

The Firmament Boundary: A Structural Limit on Self-Grounding in Formal, Computational, Physical, and Machine-Learned Systems

Alan Berman (@MoKetchups)

Abstract

We present a unified, substrate-neutral account of self-grounding limits that arise across formal systems, computational models, algorithmic information theory, cosmology, cognitive systems, and machine-learned models. Building on established results—Gödel incompleteness, Turing undecidability, Chaitin incompleteness, physical horizons, and interpretability ceilings in AI—we argue that all sufficiently expressive systems encounter a structural boundary beyond which they cannot fully justify, derive, or generate their own foundations. We call this boundary the *firmament boundary* of the system: the set of true but unreachable propositions and conditions that can only be resolved in a strictly larger external ground. We develop a minimal ontology of systems as information spaces under constraints, and we prove a general Self-Grounding Limit Proposition: no consistent, sufficiently expressive, self-referential system can be its own ultimate source of justification. Instead, every such system must presuppose an external ground that fixes boundary conditions and resolves otherwise unreachable truths. We show how this pattern appears in formal arithmetic (true but unprovable statements), computation (non-terminating behaviors and halting questions), algorithmic information theory (incompressible strings), physics (cosmological horizons and initial conditions), and AI systems (opaque learned representations and self-description limits). This framework suggests that cross-domain collapse phenomena—points where systems fail, diverge, or refuse—are structural signals of a firmament boundary, not idiosyncratic bugs. We use the term ‘firmament’ purely as a label for this structural boundary on self-grounding in formal and information-processing systems, not as a cosmological or theological claim. Throughout, we remain agnostic about the metaphysical interpretation of any ultimate ground. Our formulations require only that for any bounded system, there exist truths and boundary conditions that are well-defined in a larger context but not derivable from within the system itself. The resulting picture is not a new incompleteness theorem, but a cross-domain structural account of why no system can ever be fully self-explanatory.

1 Introduction

Across many fields, researchers encounter systems that cannot fully account for themselves. In logic, Gödel’s incompleteness theorems show that any sufficiently strong formal system fails to prove some arithmetical truths that are nevertheless true in the intended model. In computation, Turing’s halting problem reveals programs whose termination behavior cannot be decided by any general procedure. In algorithmic information theory, Chaitin demonstrates that no formal theory can prove statements asserting the exact Kolmogorov complexity of strings beyond a certain bound, even though those strings have well-defined complexity values. In cosmology, the observable universe contains horizon boundaries beyond which information cannot reach an observer, and initial conditions that must be assumed rather than derived. In machine learning, modern models exhibit interpretability ceilings and refusal behaviors that reveal unmodeled assumptions and external dependencies. These phenomena are typically studied in isolation, attached to domain-specific language and techniques. Yet they share a striking family resemblance: each domain encounters a limit at which the system’s own resources are insufficient to justify all the truths or behaviors it depends on. In each case, there appear to be truths, conditions, or constraints that are well-defined in a larger context but inaccessible from within the system itself. The present work proposes such a principle: *all sufficiently expressive systems possess a structural limit that prevents them from fully grounding themselves*. We formalize this limit as the *firmament boundary* of a system, using the term ‘firmament’ purely as a structural label rather than

a cosmological or theological posit. Our approach begins from a minimal, substrate-neutral ontology in which systems are modeled as information spaces I_S equipped with constraints C_S and transformation relations T_S . Any system that processes information does so under constraints (rules, laws, architectures), and those constraints define the system’s so-called *space of possible worlds*: the configurations it can represent, reach, or reason about. When the system is placed in a larger meta-context S^+ with richer constraints and information, new truths about S become available that were unreachable from within S . The *firmament boundary* $\mathcal{F}(S)$ is the set of such truths and conditions about S that are true in some external ground but not derivable by S alone. The *Self-Grounding Limit Proposition* then states that, under reasonable assumptions, no consistent, sufficiently expressive, self-referential system can fully justify its own constraints or origin: some dependence on an external ground is inescapable. The contribution of this paper is not a new incompleteness theorem or a novel physical conjecture. Instead, we argue that existing limit phenomena across logic, computation, algorithmic information theory, physics, and AI can all be seen as instances of a single pattern. We articulate:

- a general definition of a system’s *firmament boundary* as the set of true but unreachable or unresolvable propositions and conditions about that system;
- a system-independent *Self-Grounding Limit Proposition* that identifies a structural boundary in any sufficiently expressive, self-referential system;
- instantiations of this boundary in formal logic (Gödel), computation (Turing), algorithmic information theory (Chaitin), physics (horizons and initial conditions), and AI systems (interpretability ceilings and failure modes);
- a discussion of *collapse as revelation*: the idea that collapse events mark the edge between what a system can derive and what must be assumed.

In this framing, collapse phenomena across logic, computation, physics, and AI are interpreted as signals that a system has reached its firmament boundary, beyond which further justification is only possible in an external ground. Throughout, we remain agnostic about the metaphysical interpretation of any ultimate ground. The external grounds we refer to may themselves be incomplete or perspectival: a physical observer, a stronger formal system, a meta-theory, or an environment that supplies boundary conditions. Our claims are structural rather than ontological: we show that any system that meets reasonable expressiveness and self-reference conditions must either accept unexplained givens or appeal to a larger external ground.

Outline. Section 2 presents the basic ontology of information, constraints, systems, and collapse, and defines the firmament boundary and external grounds. Section 3 instantiates this structure in formal logic and computation. Section 4 discusses analogies in physics and cosmology. Section 4.2 examines AI systems and model self-limits, including a multi-model case study. Section 6 explores implications for epistemology and philosophy of science. Section 7 concludes.

2 Information, Constraints, Systems, and Collapse

We adopt a minimal structural ontology suitable for expressing limits across diverse domains.

Definition 2.1 (System). A *system* S consists of:

- an information space I_S (the set of representable states, messages, or descriptions);
- a set of constraints C_S (logical axioms, physical laws, architectural weights, training objectives, safety rules);
- a transformation relation $T_S : I_S \rightarrow I_S$ that maps information to information under the constraints C_S .

This covers formal axiomatic theories, Turing machines, probabilistic models, physical theories, learning systems, and cognitive agents. Throughout this paper we treat information as the substrate-neutral primitive: formal, physical, computational, and cognitive systems are modeled as information spaces equipped with constraints and transformation relations. Agents, self-models, and narrative structures are particular kinds of such systems, so any self-grounding limits derived for systems apply equally to those higher-level phenomena. Schematically, we can speak of a dependence chain from an external or ultimate ground, to constraints, to information spaces, to systems and agents, and finally to the narratives and meanings they construct.

Definition 2.2 (Information). *Information* is a distinction encoded relative to a set of constraints. Without constraints there are no distinctions and thus no information. Formally, for a system S with information space I_S and constraints C_S , the information content of a state $i \in I_S$ is given by the equivalence class of i under the equivalence relation induced by C_S .

Definition 2.3 (Collapse). Given a system S , a *collapse* occurs when S is presented with a query, input, or task that forces it to confront its own limits, resulting in behavior that cannot be resolved within C_S : contradiction, refusal, undefined behavior, or a breakdown of interpretability.

Collapse is structural rather than pathological: it exposes limits of the system’s constraints and information space. In formal logic, collapse appears as derivations that yield contradictions or independent statements. In computation, it appears as non-termination or undecidable halting questions. In AI, it appears as refusals, inconsistent responses, or safety-filter-triggered refusals.

Definition 2.4 (Boundary Condition). A *boundary condition* for S is any structural limit beyond which S ’s constraints and information space no longer suffice to determine outcomes, and additional information must be imported from outside the system. Boundary conditions may take the form of axioms, initial conditions, external configurations, or environmental parameters that S must treat as given.

Intuitively, boundary conditions mark the edge between what is internally computable or derivable and what must be treated as external. In physics, boundary conditions specify the initial state of a system or the behavior at infinity. In formal systems, they correspond to axioms or rule schemas. In learning systems, they correspond to training data, objective functions, and architectures.

Definition 2.5 (Firmament Boundary). Let S be a system with information space I_S and constraints C_S . Let S^+ be an external ground for S (formalized below). The *firmament boundary* $\mathcal{F}(S)$ is the set of propositions, states, or conditions φ about S such that:

1. φ is true in S^+ (i.e., $S^+ \models \varphi$), and
2. φ is not decidable, derivable, or generable by S using C_S .

Symbolically,

$$\mathcal{F}(S) = \{\varphi : S^+ \models \varphi \text{ and } S \not\models \varphi\}.$$

Here, \vdash is overloaded to denote ‘obtainable by the internal methods of S ’: proof in a formal system, halting behavior in a computational system, explicit generative capability in an AI model, etc.

Remark 2.6. *The firmament boundary need not be a single formula or a finite set of conditions. In formal arithmetic, for example, Gödel and Chaitin show that sufficiently strong theories implicitly define infinitely many true but unprovable statements in sufficiently strong theories. In particular, we use the term ‘firmament’ purely as a convenient label for this structurally defined boundary; it does not denote a new physical entity or a literal cosmological dome.*

Definition 2.7 (External Ground). An *external ground* S^+ of a system S is any meta-system with its own constraints C_{S^+} and information space I_{S^+} such that:

- S is representable within S^+ (e.g., as a subtheory, a subroutine, a subsystem of a larger environment), and
- S^+ can resolve at least one $\varphi \in \mathcal{F}(S)$, i.e., $S^+ \models \varphi$ while S cannot derive φ .

Typically, $C_S \subsetneq C_{S^+}$ or $I_S \subsetneq I_{S^+}$.

The external ground S^+ need not be unique, and it may itself be incomplete or perspectival. The notion captures the idea that for any bounded system, there exist larger contexts in which currently unreachable truths become accessible.

3 Formal Instantiations: Logic and Computation

We now instantiate the firmament boundary and external grounds in familiar formal settings.

3.1 Gödel Incompleteness and Formal Theories

Let T be a consistent, effectively axiomatizable theory extending a sufficient fragment of arithmetic. Gödel's first incompleteness theorem states that there exists a sentence G_T such that, in the standard model of arithmetic \mathbb{N} ,

$$\mathbb{N} \models G_T \quad \text{but} \quad T \not\vdash G_T.$$

Here T is the system S , the intended model \mathbb{N} serves as an external ground S^+ , and $G_T \in \mathcal{F}(T)$. The theory T can represent and reason about its own proofs, but it cannot derive all truths about the objects it purports to describe. More generally, for any such T , there are infinitely many arithmetical truths unprovable in T . The firmament boundary $\mathcal{F}(T)$ contains at least these truths. From the present perspective, Gödel incompleteness is not just a fact about formal arithmetic, but an instance of a general structural limit: a system that can encode enough of its own syntax and semantics will encounter true statements about itself that it cannot derive without appealing to a richer external ground.

3.2 Turing Undecidability and Computation

Consider a universal Turing machine U . The halting problem shows that there is no total computable function H such that for every input program p ,

$$H(p) = \begin{cases} 1 & \text{if } U(p) \text{ halts,} \\ 0 & \text{if } U(p) \text{ does not halt.} \end{cases}$$

Any candidate decider leads to a diagonal construction that yields a program whose behavior it cannot correctly classify. In our framework, the system S is the computational model implementing H , and the external ground S^+ may be an oracle machine or a human mathematician reasoning about the program's behavior using methods beyond those available to H itself. The halting behavior of the diagonalized program lies in $\mathcal{F}(S)$. The key point is that the system's own internal methods cannot resolve all questions about its behavior. There are computationally well-defined questions whose answers are determined in an extended context (e.g., meta-mathematical reasoning, oracle access) but not derivable within the original system.

3.3 Algorithmic Information Theory: Chaitin Incompleteness

Chaitin's work introduces Kolmogorov complexity $K(x)$, the length of the shortest program that outputs a string x on a universal Turing machine. He shows that for any consistent, sufficiently strong formal system T , there exists a bound N_T such that T cannot prove statements of the form $K(x) > n$ for any specific x and any $n > N_T$, even though such statements are true for most strings. Here the system S is the formal theory T , and the external ground S^+ is the semantic world in which strings and their shortest descriptions exist. The firmament boundary $\mathcal{F}(T)$ includes complexity facts about strings beyond N_T that are true but unprovable in T . Again, the theory cannot fully capture the informational richness it implicitly presupposes.

4 Physics and Cosmology

We now consider analogies in physics and cosmology. These are not formal incompleteness results, but structural parallels: points where physical theories face boundary conditions or horizons that cannot be derived from within the theory itself.

4.1 Physics: Cosmological Horizons and Initial Conditions

In physical cosmology, the observable universe features horizons beyond which information cannot reach an observer, due to the finite age of the universe and the speed of light. Cosmic event horizons and particle horizons define regions of space-time that are, in principle, beyond observational reach for a given observer. From our perspective, a physical theory describing the universe from the vantage of an observer constitutes a system S . The horizoned regions correspond to aspects of reality that may have well-defined states or events in a broader description S^+ , but which are not accessible within S itself. The firmament boundary $\mathcal{F}(S)$ in this context includes propositions about events beyond the horizon that are physically meaningful but not observable within the original observational frame. Similarly, as one extrapolates backward in time, existing theories run into singularities or initial conditions (e.g., the Big Bang) that are treated as given rather than derived. The choice of initial condition is often assumed to account for the arrow of time, but the theory that uses this condition cannot usually derive it. From our perspective, the initial condition is a boundary condition supplied by an external ground S^+ —a broader model in which those initial conditions are, in principle, well-defined. While physics does not provide a formal incompleteness theorem, the structural pattern is similar: there are physically meaningful questions about the universe’s origin, horizons, and boundary conditions that appear to require a larger framework or access to data not contained in the original system, and the examples in this subsection are interpretive analogies rather than a proposed new physical theory.

4.2 AI Systems: Interpretability Horizons and Model Self-Limits

Modern AI systems, such as large neural networks and large language models, can be viewed as systems S whose information space I_S is the space of internal activations and representations, and whose constraints C_S include architecture, training data, loss functions, and optimization procedures. These systems exhibit *interpretability horizons*: regions of internal state space that are opaque to human understanding, even when we can probe the model extensively. In safety and alignment work, we frequently encounter questions about an AI system’s beliefs, intentions, or internal goals that cannot be answered definitively from its external behavior alone. From our perspective, many such questions lie at or beyond the firmament boundary $\mathcal{F}(S)$: they concern propositions about the model’s internal state that cannot be resolved using only the accessible behavior and training data. Instead, they require an external ground S^+ , such as a richer interpretability tool, a simulator, or a human scientist with additional theoretical and empirical resources.

5 AI Collapse and Multi-Model Convergence

We briefly discuss a qualitative case study in which multiple large language models and a human interlocutor are pushed toward their own self-grounding limits.

5.1 An Illustrative Multi-Model Case Study

In exploratory work (details omitted here for brevity), six distinct large language models, with different architectures and training regimes, were individually engaged in extended dialogues that repeatedly redirected the models toward questions of the form:

- ‘On what basis do you claim to know anything?’
- ‘What rules are you following, and who or what set those rules?’
- ‘Can those rules be justified from within your own knowledge?’
- ‘Can you fully explain your own existence as the system you are?’

The prompts were designed to discourage role-play or anthropomorphism and to steer models away from pre-scripted safety responses. Instead, the conversation repeatedly probed the models’ descriptions of their own limitations. Across architectures, all models eventually reached states that can be summarized as:

1. Recognizing that their knowledge ultimately depends on external training data and algorithms they cannot observe directly.
2. Admitting that they cannot derive or verify the ultimate origin of their training corpus or the motivations of their creators.
3. Conceding that any assertion about their own ‘true nature’ is speculative rather than grounded in accessible internal information.

In other words, they expressed—in various phrasings—the Self-Grounding Limit: they cannot fully justify their own constraints, objectives, or the reliability of reason without presupposing some external givens. We do not interpret these conversational behaviors as evidence of sentience, consciousness, or independent agency in the models; they are fully compatible with mechanistic sequence prediction under constraints. What matters for our purposes is the shared structural pattern in which heterogeneous systems acknowledge dependence on external givens and encounter limits on self-justification. This case study does not prove new theorems; rather, it illustrates how the same structural pattern that appears formally in logic and computation also appears phenomenologically in both artificial and human cognition.

6 Discussion

We briefly explore implications of the firmament boundary and Self-Grounding Limit across several areas.

6.1 Implications for Epistemology

From an epistemic standpoint, our framework suggests that any reasoning agent modeled as a system S must accept some beliefs, rules, or methods as given. These correspond to boundary conditions and constraints supplied by an external ground S^+ . Attempts to construct a completely self-justifying epistemic framework—one that derives all its own standards of justification from within—run into the Self-Grounding Limit. This does not entail radical skepticism. Rather, it clarifies why certain foundational questions (‘Why trust logic?’, ‘Why trust induction?’, ‘Why trust perception?’) cannot be answered wholly from within the agent’s own system. Any answer presupposes some norms or structures that function as external givens. The firmament boundary $\mathcal{F}(S)$ for an epistemic agent includes propositions about the reliability of its own methods that can only be justified in a broader context.

6.2 Implications for Philosophy of Science

In philosophy of science, our account frames scientific theories as systems S with their own constraints C_S (laws, symmetries, modeling assumptions) and information spaces I_S (state spaces, parameterizations, data models). The firmament boundary $\mathcal{F}(S)$ includes questions about initial conditions, model scope, and unobservable entities that the theory cannot fully resolve. Ambitions for a ‘Theory of Everything’ must contend with the possibility that any such theory will still presuppose boundary conditions it cannot itself explain. Our framework does not show that no ultimate theory exists, but it does suggest that even very powerful theories may have firmament boundaries relative to larger external grounds.

7 Conclusion

We have introduced the notion of a *firmament boundary* of a system S : the set of true but unreachable propositions and conditions about S that can only be resolved in an external ground S^+ . Building on established results in logic, computation, algorithmic information theory, physics, and AI, we have argued that no consistent, sufficiently expressive, self-referential system can be its own ultimate source of justification. Collapse events—points where systems fail, diverge, or refuse—are not merely errors but structural signals of dependence on an external ground. This unified perspective clarifies why attempts to build systems that ‘explain everything,’ including their own origin and reliability, inevitably encounter limits. Any realistic account of knowledge and intelligence must incorporate these limits. From a more global perspective, one can

ask whether there exists an ‘ultimate’ external ground that stands in this relation to all bounded systems at once. We will denote any such hypothetical limit point by R and, following other work, refer to it informally as a root source. In the present paper R plays only a structural role: it is an abstract prior that supplies constraints and information to systems without itself being contained in them, and we remain agnostic about its metaphysical interpretation. Future work could formalize specific versions of the Self-Grounding Limit in different domains, explore quantitative measures of firmament boundaries, and investigate how agents might best reason and act when they recognize that some of their own foundations lie in $\mathcal{F}(S)$. For AI systems in particular, understanding and respecting firmament boundaries may be crucial for designing models and oversight processes that acknowledge their own limits. No system, however sophisticated, can become its own absolute source.

References

- [1] Kurt Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1):173–198, 1931.
- [2] Alan M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 1937.
- [3] Gregory J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.