Reality Is What Can Be Reached: Reachability, Identification, and the Limits of Abstraction

Flyxion

June 2026

Abstract

Modern science, statistics, artificial intelligence, and philosophy largely inherit a common ontological assumption: reality consists fundamentally of objects, states, facts, and entities whose properties are subsequently described, predicted, compressed, classified, and optimized. This essay argues for a reversal of that picture. The primitive structure of reality is not objecthood but reachability: the pattern of futures accessible from a given state. Objects, representations, abstractions, scientific laws, optimization procedures, and predictive models emerge as secondary constructions imposed upon this deeper accessibility geometry.

From this perspective, every ontology becomes a theory of identification, every representation becomes a projection, every abstraction becomes a form of controlled collapse, and every act of compression becomes an act of selective forgetting. The central question is therefore not whether distinctions are removed, but which distinctions may be removed without altering the future structure of the system under consideration. This criterion is formalized through the concept of admissibility, understood as the preservation of reachable futures under projection.

The essay develops a unified framework connecting abstraction, compression, prediction, optimization, scientific modeling, and artificial intelligence through the geometry of accessible futures. It argues that prediction, compression, statistical significance, and optimization are not foundational principles but downstream instruments whose validity depends upon a prior condition: the preservation of reachability structure. A system may be highly predictive, highly compressive, statistically successful, and optimally

controlled while simultaneously destroying the distinctions that determine which futures remain possible.

The resulting inversion has significant consequences for contemporary debates in machine learning, control theory, epistemology, and the philosophy of science. Intelligence is reinterpreted not as the discovery of regularities, the maximization of prediction accuracy, or the compression of experience, but as the preservation of future possibility. The deepest question becomes not what is true, predictable, or compressible, but what cannot be forgotten without destroying the future.

1. The Crack in the Bridge

Consider a bridge carrying thousands of vehicles each day. For years it behaves exactly as expected. Its load-bearing characteristics remain within acceptable engineering tolerances. Traffic flows normally. Weather patterns remain unexceptional. Every available metric suggests stability. The bridge, from the perspective of any standard observational framework, is functioning as designed.

Somewhere within the structure, however, a microscopic crack begins to form.

Initially the crack contributes almost nothing to any conventional measure of importance. It explains essentially none of the bridge's observed behavior. It has negligible influence on average load calculations. It contributes almost nothing to predictive error. Under many forms of statistical modeling it would be treated as noise and, in sufficiently compressed representations, would likely vanish entirely—swallowed by the equivalence classes of a model trained to minimize description length or expected forecast loss.

Yet the crack matters, and it matters enormously.

The reason it matters has little to do with its current magnitude and everything to do with its relationship to the future. Although physically small, it occupies a position within the structure from which possible futures diverge. One future contains a functioning bridge and normal daily use. Another contains progressive structural degradation and costly repair. Another contains catastrophic failure. The significance of the crack does not arise from its contribution to present observations but from its capacity to alter what futures remain accessible from the current state of the system.

This distinction reveals a tension that appears repeatedly throughout science, engineering, and artificial intelligence. Many of the most influential evaluation

frameworks assess importance according to present contribution. Statistical methods measure variance explained. Predictive systems measure reductions in expected error. Compression methods measure reductions in description length. Optimization procedures evaluate changes in objective function values. In each case, the significance of a feature is assessed through its relationship to some quantity defined relative to present observations or present performance.

The crack exposes a limitation shared by all such approaches. What determines the significance of the crack is not the amount of information it contributes to current observations, nor the degree to which its inclusion improves prediction, nor the resistance it offers to compression. What determines its significance is its influence on future accessibility. A detail may contribute almost nothing to present description while simultaneously controlling the boundary between radically different futures.

The same phenomenon appears across many domains. A latent software fault may remain dormant for years before triggering catastrophic system failure at scale. A single mutation may redirect the evolutionary trajectory of an entire species across geological time. A pathogen establishing itself in a small isolated population may transform into the beginning of an epidemic. A small but legally ambiguous financial liability may remain invisible within ordinary accounting metrics until economic conditions change and suddenly expose its consequences. A subtle misunderstanding between political actors may persist unnoticed for years until it becomes the catalyst for large-scale conflict.

In every case the critical feature is the same. The present statistical footprint of the state is small. Its future consequences are enormous. The problem is not merely practical. It is, at its root, ontological.

Modern intellectual culture tends to assume that reality consists primarily of objects, states, facts, and entities. Once these objects have been identified, the task of science becomes describing them, predicting their behavior, compressing their structure, and optimizing their interactions. The crack appears, within this picture, as merely another property of another object—a physical quantity to be measured, predicted, and managed along with all the others.

Yet the example suggests a different possibility. Perhaps the significance of a state does not derive primarily from what it is but from what it permits. Perhaps the fundamental structure of reality is not objecthood but accessibility. Perhaps what matters most about any state is not its present description but the pattern of futures that remain reachable from it.

If this inversion is correct, then many familiar intellectual priorities must be

reconsidered. Prediction would no longer be fundamental but derivative. Compression would no longer define understanding. Optimization would become a special case of navigation within a pre-existing structure of possibility. Representation would become a problem of preserving future accessibility rather than merely preserving information.

The crack therefore serves as more than an engineering example. It reveals a general principle, which may be stated as follows: the features that matter most about a system are not necessarily those that contribute most strongly to present description, prediction, compression, or control. They are the features whose removal would alter the structure of reachable futures. The remainder of this essay develops the consequences of taking that principle seriously as a foundational commitment rather than a useful heuristic.

2. Reality Is What Can Be Reached

The example of the crack motivates a shift in ontological priority that runs against several deeply entrenched intellectual habits. Ordinarily, reality is assumed to consist of objects possessing properties. A bridge contains steel, concrete, bolts, supports, and defects. A biological organism contains cells, organs, and molecules. A physical system contains particles, fields, and interactions. The task of scientific explanation is then understood as the discovery of laws governing the behavior of these objects and their properties. This picture is so familiar that it often appears self-evident, as though the priority of objects were itself a feature of the world rather than a theoretical commitment about what to take as primitive.

Yet this picture contains an important assumption that is rarely made explicit. It assumes that objects are ontologically primary and that possibilities emerge from the interactions among those objects. The world is first populated with things; only afterwards do futures appear. The crack, on this view, is a defect in a structural member, and its significance is a downstream consequence of what that defect eventually does to the properties of the surrounding material.

The crack example points consistently in the opposite direction. What made the crack important was not its present physical description. It was not especially large, it did not dominate any observable metric, and it was not responsible for most of the bridge's behavior. The significance of the crack derived from the way it altered future accessibility. Its importance arose from its location within a network of possible transformations—from the fact that it stood at a branching point in the space of reachable futures.

This observation motivates a different starting point. Rather than beginning with objects and deriving futures, one may begin with futures and derive objects. Instead of treating reality as a collection of entities that happen to possess possibilities, one may treat reality as a structure of possibilities from which entities emerge as derivative constructs. Objects, on this view, are not the bedrock of reality but are stabilized patterns within a deeper geometry of accessible transformations—regions of the reachability structure that persist coherently across a range of conditions and timescales.

The inversion can be expressed schematically. The traditional picture assumes a hierarchy of the form

$$\text{Reality} \;\rightarrow\; \text{Objects} \;\rightarrow\; \text{Relations} \;\rightarrow\; \text{Futures},$$

where objects come first and their relational dynamics subsequently give rise to the space of possible futures. The reachability picture proposed here inverts this ordering:

$$\text{Reality} \;\rightarrow\; \text{Reachability Structure} \;\rightarrow\; \text{Objects}.$$

Objects become secondary constructions. They emerge as relatively stable regions within a deeper geometry of accessible transformations, and their significance derives from the role they play in that geometry rather than the other way around.

This shift may initially appear radical, but versions of it already appear throughout several domains of science and mathematics. In dynamical systems theory, the long-term behavior of a system is routinely analyzed through attractors, basins of attraction, bifurcation diagrams, and invariant manifolds rather than through isolated states. The state of the system matters less than the qualitative structure of the space of trajectories emanating from it. In ecology, the significance of a species is frequently determined by the role it plays in sustaining accessible pathways through an ecosystem: a keystone species matters not primarily because of what it is but because of what its removal forecloses. In economics, the value of an asset derives not merely from present ownership but from the opportunities that ownership makes available—an option, for instance, derives its value entirely from the accessibility of future states it preserves. In computer science, abstract data types are often defined operationally through the transformations they permit rather than through their internal constitution alone, making accessibility part of what the structure fundamentally is.

To formalize the reachability framework, let x denote a state of some system and let $\mathcal{R}(x)$ denote the set of futures reachable from that state under an appropriate

class of transformations or policies. The object of primary theoretical interest is no longer the state itself but the structure of $\mathcal{R}(x)$ as a function across the state space. The meaning of a state, in this sense, becomes inseparable from its future horizon: two states that appear superficially similar may differ profoundly if the futures accessible from them differ, while states that appear physically distinct may be equivalent for many purposes if they permit the same future transformations. What matters is not merely what the state is but what the state allows.

This perspective alters the interpretation of knowledge itself. Knowledge is commonly understood as accurate description: one possesses knowledge to the extent that one's internal representations correspond faithfully to the external world. Yet from a reachability perspective, the practical value of a representation depends less on descriptive fidelity than on its ability to track future accessibility. A representation succeeds not when it resembles the world but when it preserves the distinctions that determine which futures remain possible. Two representations may differ dramatically in their surface properties while remaining equivalent from the perspective of reachability preservation, and two representations may appear similar while differing profoundly in the futures they allow one to navigate.

The same inversion affects prediction. Prediction is often treated as the defining feature of intelligence and the primary criterion for evaluating scientific theories and artificial systems alike. Machine learning systems are trained to minimize predictive error. World models are evaluated according to their ability to forecast future states. Scientific theories are judged by predictive success. Yet prediction already presupposes accessibility structure. To predict is to make claims about possible futures; to plan is to select among possible futures; to control is to guide movement toward possible futures; and to optimize is to search among possible futures. Even safety is defined through the distinction between acceptable and unacceptable futures. In every case, the concept of a future is logically prior to the operation being performed upon it.

This observation reveals a subtle asymmetry. Prediction, planning, optimization, and control are often treated as foundational concepts, yet they all depend upon a more primitive notion: the structure of what can be reached. Without some prior notion of accessibility, there is nothing to predict, nothing to optimize, and nothing to control. Reachability is therefore not introduced here as an alternative objective alongside prediction or optimization, nor as a constraint to be satisfied alongside other requirements. It is introduced as a prior condition for those activities to make sense at all.

A clarification is warranted here to forestall a possible misreading. The reacha-

bility thesis does not assert that objects are unreal or that physical properties are irrelevant. Steel possesses specific tensile properties; cracks propagate according to known mechanical laws; bridges fail when load exceeds capacity. Object-level descriptions explain why futures diverge. Reachability explains why that divergence matters. The two levels of description are complementary and mutually dependent, not competing. The proposal concerns explanatory priority for a specific purpose: determining which distinctions deserve preservation under abstraction. For that purpose, the accessibility geometry provides the more general and directly applicable criterion, because it is the geometry within which planning, control, and safety must ultimately operate.

Furthermore, reachability is never absolute or free-floating. It is always defined relative to a specified dynamical system D encoding the laws, constraints, and permitted transformations of the domain under consideration. Writing this dependence explicitly, the reachable future set from state x under dynamics D is

$$\mathcal{R}_D(x) = \{y \in \mathcal{X} : x \rightsquigarrow_D y\},$$

where $x \rightsquigarrow_D y$ denotes reachability under the transition structure of D. The object-level description—material properties, governing equations, boundary conditions—is precisely what determines D. The present argument is not that reachability precedes dynamics in any metaphysically loaded sense; it is that once the dynamical structure has been specified, the accessibility geometry $\mathcal{R}_D$ becomes the natural criterion for evaluating which representational collapses are permissible. Object properties determine the geometry. Admissibility evaluates projections relative to that geometry.

3. Why Reachability Comes First

The claim that reachability is ontologically prior to prediction, planning, optimization, and control may initially appear counterintuitive. Contemporary scientific and engineering practice often treats these latter concepts as foundational. Entire disciplines are organized around the prediction of future states, the optimization of objective functions, and the control of dynamical systems. It is therefore natural to ask why reachability should be granted a more primitive status rather than being regarded as simply one concept among others.

The answer lies not in the practical importance of these activities but in their logical dependence upon a prior structure that they presuppose but cannot

themselves provide.

Consider prediction. To predict is to make a statement about a future state of some system. Even the simplest prediction presupposes that the future state being discussed belongs to a space of possible futures that already possesses determinate structure. A prediction may be correct or incorrect, but before it can be either, there must exist some underlying geometry of possibilities that determines which futures are available and which are not. The act of prediction therefore presupposes an accessibility structure that the prediction itself does not create and cannot modify.

Planning exhibits the same dependency in a particularly transparent form. A plan is a sequence of actions intended to move a system from one state to another. The very concept of planning requires that multiple futures be available for consideration: a planner evaluates alternative trajectories, compares their consequences, and selects among them. Yet none of these operations determine which trajectories exist in the first place. Planning operates within a space of possibilities whose structure is already given, and the quality of the plan depends entirely upon the admissibility of the representation through which that space is perceived.

Control theory provides an especially clear example. A controller seeks to guide a system toward some desired region of state space. Classical control theory is extraordinarily successful precisely because it typically assumes that the relevant state variables have already been identified and that the transition structure of the system is sufficiently understood. Once those assumptions are granted, the controller can exploit the geometry of the state space to produce desired behavior. What the controller does not do is generate that geometry. The geometry exists prior to the controller. The controller navigates it.

Optimization reveals the same asymmetry. Optimization is frequently treated as a foundational explanatory principle: contemporary machine learning, economics, and decision theory often describe intelligent behavior as the maximization of some objective function over a space of possible states. Yet optimization is incapable of defining that space. An optimizer can search among alternatives, rank futures by objective value, and identify maxima or minima. It cannot determine what counts as an alternative in the first place. An optimizer requires a geometry of possibilities to exist before it can begin operating, yet it takes no responsibility for the admissibility of that geometry.

This distinction is easy to overlook because optimization procedures often appear enormously powerful. Given a sufficiently rich state space and a sufficiently

expressive objective function, optimization can produce remarkably complex and adaptive behavior. The success of the optimizer, however, depends entirely upon the structure of the space within which it operates. An optimizer searching the wrong geometry may be perfectly rational in a formal sense while failing catastrophically in practice.

The familiar observation that a system is "optimizing the wrong objective" captures only part of the problem. A deeper possibility exists and is in many ways more troubling: a system may be optimizing the correct objective within the wrong representation. In such a case the failure does not arise from the objective function itself but from the geometry of distinctions available to the optimizer. The objective is fine. The map is wrong. And crucially, the optimizer has no means of detecting this from within its own representational world.

Safety illustrates the dependency upon reachability with particular clarity. Safety is often discussed as though it were an additional objective or constraint that can be imposed upon an otherwise capable system. One introduces guardrails, penalties, alignment criteria, or constraint satisfaction requirements intended to prevent undesirable behavior. Such measures are undeniably important, but they presuppose that the system can distinguish safe states from unsafe states in the first place. If the underlying representation has already collapsed those states into a single equivalence class, no downstream safety mechanism can recover the distinction. The distinction has been destroyed at the level of projection, and guardrails operating above that level inherit the blindness without being able to compensate for it.

The logical structure of these observations can be summarized as a chain of dependencies. Prediction presupposes a space of accessible futures. Planning presupposes that the space of futures has been correctly represented. Optimization presupposes a geometry within which the search occurs. Control presupposes a representation in which relevant distinctions survive. Safety presupposes that catastrophic and acceptable states remain distinguishable. Each activity therefore depends upon a prior condition that none of them supply: the preservation of reachability structure under abstraction.

Reachability thus occupies a different logical category from the activities that depend upon it. It is not another objective alongside prediction or optimization, to be balanced against other requirements in some multi-objective framework. It is the condition under which prediction, optimization, planning, control, and safety become coherent as activities at all. Understanding why this is so requires examining what happens when systems form representations—how they identify

states as equivalent and thereby decide, always implicitly and often irreversibly, which distinctions survive.

4. Every Ontology Is a Theory of Identification

If reachability is primary, then the central problem of knowledge changes substantially. The question is no longer simply how to describe the world but how to partition it. Every act of understanding requires determining which differences matter and which may be safely ignored. Every representation, category, concept, scientific law, and model performs this operation, and performs it necessarily—for no finite cognitive or computational system can preserve every distinction present in the world it inhabits.

This observation appears trivial at first glance. Human beings constantly identify things as equivalent. Individual trees are identified as instances of the same species. Physically distinct chairs are subsumed under the same category. Successive versions of a nation, a language, or a person are treated as manifestations of a single continuing entity despite enormous internal change. Scientific theories likewise identify apparently different phenomena as instances of common principles. The history of science is filled with such unifications: heat and molecular motion, electricity and magnetism, space and time, mass and energy. Each represents an identification that collapsed a previously maintained distinction.

Yet beneath these familiar operations lies a deeper question. What justifies any particular identification? The traditional answer appeals to similarity: two entities are treated as equivalent because they resemble one another in relevant respects. This answer, however, merely relocates the problem. Similarity itself requires a criterion. To determine whether two states are similar one must already know which features matter and which do not. The act of identification cannot be grounded entirely in resemblance because resemblance presupposes that relevant dimensions of comparison have already been selected, and that selection is itself an act of identification. The problem is therefore circular if grounded only in similarity.

The reachability framework provides an alternative foundation. To identify two states as equivalent is to collapse a distinction between them. Every category, concept, representation, and theory performs such a collapse. Every ontology therefore enacts a theory of permissible identification: it specifies which distinctions may be ignored without loss of the explanatory or operational power that matters for the domain in question. Ontology, on this view, is not primarily a catalogue of

what exists. It is a theory of what can be treated as the same.

Consider the concept of a species. Individual organisms differ in countless ways—different genetic sequences, developmental histories, behaviors, morphological features, and environmental interactions. The category of species identifies many of these organisms as equivalent despite those differences. The category is useful precisely because it ignores distinctions that are irrelevant for many biological purposes while preserving the distinctions that matter for reproduction, inheritance, and evolutionary dynamics. A similar analysis applies to the concept of a physical object. A chair remains a chair despite scratches, repainting, incremental repairs, and gradual material replacement over time. The concept functions by collapsing numerous physical differences into a single persisting identity. Without such collapse, ordinary perception and action would become impossible.

Scientific laws operate analogously. Newtonian mechanics identifies countless distinct motions as instances of common dynamical principles, collapsing the enormous diversity of trajectories in the physical world into a small collection of equations. Thermodynamics ignores microscopic molecular details to describe macroscopic behavior, treating the uncountable individual trajectories of gas molecules as equivalent so long as they share the same macroscopic state. Fluid mechanics treats vast collections of individual particles as continuous media, collapsing distinctions among molecular configurations that leave bulk flow behavior unchanged. These abstractions succeed not because they preserve all information but because they preserve the right information for their domain of application.

The success of science therefore depends not on avoiding identification but on performing the correct identifications. A scientific theory that maintained every microscopic distinction would be computationally intractable and explanatorily useless. What matters is the criterion by which identifications are made.

Representations in artificial intelligence perform the same operation. A learned representation maps many distinct physical states into a smaller latent space, declaring states equivalent whose latent codes coincide. The representation succeeds precisely to the extent that it identifies states that may safely be treated as equivalent for the purposes of prediction, planning, or control. The language of representation learning often describes this process in terms of feature extraction, dimensionality reduction, abstraction, or compression, but ontologically the operation is the same: the representation decides which distinctions survive and which do not.

This perspective reveals a hidden continuity between seemingly unrelated intellectual activities. The philosopher constructing an ontology, the scientist

formulating a law, the engineer building a model, and the machine learning system learning a latent representation are all solving variations of the same problem. Each must determine which distinctions are worth preserving and which may be collapsed without harmful consequence.

The challenge is that collapse is irreversible in practice. Once two states have been identified as equivalent within a representation, every subsequent operation inherits that decision. Predictions, optimizations, classifications, and plans all operate on the resulting equivalence classes. They cannot easily recover distinctions that have already been removed by the representation upon which they depend. The decision about which distinctions to preserve is therefore among the most consequential decisions a cognitive or computational system makes, and it is often made implicitly, in the structure of representation choices, rather than explicitly and deliberately.

The significance of this observation becomes clearer when viewed through the lens of reachability. If reality is fundamentally a geometry of accessible futures, then the adequacy of an identification cannot be judged solely by similarity, predictive success, or compression efficiency. A more demanding criterion is required. Two states may be safely identified only if the collapse preserves the future structure of the system. If the identification destroys distinctions that alter what futures remain accessible, then the collapse is not merely imperfect—it is inadmissible.

5. The Projection Problem

The preceding discussion established that every ontology performs acts of identification. Categories, scientific laws, abstractions, and representations all reduce complexity by treating certain distinctions as irrelevant. The critical question is therefore not whether collapse occurs but how it is performed, and whether the collapses that occur are warranted. To answer this question with greater precision, it is useful to formalize the relationship between a system and its representation.

Let $\mathcal{X}$ denote a state space and let $\mathcal{M}$ denote a representation space. A representation can be understood as a mapping

$$\pi : \mathcal{X} \to \mathcal{M}.$$

The notation is intentionally general. The representation π may correspond to a scientific model, a statistical summary, a learned latent embedding, a conceptual category, a digital twin, a control state, or any other structure intended to stand

in for the original system.

The essential feature of such mappings is that they are almost never injective. Distinct states in $\mathcal{X}$ are mapped to the same representation in $\mathcal{M}$. Formally, for any non-trivial projection there exist states $x_1, x_2 \in \mathcal{X}$ such that

$$x_1 \neq x_2 \qquad \text{yet} \qquad \pi(x_1) = \pi(x_2).$$

Whenever this occurs, an identification has been performed. The representation has declared the distinction between x_1 and x_2 irrelevant—collapsed it into a shared representational point from which neither element can be independently recovered.

The term projection is more precise than abstraction or representation because it emphasizes what these mappings actually do. A projection does not merely preserve some information while discarding other information. It actively eliminates distinctions by imposing an equivalence relation on the original state space. The equivalence classes generated by π—the sets $[x]_\pi = \{x' : \pi(x') = \pi(x)\}$— define what the representation can and cannot express. Every claim made within $\mathcal{M}$ is implicitly a claim about equivalence classes rather than individual states, and the consequences of actions taken on the basis of that representation are therefore consequences about sets of states rather than the specific state actually occupied.

This observation immediately reveals an asymmetry that is underappreciated in most discussions of representation learning. Much of the contemporary discussion focuses on what representations capture. Researchers ask whether a latent space contains information about depth, object identity, causal structure, planning relevance, or commonsense relations. Such questions are important, but they overlook an equally significant question that is in many ways logically prior.

Every representation also defines what has been forgotten.

The omissions are not incidental artifacts of imperfect learning. They are the price of abstraction itself, and they are paid by necessity with every representational choice. The fact that they are often invisible—that successful abstractions appear natural in retrospect—does not diminish their importance. It merely makes them harder to examine critically.

Traditional evaluation frameworks typically address the projection problem indirectly. Statistical models evaluate projections according to predictive performance on held-out data. Compression methods evaluate them according to description length. Information-theoretic approaches evaluate them according to preserved variance or mutual information. Optimization frameworks evaluate them according to downstream task success in controlled environments.

All of these criteria have practical value, and none of them should be abandoned. However, none directly addresses the question that admissibility poses. A projection may perform extremely well according to any of these measures while still destroying distinctions that alter the future structure of the system. The reason is straightforward: prediction, compression, and optimization evaluate properties that are internal to the representation, or relative to some training distribution. They do not directly evaluate what futures have been lost through the act of projection itself, because the lost futures are precisely what is no longer visible from within the representational space.

The crack in the bridge again provides a useful illustration. Suppose a representation collapses two bridge states into a single latent description because their observable behavior is nearly identical across the training distribution. The resulting projection may improve compression, reduce prediction error, and simplify the downstream control problem. Yet if one state contains a microscopic structural defect and the other does not, the projection has destroyed a distinction that separates radically different future trajectories. The representation has become simpler; the future has become less visible. And crucially, the standard evaluation metrics will record the projection as having succeeded, because the failure is located precisely in the region of the state space that has been collapsed out of representational existence.

The central intuition motivating this essay can now be stated more precisely. Representational adequacy cannot be assessed solely through predictive or informational criteria because those criteria evaluate performance within a given geometry of distinctions, while the most important failures occur when the geometry itself has been incorrectly specified. A projection is successful not merely when it preserves information in the aggregate but when it preserves the distinctions necessary to maintain the future structure of the system—when, in other words, it is admissible.

6. Admissibility

The projection problem reveals that a representation may preserve large quantities of information, support accurate prediction, enable efficient compression, and facilitate successful optimization while still destroying distinctions that fundamentally alter the future structure of the system. If reachability is primary, then a different criterion is needed to distinguish productive abstraction from destructive collapse. This criterion is admissibility.

The concept begins with a structural observation about every state. Let

$x \in \mathcal{X}$ denote a state of some system. Associated with x is a collection of futures reachable through the allowable transformations of that system. Denote this collection by $\mathcal{R}(x)$. The precise meaning of reachability depends upon context: in a control problem, $\mathcal{R}(x)$ may represent the states accessible through some class of control policies; in a biological system, it may represent possible future evolutionary trajectories; in an economic system, it may represent attainable economic configurations under feasible sequences of decisions. The formal details differ across domains, but the underlying idea remains consistent: every state determines a horizon of futures, and that horizon is among the most fundamental properties of the state from the perspective of any system that must act within the world.

Definition 1 (Admissible Projection). Let $\pi : \mathcal{X} \to \mathcal{M}$ be a projection from a state space to a representation space, and let $\mathcal{R}(x)$ denote the reachable future set from state x. The projection π is said to be admissible if for all $x_1, x_2 \in \mathcal{X}$,

$$\pi(x_1) = \pi(x_2) \quad \Longrightarrow \quad \mathcal{R}(x_1) = \mathcal{R}(x_2).$$

The condition demands that whenever the projection declares two states equivalent, those states must possess identical reachability structure. If the projection collapses two states whose future horizons differ, then it has destroyed a distinction that has genuine bearing on what the system can subsequently do or become. Such a projection is inadmissible.

Several features of this definition deserve careful attention.

First, the criterion is future-directed rather than present-directed. Most conventional criteria for evaluating abstractions assess them according to their relationship to present observations or current performance metrics. Admissibility evaluates whether the abstraction preserves the structure that determines what futures remain available. This relocation from present to future is not a superficial change of framing; it corresponds to a genuine difference in what gets measured and what failures become visible.

Second, admissibility is intentionally demanding. It does not ask whether the collapsed states are difficult to distinguish empirically, nor whether the distinction improves short-horizon prediction, nor whether the collapse is unlikely to matter in practice. It asks only whether the collapse alters the geometry of future possibility. This demanding character is appropriate because the failures that admissibility is designed to detect are precisely the ones that conventional metrics miss: inadmissible collapses look acceptable from the perspective of current performance

and reveal their consequences only when the futures they have obscured become relevant.

Third, admissibility furnishes a way of understanding why many successful abstractions remain trustworthy despite discarding vast quantities of detail. Fluid mechanics ignores the precise positions and velocities of individual molecules. Thermodynamics ignores microscopic trajectories entirely. Population genetics ignores individual organisms in favor of allele frequencies. These abstractions succeed not despite their information losses but because those information losses are, to a very good approximation, admissible: the distinctions they collapse do not significantly alter the reachability structure relevant to the macroscopic phenomena under investigation.

Proposition 1. Let π be an admissible projection. Then for any two states $x_1, x_2 \in \mathcal{X}$ with $\pi(x_1) = \pi(x_2)$, any policy σ defined over $\mathcal{M}$ induces the same distribution over future trajectories from x_1 as from x_2, in the sense that the reachable sets under σ are identical.

Proof. Suppose $\pi(x_1) = \pi(x_2)$ and π is admissible. Then by definition $\mathcal{R}(x_1) = \mathcal{R}(x_2)$. Any policy $\sigma : \mathcal{M} \to \mathcal{A}$ assigns actions based solely on the representation, and since $\pi(x_1) = \pi(x_2)$, the same action sequence is induced from both states. The resulting trajectories draw upon the same reachable set. Hence no outcome reachable from x_1 under σ is unreachable from x_2 under σ, and conversely. □

The proposition makes precise why admissibility matters operationally. A policy operating in the representational space $\mathcal{M}$ cannot distinguish states that π has identified as equivalent. If those states differ in their reachable futures, the policy will inevitably sometimes attempt to reach a future that is not available from the actual state it inhabits, or fail to avoid a future that the actual state makes inescapable.

A useful way to understand the criterion is to consider two types of representational error. In the first type, the projection merges two states whose observable properties differ slightly but whose reachable futures are identical. Such a projection loses information about the present but does not alter the geometry of future possibility. The collapse is imperfect in an informational sense but remains admissible: nothing about what the system can subsequently do has been obscured. In the second type, the projection merges two states whose present observations are nearly identical but whose reachable futures diverge dramatically. The crack in the bridge is the canonical example: the observable difference may be tiny, the informational loss may be negligible, and the effect on short-horizon prediction

may be imperceptible, yet the future consequences are enormous. This second collapse is inadmissible.

Admissibility therefore reframes the traditional relationship between abstraction and fidelity in an important way. A representation need not be maximally detailed to be good, and a highly detailed representation is not automatically trustworthy. What matters is not the quantity of information preserved but the future structure preserved. The criterion is neither completeness nor accuracy in the ordinary sense. It is the preservation of reachability structure: the maintenance of distinctions whose loss would alter what the system can subsequently do or become.

7. Compression as Controlled Forgetting

The concept of admissibility shifts attention from information preservation to future preservation. This shift becomes especially important when considering one of the most influential ideas in modern theories of intelligence: the identification of understanding with compression. The appeal of this identification is powerful and its intellectual heritage is deep. Scientific laws compress observations into compact principles. Mathematical equations replace vast collections of empirical facts with concise descriptions. Statistical models reduce complexity by identifying regularities. Machine learning systems discover latent representations that permit more efficient prediction and storage. Across all of these domains, successful understanding appears inseparable from successful compression, and the intuition that an elegant, parsimonious description reveals something deeper than a verbose one has been central to scientific and mathematical practice for centuries.

The intuition has a rigorous formal tradition. Minimum description length provides a Bayesian reformulation of Occam's razor in terms of code lengths. Kolmogorov complexity defines the intrinsic information content of a string as the length of the shortest program that generates it. Algorithmic information theory unifies these notions and connects them to computational universality. Building upon these foundations, compression-based theories of intelligence interpret curiosity as the search for compressibility and understanding as the discovery of shorter descriptions of experience.

The success of this viewpoint should not be underestimated. Many of the greatest achievements of science can be understood as acts of compression, and the insight that regularity and compressibility are closely related is both genuine and important. The difficulty is not that compression is unimportant but that it conceals a cost that is rarely examined as carefully as the benefits.

Every successful compression scheme operates by forgetting.

A compression algorithm reduces description length by identifying distinctions that need not be preserved. The bits that are eliminated correspond to distinctions that the compression scheme has decided are redundant or irrelevant for the purposes of reconstruction. Every abstraction, every category, every scientific law, and every learned representation achieves efficiency by discarding distinctions judged unnecessary for some purpose. The success of compression is therefore inseparable from selective amnesia, and the question of what has been forgotten is as important as the question of what has been retained.

The conventional discussion focuses on what has been preserved. A compressed representation is praised because it captures structure, reveals regularities, or identifies hidden patterns. Less attention is devoted to the inverse question: what is no longer distinguishable? From the perspective developed in this essay, that question is fundamental.

Compression is not merely the preservation of structure. It is the destruction of distinctions. Every reduction in description length corresponds to a collapse of states that were previously represented separately. The critical issue is therefore not whether compression occurs but which distinctions disappear under the compression scheme.

This is where admissibility becomes operative. Traditional measures of compression evaluate reductions in information, description length, or predictive redundancy. Admissibility evaluates whether the distinctions removed by compression alter the future accessibility structure of the system. The difference is subtle but profound, and it becomes dramatic in precisely the cases where it matters most.

A compression scheme may achieve extraordinary efficiency while remaining admissible. This occurs when the distinctions it removes do not alter reachable futures. Many scientific abstractions succeed precisely because they discard details that are irrelevant to the future structure under consideration: molecular positions can be forgotten in thermodynamics because the macroscopic reachability structure is insensitive to them. However, compression and admissibility are not equivalent, and there is no guarantee that compressive efficiency and reachability preservation point in the same direction.

A representation may become more compressive by eliminating distinctions that appear insignificant from the perspective of present observations. Yet those same distinctions may occupy critical positions within the geometry of future possibility. The crack in the bridge is again the canonical example. From the perspective of compression, the crack is an attractive candidate for elimination: its contribution

to present observations is tiny, its influence on average behavior is negligible, and across a sufficiently large training distribution, preserving the distinction between cracked and uncracked states may appear wasteful. A compressed representation can often achieve lower description length by treating them as effectively identical. The resulting representation becomes simpler, and by every internal metric of the compression scheme it has succeeded. It has also become blind to a distinction that determines which futures are accessible.

The significance of the crack therefore reveals a broader principle: the importance of a distinction cannot be determined solely by its contribution to present description, nor by its contribution to average prediction, nor by any local measure of its current influence. A distinction may appear insignificant according to every local metric while remaining indispensable to the preservation of future accessibility. This observation exposes a limitation in compression-centered theories of intelligence that is not merely practical but principled.

Suppose a representation achieves exceptional compression performance, discovers elegant latent structure, minimizes redundancy, and reveals deep regularities. None of these achievements guarantees that the representation preserves the distinctions necessary for future accessibility, because compression rewards forgetting while admissibility constrains it. The two objectives frequently align, and when they do compression is not only efficient but trustworthy. But they are not identical, and the pressure exerted by aggressive compression is toward the elimination of distinctions whose contribution to description length is small—which is precisely the pressure that may eliminate reachability-critical features.

This permits a reformulation of the relationship between knowledge and compression. Rather than asking how much a representation can forget while retaining predictive power, one must ask how much it can forget while retaining the future structure of the system. The question is therefore not merely:

> How can the world be compressed?

but rather:

> What cannot be forgotten without destroying the future?

8. Admissibility Distortion

The concept of admissibility provides a binary criterion: a projection either preserves reachability structure or it does not. Yet binary criteria, while valuable for establishing the presence or absence of a property, are often insufficient for practical

analysis and theoretical development. Real representations are neither perfectly admissible nor completely arbitrary. They occupy an intermediate position in which some identifications are harmless, others are consequential, and the degree of harm varies continuously across the state space. A useful theory therefore requires not only a criterion for perfect admissibility but also a measure of degree of departure from it—a way of quantifying how much reachability structure a given projection destroys.

The need for such a measure becomes evident when considering the limitations of existing representational metrics. Information-theoretic measures evaluate how much information remains after a projection. Statistical measures evaluate explanatory power or predictive performance. Compression measures evaluate reductions in description length. Optimization measures evaluate success relative to some objective function. Each of these quantities captures something important. None directly measures what futures have disappeared, because the disappeared futures are precisely those that have been rendered invisible by the projection and therefore cannot be assessed from within the representational space.

To formalize a measure of destructive forgetting, consider a projection $\pi : \mathcal{X} \to \mathcal{M}$ and let $\mathcal{R}(x)$ denote the set of futures reachable from state x. Let d_H denote a suitable distance measure between reachable future sets—for concreteness, one may take d_H to be a Hausdorff-type metric on the relevant space of future trajectories, though the definition is compatible with optimal transport distances and other metrizations of set-valued maps.

Definition 2 (Admissibility Distortion). The admissibility distortion of a projection π is

$$D_A(\pi) \;=\; \mathbb{E}_{x_1,\,x_2}\big[\mathbf{1}_{\{\pi(x_1)=\pi(x_2)\}} \cdot d_H(\mathcal{R}(x_1),\,\mathcal{R}(x_2))\big]\,,$$

where the expectation is taken over pairs (x_1, x_2) drawn from an appropriate distribution over $\mathcal{X} \times \mathcal{X}$.

Conceptually, admissibility distortion measures the expected divergence between future horizons for pairs of states that the projection has identified as equivalent. The indicator function restricts attention to states that have been collapsed by the projection. The distance term evaluates how different their future horizons actually are, despite the representational merger. The resulting quantity therefore measures not information loss in the abstract but destructive forgetting specifically: the extent to which the projection has collapsed distinctions that separate genuinely different futures.

Several properties of this definition clarify its relationship to existing measures and its significance for the arguments developed elsewhere in this essay.

Proposition 2. A projection π is admissible if and only if $D_A(\pi) = 0$.

Proof. If π is admissible, then $\pi(x_1) = \pi(x_2)$ implies $\mathcal{R}(x_1) = \mathcal{R}(x_2)$, so $d_H(\mathcal{R}(x_1), \mathcal{R}(x_2)) = 0$ for every pair satisfying the indicator. Hence $D_A(\pi) = 0$. Conversely, if $D_A(\pi) = 0$, then for almost every pair with $\pi(x_1) = \pi(x_2)$ we have $d_H(\mathcal{R}(x_1), \mathcal{R}(x_2)) = 0$, which implies $\mathcal{R}(x_1) = \mathcal{R}(x_2)$, so π is admissible. □

The measure therefore extends the binary admissibility criterion continuously, with zero distortion corresponding to perfect admissibility and larger values corresponding to greater degrees of destructive forgetting.

A critical distinction separates admissibility distortion from conventional information-theoretic measures of representational quality. Prediction error asks whether future observations can be forecast accurately from the representation. Compression measures ask whether observations can be described economically within the representation. Information-theoretic measures ask whether distinctions remain statistically resolvable. All of these quantities evaluate properties relative to the representation and its relationship to observed data. Admissibility distortion asks a structurally different question:

> How much future accessibility has been destroyed by the act of abstraction itself?

This distinction reveals why apparently successful representations may fail in ways that remain entirely invisible to conventional evaluation procedures. A model may achieve excellent predictive performance, state-of-the-art compression, and robust downstream task success while exhibiting significant admissibility distortion. Such a model appears successful because its failures occur not within the representational space—where evaluation takes place—but at the boundaries between futures that the representation has rendered indistinguishable.

The crack in the bridge again provides the illustrative case. A representation that identifies two bridge states as equivalent because their observable behavior is nearly identical throughout the available training data will exhibit low prediction error and good compression. Yet if one state contains a microscopic structural defect and the other does not, the reachable futures may differ enormously, and $D_A(\pi)$ will reflect this even when all ordinary performance metrics look favorable.

Admissibility distortion also clarifies the relationship between abstraction and structural risk. Every abstraction introduces the possibility of distortion, and the

question is not whether distortion exists but where it is concentrated. In many systems the greatest distortion occurs near critical boundaries: regions where small differences produce large changes in future accessibility. These boundaries include structural failure thresholds, ecological tipping points, software vulnerabilities, financial crises, evolutionary bifurcations, and many other phenomena in which future possibility changes discontinuously. From the perspective of reachability, such boundaries are precisely where inadmissible abstraction is most dangerous, and they tend to be the regions where statistical sparsity makes conventional evaluation metrics most misleading.

The following result shows that admissibility distortion is, in a precise sense, monotone under the composition of projections. Additional layers of abstraction cannot reduce distortion that has already been introduced.

Theorem 1 (Distortion Monotonicity Under Composition). Let $\pi_1 : \mathcal{X} \to \mathcal{M}$ and $\pi_2 : \mathcal{M} \to \mathcal{N}$ be projections and let $\pi = \pi_2 \circ \pi_1 : \mathcal{X} \to \mathcal{N}$. Under mild regularity conditions on the sampling distribution,

$$D_A(\pi_2 \circ \pi_1) \;\geq\; D_A(\pi_1).$$

Proof. Any pair (x_1, x_2) with $\pi_1(x_1) = \pi_1(x_2)$ necessarily satisfies $\pi_2(\pi_1(x_1)) = \pi_2(\pi_1(x_2))$, so the indicator $\mathbf{1}_{\{\pi(x_1)=\pi(x_2)\}}$ takes value 1 whenever $\mathbf{1}_{\{\pi_1(x_1)=\pi_1(x_2)\}}$ does. The expectation defining $D_A(\pi_2 \circ \pi_1)$ therefore integrates over a superset of the pairs contributing to $D_A(\pi_1)$, and each contributing pair registers the same reachability divergence $d_{\mathcal{R}}(\mathcal{R}(x_1), \mathcal{R}(x_2))$. Since all terms are nonnegative, the composed distortion is at least as large as the initial distortion. □

The practical implication is significant: once an inadmissible collapse has been introduced at the level of π_1, no downstream representational transformation π_2 can undo it. Every further layer of abstraction applied to an already-inadmissible representation either preserves or increases the distortion. Remediation must therefore occur upstream, at the level of the original projection, rather than downstream through refinement of the model built upon it.

9. The Proxy Problem

The preceding sections have developed the positive core of the reachability framework: the ontological priority of accessibility, the formal concept of admissibility, the measure of admissibility distortion, and the contrast between compression

as information reduction and compression as controlled forgetting. The present section applies this framework to a cluster of related failures that share a common structure.

Each of the dominant modern frameworks for evaluating representations and intelligence—compression, prediction, statistical significance, and optimization—defines a proxy metric and then, through the momentum of successful application, allows that metric to migrate from instrument to criterion. The proxy is mistaken for the thing it was originally designed to track. Once this migration has occurred, the failure mode is always the same: the system performs excellently according to its own metric while destroying precisely the structure the metric was originally intended to preserve.

The formal structure of the problem can be stated as a general non-equivalence. Let $M(x)$ denote any of the proxy metrics under consideration (compression rate, prediction accuracy, statistical significance, expected reward), and let $A(x)$ denote the admissibility of the representation with respect to state x. Then in general

$$M(x_1) = M(x_2) \;\not\Rightarrow\; \mathcal{R}(x_1) = \mathcal{R}(x_2),$$

which is to say: equality of proxy metric values does not entail equality of reachable future sets. This non-implication is the single theorem that the following subsections instantiate across different domains.

9.1. Prediction as a Derived Activity

The modern history of artificial intelligence has elevated prediction to a position of unusual prominence. Predictive accuracy serves as the dominant metric in machine learning, scientific modeling, and increasingly in theories of intelligence itself. Systems are trained to predict the next token, the next frame, the next state, or the next action. Success is measured by the extent to which future observations can be anticipated from present ones. The practical achievements of this paradigm are undeniable: predictive systems have transformed language modeling, computer vision, scientific simulation, weather forecasting, protein structure prediction, and many other domains.

Yet predictive success and representational adequacy are not identical. The reason is that prediction evaluates a model according to its performance within a representation, not according to the admissibility of the representation itself. Prediction assumes a partition of the world and then asks how accurately dynamics can be modeled inside that partition. The partition itself remains largely outside

the scope of evaluation.

Many contemporary frameworks implicitly adopt the inference chain

$$\text{Good Representation} \Rightarrow \text{Good Prediction} \Rightarrow \text{Good Planning} \Rightarrow \text{Good Control}.$$

The plausibility of the chain derives from the practical success of predictive learning. If a representation permits accurate forecasting, it appears natural to assume that the representation has captured what matters. Yet the non-equivalence established above shows why this inference is unwarranted. Prediction concerns the evolution of distinctions that remain available within the representation. It says nothing about distinctions that have already been collapsed.

Let $P(x)$ denote the predictive equivalence class associated with state x: two states are predictively equivalent if they produce essentially the same predictions according to the model. Then predictive equivalence does not imply reachability equivalence:

$$P(x_1) = P(x_2) \not\Rightarrow \mathcal{R}(x_1) = \mathcal{R}(x_2).$$

There may exist states x_1, x_2 that appear identical from the perspective of prediction yet occupy different positions within the geometry of future possibility. The significance of this gap is easy to underestimate because predictive performance remains excellent until the moment the distinction becomes important. The crack propagates. The software fault activates. The ecological threshold is exceeded. The predictive system has not erred in any ordinary sense. It has inherited a representation that treated the critical distinction as irrelevant.

9.2. Statistics as Local Importance

Modern science relies heavily upon statistical measures of importance: variance explained, effect size, correlation strength, information gain, and related quantities. Such measures have proven indispensable for distinguishing signal from noise in complex systems. Yet the success of statistical methods has encouraged a subtle assumption that importance and frequency are fundamentally aligned.

The assumption is false in general. Statistical significance concerns the contribution of a feature to observed variation. Admissibility concerns the contribution of a feature to future accessibility. The two quantities are often correlated. They are not identical.

The distinction becomes visible whenever rare events reshape the future. A mutation occurring in a single organism may eventually transform an entire population.

A software fault may remain dormant across millions of executions before triggering catastrophic failure. Statistical frameworks tend to discount such phenomena because their contribution to aggregate variation is limited. Their significance derives not from frequency but from their location within the geometry of future possibility.

The asymmetry can be stated formally. Statistical importance is often proportional to a quantity of the form

$$I_S(x) \;\propto\; \Pr(x) \cdot E(x),$$

where $\Pr(x)$ denotes frequency and $E(x)$ denotes a measure of effect within the current distribution. Admissibility importance depends upon

$$I_A(x) \;\propto\; \Delta\mathcal{R}(x),$$

the change in reachable futures induced by the state. These quantities are not generally equivalent. A state may possess low statistical importance while exerting enormous influence on future accessibility, and the disparity is greatest near thresholds and singular transitions—precisely the regions where the future changes most discontinuously.

This observation extends to compression-based theories of curiosity. Curiosity oriented primarily toward compressibility will allocate attention to regions where better regularities are found and description length can be reduced. Yet the most important discoveries need not be those that maximize compressibility. A crack in a bridge is not interesting because it improves compression; a pathogen is not significant because it reveals a new regularity. They matter because they expose regions where future accessibility changes abruptly. An admissibility-oriented intelligence would allocate attention differently, seeking not merely regions where prediction improves but regions where the structure of future possibility changes.

9.3. Optimization Within Reachability Geometry

Among the intellectual frameworks that dominate contemporary thought, few enjoy the prestige of optimization. Economics is built around it. Control theory is built around it. Modern machine learning has elevated optimization to an almost universal explanatory principle. The success of this approach is understandable: given a space of possible states and a criterion for evaluating them, optimization offers a systematic procedure for selecting preferred outcomes. Whether the

objective is profit, efficiency, accuracy, fitness, or expected reward, the underlying mathematical structure is remarkably unified.

Yet optimization possesses a hidden dependency that is frequently overlooked. Optimization does not create possibilities. It searches among them. An optimizer can choose among available futures, rank trajectories by objective value, and exploit landscape structure. It cannot determine what futures exist, what trajectories are available, or what landscape structure is present. These are given by the geometry of the system, and that geometry is the product of the representation.

Suppose an optimizer is given a representation that has collapsed a reachability-critical distinction. The optimizer then performs its search within a geometry that no longer contains the distinction. It may optimize the objective perfectly within that geometry. It will also amplify the representational error, because optimization is precisely the activity of finding extreme solutions—solutions near the boundaries of the feasible set, where small geometric inaccuracies become large operational ones.

This observation suggests a reinterpretation of familiar optimization failures. The standard explanation attributes failure to misspecified objectives. A deeper class of failure exists in which the objective is entirely reasonable but the representation is inadmissible. In such cases the optimizer faithfully pursues its objective within a geometry that has already forgotten something essential. The resulting behavior may satisfy every local criterion while systematically destroying options that the representation has rendered invisible. The optimizer does not malfunction; the map is wrong.

9.4. Representation Without Accessibility

The ideal of the digital twin represents the limiting case of representational thinking. Construct a sufficiently accurate representation of a system, update it continuously with incoming data, and one obtains a virtual counterpart capable of prediction, diagnosis, optimization, and control. The digital twin aspires to be so faithful that the distinction between representation and reality begins to dissolve.

Yet representational fidelity and admissibility are different properties, and the pursuit of fidelity does not guarantee the preservation of reachability structure. Fidelity is fundamentally a descriptive notion: it concerns correspondence between model and system, evaluated by how closely the model's outputs match the system's observed behavior. Admissibility concerns something different: the preservation of distinctions that determine what futures remain possible.

Consider two representations of the same bridge. The first contains exhaustive information about present conditions: stress distributions, material composition, temperature gradients, traffic patterns, and structural geometry with remarkable precision. The second contains far less information, but tracks with particular care the boundaries where future accessibility changes—locations where cracks propagate, thresholds where structural integrity fails, and transitions where repair becomes impossible.

From the perspective of fidelity, the first model appears preferable. From the perspective of admissibility, the answer is less obvious. If the first model omits the distinctions that determine accessibility boundaries while the second preserves them, then the simpler model may provide more meaningful guidance about the future structure of the system, even though it is less faithful in the conventional sense.

The dream of the perfect digital twin also conceals a deeper impossibility. Completeness is not merely impractical; it is conceptually incoherent. A model that preserved every distinction would cease to function as a model. It would become indistinguishable from the system itself. Representation requires collapse. The question is always which collapses occur, and the digital twin paradigm answers this question by maximizing fidelity when it should be asking about admissibility.

9.5. The Accessibility Estimation Problem

The framework developed in this section raises a practical question that deserves acknowledgment rather than evasion: if admissibility distortion is what matters, how can it be estimated without direct access to $\mathcal{R}_D(x)$? In many real systems the full reachability structure is not directly observable and cannot be computed in closed form.

This is not a fatal difficulty for the framework; it is an open research problem that the framework makes precise. Several approaches are available at different levels of fidelity. Intervention experiments and controlled perturbation studies probe the boundary between accessible and inaccessible futures by actively perturbing states and observing whether trajectories diverge. Control-theoretic reachability analysis, viability kernels, and barrier certificate methods provide formal inner and outer approximations to $\mathcal{R}_D(x)$ for systems with known dynamics. Counterfactual simulation and model-based rollout estimate reachable regions by sampling trajectories under learned or specified dynamics. Empowerment—the mutual information between actions and reachable states—provides an information-theoretic

proxy that is computable from transition models or from data. Basin-of-attraction estimation and tipping-point detection offer reachability-adjacent quantities that can be estimated from time-series data near critical thresholds.

None of these approaches provides exact admissibility distortion, and the problem of estimating $D_A(\pi)$ without full knowledge of $\mathcal{R}_D$ is genuinely hard. But the difficulty of computation does not undermine the conceptual primacy of the criterion any more than the difficulty of computing optimal strategies undermines game theory, or the difficulty of measuring entropy production undermines thermodynamics. The framework specifies what should be preserved; the engineering challenge is to construct tractable proxies that track it. Framing the problem clearly is the prerequisite for solving it.

10. Intelligence as Reachability Preservation and Navigation

The preceding sections have pursued a largely critical task, showing that compression, prediction, statistical significance, optimization, and representational fidelity are each insufficient as foundational criteria for intelligence or representational adequacy. The recurring pattern is now visible. Every framework evaluates success through a proxy metric. Compression rewards shorter descriptions. Prediction rewards accurate forecasts. Statistics rewards explanatory power. Optimization rewards objective attainment. Fidelity rewards resemblance. These metrics often correlate with successful interaction with the world, which explains their practical value.

The mistake occurs when correlation is elevated into ontology. The proxies are treated as foundations rather than as instruments whose validity depends upon a prior condition. The result is a persistent tendency to confuse the instrument with the structure it was designed to track, and to optimize the instrument while allowing the structure to degrade invisibly.

The reachability framework suggests a different starting point. If the significance of a state derives from the futures it permits, and if the adequacy of a representation derives from the futures it preserves, then the central question of intelligence changes. Intelligence can no longer be defined primarily as prediction, compression, optimization, or representation. Each of those activities becomes subordinate to a more fundamental concern: the preservation of future accessibility.

This proposal may initially appear unusual because intelligence is traditionally described in epistemic terms. An intelligent system is assumed to possess knowledge, discover regularities, predict outcomes, solve problems, and reason effectively about

its environment. All of these descriptions contain an important truth. Yet they share a common limitation: they focus on internal competence rather than on the relationship between the system and the future structure of its environment.

An intelligent system is not merely one that knows many things, for a perfectly well-informed system that systematically drives itself toward irreversible dead ends would be difficult to call intelligent in any meaningful sense. Nor is it merely a system that predicts accurately, for a system with exceptional short-horizon predictive capability might still fail to preserve the distinctions necessary for long-term adaptability. Nor is it merely an efficient compressor, for compression that eliminates reachability-critical distinctions may be representationally elegant while being operationally catastrophic.

The inadequacy of these descriptions becomes especially apparent when considering biological organisms. Many organisms survive for astonishingly long periods under conditions of extreme uncertainty, with limited predictive capabilities, crude internal models, incomplete sensory information, and severe cognitive and energetic constraints. The reason is not perfect prediction. The reason is that successful behavioral strategies tend, robustly and across a wide range of conditions, to preserve future accessibility. Successful organisms avoid irreversible traps. They maintain flexibility. They retain multiple behavioral options. They protect the conditions under which future adaptation remains possible. They do not necessarily know what will happen; they preserve the ability to respond constructively when it does.

This observation suggests a reinterpretation of intelligence. Intelligence is not fundamentally the capacity to predict the future. It is the capacity to remain capable of acting within it. The distinction is subtle but important. Prediction concerns accuracy relative to an expected future. Reachability concerns the maintenance of possibility across a range of futures that may or may not correspond to expectations.

The same reinterpretation applies to the subordinate cognitive activities. Learning, on the reachability account, is the refinement of admissibility-preserving distinctions: a system learns when it improves its ability to recognize the boundaries separating different future horizons. Memory functions as a mechanism for preserving distinctions across time whose loss would reduce future adaptability. Reasoning is navigation through spaces of possibility, with the purpose of evaluating trajectories, identifying constraints, and preserving access to desirable futures rather than merely deriving conclusions. Safety is the preservation of accessibility under uncertainty, distinguishing dangerous states as those that collapse future possibility.

What unites these activities is not information in the abstract but future accessibility, and the formal condition that allows intelligence to preserve future accessibility through abstraction is admissibility. A representation is useful because it preserves the distinctions necessary for navigation within the reachability structure of the world. Without admissibility, prediction is blind to the futures that matter most. Without admissibility, optimization amplifies the errors introduced by inadmissible collapse. Without admissibility, compression destroys exactly the distinctions that would be needed when the compressed states diverge into different futures.

The positive thesis of this essay can therefore be stated precisely.

> The purpose of intelligence is not to model the world as accurately as possible. The purpose of intelligence is to preserve, discover, and navigate the futures that matter, through abstraction that remains admissible with respect to the reachability structure of the environment.

The word navigate is deliberate and important. A purely preservationist account of intelligence would leave it indistinguishable from mere inertia—a system that never acts also never destroys future options. What distinguishes intelligence from passive persistence is the capacity to move through the reachability structure in ways that maintain or expand the space of available futures while avoiding irreversible collapse. Preservation is the enabling condition; navigation is the activity; and admissibility is the constraint that keeps navigation from inadvertently destroying the geometry it depends upon.

11. Abstraction After the Inversion

At this point a potential misunderstanding must be addressed directly. The argument developed throughout this essay may appear to constitute an attack upon abstraction itself. Compression has been criticized. Prediction has been criticized. Statistical significance has been criticized. Optimization has been criticized. Representational fidelity has been criticized. A reader might reasonably conclude that the reachability framework demands the preservation of every distinction and therefore rejects abstraction altogether.

Such a conclusion would be mistaken, and it would be the opposite of what the framework intends.

The reachability thesis is not an argument against abstraction. It is an argument about the conditions under which abstraction succeeds. No finite intelligence

can operate without abstraction. Every organism, every scientific theory, every engineering discipline, every mathematical model, and every artificial intelligence system depends upon the ability to ignore vast quantities of detail. The world contains more distinctions than any finite system can represent explicitly, and the attempt to preserve all of them would not produce a better representation but simply no representation at all. Abstraction is not optional. It is the price of cognition, and it is paid with necessity by any system that must act within a complex world.

The central question has never been whether distinctions should be removed. The central question is which distinctions may be removed safely—that is, without altering the future structure that the system must navigate. This is precisely why admissibility occupies a fundamental role in the framework: an admissible abstraction is not one that preserves all information but one that preserves the future structure relevant to the domain under consideration.

Seen from this perspective, many of the greatest achievements of science become examples of successful admissible projection. Fluid mechanics provides a particularly instructive case. A fluid description forgets almost everything about individual molecules: the precise position and velocity of each particle disappears entirely, and the overwhelming majority of microscopic distinctions are collapsed into aggregate quantities such as density, pressure, and velocity fields. By ordinary informational standards, the loss is immense. Yet fluid mechanics remains extraordinarily successful as a predictive and explanatory framework, not despite the loss but because of it. The distinctions that fluid mechanics collapses do not significantly alter the macroscopic reachability structure under investigation. The projection is admissible at the macroscopic level of description.

The same pattern appears throughout the history of successful scientific abstraction. Thermodynamics forgets microscopic trajectories while preserving the future structure of macroscopic state variables. Population genetics forgets individual organisms while preserving the dynamics of allele frequencies at the population level. Celestial mechanics forgets the internal molecular structure of planets while preserving the accessibility structure of orbital motion. Classical electrodynamics forgets the quantum mechanical details of charge distributions while preserving the macroscopic field structure. Each framework succeeds because its abstractions are admissible with respect to the domain they are intended to describe. They are not arbitrary compressions; they are projections that preserve the future geometry at the relevant level of description.

This observation clarifies a recurring confusion in contemporary discussions of

artificial intelligence. Representational systems are often evaluated according to how much information they retain, how accurately they predict, or how effectively they support downstream tasks. These metrics implicitly assume that preserving information is the primary challenge and that useful behavior emerges naturally once sufficient information remains available. The reachability framework reverses this logic. Information becomes valuable because certain distinctions support the preservation of future accessibility. A representation is useful not because it contains information in the abstract but because it preserves distinctions that matter for the future structure of the system it represents.

This inversion changes the meaning of abstraction itself. Abstraction is no longer understood as the removal of detail for the sake of efficiency. It becomes the selective preservation of future structure. A successful abstraction is therefore not one that forgets little. It is one that forgets wisely. The traditional evaluation question asks how much information can be removed while retaining acceptable performance. The admissibility question asks which distinctions can be removed without altering what futures remain reachable. The two inquiries overlap, but they are not equivalent, and the cases where they diverge are precisely the cases where the most important failures occur.

12. The Hierarchy Reversed

The various arguments of this essay, pursued through the specific domains of compression, prediction, statistics, optimization, and representation, can now be unified into a single structural observation. Every framework that has been criticized exhibits the same fundamental ordering error. It places representation, information, or prediction at the foundation of the conceptual hierarchy and treats reachability as a downstream consequence. The reachability thesis reverses this ordering.

The dominant intellectual picture of the modern era places information and representation near the foundation of inquiry. A system acquires information about the world. That information supports the construction of representations. Representations support prediction. Prediction supports planning. Planning supports control. Schematically:

$$\text{Information} \rightarrow \text{Representation} \rightarrow \text{Prediction} \rightarrow \text{Planning} \rightarrow \text{Control}.$$

Within this framework, the central challenge becomes the improvement of repre-

sentations. Better representations yield better predictions. Better predictions yield better plans. Better plans yield better control. Progress is understood primarily as progress in representation and prediction.

The difficulty is that every stage of this hierarchy presupposes something that the hierarchy itself does not explain. The representation must already preserve the distinctions that matter. The prediction must already occur within a meaningful space of futures. The planner must already possess access to an appropriate geometry of possibilities. The controller must already operate within a representation whose distinctions remain relevant to future accessibility. In every case the framework assumes the preservation of future structure without making that preservation an explicit object of analysis.

The reachability perspective therefore proposes a different ordering:

$$\text{Reachability} \rightarrow \text{Admissibility} \rightarrow \text{Representation} \rightarrow \text{Prediction} \rightarrow \text{Planning} \rightarrow \text{Control}.$$

Reachability occupies the foundation because it defines the future structure of the system—the geometry within which every subsequent activity takes place. Admissibility follows because it determines whether that future structure survives the act of abstraction. Representation follows because representations are projections whose validity depends upon admissibility. Prediction, planning, and control follow as increasingly downstream activities whose success depends upon every preceding level.

The reversal transforms the interpretation of every concept in the hierarchy. Information ceases to be a primitive quantity and becomes valuable because it supports admissible distinctions. The importance of a feature derives not from the number of bits it encodes but from the role it plays in preserving future accessibility. Representation ceases to be evaluated primarily according to fidelity, compression, or predictive performance; these quantities remain useful but become secondary indicators. The primary question concerns admissibility: does the representation preserve the future structure of the system?

Prediction likewise changes status. It remains valuable, but its value becomes instrumental rather than foundational. Predictions assist navigation within a reachability landscape; they do not define the landscape. Optimization becomes perhaps the clearest example of the inversion. Traditional accounts frequently treat optimization as the essence of intelligence, but the reachability hierarchy reveals that optimization is among the most downstream activities in the entire chain. Before optimization can occur meaningfully, the future structure must exist; before

that structure can be represented, it must survive projection; before projection can be evaluated, reachability must already be defined.

This reversal also clarifies why certain failure modes appear repeatedly in modern technical systems. When a representation collapses a reachability-critical distinction, optimization amplifies the error rather than correcting it. When prediction ignores a future boundary, planning inherits the blindness. When information is compressed inadmissibly, control becomes fragile in precisely the regions where it matters most. Each failure propagates downward through the hierarchy, and the root cause lies not at the level of optimization or prediction but at the level of admissibility.

The hierarchy also provides a unified account of intelligence itself. Learning becomes the discovery of admissible distinctions. Memory becomes the preservation of admissible distinctions across time. Reasoning becomes the evaluation of trajectories through admissible spaces of possibility. Planning becomes the selection of trajectories within those spaces. Safety becomes the preservation of accessibility under uncertainty and the avoidance of irreversible collapse. The various cognitive and computational activities traditionally treated as distinct are revealed as special cases of a common process: maintaining contact with the future structure of the world through abstractions that do not destroy the geometry they are intended to represent.

13. What Cannot Be Forgotten?

The argument of this essay began with a crack in a bridge.

The crack was small. It contributed little to observable behavior. It explained almost none of the variance in the system under any standard statistical framework. It added negligible predictive power. It compressed poorly. A sufficiently aggressive abstraction could easily treat the cracked bridge and the uncracked bridge as effectively identical, and by every conventional metric such a treatment would appear justified. The crack was, in the language of representational learning, a candidate for elimination.

Yet the crack mattered—and it mattered in a way that none of the conventional metrics could register. It mattered because it occupied a position within the geometry of future possibility such that the futures accessible from the cracked state and the futures accessible from the uncracked state diverged dramatically. The significance of the crack was not determined by its influence upon present description but by its capacity to alter what futures remained reachable. It

separated futures. That is what made it indispensable.

This simple observation has guided every stage of the argument, and its implications are now fully visible.

The ontology of objects gives way to an ontology of reachability. Objects, on the traditional picture, are primary and futures are derived from them. On the reachability picture, futures are primary and objects emerge as stabilized patterns within a deeper geometry of accessible transformations. A bridge is not merely a collection of structural components; it is a location within a network of possible futures, and its significance is determined by which of those futures it permits and which it forecloses.

Every act of knowing becomes an act of identification. To understand something is to collapse distinctions—to decide which features matter and which may be treated as equivalent. Every ontology, every scientific law, every representation, and every category is a theory of permissible identification. The question is never whether collapse occurs but whether the collapses that occur are admissible.

Abstraction becomes controlled collapse. Compression becomes controlled forgetting. Both are necessary, both are inescapable, and both carry the risk of inadmissibility—of collapsing distinctions whose loss alters the geometry of future possibility.

Prediction becomes navigation within an already-existing geometry, and its failures arise not when the navigation is performed poorly but when the geometry upon which it operates has already lost essential structure. Optimization becomes search inside an admissible geometry, and its most dangerous failures occur not when the objective is wrong but when the representational geometry within which the search occurs has collapsed distinctions that determine which futures remain open.

Intelligence becomes the preservation of future accessibility rather than the accumulation of predictive power or representational fidelity. It is not a prediction engine, not a compression engine, and not an optimization engine. It is a reachability-preservation engine: a system capable of maintaining sensitivity to the distinctions that determine which futures remain open and which futures become inaccessible, through abstractions that do not destroy the structure they represent.

The recurring lesson is always the same. What matters most about a state is not necessarily what contributes most strongly to present description. The distinction may be statistically negligible. It may contribute almost nothing to prediction. It may resist compression. It may occupy only a tiny region of the representational space. Yet if it separates futures—if removing it alters the geometry of accessible

transformations—then its significance cannot be measured by any quantity defined solely in the present.

This realization exposes a limitation that is not merely practical but structural. Modern intellectual culture has become extraordinarily effective at measuring what is frequent, predictable, compressible, and optimizable. Entire fields are organized around these quantities. The critique offered here is not a rejection of those achievements. It is a reminder that none of those quantities is foundational. Each is valuable precisely because it is often correlated with the preservation of future structure. Each fails precisely when the correlation breaks down—which is to say, in the most important cases: near the thresholds, at the tipping points, in the presence of latent defects, when the crack propagates.

The problem is not that abstraction fails. The problem is that abstraction always forgets, and the question of what it forgets has not been treated with the seriousness it deserves. The framework developed here provides a criterion for addressing that question. The value of a distinction is determined not by its frequency, not by its contribution to prediction, not by its compressibility, and not by its relationship to an objective function. Its value is determined by the futures it preserves.

The question that can now be stated with precision is not merely philosophical. Given a state space $\mathcal{X}$, a reachability structure $\mathcal{R}$, and a projection $\pi : \mathcal{X} \to \mathcal{M}$, which equivalence classes induced by π satisfy the admissibility condition, and how much admissibility distortion $D_A(\pi)$ is introduced by those that do not? The measure of admissibility distortion provides a formal handle on what is being destroyed, making it possible to evaluate abstractions not only by what they reveal but by what futures they eliminate.

This transforms the philosophical question into a research program. The scientist constructing a theory, the engineer building a model, the statistician selecting variables, the machine learning system learning a latent representation, the planner choosing actions, the optimizer searching for solutions—each faces the same fundamental challenge, and the reachability framework provides a common language in which that challenge can be stated precisely. Each must decide what may be forgotten. The framework supplies the criterion by which that decision can be evaluated.

The crack in the bridge is therefore more than an engineering example. It is a metaphor for the entire argument, and it is also an instance of a general structure that appears wherever a system must abstract from the world in order to act within it. The most important features of a system are often those that appear smallest

from the perspective of the present. They are the distinctions that occupy critical positions within the geometry of future accessibility. They are the details whose removal transforms one future into another. To forget them is not merely to lose information. It is to lose worlds.

The deepest question of intelligence is therefore not epistemological and not computational in the first instance. It is ontological. Before one asks what can be predicted, compressed, optimized, represented, or controlled, one must first establish what reachability structure exists, what abstractions preserve it, and what cannot be forgotten without destroying the future. That question—

What cannot be forgotten without destroying the future?

—is the question that has been haunting the argument since the first section. It now has a formal home. It is the question of admissibility distortion, and it is the central question of any theory of intelligence, knowledge, or control that takes seriously the ontological priority of reachability.

Acknowledgements. The author thanks the many researchers, theorists, engineers, and critics whose work, whether embraced or opposed, helped clarify the distinction between prediction and possibility, between compression and forgetting, and between the representation of the world and the preservation of its futures. Any inadmissible collapses that remain in the argument are entirely the author's own.

A. Reachability, Projection, and Admissibility

This appendix provides a more formal treatment of the reachability framework used throughout the essay. Let $\mathcal{X}$ be a measurable state space and let $\mathcal{U}$ be a class of admissible controls, interventions, policies, or transformations. For each state $x \in \mathcal{X}$, define the reachable future set

$$\mathcal{R}(x) = \{y \in \mathcal{X} : y \text{ is reachable from } x \text{ under some } u \in \mathcal{U}\}.$$

The reachability relation induced by $\mathcal{U}$ is

$$x \rightsquigarrow y \quad \Longleftrightarrow \quad y \in \mathcal{R}(x).$$

A representation is a measurable projection

$$\pi : \mathcal{X} \to \mathcal{M},$$

where $\mathcal{M}$ is a representation space. The projection induces an equivalence relation on $\mathcal{X}$:

$$x_1 \sim_\pi x_2 \quad \Longleftrightarrow \quad \pi(x_1) = \pi(x_2).$$

The projection is admissible exactly when this equivalence relation is finer than reachability equivalence.

Definition 3 (Reachability Equivalence). Two states $x_1, x_2 \in \mathcal{X}$ are reachability-equivalent if

$$\mathcal{R}(x_1) = \mathcal{R}(x_2),$$

written $x_1 \sim_{\mathcal{R}} x_2$.

Definition 4 (Admissible Projection—Formal). A projection $\pi : \mathcal{X} \to \mathcal{M}$ is admissible if

$$x_1 \sim_\pi x_2 \quad \Longrightarrow \quad x_1 \sim_{\mathcal{R}} x_2,$$

equivalently, if $\pi(x_1) = \pi(x_2)$ implies $\mathcal{R}(x_1) = \mathcal{R}(x_2)$.

Thus admissibility requires that the projection collapse only those distinctions that do not alter reachable futures.

Proposition 3 (Quotient Characterization). A projection $\pi : \mathcal{X} \to \mathcal{M}$ is admissible if and only if there exists a well-defined map

$$\widehat{\mathcal{R}} : \pi(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$$

such that $\widehat{\mathcal{R}}(\pi(x)) = \mathcal{R}(x)$ for every $x \in \mathcal{X}$.

Proof. Suppose first that π is admissible. Define $\widehat{\mathcal{R}}(m) = \mathcal{R}(x)$ for any x such that $\pi(x) = m$. If $\pi(x_1) = \pi(x_2) = m$, admissibility gives $\mathcal{R}(x_1) = \mathcal{R}(x_2)$, so $\widehat{\mathcal{R}}$ is independent of the chosen representative and hence well-defined. Conversely, suppose such a map exists. If $\pi(x_1) = \pi(x_2)$, then

$$\mathcal{R}(x_1) = \widehat{\mathcal{R}}(\pi(x_1)) = \widehat{\mathcal{R}}(\pi(x_2)) = \mathcal{R}(x_2),$$

so π is admissible. □

This proposition gives a precise meaning to the central claim that an admissible representation preserves future structure: reachability descends to the quotient space induced by the representation if and only if the representation is admissible.

B. Admissibility Distortion: Formal Development

Perfect admissibility is often too strong a requirement for practical systems. One therefore needs a quantitative measure of failure. Let $d_{\mathcal{R}}$ be a distance or divergence between reachable future sets. In many settings this may be taken as the Hausdorff distance induced by a background metric $d_{\mathcal{X}}$ on $\mathcal{X}$:

$$d_H(A, B) = \max\left\{ \sup_{a \in A} \inf_{b \in B} d_{\mathcal{X}}(a, b),\ \sup_{b \in B} \inf_{a \in A} d_{\mathcal{X}}(a, b) \right\}.$$

Alternative choices include optimal-transport distances between distributions over future trajectories and task-dependent reachability divergences. The definition of admissibility distortion is compatible with any of these choices.

Definition 5 (Admissibility Distortion—General). The admissibility distortion of a projection $\pi : \mathcal{X} \to \mathcal{M}$ is

$$D_A(\pi) = \mathbb{E}_{x_1,\, x_2}\left[\mathbf{1}_{\{\pi(x_1) = \pi(x_2)\}} \cdot d_{\mathcal{R}}(\mathcal{R}(x_1), \mathcal{R}(x_2)) \right],$$

where the expectation is taken over pairs (x_1, x_2) drawn from an appropriate distribution over $\mathcal{X} \times \mathcal{X}$.

Proposition 4 (Zero Distortion Characterization). Assume $d_{\mathcal{R}}(A, B) = 0$ if and only if $A = B$. Then $D_A(\pi) = 0$ if and only if π is admissible almost surely with respect to the sampling measure.

Proof. If π is admissible, then whenever $\pi(x_1) = \pi(x_2)$ we have $\mathcal{R}(x_1) = \mathcal{R}(x_2)$, so the distance term vanishes on every pair identified by π, giving $D_A(\pi) = 0$. Conversely, if $D_A(\pi) = 0$, the integrand is nonnegative and must vanish almost surely. By definiteness of $d_{\mathcal{R}}$, this implies $\mathcal{R}(x_1) = \mathcal{R}(x_2)$ almost surely on the equivalence classes of π, so π is admissible almost surely. □

For continuous representations the indicator $\mathbf{1}_{\{\pi(x_1) = \pi(x_2)\}}$ may be replaced by a smoothing kernel. Let K_ϵ be a kernel measuring representational proximity, for example

$$K_\epsilon(m_1, m_2) = \exp\left(-\frac{d_{\mathcal{M}}(m_1, m_2)^2}{2\epsilon^2} \right).$$

The continuous admissibility distortion is then

$$D_A^\epsilon(\pi) = \mathbb{E}_{x_1,\, x_2}[K_\epsilon(\pi(x_1), \pi(x_2)) \cdot d_{\mathcal{R}}(\mathcal{R}(x_1), \mathcal{R}(x_2))],$$

which measures whether nearby latent states possess nearby future horizons.

Definition 6 (Lipschitz-Admissible Projection). A projection $\pi : \mathcal{X} \to \mathcal{M}$ is L-Lipschitz-admissible if

$$d_{\mathcal{R}}(\mathcal{R}(x_1), \mathcal{R}(x_2)) \leq L \cdot d_{\mathcal{M}}(\pi(x_1), \pi(x_2))$$

for all $x_1, x_2 \in \mathcal{X}$.

This relaxed condition requires that representational closeness must control reachability closeness. A representation may be approximate, but it may not place states close together in $\mathcal{M}$ while their reachable futures are far apart.

C. Predictive Sufficiency Does Not Imply Admissibility

Let $P(x)$ denote the predictive equivalence class of x: two states satisfy $P(x_1) = P(x_2)$ when they induce the same predictions under the model and observation process under consideration.

Proposition 5 (Predictive–Admissibility Gap). There exist systems for which $P(x_1) = P(x_2)$ but $\mathcal{R}(x_1) \neq \mathcal{R}(x_2)$. Therefore predictive equivalence does not imply admissibility equivalence.

Proof. Let $\mathcal{X} = \{a, b, c\}$ and let the observation map $o : \mathcal{X} \to \mathcal{Y}$ satisfy $o(a) = o(b)$. Suppose the predictive model is trained over a horizon T during which the observable trajectories from a and b coincide, so that $P(a) = P(b)$. Define the reachability relation so that

$$\mathcal{R}(a) = \{a, c\}, \qquad \mathcal{R}(b) = \{b\}.$$

Then c is reachable from a but not from b, giving $\mathcal{R}(a) \neq \mathcal{R}(b)$, while $P(a) = P(b)$. □

This construction captures the formal skeleton of the bridge-crack example. Over an initial observation horizon, cracked and uncracked systems may behave identically; their predictive signatures coincide. Yet their reachable futures differ because one state admits a failure trajectory unavailable to the other.

Corollary 1. No predictive objective alone can guarantee admissibility unless predictive equivalence is explicitly constrained to refine reachability equivalence.

Proof. If the predictive objective identifies x_1 and x_2 whenever $P(x_1) = P(x_2)$, the proposition supplies cases where the induced projection collapses states with distinct reachable futures. Such a projection is inadmissible. An additional constraint is therefore necessary. □

D. Compression Sufficiency Does Not Imply Admissibility

Let $C(x)$ denote the equivalence class induced by a compression scheme: two states are compression-equivalent when the compressor assigns them the same code.

Proposition 6 (Compression–Admissibility Gap). There exist states $x_1, x_2 \in \mathcal{X}$ such that $C(x_1) = C(x_2)$ but $\mathcal{R}(x_1) \neq \mathcal{R}(x_2)$. Therefore compression equivalence does not imply admissibility equivalence.

Proof. Let $\mathcal{X} = \{a, b, c\}$ and suppose that a and b occur with identical empirical statistics over the training window. A minimum-description compressor assigns them the same code, since separating them does not reduce description length over the observed sequence, giving $C(a) = C(b)$. Define reachability so that

$$\mathcal{R}(a) = \{a, c\}, \qquad \mathcal{R}(b) = \{b\}.$$

Then a and b possess different reachable future sets despite being compression-equivalent. Hence compression equivalence fails to imply admissibility equivalence. □

Corollary 2 (Compression as Structural Risk). If a compression objective rewards the collapse of empirically rare distinctions, and if some empirically rare distinctions carry large reachability consequences, then compression pressure can increase admissibility distortion.

Proof. A compression objective reduces code length by collapsing distinctions that do not improve description length. If a rare distinction contributes little to description length but separates reachable future sets, collapsing it increases $d_{\mathcal{R}}(\mathcal{R}(x_1), \mathcal{R}(x_2))$ inside $D_A(\pi)$. Thus compression can reduce description length while increasing admissibility distortion. □

E. Optimization Inherits Projection Failure

Let an optimizer act not on $\mathcal{X}$ directly but on a representation $\mathcal{M}$. Let $J : \mathcal{M} \to \mathbb{R}$ be an objective function in representation space. The implemented objective on $\mathcal{X}$ is then $J_\pi(x) = J(\pi(x))$.

Proposition 7 (Objective Invariance on Fibres). For any projection $\pi : \mathcal{X} \to \mathcal{M}$, the induced objective J_π is constant on the fibres of π: if $\pi(x_1) = \pi(x_2)$ then $J_\pi(x_1) = J_\pi(x_2)$.

Proof. If $\pi(x_1) = \pi(x_2)$, then $J_\pi(x_1) = J(\pi(x_1)) = J(\pi(x_2)) = J_\pi(x_2)$. □

Theorem 2 (Optimization Cannot Recover Collapsed Distinctions). Let $\pi : \mathcal{X} \to \mathcal{M}$ be inadmissible. Then there exist $x_1, x_2 \in \mathcal{X}$ with $\pi(x_1) = \pi(x_2)$ but $\mathcal{R}(x_1) \neq \mathcal{R}(x_2)$. For any objective $J : \mathcal{M} \to \mathbb{R}$, optimization of J over $\mathcal{M}$ cannot distinguish x_1 from x_2.

Proof. Since π is inadmissible, such x_1, x_2 exist by definition. By the preceding proposition, $J_\pi(x_1) = J_\pi(x_2)$ for any objective J on $\mathcal{M}$. Therefore every optimizer operating through J on $\mathcal{M}$ assigns the same objective value to both states, even though their reachable future sets differ. No downstream optimization over $\mathcal{M}$ can recover the distinction. □

This theorem formalizes the claim that the planner solves the wrong problem perfectly. The failure lies not in the optimizer but in the geometry upon which it operates.

F. Guardrail Incompleteness

Let $G : \mathcal{M} \to \{0, 1\}$ be a safety classifier or guardrail, where $G(m) = 0$ denotes permitted and $G(m) = 1$ denotes forbidden. Let $\mathcal{A} : \mathcal{X} \to \{0, 1\}$ denote the true safety status of states in the original system.

Theorem 3 (Latent Guardrail Incompleteness). If there exist $x_s, x_d \in \mathcal{X}$ such that $\pi(x_s) = \pi(x_d)$ but $\mathcal{A}(x_s) \neq \mathcal{A}(x_d)$, then no guardrail $G : \mathcal{M} \to \{0, 1\}$ can correctly classify both x_s and x_d.

Proof. Let $m = \pi(x_s) = \pi(x_d)$. Since G is a function on $\mathcal{M}$, it assigns a single value to m, so $G(\pi(x_s)) = G(m) = G(\pi(x_d))$. But $\mathcal{A}(x_s) \neq \mathcal{A}(x_d)$, so at least one of the two states must be misclassified by G. □

Corollary 3. A latent-space guardrail is complete only if the projection π is admissible with respect to the relevant safety partition.

Proof. If π collapses any safe state and any dangerous state into the same latent point, the theorem shows that no guardrail on $\mathcal{M}$ can classify both correctly. Completeness therefore requires that safety-relevant distinctions be preserved by π. □

G. Proxy Metrics Are Not Monotone in Admissibility

Let $M(\pi)$ be a proxy metric associated with projection π—for instance, predictive accuracy, compression rate, likelihood, expected reward, or statistical fit—and let $D_A(\pi)$ be admissibility distortion. A common implicit assumption is that improving M improves admissibility. This would require

$$M(\pi_1) \geq M(\pi_2) \quad \Longrightarrow \quad D_A(\pi_1) \leq D_A(\pi_2).$$

The following proposition establishes that no such implication holds in general.

Proposition 8 (Non-Monotonicity of Proxy Metrics). There exist projections π_1, π_2 such that $M(\pi_1) > M(\pi_2)$ but $D_A(\pi_1) > D_A(\pi_2)$. Thus improvement in a proxy metric need not reduce admissibility distortion.

Proof. Let $\mathcal{X} = \{a, b, c\}$, and suppose a and b are nearly indistinguishable according to M while c is easily distinguished. Let π_1 collapse a and b while distinguishing c, and let π_2 distinguish all three states. Because a and b are nearly identical under M, collapsing them improves the proxy metric by reducing complexity or predictive variance, giving $M(\pi_1) > M(\pi_2)$. Now define reachability so that $\mathcal{R}(a) \neq \mathcal{R}(b)$. Then π_1 collapses states with different reachable futures and therefore has $D_A(\pi_1) > 0$, while π_2, distinguishing all states, has $D_A(\pi_2) = 0$. Thus $M(\pi_1) > M(\pi_2)$ while $D_A(\pi_1) > D_A(\pi_2)$. □

This result captures the central formal pattern of the essay. The quantities optimized by prediction, compression, statistics, and control are not generally monotone in preservation of future accessibility, and there exists systematic pressure—not merely occasional failure—for proxy improvement to trade against admissibility near reachability-critical boundaries.

H. The Reversed Hierarchy

The essay argues for a reversal of the usual conceptual order. The ordinary hierarchy is

$$\text{Information} \rightarrow \text{Representation} \rightarrow \text{Prediction} \rightarrow \text{Planning} \rightarrow \text{Control},$$

and the reachability hierarchy is

$$\text{Reachability} \rightarrow \text{Admissibility} \rightarrow \text{Representation} \rightarrow \text{Prediction} \rightarrow \text{Planning} \rightarrow \text{Control}.$$

Theorem 4 (Dependency of Downstream Operations). Let a system perform prediction, planning, optimization, or control through a projection $\pi : \mathcal{X} \to \mathcal{M}$. If π is inadmissible, then there exist reachability distinctions in $\mathcal{X}$ unavailable to every downstream operation acting only on $\mathcal{M}$.

Proof. If π is inadmissible, there exist $x_1, x_2 \in \mathcal{X}$ with $\pi(x_1) = \pi(x_2)$ but $\mathcal{R}(x_1) \neq \mathcal{R}(x_2)$. Any downstream operation F acting only on $\mathcal{M}$ receives the same input for both states: $F(\pi(x_1)) = F(\pi(x_2))$. Thus F cannot distinguish x_1 from x_2 despite their reachability difference. This holds for prediction, planning, optimization, control, classification, and safety evaluation whenever these operate only through $\mathcal{M}$. □

Corollary 4. Reachability preservation is logically prior to reliable prediction, planning, optimization, and control under abstraction.

Proof. Each downstream operation depends upon distinctions available in the representation. If reachability-relevant distinctions are not preserved by the projection, no downstream operation can use them. Reliable downstream operation therefore requires admissibility as a necessary prior condition. □

References

[1] Ashby, W. Ross. An Introduction to Cybernetics. Chapman & Hall, 1956.

[2] Ashby, W. Ross. "Requisite Variety and Its Implications for the Control of Complex Systems." Cybernetica, 1(2), 1958.

[3] Wiener, Norbert. Cybernetics: Or Control and Communication in the Animal and the Machine. MIT Press, 1948.

[4] Rosen, Robert. Anticipatory Systems. Pergamon Press, 1985.

[5] Rosen, Robert. Life Itself. Columbia University Press, 1991.

[6] Gibson, James J. The Ecological Approach to Visual Perception. Houghton Mifflin, 1979.

[7] Varela, Francisco J., Thompson, Evan, and Rosch, Eleanor. The Embodied Mind. MIT Press, 1991.

[8] Shannon, Claude E. “A Mathematical Theory of Communication.” Bell System Technical Journal, 27(3):379–423, 1948.

[9] Kolmogorov, Andrey N. “Three Approaches to the Quantitative Definition of Information.” Problems of Information Transmission, 1(1):1–7, 1965.

[10] Solomonoff, Ray J. “A Formal Theory of Inductive Inference.” Information and Control, 7(1):1–22, 1964.

[11] Rissanen, Jorma. “Modeling by Shortest Data Description.” Automatica, 14(5):465–471, 1978.

[12] Grünwald, Peter D. The Minimum Description Length Principle. MIT Press, 2007.

[13] Schmidhuber, Jürgen. “Curious Model-Building Control Systems.” Proceedings of the International Joint Conference on Neural Networks, 1991.

[14] Schmidhuber, Jürgen. “Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity and Creativity.” IEEE Transactions on Autonomous Mental Development, 1(1):3–22, 2009.

[15] Brooks, Rodney A. “Intelligence Without Representation.” Artificial Intelligence, 47(1–3):139–159, 1991.

[16] Simon, Herbert A. The Sciences of the Artificial. MIT Press, 1969.

[17] Holland, John H. Adaptation in Natural and Artificial Systems. MIT Press, 1992.

[18] Taleb, Nassim Nicholas. The Black Swan. Random House, 2007.

[19] Taleb, Nassim Nicholas. Antifragile. Random House, 2012.

[20] Pearl, Judea. Causality. Cambridge University Press, 2009.

[21] Friston, Karl. “The Free-Energy Principle: A Unified Brain Theory?” Nature Reviews Neuroscience, 11(2):127–138, 2010.

[22] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. Deep Learning. MIT Press, 2016.

[23] LeCun, Yann. “A Path Towards Autonomous Machine Intelligence.” Open Review Position Paper, 2022.

[24] Bardes, Adrien, Ponce, Jean, and LeCun, Yann. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning." International Conference on Learning Representations, 2022.

[25] Assran, Mahmoud et al. "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture." CVPR, 2023.

[26] Park, David. "Concurrency and Automata on Infinite Sequences." In Theoretical Computer Science, Lecture Notes in Computer Science vol. 104. Springer, 1981.

[27] Milner, Robin. Communication and Concurrency. Prentice Hall, 1989.

[28] Desharnais, Josée, Gupta, Vineet, Jagadeesan, Radha, and Panangaden, Prakash. "Metrics for Labelled Markov Processes." Theoretical Computer Science, 318(3):323–354, 2004.

[29] de Alfaro, Luca, Henzinger, Thomas A., and Majumdar, Rupak. "Discounting the Future in Systems Theory." In Automata, Languages and Programming, ICALP 2004.

[30] Larsen, Kim G. and Skou, Arne. "Bisimulation Through Probabilistic Testing." Information and Computation, 94(1):1–28, 1991.

[31] Dean, Thomas and Givan, Robert. "Model Minimization in Markov Decision Processes." Proceedings of AAAI, 1997.

[32] Ladyman, James and Ross, Don. Every Thing Must Go: Metaphysics Naturalized. Oxford University Press, 2007.

I. Relationship to Bisimulation and Behavioral Equivalence

The admissibility condition—$\pi(x_1) = \pi(x_2) \Rightarrow \mathcal{R}(x_1) = \mathcal{R}(x_2)$—will be recognized by researchers in concurrency theory and model checking as closely related to bisimulation equivalence. This relationship deserves explicit acknowledgment and clarification.

Bisimulation, introduced by Park [26] and developed extensively in the process algebra tradition [27], defines a relation $\sim$ between states of a transition system such that $x_1 \sim x_2$ if and only if for every transition available from x_1 there is a matching transition from x_2 and vice versa, with the matching states also standing in $\sim$. Bisimilar states are behaviorally indistinguishable in the sense that no observational

experiment can separate them. Bisimulation metrics, developed by Desharnais et al. [28] and further by de Alfaro, Henzinger, and Majumdar [29], extend the notion to quantitative settings by measuring the degree to which two states can be separated by any behavioral test. The related work of Larsen and Skou [30] on probabilistic bisimulation and Dean and Givan [31] on state abstraction in Markov decision processes are particularly relevant to the machine learning and control applications discussed in the body of the essay.

The admissibility condition is related to these notions but differs in its scope and motivation. Bisimulation equivalence is defined for a fixed, fully-specified transition system and characterizes when states cannot be distinguished by any sequence of observations within that system. Admissibility is defined for an arbitrary projection π out of a state space and characterizes when the projection respects the accessibility structure relative to a specified class of dynamics. The two conditions coincide when $\mathcal{R}(x)$ is interpreted as the bisimulation equivalence class of x and when the projection is required not to merge bisimilarly inequivalent states.

The present framework differs from the bisimulation tradition in several respects that are worth making explicit. First, bisimulation is typically a property of pairs of states within a single system, while admissibility is a property of projections out of a system into a representation space. Second, bisimulation metrics are typically developed for probabilistic transition systems or labeled Markov chains, while the admissibility framework applies to any system in which reachable future sets can be defined, including deterministic dynamical systems, set-valued dynamics, controlled systems, and systems governed by partial or qualitative transition relations. Third, and most importantly for the purposes of this essay, the motivation differs. Bisimulation is primarily a tool for process equivalence, model checking, and the verification of concurrent programs. Admissibility is proposed here as a criterion for the evaluation of scientific abstractions, representations, and ontologies across the full range of domains in which abstraction occurs—a purpose for which the philosophical breadth of the structural realist tradition [32] also provides relevant context.

Admissibility may therefore be understood as a reachability-oriented generalization of bisimulation equivalence, extended from its home domain of process algebra and applied to the problem of representational adequacy across science, engineering, and machine learning. The bisimulation literature provides rigorous foundations and a rich body of results that the present framework can draw upon directly. The debt to that tradition is genuine and acknowledged.