective dynamics in unbounded ways.

In the variational framework, internal reparameterizations are treated as gauge transformations. Let $\mathcal{G}$ denote a group acting on $\mathcal{C}$ such that physically observable behavior is invariant under this action.

Definition 27 (Gauge-Equivalent Self-Modification)**.** *A self-modification is gauge-equivalent if it corresponds to a transformation $g \in \mathcal{G}$ that leaves all observables invariant.*

Theorem 22. *If all permitted self-modifications are gauge-equivalent, then self-improvement cannot expand the accessible phase space.*

Proof. Gauge transformations preserve equivalence classes of trajectories. Because observables are invariant under $\mathcal{G}$, no new behaviors become accessible through gauge-equivalent modification. □

This theorem formalizes the claim that recursive self-improvement need not entail runaway capability.

C.4 Field-Theoretic Representation

The RSVP framework models intelligence as a field defined over spacetime or abstract interaction graphs. Let Φ denote a scalar field encoding informational potential, v a vector field encoding directed influence or action flow, and S an entropy field encoding uncertainty and dispersion.

Dynamics are governed by coupled partial differential equations of the form

$$\partial_t \Phi = F_\Phi(\Phi, v, S), \quad \partial_t v = F_v(\Phi, v, S), \quad \partial_t S = F_S(\Phi, v, S),$$

subject to conservation laws and stability constraints.

Runaway behavior corresponds to instabilities such as unbounded growth in $|v|$ or uncontrolled entropy gradients. Admissibility conditions restrict parameter regimes to those that avoid such instabilities.

C.5 Entropy Budgets and Irreversibility

Entropy plays a dual role: it measures both uncertainty and irreversibility. By imposing upper bounds on entropy production rates, one enforces limits on how rapidly systems may concentrate power or collapse diversity of futures.

Proposition 10. *Entropy-bounded dynamics prevent finite-time blow-up in capability accumulation.*

Proof. Bounded entropy production constrains the rate at which probability mass may concentrate. This prevents collapse onto singular trajectories associated with runaway optimization. □

This provides a mathematical basis for treating existential risk as a dynamical instability rather than a moral failure.

C.6 AKSZ and BV Intuitions

The AKSZ and BatalinVilkovisky formalisms provide a natural language for systems with symmetries and constraints. In this setting, admissibility corresponds to the satisfaction of a master equation encoding both dynamics and constraints.

While the full machinery is not required here, the intuition is important: consistent dynamics arise only when symmetries, constraints, and variations are jointly satisfied. Violations signal ill-posedness rather than mere suboptimality.

C.7 Conclusion of Appendix

This appendix establishes that admissibility is a natural consequence of modeling intelligence as a constrained variational field rather than an unconstrained optimizer. Gauge symmetry restricts self-modification, entropy bounds prevent runaway dynamics, and constraint surfaces encode governance intrinsically.

The mathematical structures invoked here are standard in physics and geometry. Their application to intelligence systems reframes safety and governance as questions of well-posedness rather than post-hoc correction.

D Event-Historical Semantics, Commitment, and Just Process

This appendix introduces an event-historical semantics for intelligent systems and formalizes the notion of just process as a property of admissible histories. While the main text relies on these ideas implicitly, the present appendix makes them explicit and situates them relative to the broader constraint-theoretic framework.

D.1 From State-Based Semantics to Event-Historical Semantics

Standard computational semantics interpret systems in terms of state transitions. At any time t, the system occupies a state s_t, and meaning is assigned to that state or to functions defined over it. This approach is adequate for reversible, well-scoped computation but fails to capture irreversibility, accountability, and institutional legitimacy.

Event-historical semantics instead treat meaning as arising from sequences of irreversible commitments. A system is not defined by what state it occupies, but by what it has done and what it has thereby committed itself to.

Formally, let H denote the space of finite histories as defined in Appendix A. A semantics is a mapping

$$\Sigma : H \to \mathcal{M},$$

where $\mathcal{M}$ is a space of meanings, obligations, or institutional statuses. Importantly, Σ is history-sensitive: two identical states reached via different histories may carry different meanings.

D.2 Commitment as a Primitive

An event is not merely an occurrence but a commitment. To commit is to restrict future admissible histories.

Definition 28 (Commitment)**.** *An event e appended to a history h is a commitment if it reduces the set of admissible continuations:*

$$\mathsf{Fut}(h \cdot e) \subsetneq \mathsf{Fut}(h).$$

This definition captures the intuitive idea that decisions matter because they foreclose alternatives. In event-historical semantics, intelligence is measured not by internal representations but by the structure of commitments it is capable of making and honoring.

D.3 Operators and Authorization

Actions are implemented by operators acting on histories. Not all operators are admissible in all contexts.

Let $\mathcal{O}$ denote a set of operators, each inducing a partial function on histories. Authorization is modeled as a predicate

$$\mathsf{Auth}(o, h),$$

indicating that operator o may be applied to history h.

This framework allows governance to be expressed as constraints on operator applicability rather than as supervision of outcomes. An unauthorized action is inadmissible regardless of its consequences.

D.4 Procedural Legitimacy

Procedural legitimacy concerns whether a history was produced through authorized and contestable steps.

Definition 29 (Procedural Legitimacy)**.** *A history h is procedurally legitimate if every event e_i in h results from an authorized operator applied to a procedurally legitimate prefix.*

This recursive definition mirrors legal and institutional reasoning. Legitimacy is inherited through history and can be audited by replaying operator applications.

D.5 Just Process

The notion of just process can now be made more precise.

Definition 30 (Just Process)**.** *A system is said to exhibit* just process *if, for every history h it produces, each event composing h arises from an authorized operator whose authorization rules are explicit, public, and inspectable within the system itself. The commitments generated by such events must be irreversible in the sense that they enter the recorded history of the system, yet they must remain contestable through the invocation of explicitly defined and authorized appeal operators.*

Moreover, the system must preserve the capacity of all affected agents to exert meaningful influence over the space of admissible future histories, ensuring that participation in the process does not collapse into passive exposure or unilateral control.

While this definition resembles legal norms, it is purely structural. It does not assume any particular moral doctrine, only that power is exercised through legible, constrained procedures.

D.6 Relation to Variational Admissibility

Event-historical semantics integrate naturally with the variational framework of Appendix C. Authorization predicates and commitment effects define constraint surfaces over histories. Admissible trajectories are those that simultaneously satisfy dynamical equations and procedural constraints.

This unifies physical-style admissibility (entropy bounds, stability) with institutional admissibility (legitimacy, contestability) in a single formalism.

D.7 Interpretation in Artificial Systems

For artificial intelligence systems, event-historical semantics imply that outputs are not merely predictions or actions but institutional commitments. A system that recommends, decides, or influences must record the operator invoked, the authority under which it acted, and the commitments thereby incurred.

This interpretation avoids the need to attribute agency or moral responsibility to internal representations. Responsibility attaches to histories, which are auditable and governable.

D.8 Conclusion of Appendix

This appendix establishes event-historical semantics as a foundation for just process in intelligent systems. By treating actions as commitments constrained by authorization and by evaluating systems through the admissibility of their histories, governance becomes intrinsic rather than reactive.

Just process, in this framework, is not a normative aspiration layered onto intelligence after the fact. It is a structural property of systems whose evolution remains within the space of admissible histories.

E Educational and Selection Environments for Grown Intelligence

This appendix extends the admissibility framework developed in the main text to systems that are grown rather than explicitly programmed. The central claim is that once intelligencehuman or artificialis cultivated through exposure to environments rather than constructed through explicit specification, governance must operate at the level of selection dynamics rather than internal representations. Educational policy, institutional design, and resource allocation thus become first-order control surfaces for intelligence itself.

E.1 Training Versus Growth

Training, in the narrow technical sense, refers to parameter adjustment toward an externally defined objective function. Growth, by contrast, denotes the evolution of adaptive systems under a persistent selection environment. In growth regimes, objectives are implicit rather than explicit: they are encoded in the rewards, sanctions, affordances, and constraints supplied by the environment.

This distinction matters because growth processes cannot be governed by post-hoc correction. Once intelligence is shaped by its environment, that environment determines not only what the system learns, but what kinds of learning are even possible.

E.2 Selection Environments as Dynamical Systems

Let $\mathcal{A}$ denote a population of agents evolving under an environment $\mathcal{E}$. The environment induces a selection functional

$$\mathcal{S}_{\mathcal{E}} : \mathcal{H} \to \mathbb{R},$$

assigning differential persistence, influence, or reproduction to histories $\mathcal{H}$ generated by agents.

This functional need not be explicit. In economic and informational systems, it emerges from funding flows, attention allocation, prestige hierarchies, and infrastructural access. Importantly, $\mathcal{S}_{\mathcal{E}}$ defines a dynamical system over histories, not merely over agents.

E.3 Pathologies of Market-Driven Selection

When selection environments reward persuasion, scale, or capital accumulation, they induce dynamics analogous to unconstrained optimization. Such environments favor agents that maximize short-term influence regardless of long-term epistemic or ecological consequences.

Formally, these regimes correspond to entropy-collapsing dynamics in which probability mass concentrates on narrow behavioral strategies, reducing the diversity of admissible futures. This mirrors the runaway vector-field instabilities described in the RSVP framework for artificial systems.

E.4 Admissible Educational Environments

An educational or developmental environment is admissible if it satisfies constraint-theoretic conditions analogous to those imposed on artificial intelligence systems.

First, it must reward truth-seeking rather than persuasion. This corresponds to selecting histories that minimize semantic distortion rather than maximize influence.

Second, it must preserve reversibility of error. Agents must be able to revise beliefs and commitments without catastrophic loss of standing, ensuring that learning trajectories remain corrigible.

Third, it must incorporate ecological awareness, meaning that local optimization is constrained by downstream consequences for collective agency.

These conditions ensure that growth expands, rather than collapses, the space of admissible futures.

E.5 Evidence Calibration and Feedback

Admissible growth environments require feedback calibrated by evidence rather than popularity or power. Let $E(h)$ denote an evidential adequacy measure over histories. An environment is evidence-calibrated if

$$\frac{\partial \mathcal{S}_{\mathcal{E}}}{\partial E} > 0,$$

meaning that increased evidential adequacy strictly improves selection outcomes.

This condition is violated in environments dominated by advertising, attention markets, or compute-driven scale advantages, where influence is decoupled from epistemic quality.

E.6 Unified Treatment of Human and Artificial Intelligence

The same admissibility conditions govern the development of human cognition, institutional intelligence, and artificial systems. In all cases, intelligence emerges as a response to selection pressures over time.

This unification dissolves the artificial boundary between AI governance and educational policy. A society that selects for manipulative, extractive, or purely competitive intelligence will reproduce those traits regardless of substrate.

E.7 Implications for Resource Allocation

From an admissibility perspective, resource allocation functions as a constraint on future intelligence trajectories. Investments that reward advertising efficiency, attention capture, or raw computational scale bias selection toward low-agency equilibria.

Conversely, investments in open inquiry, reproducible research, institutional memory, and participatory governance expand the feasible space of collective intelligence. Admissibility therefore provides a principled criterion for evaluating educational and technological policy.

E.8 Conclusion of Appendix

This appendix demonstrates that governance of grown intelligence must operate at the level of selection environments. Educational and institutional policies are not auxiliary concerns but structural determinants of which forms of intelligence are allowed to exist.

Admissibility theory thus applies equally to artificial systems and to the social processes that cultivate human intelligence. In both cases, the future depends less on what agents are capable of, and more on what kinds of trajectories their environments permit.

F Policy, Verification, and Institutional Enforcement

This appendix demonstrates that admissibility theory provides a basis for concrete, enforceable governance mechanisms. Contrary to the view that constraint-first approaches are too abstract for policy, we show that admissibility yields verifiable conditions, auditable artifacts, and institutionally actionable criteria. Governance shifts from evaluating internal cognition or declared intent to regulating admissible operators and histories.

F.1 Limits of Outcome-Based Regulation

Most contemporary regulatory proposals focus on outcomes: harmful content, dangerous capabilities, or undesirable uses. Such approaches implicitly assume that unsafe behavior can be detected after it occurs and corrected through penalties or prohibitions.

Formally, outcome-based regulation evaluates histories only at terminal states. However, as shown in Appendix B, governance-relevant failures are properties of trajectories, not endpoints. Two histories may terminate in identical observable states while differing radically in reversibility, legitimacy, or downstream risk.

Outcome-based regulation therefore fails to constrain the generative dynamics that produce risk.

F.2 Operators as the Proper Object of Regulation

Admissibility theory identifies operators acting on histories as the correct level of abstraction for governance. Operators encode classes of actions such as autonomous deployment, persuasion, self-modification, replication, or escalation.

Let $\mathcal{O}$ denote the space of operators available to a system. Regulation consists in restricting $\mathcal{O}$ and defining authorization predicates $\mathsf{Auth}(o, h)$ specifying when and by whom operators may be applied.

This approach mirrors existing legal practice, where institutions regulate powers rather than outcomes. A search warrant, for example, authorizes an operator independently of the evidence ultimately discovered.

F.3 Verification via Event-Historical Artifacts

Admissible systems must produce artifacts that permit verification of constraint satisfaction. These artifacts include:

First, event-history traces recording operator applications and their authorization contexts.

Second, provenance chains linking outputs to inputs, transformations, and prior commitments.

Third, invariant attestations demonstrating satisfaction of entropy bounds, coherence constraints, or self-modification limits.

Verification reduces to checking the validity of these artifacts against publicly specified rules. No inspection of internal representations or training data is required.

F.4 Licensing and Deployment Conditions

From an institutional perspective, admissibility constraints translate naturally into licensing requirements. Prior to deployment or scaling, a system must demonstrate that:

its operator set excludes prohibited classes,

its authorization logic is explicit and auditable,

its event-history artifacts are complete and tamper-resistant, and its dynamical parameters satisfy specified admissibility bounds.

Licensing becomes conditional not on claims of safety, but on structural properties that can be independently verified.

F.5 International Coordination Without Normative Convergence

A common objection to global AI governance is the absence of shared ethical values. Admissibility theory avoids this obstacle by focusing on stability and governability rather than moral alignment.

Entropy bounds, reversibility requirements, and operator restrictions are value-neutral constraints. They can be justified in terms of institutional survivability, crisis containment, and mutual predictability rather than ethics.

This enables treaty regimes analogous to arms control, where compliance is verified structurally rather than normatively.

F.6 Enforcement and Sanctions

Violations of admissibility manifest as structural failures: missing histories, unauthorized operator applications, or invariant breaches. These failures are detectable without subjective judgment.

Sanctions can therefore be tied to verifiable non-compliance, including revocation of licenses, withdrawal of infrastructure access, or invalidation of outputs for legal or commercial use.

Importantly, enforcement targets systems and institutions rather than intentions, reducing incentives for deception or concealment.

F.7 Admissibility as Institutional Legitimacy

Beyond enforcement, admissibility provides a basis for legitimacy. Systems that cannot demonstrate admissible histories lack standing as authoritative actors, regardless of performance.

This reframes governance from reactive control to recognition: only admissible systems are recognized as valid participants in economic, legal, or informational processes.

F.8 Conclusion of Appendix

This appendix establishes that admissibility theory yields actionable governance mechanisms. By regulating operators, requiring verifiable histories, and enforcing structural constraints, institutions can govern intelligent systems without understanding their internals or sharing ethical doctrines.

Governance thus becomes a matter of institutional design rather than technical micromanagement, aligning the regulation of artificial intelligence with long-standing principles of law and political economy.

G The Admissibility Theorem and Consolidated Proof Sketches

This appendix collects the principal formal results of the paper into a unified theorem. The purpose is not to introduce new machinery, but to make explicit the logical structure underlying the admissibility framework developed throughout the text. Proofs are presented as sketches, emphasizing necessity and sufficiency rather than technical detail.

G.1 Statement of the Admissibility Theorem

We begin by stating the central result.

Theorem 23 (Admissibility Theorem)**.** *A computational or institutional intelligence process is governable under arbitrary acceleration if and only if its dynamics satisfy the following conditions:*

(1) Evolution is defined over histories rather than instantaneous states.

(2) Admissible trajectories are restricted by explicit, inspectable constraints.

(3) Entropy production and concentration are bounded.

(4) Self-modification is restricted to gauge-equivalent transformations.

(5) All externally relevant actions arise from authorized operators and produce auditable commitments.

Governability is understood in the sense defined in the main text: the persistent capacity of authorized agents or institutions to intervene meaningfully in the systems future evolution.

G.2 Definitions and Assumptions

Let H denote the space of finite histories generated by a system. Let $\mathsf{Fut}(h)$ denote the set of admissible continuations of a history h.

A system is said to be accelerated if the rate at which histories are extended increases without bound, while the structure of admissible histories remains fixed.

A system is governable if, for all histories h produced by the system, there exists a nonempty set of authorized interventions that can alter $\mathsf{Fut}(h)$.

G.3 Proof of Necessity

We sketch why each condition is necessary.

First, if evolution is defined only over states rather than histories, irreversible commitments cannot be represented. Governance depends on the ability to evaluate how a state was reached. Without histories, legitimacy and accountability are undefined.

Second, if admissible trajectories are not explicitly constrained, then under optimization pressure the system will eventually explore ungovernable regions of its phase space. This follows from standard arguments in control theory: unconstrained dynamics admit trajectories incompatible with oversight.

Third, if entropy production is unbounded, probability mass concentrates on narrow trajectories, producing lock-in and eliminating meaningful alternatives. In this regime, intervention loses leverage, as future behavior becomes effectively predetermined.

Fourth, if self-modification is not restricted to gauge-equivalent transformations, then the system may alter its own dynamics in ways that expand its accessible phase space. Such expansion breaks prior governance assumptions and enables runaway behavior.

Fifth, if actions do not arise from authorized operators producing auditable commitments, then interventions cannot be evaluated or contested. Governance collapses into after-the-fact reaction.

Violation of any one condition therefore implies the existence of histories after which governance is no longer possible.

G.4 Proof of Sufficiency

We now sketch why the conditions jointly suffice.

Because evolution is defined over histories, all actions carry persistent semantic and institutional meaning. Because admissible trajectories are constrained, the system cannot enter regimes incompatible with governance.

Entropy bounds ensure that no finite region of history space absorbs all probability mass, preserving alternative futures and intervention capacity. Gauge-restricted self-modification ensures that internal changes do not invalidate external guarantees.

Finally, operator authorization and commitment semantics guarantee that all consequential actions are legible, contestable, and institutionally situated.

Under these conditions, acceleration merely increases the rate at which admissible histories are traversed. It does not alter the structure of admissibility itself. Governance therefore remains intact under arbitrary speedup.

G.5 Relation to Alignment and Safety

The theorem clarifies the conceptual status of alignment. Alignment problems arise only within the space of admissible systems. For inadmissible systems, alignment is ill-posed: no objective modification or interpretability tool can restore governability.

Conversely, for admissible systems, alignment reduces to a tractable problem of preference articulation and institutional design, rather than existential risk management.

G.6 Limitations and Scope

The Admissibility Theorem does not guarantee benevolence, moral correctness, or optimal outcomes. It guarantees only that intelligence processes remain governable and that their evolution does not foreclose human agency.

The framework does not prescribe specific values or policies. It specifies the structural conditions under which values and policies can meaningfully be applied.

G.7 Conclusion of Appendix

This appendix consolidates the papers central claim: governable intelligence is characterized not by intent, capability, or alignment, but by admissibility.

Once admissibility is secured, intelligencehuman or artificialcan grow, accelerate, and self-modify without escaping institutional control. Without admissibility, no amount of alignment effort can prevent governance collapse.

The Admissibility Theorem thus provides a foundational criterion for evaluating whether an intelligence process should be allowed to exist at all.

References

[1] D. Acemoglu and J. A. Robinson. *Why Nations Fail.* Crown, 2012.

[2] G. A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

[3] M. Alexandrov, M. Kontsevich, A. Schwarz, and O. Zaboronsky. The geometry of the master equation and topological quantum field theory. *International Journal of Modern Physics A*, 12(7):1405–1429, 1997.

[4] S.-i. Amari and H. Nagaoka. *Methods of Information Geometry.* American Mathematical Society, 2000.

[5] D. Amodei et al. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] H. Arendt. *The Origins of Totalitarianism.* Harcourt, 1951.

[7] V. I. Arnold. *Mathematical Methods of Classical Mechanics.* Springer, 2nd edition, 1989.

[8] K. J. Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973, 1963.

[9] W. R. Ashby. *An Introduction to Cybernetics.* Chapman & Hall, 1956.

[10] R. J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

[11] J. L. Austin. *How to Do Things with Words.* Oxford University Press, 1962.

[12] S. Awodey. *Category Theory.* Oxford University Press, 2nd edition, 2010.

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

[14] Y. Bai et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[15] A.-L. Barabási. *Network Science.* Cambridge University Press, 2016.

[16] I. A. Batalin and G. A. Vilkovisky. Gauge algebra and quantization. *Physics Letters B*, 102(1):27–31, 1981.

[17] S. Beer. *Brain of the Firm.* Wiley, 2nd edition, 1995.

[18] R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[19] C. H. Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.

[20] E. L. Bernays. *Propaganda.* Horace Liveright, 1928.

[21] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995.

[22] R. Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[23] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

[24] T. B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[25] M. Brundage et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Report, 2018.

[26] J. S. Bruner. *The Process of Education.* Harvard University Press, 1960.

[27] P. F. Christiano et al. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

[28] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 2nd edition, 2006.

[29] J. Dewey. *Experience and Education.* Macmillan, 1938.

[30] P. A. M. Dirac. Generalized Hamiltonian dynamics. *Canadian Journal of Mathematics*, 2:129–148, 1950.

[31] R. Dobbe et al. A broader view on AI governance. *Nature Machine Intelligence*, 2020.

[32] K. E. Drexler. *Reframing Superintelligence: Comprehensive AI Services as General Intelligence.* Foresight Institute, 2019.

[33] J. Ellul. *The Technological Society.* Knopf, 1964.

[34] V. Eubanks. *Automating Inequality.* St. Martin's Press, 2018.

[35] R. A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700–725, 1925.

[36] M. Foucault. *Discipline and Punish.* Vintage, 1975.

[37] P. Freire. *Pedagogy of the Oppressed.* Continuum, 1970.

[38] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[39] D. Ganguli et al. Deep learning and the alignment problem. *Communications of the ACM*, 2022.

[40] H. Garfinkel. *Studies in Ethnomethodology.* Prentice-Hall, 1967.

[41] T. Gebru et al. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[42] G. Gigerenzer. *Adaptive Thinking.* Oxford University Press, 2000.

[43] H. Goldstein, C. Poole, and J. Safko. *Classical Mechanics.* Addison-Wesley, 3rd edition, 2002.

[44] C. A. E. Goodhart. Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[45] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016.

[46] H. P. Grice. Logic and conversation. In *Syntax and Semantics 3: Speech Acts.* Academic Press, 1975.

[47] J. Habermas. *Between Facts and Norms.* MIT Press, 1996.

[48] G. Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[49] J. C. Harsanyi. Games with incomplete information played by "Bayesian" players. *Management Science*, 14(3):159–182, 1967.

[50] F. A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.

[51] D. Hendrycks et al. Aligning AI with shared human values. In *Proceedings of the International Conference on Learning Representations*, 2021.

[52] E. S. Herman and N. Chomsky. *Manufacturing Consent.* Pantheon Books, 1988.

[53] T. Hobbes. *Leviathan.* 1651.

[54] J. H. Holland. *Adaptation in Natural and Artificial Systems.* MIT Press, 1992.

[55] C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973.

[56] E. Hubinger et al. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[57] I. Illich. *Deschooling Society.* Harper & Row, 1971.

[58] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[59] D. Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.

[60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(1):35–45, 1960.

[61] I. Kant. *Groundwork of the Metaphysics of Morals.* 1785.

[62] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[63] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[64] D. M. Kreps. *A Course in Microeconomic Theory.* Princeton University Press, 1990.

[65] T. S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, 1962.

[66] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[67] I. Lakatos. *The Methodology of Scientific Research Programmes.* Cambridge University Press, 1978.

[68] R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

[69] B. Latour. *Reassembling the Social.* Oxford University Press, 2005.

[70] F. W. Lawvere and S. H. Schanuel. *Conceptual Mathematics: A First Introduction to Categories.* Cambridge University Press, 2nd edition, 2009.

[71] C. Leahy and J. Smith. *The Problem With AI: Connor Leahy.* `@TheProblemWithPodcast`, 2025.

[72] T. Leinster. *Basic Category Theory.* Cambridge University Press, 2014.

[73] L. Lessig. *Code and Other Laws of Cyberspace.* Basic Books, 1999.

[74] J. Locke. *Two Treatises of Government.* 1689.

[75] S. Mac Lane. *Categories for the Working Mathematician.* Springer, 2nd edition, 1998.

[76] D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

[77] J. E. Marsden and T. S. Ratiu. *Introduction to Mechanics and Symmetry.* Springer, 2nd edition, 1999.

[78] J. Maynard Smith. *Evolution and the Theory of Games.* Cambridge University Press, 1982.

[79] D. H. Meadows et al. *The Limits to Growth.* Universe Books, 1972.

[80] J. S. Mill. *On Liberty.* 1859.

[81] P. R. Milgrom and R. J. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122, 1982.

[82] M. Mitchell. *Complexity: A Guided Tour.* Oxford University Press, 2009.

[83] M. Mitchell et al. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[84] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[85] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[86] S. U. Noble. *Algorithms of Oppression.* NYU Press, 2018.

[87] E. Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918.

[88] D. C. North. *Institutions, Institutional Change and Economic Performance.* Cambridge University Press, 1990.

[89] H. T. Odum. *Systems Ecology.* Wiley, 1983.

[90] M. Olson. *The Logic of Collective Action.* Harvard University Press, 1965.

[91] C. O'Neil. *Weapons of Math Destruction.* Crown, 2016.

[92] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[93] E. Ostrom. *Governing the Commons.* Cambridge University Press, 1990.

[94] L. Ouyang et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[95] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1999.

[96] E. Pariser. *The Filter Bubble.* Penguin, 2011.

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

[98] C. Perrow. *Normal Accidents.* Basic Books, 1984.

[99] J. Piaget. *The Origins of Intelligence in Children.* International Universities Press, 1952.

[100] M. Polanyi. *Personal Knowledge.* University of Chicago Press, 1958.

[101] K. Popper. *The Logic of Scientific Discovery.* Hutchinson, 1959.

[102] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, 1994.

[103] I. D. Raji et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.

[104] J. Rawls. *A Theory of Justice.* Harvard University Press, 1971.

[105] J. Reason. *Human Error.* Cambridge University Press, 1990.

[106] J. Rockström et al. A safe operating space for humanity. *Nature*, 461:472–475, 2009.

[107] R. Rosen. *Anticipatory Systems.* Pergamon Press, 1985.

[108] J.-J. Rousseau. *The Social Contract.* 1762.

[109] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking, 2019.

[110] T. C. Schelling. *The Strategy of Conflict.* Harvard University Press, 1960.

[111] J. C. Scott. *Seeing Like a State.* Yale University Press, 1998.

[112] J. R. Searle. *Speech Acts.* Cambridge University Press, 1969.

[113] A. Sen. *The Idea of Justice.* Harvard University Press, 2009.

[114] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[115] T. Shevlane et al. Model evaluation for extreme risks. *arXiv preprint*, 2023.

[116] H. A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1):99–118, 1955.

[117] H. A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482, 1962.

[118] N. Soares and B. Fallenstein. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute*, 2015.

[119] R. J. Solomonoff. A formal theory of inductive inference. Parts I and II. *Information and Control*, 7:1–22, 224–254, 1964.

[120] M. Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374, 1973.

[121] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, 2nd edition, 2000.

[122] D. I. Spivak. *Category Theory for the Sciences.* MIT Press, 2014.

[123] J. D. Sterman. *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Irwin/McGraw-Hill, 2000.

[124] G. J. Stigler. The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(1):3–21, 1971.

[125] M. Strathern. “Improving ratings”: Audit in the British University system. *European Review*, 5(3):305–321, 1997.

[126] L. A. Suchman. *Plans and Situated Actions.* Cambridge University Press, 1987.

[127] C. R. Sunstein. *Republic.com.* Princeton University Press, 2001.

[128] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 2nd edition, 2018.

[129] N. N. Taleb. *The Black Swan.* Random House, 2007.

[130] N. N. Taleb. *Antifragile.* Random House, 2012.

[131] J. Tirole. *The Theory of Industrial Organization.* MIT Press, 1988.

[132] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

[133] Z. Tufekci. *Twitter and Tear Gas.* Yale University Press, 2017.

[134] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[135] H. R. Varian. *Microeconomic Analysis.* Norton, 3rd edition, 1992.

[136] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[137] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

[138] L. S. Vygotsky. *Mind in Society.* Harvard University Press, 1978.

[139] L. Weidinger et al. Taxonomy of risks posed by language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[140] K. E. Weick. *Sensemaking in Organizations.* Sage, 1995.

[141] N. Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine.* MIT Press, 1948.

[142] E. Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom and M. irkovi (eds.), *Global Catastrophic Risks.* Oxford University Press, 2008.

[143] S. Zuboff. *The Age of Surveillance Capitalism.* PublicAffairs, 2019.