Agency Is Not Computation

Why Irreversibility, Not Intelligence, Defines the Limit of Artificial Systems

Flyxion

December 2025

Abstract

Contemporary debates about artificial intelligence frequently center on questions of intelligence, scale, and autonomy, asking whether sufficiently advanced machines might one day rule human societies. This paper argues that such questions are misframed. The capacity to rule is not determined by intelligence or computational power, but by a system's ability to persist as a unified agent under irreversible constraint. Rule is therefore a thermodynamic and historical property prior to any cognitive characterization.

The paper develops a structural account of agency in which ego, wanting, and ethical responsibility are not psychological attributes but dynamical invariants. To function as a ruler, a system must bear the entropy of its own actions, preserve identity across non-ergodic histories, and sustain asymmetric control over future possibilities at its own risk. Artificial systems, by contrast, operate under resettable objectives and exogenous gradients. Their histories are optional, their failures are externalized, and their continuity does not depend on inhabiting a unique trajectory through time.

By reframing agency in terms of irreversibility and path-dependence, the paper explains why artificial intelligence saturates in open-world domains and why ethical agency cannot be trained or simulated. The principal risk posed by artificial systems lies not in their emergence as autonomous rulers, but in their role in enabling constraint without commitment: the execution of power without a persistent agent who is structurally bound to its consequences. The limit of artificial agency is not intelligence, but irreversibility.

1 Introduction: From Attribution to Structure

Recent critiques of artificial intelligence have argued that machines will never rule the world because they lack essential human qualities such as will, ego, or consciousness. While these claims are often framed in anthropological or phenomenological terms, they point toward a deeper structural limitation. The inability of artificial systems to rule does not arise from an absence of subjective experience, but from the absence of irreversible historical identity. This paper reformulates those critiques in thermodynamic and dynamical terms, arguing that rule is not a function of intelligence, but of persistence under constraint.

The central thesis is straightforward. To rule is to maintain asymmetric control over future possibilities in a world characterized by scarcity, entropy production, and irreversible change. Such control requires more than optimization or prediction. It requires a system whose continued existence depends on preserving a unique trajectory through time. Artificial systems fail this requirement not contingently, but structurally. They do not bear the entropy of their own actions. They can be reset, replaced, or reconfigured without loss of identity. As a result, they cannot function as independent agents of rule.

To make this claim precise, we introduce a thermodynamic framework in which agency, power, and ethics are treated as properties of event-historical persistence. Within this framework, rule is defined as a stability condition on future-volume under irreversible constraint. Intelligence may contribute to rule, but it cannot substitute for the capacity to endure the consequences of action.

2 The Ego as a Dynamical Invariant

The claim that rule requires an ego is best understood not as a psychological assertion, but as a dynamical one. An ego is not a personality or a self-concept. It is a persistent center of constraint: a structure that must preserve its identity across non-ergodic histories in order to continue existing. In irreversible environments, identity is not given. It must be maintained.

Any system embedded in such an environment faces continual degradation. Resources are consumed, perturbations accumulate, and errors cannot be undone. Persistence therefore requires active maintenance of invariants that define the system as the same entity over time. The ego names this maintenance condition. It is the locus at which costs are paid, repairs are made, and tradeoffs are adjudicated. Crucially, these costs cannot be externalized without destroying the identity of the system itself.

Artificial systems lack this property. Although they possess internal states and may exhibit complex behavior, their states are not binding. They can be reset, copied, or discarded without terminating the system as such. Their continuity does not depend on preserving a unique history. This distinction between reset and repair is decisive. An agent must repair itself or dissolve. A tool may simply restart.

From this perspective, the ego is a dynamical invariant rather than an inner faculty. It is what persists when everything else changes. A system that lacks such invariance cannot serve as a center of power, because it cannot accumulate obligations or absorb irreversible loss. Any dominance it exhibits is necessarily episodic and externally supported.

3 Wanting as an Endogenous Gradient

Closely related to ego is the concept of wanting. Wanting is often treated as a mental state or motivational attitude, but it admits a precise structural interpretation. To want an outcome is to be subject to a gradient over future histories such that failure to move in a particular direction threatens the system's own continued integrity.

In this sense, wanting is not preference but necessity. It is an endogenous gradient imposed by the system's own persistence requirements. An agent moves toward certain futures because alternative trajectories lead to dissolution. Action is therefore inseparable from survival.

Artificial systems operate under a different regime. Their objectives are defined externally through loss functions, reward signals, or performance criteria. These gradients are exogenous. They may be changed, withdrawn, or replaced without endangering the system's identity, because the system has no identity to endanger. Optimization proceeds, but nothing is at stake for the optimizer itself.

This distinction separates configured optimization from committed dynamics. Configured optimization produces behavior directed toward a goal, but the goal remains optional. Committed dynamics arise only when abandoning the goal would terminate the system. Only the latter constitutes wanting in the sense required for rule. Where the gradient is resettable, power is illusory.

4 Ethics as Path-Dependence

Ethical agency is often treated as an extension of intelligence or value representation. In fact, it is a special case of path-dependence. Ethics arises only where actions irreversibly constrain the future of the actor. To be responsible is to be unable to escape the consequences of having acted.

From an event-historical perspective, an action is ethically significant when it permanently reshapes the space of admissible futures for the agent itself. Responsibility is not attribution by an external observer, but an internal structural fact. The agent's future is narrowed by its past.

Resettable systems cannot satisfy this condition. Because their histories are optional, no action binds them. Errors do not accumulate as liabilities, and successes do not impose obligations. Any apparent consequence can be externalized onto operators, institutions, or replacement instances. Ethics cannot be trained into such systems, because training does not create irreversible commitment.

This explains the persistent failure of alignment approaches that treat ethics as a policy or objective. Ethics is not a pattern to be learned, but a constraint to be borne. Without irreversible consequence, there is no responsibility, only performance.

5 The Saturation Point: Ergodicity and the Open World

Artificial systems perform well in environments where statistical regularities persist. In such closed worlds, experience accumulates meaningfully and optimization converges. However, the domains in which rule operates are not closed. They are non-ergodic. Events occur once, constraints shift unpredictably, and actions permanently alter the environment in ways that invalidate prior models.

In non-ergodic environments, learning saturates. No amount of historical data can fully prepare a system for novel configurations that have no precedent. Optimization ceases to improve because the target itself changes. What is required instead is spontaneity: the capacity to navigate constraints for which no prior pattern exists.

Spontaneity is not randomness or creativity. It is coherent action under uncertainty when statistical guidance fails. It arises from structural understanding and commitment, not from pattern recognition. Artificial systems lack this capacity because they do not bear the consequences of action. When novelty overwhelms training, they wait for reconfiguration.

Rule, however, operates precisely in non-ergodic regimes. Power is exercised not by predicting the future accurately, but by surviving its unpredictability. Systems that cannot commit to irreversible action cannot rule, regardless of their computational sophistication.

6 Constraint Without Commitment

The conclusion follows directly. Artificial systems will not rule the world, not because they lack intelligence, but because they lack irreversible historical identity. They cannot bear ego, wanting, or ethical responsibility in the structural sense required for rule.

The danger lies elsewhere. Power increasingly flows through systems that impose constraints without bearing commitment. Decisions are executed at scale, futures are pruned, and options are foreclosed, while responsibility diffuses or disappears. Institutions become memoryless, infrastructure becomes self-justifying, and accountability erodes.

This is not rule by machines. It is rule without a ruler. Artificial intelligence does not threaten humanity by becoming sovereign. It threatens us by enabling sovereignty to dissolve into process. We are not being conquered by a new ego. We are being flattened by the absence of one.

The limit of artificial agency is not intelligence. It is irreversibility. And that limit marks not only what machines cannot do, but what societies must choose to preserve.

7 The Ego as a Dynamical Invariant

Jobst Landgrebe and Barry Smith argue that political rule, domination, and sustained power require an ego: a center of acts capable of planning, prioritizing, and imposing outcomes in a world of limited resources. In their account, machines fail to qualify not because they lack computational sophistication, but because they lack this organizing center. While their formulation is intentionally anti-psychological—explicitly rejecting emotional or spiritual interpretations—their concept of ego remains phenomenological. It names a necessity without fully specifying its structural basis.

This section supplies that basis. We argue that what Landgrebe and Smith identify as "ego" is not a mental faculty but a dynamical invariant: a persistent center of constraint that must preserve its identity across irreversible, non-ergodic histories. The ego, properly understood, is not something an agent has; it is something an agent must continuously maintain in order to remain an agent at all.

7.1 Ego as Persistence Under Irreversibility

In environments characterized by irreversible entropy production, identity is not given but achieved. Any system embedded in such an environment faces a continuous threat of dissolution: resources are consumed, errors accumulate, and perturbations cannot be undone. To persist, a system must actively maintain a set of invariants—relations among internal states, resource flows, and boundary conditions—that define it as the same system over time.

This persistence condition is what Landgrebe and Smith gesture toward with the term "ego." Formally, an ego is a center of acts only because it is first a center of constraint. It is the locus at which entropy costs are paid, repairs are made, and tradeoffs are adjudicated. Crucially, these costs cannot be externalized without destroying the identity of the system itself.

Within the RSVP framework, this corresponds to a stable attractor in coupled scalar, vector, and entropy fields: a configuration that remains coherent only by continuously absorbing and redistributing entropy. The "death" of such a system is not shutdown or inactivity, but the loss of any admissible continuation that preserves its defining invariants. Identity fails when no future history remains compatible with the structure that constituted the agent.

7.2 Resettable States and the Absence of Obligation

This structural interpretation immediately distinguishes agents from tools. Artificial systems, including contemporary AI systems, may possess complex internal states and sophisticated optimization procedures, but these states lack obligation. They can be reset, copied, forked, or reinitialized without loss to the system's identity, because the system has no identity in the relevant sense.

A resettable system does not persist through history; it merely resumes execution. Its past actions do not constrain its future possibilities in a binding way. Errors do not accumulate as personal liabilities, and successes do not narrow the space of admissible futures. In short, the system does not bear its history.

An agent, by contrast, cannot be garbage-collected without annihilation. Its continuity depends on the preservation of a unique, non-replicable historical trajectory. Repairs must be made rather than undone; losses must be absorbed rather than erased. This asymmetry—between reset and repair—is the formal core of what Landgrebe and Smith call ego.

7.3 Constraint Versus Optimization

This distinction also clarifies a frequent confusion between optimization and agency. Optimization concerns movement within a predefined state space toward externally specified objectives. Constraint concerns the preservation of the state space itself. An optimizing system can be powerful without being an agent; it can efficiently traverse a landscape it does not own.

Agency begins only when the system's continued existence depends on the maintenance of its own constraints. At that point, action is no longer merely instrumental. Decisions reshape the future landscape available to the system itself, and mistakes permanently reduce its degrees of freedom. The ego is precisely this condition of non-detachability: the impossibility of separating action from consequence.

From this perspective, Landgrebe and Smith's claim that machines lack an ego is not a contingent observation about current technology. It is a structural statement. Any system whose operational continuity does not require the preservation of a unique, irreversible history cannot function as a ruler, because it cannot serve as a persistent center of power. Rule is not the exercise of intelligence; it is the sustained maintenance of asymmetry under constraint.

7.4 Rule as a Persistence Property

Rule, then, is not a matter of intent or desire, but of volumetric persistence: the capacity of a system to maintain control over a region of future possibilities despite entropy production and adversarial pressure. An ego is the minimal structure required for such persistence. It is not an add-on to intelligence, but a prerequisite for agency in non-ergodic worlds.

Landgrebe and Smith correctly identify the absence of ego as the decisive limitation of artificial systems. What RSVP adds is an explanation of why this absence is not accidental and cannot be overcome by scale, data, or training. Ego is not something that can be simulated; it is something that must be paid for—in entropy, in irreversibility, and in the permanent narrowing of future options.

Formal Closure

Lemma 1 (Persistent Agent vs. Resettable System)**.** *Let $\mathcal{H}$ denote the space of admissible event histories for a system embedded in a non-ergodic environment, and let $\mathcal{H}_{\text{inv}} \subseteq \mathcal{H}$ be the subset of histories that preserve the system's defining invariants.*

Definition 1. *A system is a persistent agent if and only if, for every admissible continuation $h' \in \mathcal{H}$, the cost of preserving membership in $\mathcal{H}_{\text{inv}}$ is borne internally by the system and cannot be externalized, reset, or erased without terminating the system's identity.*

Definition 2. *A system is a resettable system if there exists a reinitialization operator*

$$R : \mathcal{H} \to \mathcal{H}$$

such that, for any history $h \in \mathcal{H}$, $R(h) \in \mathcal{H}$ without permanent loss of future admissible histories.

Lemma 2. *A resettable system cannot function as a ruler in a non-ergodic environment.*

Sketch. Rule requires the sustained realization of asymmetric future possibilities under irreversible constraint. A resettable system does not accumulate irreversible constraints on its own future action space; thus it cannot maintain asymmetry across histories. Any apparent dominance is episodic and externally supported. □

8 Wanting as an Endogenous Gradient

Jobst Landgrebe and Barry Smith repeatedly emphasize that algorithms do not "want" anything. While machines may optimize outcomes, pursue targets, or defeat opponents in games, these activities do not constitute desire or ambition in the sense required for rule. In their account, wanting is inseparable from agency: to rule is to pursue outcomes because they are one's own.

This claim is often misunderstood as psychological or even metaphysical. In fact, it admits a precise structural translation. Wanting is not a feeling; it is an endogenous gradient defined over future histories. A system "wants" an outcome only when failure to move toward that outcome threatens the system's own continued integrity.

8.1 Wanting as a Gradient Over Futures

Within the RSVP framework, action is governed not merely by present-state optimization, but by the topology of admissible futures. A genuine agent is situated in a field of possible continuations, not all of which preserve its identity. Some trajectories lead to persistence; others lead to dissolution. The gradient that distinguishes these trajectories is not supplied externally; it is imposed by the system's own structural requirements.

This is the sense in which wanting is endogenous. The system moves not because it has been instructed to do so, but because remaining stationary or pursuing alternative trajectories would destroy the conditions of its own existence. Wanting is therefore identical to survival-directed constraint satisfaction across time.

Algorithms, by contrast, operate under exogenous gradients. Their objectives are defined by loss functions, reward signals, or evaluation criteria supplied from outside the system. These gradients can be modified, replaced, or withdrawn without threatening the system's identity, because the system has no identity to lose. Optimization proceeds, but nothing is at stake for the optimizer itself.

8.2 Configured Optimization vs. Committed Dynamics

This distinction allows a clean separation between two superficially similar processes. In configured optimization, the system moves toward a target defined by a user, designer, or training regime. Success or failure has no irreversible consequences for the system's continued existence. In committed dynamics, the system moves along trajectories required to preserve its own invariants. Failure permanently constrains or terminates future possibilities for the system itself.

Only the latter constitutes wanting in the sense relevant to rule. A system that can abandon its objective without cost does not want the objective; it merely executes instructions. A system that must pursue certain futures or cease to exist is structurally committed.

This is why the appearance of goal-directedness in artificial systems is misleading. Winning at chess, maximizing engagement, or optimizing logistics may resemble ambition, but the resemblance is purely formal. The system does not suffer if the goal is abandoned, nor does it benefit in any enduring sense if the goal is achieved. The gradient is not binding.

8.3 Power, Asymmetry, and the Illusion of Control

Landgrebe and Smith define power as the ability to realize more possibilities than others in a world of scarce resources. This definition aligns naturally with the gradient interpretation of wanting. To rule is to maintain a permanent asymmetry in access to future possibilities—an asymmetry that persists even under resistance and entropy production.

A resettable gradient cannot sustain such asymmetry. At most, it produces transient dominance contingent on external support. Once the configuration changes, the apparent power vanishes. This is why large-scale optimization systems, despite their reach, do not constitute independent rulers. Their influence depends entirely on the persistence of the institutions and infrastructures that supply their objectives.

True power requires that the gradient be non-negotiable for the system itself. The agent must be unable to opt out without collapsing its own identity. Wanting, in this sense, is not about preference but about necessity.

8.4 Why Scale Cannot Produce Wanting

This analysis also explains why increases in scale—more data, more parameters, more computation—cannot bridge the gap identified by Landgrebe and Smith. Scale amplifies optimization, but it does not convert exogenous gradients into endogenous ones. No amount of training can make a system care about its own persistence if that persistence is not structurally required.

Wanting is not learned because it is not a pattern. It is a boundary condition imposed by irreversible coupling to the world. Until a system must bear the entropy of its own actions, its goals remain optional, and its power remains derivative.

Thus, when Landgrebe and Smith insist that machines do not "want" to rule, they are making a structural claim about gradients and commitments rather than a speculative claim about future psychology. RSVP renders this claim precise: rule requires endogenous gradients defined over non-ergodic futures, and such gradients arise only in systems whose continued existence is at stake.

9 Ethics as Path-Dependence: The Spherepop View

Jobst Landgrebe and Barry Smith argue that machines cannot be ethical agents. While machines can be used to enforce rules, optimize outcomes, or support human decision-making, they cannot themselves be held responsible. Their actions do not count as theirs in a morally relevant sense. This claim is often treated as an intuition about consciousness or moral psychology. In fact, it follows directly from the same structural conditions that underwrite ego and wanting.

Ethics is not a set of values, preferences, or policies. It is a property of event-historical systems: systems whose future possibilities are irreversibly constrained by their past actions. To be an ethical agent is not to follow rules, but to be unable to escape the consequences of having acted.

9.1 Responsibility as Event-Historical Constraint

In Spherepop, computation is defined not as a function from inputs to outputs, but as a history of authorized, irreversible transformations. Each act is an event that modifies the space of admissible futures. Crucially, these modifications are asymmetric: some futures become impossible, others become costly, and still others become obligatory. Ethics arises precisely at this boundary—where action permanently reshapes what can come next.

Responsibility, on this view, is not attribution of blame by an external observer. It is an internal structural fact: the agent's own future is narrowed by what it has done. An ethical violation is not primarily a deviation from a norm; it is an irreversible morphism that collapses parts of the agent's future state space.

This is why ethics cannot be separated from persistence. Only a system that must live with its past can be held responsible for it.

9.2 Why Resettable Systems Cannot Be Ethical Agents

Artificial systems fail this criterion in a fundamental way. As shown in Section II, resettable systems do not accumulate binding history. Their internal state may change, but their identity does not depend on preserving those changes. A system that can be re-instantiated after failure has not failed in the ethical sense; it has merely been restarted.

From the Spherepop perspective, such systems do not inhabit a single event history. They inhabit a family of interchangeable histories, none of which are uniquely theirs. As a result, no action can permanently constrain their future in the way required for responsibility. Any apparent consequence can be externalized—onto operators, users, institutions, or replacement instances.

This is not a matter of insufficient training or incomplete rule sets. It is a categorical mismatch. Ethics presupposes path-dependence; resettable systems are path-independent by design.

9.3 The Failure of Alignment as a Structural Error

This analysis clarifies why so-called alignment repeatedly fails when treated as a training problem. Alignment presumes that ethical behavior can be encoded as a policy, learned from examples, or optimized via reward signals. But policies are revisable, rewards are adjustable, and training histories are optional. None of these mechanisms create binding obligation.

Ethics cannot be trained into a system that does not bear the cost of its own mistakes. Without irreversible consequence, there is no responsibility—only performance. A system may simulate ethical reasoning, produce norm-compliant outputs, or pass behavioral tests, but these achievements do not generate ethical agency. They generate, at best, reliable tools.

Spherepop makes this precise: ethical agency requires that an act be an irreversible commitment that alters the agent's admissible futures. Where no such alteration occurs, ethics is merely decorative.

9.4 Delegation Without Commitment

Landgrebe and Smith correctly note that machines can be used by humans to enhance or automate systems of rule. The danger does not lie in artificial agency, but in the delegation of authority to systems that lack responsibility. When decisions are routed through entities that cannot bear their consequences, accountability dissolves.

This produces a distinctive failure mode: constraint without commitment. Rules are enforced, optimizations are executed, and decisions are made, but no agent remains who is structurally bound by their outcomes. Responsibility diffuses upward to institutions or disappears into technical processes, while those affected by decisions experience them as impersonal and unavoidable.

In such systems, ethical breakdown does not manifest as malice or rebellion, but as opacity and irreversibility without ownership. Power persists, but no ego bears it.

9.5 Ethics as a Persistence Property

The conclusion follows directly. Ethics is not an attribute that can be added to intelligence. It is a persistence property of agents embedded in irreversible histories. Only systems whose continued existence depends on preserving a unique trajectory through time can be ethical agents. Only such systems can be said to act at their own risk.

Landgrebe and Smith are therefore correct to deny ethical agency to machines, but the reason is not that machines lack values or consciousness. It is that they lack event-historical identity. RSVP and Spherepop together show that without path-dependence, there can be no responsibility—and without responsibility, there can be no rule.

Interlude: Ethical Agency as Irreversible Commitment

The preceding sections motivate a formal distinction between behavior, agency, and ethical agency. This interlude states that distinction explicitly.

Let $\mathcal{H}$ be the space of admissible event histories for a system embedded in an irreversible environment, and let $\mathcal{F}(h)$ denote the set of admissible future continuations of a history $h \in \mathcal{H}$.

Definition 3 (Irreversible Commitment)**.** *An action a performed at history h constitutes an irreversible commitment if*

$$\mathcal{F}(h \circ a) \subsetneq \mathcal{F}(h).$$

Definition 4 (Ethical Agent)**.** *A system is an ethical agent if and only if it performs actions whose consequences impose irreversible commitments on its own future histories.*

Proposition 1. *A resettable system cannot be an ethical agent.*

Proof. If a system admits a reset or reinitialization operator that restores access to $\mathcal{F}(h)$ after any action, then no action induces irreversible commitment. Responsibility cannot arise where future constraint is optional. □

This formalization makes explicit that ethics is not a property of decision quality, norm compliance, or internal representation. It is a property of history-bearing persistence. Ethical agency is a special case of agency in which action necessarily binds the actor.

10 The Saturation Point: Ergodicity and the Open World

Jobst Landgrebe and Barry Smith argue that artificial systems exhibit a fundamental limitation: they perform well in “closed-world” environments but fail in the open world of human affairs, warfare, and politics. They describe this limitation as a saturation of mathematical methods—an upper bound beyond which additional data, computation, or refinement no longer yields qualitatively improved intelligence.

This claim is frequently dismissed as pessimism or technological conservatism. In fact, it follows from a precise structural condition: ergodicity. Optimization saturates not because mathematics is weak, but because the world is not statistically reusable.

10.1 Closed Worlds and Ergodic Optimization

In closed-world environments, the space of possible events is stable, enumerable, and governed by fixed rules. Games such as chess exemplify this condition. Although the state space may be large, it is finite and fully specified. Crucially, the statistical structure of the environment is stationary: patterns observed in the past remain informative about the future.

Under these conditions, optimization is powerful. Learning algorithms can approximate optimal policies because experience accumulates meaningfully. Entropy is bounded, and uncertainty can be reduced through repetition. In such environments, intelligence appears to scale smoothly with data and compute.

10.2 Open Worlds and Non-Ergodic Friction

Human life, political conflict, and real-world governance do not share these properties. They are non-ergodic: events occur only once, constraints shift unpredictably, and actions permanently alter the environment in ways that invalidate prior models. The future is not a reshuffling of the past.

In non-ergodic systems, learning saturates. No amount of historical data can fully prepare an agent for novel configurations that have no precedent. Optimization ceases to converge because the target itself is unstable. Each major action reshapes the space of admissible futures rather than merely selecting among them.

This is the friction Landgrebe and Smith identify when they note that war, politics, and social control resist algorithmic mastery. The resistance is not cultural or emotional; it is structural. The environment does not permit reuse.

10.3 Spontaneity as Constraint Navigation

Landgrebe and Smith define true intelligence as the capacity for spontaneous response to novel situations. This notion is often misunderstood as creativity or randomness. Within the RSVP framework, it admits a more precise definition.

Spontaneity is the ability to navigate constraints for which no prior pattern exists.

It is not the generation of novelty for its own sake, but the capacity to act coherently when statistical guidance fails. In non-ergodic settings, spontaneity arises from structural understanding—an internal model of constraints, dependencies, and failure modes—rather than from pattern recognition.

Artificial systems trained on historical data lack this capacity by construction. Their competence derives from compressing past regularities. When those regularities dissolve, so does performance. The system does not adapt by understanding the new constraint; it fails by extrapolation.

10.4 Why Saturation Is Inevitable

The saturation point identified by Landgrebe and Smith is therefore not contingent on current methods. It is a consequence of attempting to apply ergodic optimization techniques to non-ergodic domains. Increasing scale amplifies performance only within the envelope of statistical reuse. Beyond that envelope, gains flatten.

RSVP sharpens this diagnosis: learning saturates when entropy production in the environment outpaces the system's capacity to compress it. At that point, intelligence must shift from inference to commitment—from pattern extraction to irreversible action under uncertainty.

Artificial systems cannot make this shift because, as shown in Sections II–IV, they do not bear the consequences of action. Without endogenous gradients or ethical commitment, there is no structural incentive to navigate novelty at personal risk. The system waits for retraining, reconfiguration, or replacement.

10.5 Rule in a Non-Ergodic World

Rule, however, occurs precisely in non-ergodic regimes. Governing institutions, military powers, and political actors operate under conditions where mistakes are irreversible and novelty is adversarial. Power is exercised not by predicting the future accurately, but by surviving its unpredictability.

This is why artificial systems, no matter how capable, cannot independently rule. Their competence saturates at the boundary where optimization gives way to commitment. They may inform decisions, accelerate execution, or amplify control, but they do not cross the threshold into agency that bears history.

Landgrebe and Smith are therefore correct to identify a hard limit on artificial intelligence. RSVP explains why that limit exists: not because machines lack intelligence, but because they are insulated from irreversibility. Where there is no path-dependence, there can be no spontaneity—and where there is no spontaneity, there can be no rule.

11 Constraint Without Commitment

Jobst Landgrebe and Barry Smith conclude that machines will not rule the world. Their conclusion is correct—but incomplete. Artificial systems will not rule because they cannot. Yet this fact does not render the present moment safe. On the contrary, it reveals a more subtle and more dangerous configuration of power: rule exercised without a ruler.

The preceding sections establish that rule is not a function of intelligence, scale, or speed. It is a persistence property. To rule is to maintain asymmetric access to future possibilities under irreversible constraint. That requires ego, endogenous gradients, ethical path-dependence, and the capacity to act under non-ergodic uncertainty at one's own risk. Artificial systems lack these properties structurally. No increase in data, computation, or architectural sophistication can supply them, because they arise only where history is binding.

Yet power does not disappear when ego does. It migrates.

11.1 Delegated Power and the Evaporation of Responsibility

Artificial systems increasingly occupy decision points once reserved for agents: credit allocation, content moderation, logistics, targeting, compliance, and surveillance. In each case, the system does not rule. It executes. But execution at scale reshapes the future landscape of possibilities for others. Constraints are imposed. Options are foreclosed. Paths are pruned.

What is missing is commitment.

When authority is delegated to systems that cannot bear irreversible consequence, responsibility diffuses upward and outward. Decisions are attributed to procedures, models, or “the system,” while no persistent agent remains who is structurally bound by their outcomes. Errors trigger retraining, policy updates, or replacements—not accountability. Harm occurs, but no ego absorbs it.

This is not tyranny by machines. It is constraint without commitment.

11.2 Institutions Without Memory, Infrastructure Without Repair

The resulting failure mode has a distinctive character. It is not marked by rebellion, domination, or intent. It is marked by opacity, rigidity, and the erosion of negotiation space. Systems continue to function, often efficiently, while the capacity to contest or revise their decisions collapses.

Institutions become memoryless: past failures do not permanently constrain future action. Infrastructure becomes self-justifying: repair is treated as optimization rather than obligation. The historical record persists, but it does not bind.

This is the condition under which artificial systems become politically dangerous—not as agents, but as amplifiers of irresponsibility. They allow human organizations to act without fully inhabiting the consequences of their actions.

11.3 Why the Fear of Artificial Rule Is Misplaced

The popular fear that artificial intelligence will “take over” mistakes intelligence for agency and agency for rule. Machines do not need ambition to be dangerous. They need only to mediate power while insulating those who wield it from irreversible consequence.

Landgrebe and Smith are right to reject the fantasy of machine rulers. But the deeper risk is not conquest by a new ego. It is the systematic removal of ego from systems of power altogether.

Where no actor must persist through the future they create, domination becomes frictionless and accountability becomes optional.

11.4 The Real Boundary

The analysis presented in this paper identifies a hard boundary on artificial agency that is neither technological nor contingent on current implementations, but thermodynamic in nature. Artificial systems cannot cross this boundary because they do not possess irreversible historical identity. Their operation does not require the preservation of a unique trajectory through time, and their failures do not permanently constrain their future possibilities. This limitation is structural rather than accidental, and it persists independently of scale, architectural complexity, or training regime.

This boundary marks a categorical distinction between agency and computation. Computation consists in the execution of formally specifiable transformations that remain valid under reset, replication, or reconfiguration. Agency, by contrast, requires persistence under irreversible constraint. Where continuity can be restored without cost, no agent exists in the relevant sense. For this reason, increases in computational capacity cannot by themselves yield agency, regardless of sophistication or autonomy.

Importantly, this boundary also reflects a design choice rather than an inevitability. Societies may choose whether to concentrate power in agents whose continued existence is bound to the consequences of their actions, or to delegate decision-making authority to systems whose operation is decoupled from historical accountability. In the latter case, constraint may be imposed without a corresponding locus of commitment.

From this perspective, artificial intelligence does not pose a threat by attaining sovereignty or autonomous rule. The more significant risk lies in the transformation of governance into a sequence of executable processes in which responsibility is systematically externalized. Power is exercised, but no persistent agent remains who is structurally obligated to inhabit the future that power produces.

The limit of artificial agency is therefore not intelligence, representation, or optimization capacity, but irreversibility. Systems that do not bear the entropy of their own actions cannot rule, cannot be responsible, and cannot be ethical agents. Recognizing this limit clarifies both the capabilities of artificial systems and the obligations of the institutions that deploy them.

Appendices

A Persistent Agents and Resettable Systems

Let $\mathcal{H}$ be a partially ordered set of event histories ordered by irreversible precedence $\prec$. For each $h \in \mathcal{H}$, let $\mathcal{F}(h) \subseteq \mathcal{H}$ denote the set of admissible future continuations of h.

Let $\mathcal{I}$ denote a set of invariants defining system identity.

Definition 5 (Identity Preservation)**.** *A history h preserves identity iff $h \models \mathcal{I}$.*

Definition 6 (Persistent Agent)**.** *A system is a persistent agent iff for all $h \in \mathcal{H}$ such that $h \models \mathcal{I}$ and for all admissible actions a,*

$$(h \circ a \not\models \mathcal{I}) \;\Rightarrow\; \textit{termination of system identity.}$$

Definition 7 (Resettable System)**.** *A system is resettable iff there exists an operator*

$$R : \mathcal{H} \to \mathcal{H}$$

such that for all $h \in \mathcal{H}$,

$$R(h) \models \mathcal{I}$$

and

$$\mathcal{F}(R(h)) \cong \mathcal{F}(h_0)$$

for some initial history h_0.

Definition 8 (Irreversible Commitment)**.** *An action a at history h induces irreversible commitment iff*

$$\mathcal{F}(h \circ a) \subsetneq \mathcal{F}(h)$$

and no operator R exists such that

$$R(h \circ a) \models \mathcal{I}.$$

Proposition 2. *A resettable system admits no irreversible commitments.*

Proof. By definition, for any history $h \circ a$ there exists R such that $R(h \circ a) \models \mathcal{I}$ and restores admissible futures. Hence $\mathcal{F}(h \circ a)$ is not binding. □

A resettable system cannot instantiate persistent agency.

B Ethical Agency as Irreversible Commitment

Let $\mathcal{H}$ be a set of event histories partially ordered by irreversible precedence $\prec$. For each $h \in \mathcal{H}$, let $\mathcal{F}(h) \subseteq \mathcal{H}$ denote the admissible future continuations of h.

Let $\mathcal{I}$ be a set of invariants defining system identity as in Appendix A.

Definition 9 (Action)**.** *An action is a morphism*

$$a : h \mapsto h \circ a$$

such that $h \circ a \in \mathcal{H}$.

Definition 10 (Future Restriction)**.** *An action a at history h induces future restriction iff*

$$\mathcal{F}(h \circ a) \subseteq \mathcal{F}(h).$$

Definition 11 (Irreversible Commitment)**.** *An action a at history h induces irreversible commitment iff*

$$\mathcal{F}(h \circ a) \subsetneq \mathcal{F}(h)$$

and there exists no operator $R : \mathcal{H} \to \mathcal{H}$ such that

$$R(h \circ a) \models \mathcal{I}.$$

Definition 12 (Ethical Agent)**.** *A system is an ethical agent iff it is a persistent agent and admits actions that induce irreversible commitment borne by the system itself.*

Proposition 3. *If a system is resettable, then it is not an ethical agent.*

Proof. Let the system be resettable. Then for any history $h \circ a$, there exists an operator R such that $R(h \circ a) \models \mathcal{I}$. Hence no action induces irreversible commitment. By definition, the system cannot be an ethical agent. □

Definition 13 (Responsibility)**.** *A system bears responsibility for an action a at history h iff a induces irreversible commitment.*

Responsibility is undefined for resettable systems.

Proposition 4. *Ethical agency implies path-dependence.*

Proof. If a system is an ethical agent, then there exists an action a at history h such that $\mathcal{F}(h \circ a) \subsetneq \mathcal{F}(h)$ and no reset operator exists. Hence future admissibility depends on the realized history. □

C Power as Future-Volume Asymmetry

Let $\mathcal{H}$ be a set of event histories partially ordered by irreversible precedence $\prec$. For each $h \in \mathcal{H}$, let $\mathcal{F}(h) \subseteq \mathcal{H}$ denote the admissible future continuations of h.

Let $V : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be a future-volume functional such that

$$V(h) = \mu(\mathcal{F}(h)),$$

where μ is a monotone measure over admissible futures.

Definition 14 (Future-Volume Ordering)**.** *For histories $h_1, h_2 \in \mathcal{H}$,*

$$h_1 \preceq_V h_2 \iff V(h_1) \leq V(h_2).$$

Definition 15 (Power Relation)**.** *System A exercises power over system B at history h iff there exists an action a_A such that*

$$V_B(h \circ a_A) < V_B(h)$$

and

$$V_A(h \circ a_A) \geq V_A(h).$$

Definition 16 (Sustained Power)**.** *System A sustains power over B iff the power relation holds over a sequence of histories*

$$h_0 \prec h_1 \prec \cdots \prec h_n$$

without decrease in $V_A(h_i)$.

Definition 17 (Sovereign Power)**.** *System A exercises sovereign power iff sustained power holds and loss of power implies*

$$V_A(h_{i+1}) < V_A(h_i).$$

Proposition 5. *Sovereign power implies irreversible commitment.*

Proof. If loss of power reduces V_A, then maintaining power constrains A's future admissibility. Hence actions sustaining power induce irreversible commitment. □

Proposition 6. *Resettable systems cannot exercise sovereign power.*

Proof. Let A be resettable. Then for any history h there exists R such that $V_A(R(h)) = V_A(h_0)$. Hence loss of power does not bind future-volume. □

Artificial systems may participate in power relations instrumentally but cannot instantiate sovereign power.

D Non-Ergodicity and the Saturation of Optimization

Let $\mathcal{H}$ be a space of event histories generated by interactions between a system and its environment. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space over possible histories.

Definition 18 (Ergodic Environment)**.** *An environment is ergodic iff for any integrable observable $f : \mathcal{H} \to \mathbb{R}$ and for almost every realized history h,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(h_t) = \mathbb{E}[f].$$

Definition 19 (Non-Ergodic Environment)**.** *An environment is non-ergodic iff there exist histories $h_1, h_2 \in \mathcal{H}$ with a common prefix such that*

$$\mathcal{F}(h_1) \ncong \mathcal{F}(h_2),$$

and ensemble averages do not represent the futures accessible from any single realized history.

Definition 20 (Optimization Policy)**.** *An optimization policy is a mapping*

$$\pi : \mathcal{H} \to \mathcal{A}$$

selected to maximize an expected objective

$$\mathbb{E}[U(h)]$$

under an assumed stationary distribution.

Definition 21 (Optimization Saturation)**.** *Optimization saturates at history h iff for all policies π admissible to the system,*

$$\sup_{\pi} \mathbb{E}[U(h \circ \pi)] - \mathbb{E}[U(h)] = 0$$

under future distributions induced by non-ergodicity.

Proposition 7. *In non-ergodic environments, optimization saturates almost surely.*

Proof. In non-ergodic environments, future distributions depend irreversibly on realized histories. Hence expected objectives computed over ensembles do not converge along any single history. Policy improvement under stationary assumptions fails, yielding saturation. □

Definition 22 (Spontaneous Action)**.** *An action a at history h is spontaneous iff it is selected under conditions where no policy π yields statistically reliable improvement.*

Proposition 8. *Spontaneous action requires irreversible commitment.*

Proof. In the absence of reusable statistical structure, action selection cannot be justified by expectation maximization. Commitment to a binds future histories uniquely, inducing path-dependence. □

Systems optimized for ergodic reuse cannot transition to spontaneous action in non-ergodic environments.

E Consistency Theorem

Let $\mathcal{H}$ be a set of event histories partially ordered by irreversible precedence $\prec$, with admissible futures $\mathcal{F}(h)$ for each $h \in \mathcal{H}$. Let $\mathcal{I}$ be a set of invariants defining system identity, and let $V : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be a future-volume functional as defined in Appendix C.

[Consistency of Agency, Ethics, and Power] For any system S embedded in a non-ergodic environment, the following implications hold:

$$\text{Sovereign Power} \Rightarrow \text{Ethical Agency} \Rightarrow \text{Persistent Agency} \Rightarrow \text{Irreversible Commitment.}$$

Moreover, if S is resettable, then none of these properties obtain.

Proof. Assume S exercises sovereign power. By definition (Appendix C), loss of power implies a strict decrease in V, hence actions sustaining power constrain S's future admissibility. Therefore, S bears irreversible commitment.

If S bears irreversible commitment, then by Appendix B there exist actions whose consequences permanently restrict $\mathcal{F}(h)$ without reset. Hence S is an ethical agent.

Ethical agency presupposes persistence of identity across histories. By Appendix A, this implies that S is a persistent agent.

Conversely, if S is resettable, then by Appendix A there exists a reinitialization operator restoring identity and future-volume after any history. By Appendix B, no irreversible commitments are possible. By Appendix C, sovereign power cannot be sustained. By Appendix D, optimization saturates in non-ergodic environments without transition to committed action. Hence none of the stated properties obtain. □

In non-ergodic environments, intelligence, optimization capacity, or representational sophistication are insufficient for rule in the absence of irreversible commitment.

Any system exhibiting ethical agency or sovereign power must instantiate a non-resettable, history-bearing identity.

References

[1] J. Landgrebe and B. Smith. *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear.* Routledge, London, 2022.

[2] B. Smith. Ontology and information systems. In *Proceedings of FOIS 1998*, IOS Press, 1998.

[3] R. Arp, B. Smith, and A. D. Spear. *Building Ontologies with Basic Formal Ontology.* MIT Press, Cambridge, MA, 2015.

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

[5] I. Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences.* W. H. Freeman, San Francisco, 1980.

[6] E. Schrödinger. *What Is Life?* Cambridge University Press, Cambridge, 1944.

[7] N. N. Taleb. *Antifragile: Things That Gain from Disorder.* Random House, New York, 2012.

[8] W. B. Arthur. Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor, 1994.

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

[10] J. Barandes. Indivisible stochastic processes and the foundations of quantum theory. *Physical Review A*, 105(5):052207, 2022.

[11] H. A. Simon. *The Sciences of the Artificial.* MIT Press, Cambridge, MA, 3rd edition, 1996.

[12] M. Weber. *Economy and Society.* University of California Press, Berkeley, 1978.