Undecidable Falsification: On the Computational Limits of Scientific Method

Chenghao Li

Abstract

Karl Popper's criterion of falsifiability is a standard answer to the demarcation problem, but it abstracts away from the computational limits of the agents who carry out falsification. This paper re-examines demarcation through the lens of recursion theory. Assuming the Weak Computational Theory of Mind—on which falsification is an algorithmic process—and introducing an "empirical embedding" of the Halting Problem, we model the collective scientific enterprise as a Universal Falsifier. We show that the hypothesis that all false empirical propositions are falsifiable in finite time would make the set of non-halting Turing machines recursively enumerable and, by Post's Theorem, the Halting Problem decidable, contradicting computability theory. Thus, there must be false empirical propositions that are in-principle falsifiable but algorithmically undecidable, yielding a "computational horizon" for scientific discovery. We distinguish this principled limit of undecidability from limits of computational complexity, arguing that the former marks a logical boundary independent of resource constraints and poses a challenge to convergent scientific realism.

Keywords: Falsifiability; Computational Theory of Mind; Halting Problem; Undecidability; Philosophy of Science

1 Introduction

Karl Popper proposed falsifiability as the criterion for demarcating science from pseudoscience (Popper, 1959). Central to Popper's logic is the asymmetry between verification and falsification: while finite observations cannot conclusively verify universal laws, a single contradictory observation can, in principle, falsify them. This logical structure assumes that the act of falsification—the recognition of a counterexample—is itself an unproblematic process. However, Popper did not fully explore whether the human capability for falsification is subject to intrinsic constraints. This paper interrogates the limits of the scientific method not from the perspective of sociology or resources, but through the lens of computability theory. Specifically, we explore the consequences of the Computational Theory of Mind (CTM) for the logic of scientific discovery.

This paper's central thesis is a conditional argument: if a weak version of CTM holds—that is, if the cognitive processes involved in identifying a counterexample are algorithmic in nature—then the scientific method inherits the logical limitations of Turing machines. We argue that there exists at least one false empirical proposition that is, in principle, undecidable for human scientists. In other words, there are false scientific hypotheses that cannot be falsified, not because of a lack of data or time, but because the process of falsifying them corresponds to a non-computable function. To demonstrate this, we construct an "empirical embedding of the halting predicate," which allows us to map the Halting Problem (Turing, 1936) onto an empirical scientific experiment. We show that assuming all false hypotheses are practically falsifiable leads to a contradiction in recursion theory (specifically, it would imply that the set `NONHALT` is recursively enumerable).

This inquiry intersects with the domain of Formal Learning Theory (FLT), which has analyzed the computational limits of inductive inference. Scholars such as Kelly, 1996 and Schulte, 2008 have demonstrated that some scientific issues are computationally unsolvable in the limit. However, the focus of this paper differs from FLT in a crucial respect. While FLT concerns the *inductive* problem of converging to the truth over an infinite data stream, our argument targets the *deductive* logic of falsification for a single proposition. We aim to show that even the "negative" act of rejecting a hypothesis is subject to computational bounds analogous to Rice's Theorem (Rice, 1953).

The paper proceeds as follows. Section 2 operationalizes the concept of "falsifiability" within a computational framework and clarifies the distinction between Strong and Weak CTM. Section 3 presents the formal argument, constructing the Universal Falsifier ($T_{\mathcal{U}}$) and proving the existence of undecidable empirical propositions via the empirical embedding lemma. Section 4 discusses the robustness of the argument: what if CTM is false? Section 5 addresses potential objections, including those arising from Gödel's Incompleteness Theorems, Searle's Chinese Room argument, and the distinction between computational complexity and undecidability. Finally, Section 6 concludes with the philosophical implications for scientific realism.

2 Core Concepts

To formalize the limits of falsification, we must first operationalize Popper's concept within a computational framework. This requires establishing definitions for falsifiability and the cognitive assumptions underlying the scientific observer.

2.1 Operational Definition of Falsifiability

In this framework, we restrict the domain of p to *empirical propositions*—that is, claims whose truth values rely on observation or physical measurement (e.g., "Light bends in a gravitational field"). We explicitly exclude purely analytic propositions (mathematics or logic) and metaphysical claims, as their falsification conditions are not empirical in the Popperian sense.

Definition 2.1 (Practical Falsifiability)**.** A scientific proposition p is *practically falsifiable* if and only if there exists an effective, computable procedure (an algorithm) that, taking p as input:

- Halts and outputs "falsified" in finite time if p is false (i.e., if a counterexample exists).
- Does not necessarily halt if p is true.

In computational terms, this defines falsifiability as *semi-decidability*. A proposition is falsifiable if the set of its counterexamples is recursively enumerable. This definition captures the asymmetry Popper emphasized: we can definitively recognize an error, but we may search forever for a confirmation without success.

2.2 Weak Computational Theory of Mind

Our argument relies on a specific premise regarding the nature of the scientific agent. We distinguish between a "strong" and a "weak" version of the Computational Theory of Mind (CTM). Strong CTM posits that the mind *is* a computer. We do not require such a strong ontological claim. Instead, we rely on:

Assumption 2.1 (Weak CTM)**.** The core cognitive operations involved in scientific falsification—specifically, the recognition of a counterexample given a hypothesis and observation—are functionally equivalent to Turing-computable processes.

This assumption is modest. It does not deny qualia or consciousness; it merely asserts that the logical process of checking whether data D contradicts hypothesis $\mathcal{H}$ is an algorithmic task. If human scientists can perform this task, and if the Weak CTM holds, then the capabilities of human scientists are bounded by the limits of Turing machines.

2.3 The Halting Problem

We recall the fundamental limit of computation. Let $\mathcal{M}$ be the set of all Turing machines and $\mathcal{W}$ be the set of all input strings. For machine $M \in \mathcal{M}$ and input $w \in \mathcal{W}$, the set $\texttt{HALT}$ is defined as:

$$\texttt{HALT} = \{\langle M, w\rangle \mid M \text{ halts on input } w\} \tag{2.1}$$

Conversely, the set `NONHALT` is the complement of `HALT`. In recursion theory, while `HALT` is recursively enumerable, `NONHALT` is not (Turing, 1936). This asymmetry is the mathematical engine of the argument that follows.

3 The Main Argument

This section establishes the formal argument. We first construct a computational model of the collective effort to falsify. Then, we introduce the "Empirical Embedding Lemma," which maps computational states to empirical observables. Finally, we derive the contradiction.

3.1 The Universal Falsifier (T_U)

Science is a collective enterprise. While individual scientists have limited cognitive resources, we model the scientific community's potential for falsification as an abstract machine.

Let $\mathcal{A} = \{A_0, A_1, A_2, \ldots\}$ be the set of all possible distinct falsification methods (algorithms or experimental protocols) definable within the framework of Weak CTM. Since each A_i corresponds to a Turing machine, $\mathcal{A}$ is enumerable. We define the "Universal Falsifier," denoted as $T_{\mathcal{U}}$, as a machine that executes all methods in $\mathcal{A}$ on a given proposition p using a dovetailing schedule (executing step 1 of A_0; then step 2 of A_0 and step 1 of A_1; and so on). The enumerability of $\mathcal{A}$ follows directly from Weak CTM. Since any cognitive process of verifying a counterexample is assumed to be computable, every valid falsification method corresponds to a Turing machine. The set of all Turing machines is countably infinite; therefore, the set of all possible scientific methods is enumerable.

Definition 3.1 (Universal Falsifier $T_{\mathcal{U}}$)**.** $T_{\mathcal{U}}$ accepts an input proposition p if and only if $\exists A_k \in \mathcal{A}$ such that A_k identifies a counterexample to p in finite time.

By this construction, a proposition p is practically falsifiable (in the sense defined in Section 2) if and only if $T_{\mathcal{U}}(p)$ halts. If p is true, $T_{\mathcal{U}}(p)$ runs forever.

3.2 Empirical Embedding

To link computation with empirical science, we must ensure that computational problems can be formulated as empirical hypotheses.

Lemma 3.1 (Empirical Embedding)**.** *For any Turing machine M and input w, there exists a constructible physical apparatus $D_{\langle M,w\rangle}$ and an associated empirical proposition $p_{\langle M,w\rangle}$ such that:*

$$p_{\langle M,w\rangle} \text{ is false} \iff M \text{ never halts on } w. \tag{3.1}$$

Proof. Consider a physical device $D_{\langle M,w \rangle}$ programmed to simulate the state transitions of M on input w. The device is constructed to emit a photon (an observable signal) if and only if the simulation reaches the `HALT` state. We define the empirical proposition $p_{\langle M,w \rangle}$ as:

"The device $D_{\langle M,w \rangle}$ will eventually emit a photon."

If M halts on w, the device emits a photon, making $p_{\langle M,w \rangle}$ true. If M never halts (i.e., $\langle M, w \rangle \in$ `NONHALT`), the device never emits a photon, making $p_{\langle M,w \rangle}$ false. Thus, the proposition's falsity corresponds directly to the machine's non-halting. □

Remark (On Physical Idealization and Empirical Status). One might object that the device $D_{\langle M,w \rangle}$ is physically unrealizable since it may require infinite energy or time to run. We respond that this *idealization* is methodological, akin to "frictionless planes" in mechanics or "reversible processes" in thermodynamics. Our goal is to establish a *principled* limit. If a proposition is undecidable even for a scientist equipped with idealized, infinite resources (as modeled by Turing machines), it is *a fortiori* undecidable for finite scientists. By removing resource constraints, we isolate the logical structure of falsification from contingent physical limitations. Despite this idealization, the proposition $p_{\langle M,w \rangle}$ remains *empirical*: its truth value depends on the behavior of a physical system in spacetime (the emission of a photon) rather than on purely formal definitions.

3.3 The Undecidability of Falsification

We now proceed to the main result.

Theorem 3.1 (Incompleteness of Falsification)**.** *Assuming Weak CTM and the Empirical Embedding Lemma, there exists at least one false scientific proposition that is not practically falsifiable.*

Proof. We proceed by reductio ad absurdum. Assume the opposite: *All false scientific propositions are practically falsifiable.* Let $\mathcal{H}$ be the hypothesis that all false propositions are falsifiable. Then, for any false p, the Universal Falsifier $T_{\mathcal{U}}(p)$ halts. Consider the set of non-halting programs, `NONHALT`. For any pair $\langle M, w \rangle$, we construct the proposition $p_{\langle M,w \rangle}$ as defined in Lemma 3.1. If $\langle M, w \rangle \in$ `NONHALT`, then $p_{\langle M,w \rangle}$ is false by construction. According to hypothesis $\mathcal{H}$, since $p_{\langle M,w \rangle}$ is false, it must be practically falsifiable. By the definition of the Universal Falsifier, this implies that $T_{\mathcal{U}}(p_{\langle M,w \rangle})$ must eventually halt, allowing us to construct a decision procedure for `NONHALT`. To determine if $\langle M, w \rangle \in$ `NONHALT`, we simply run $T_{\mathcal{U}}$ on input $p_{\langle M,w \rangle}$. If $T_{\mathcal{U}}$ halts, we know $p_{\langle M,w \rangle}$ has been falsified, so $p_{\langle M,w \rangle}$ is false, implying M never halts. Consequently, `NONHALT` would be recursively enumerable. However, by Post's Theorem, a set is recursive (decidable) if and only if both it and its complement are recursively enumerable. Since `HALT` is recursively enumerable, if `NONHALT` were also recursively enumerable, then `HALT` would be recursive, contradicting the undecidability of the Halting Problem (Turing, 1936). Therefore, $\mathcal{H}$ must be false: there exists at least one proposition p such that p is false (and thus has a counterexample in principle), but $T_{\mathcal{U}}(p)$ never halts. □

4 Discussion

The argument in Section 3 is conditional: *if* the cognitive processes underlying falsification are computable (Weak CTM), *then* the scientific method is bounded by undecidability. It is therefore necessary to address the status of this antecedent.

If Weak CTM is false, the implication does not hold. The falsity of Weak CTM would imply that the human mind possesses *hyper-computational* capabilities—that is, the ability to solve problems strictly harder than the Halting Problem (e.g., deciding membership in sets higher in the arithmetical hierarchy). While some philosophers have argued for this possibility (Lucas, 1961; Penrose, 1989), it remains a controversial minority position within cognitive science and the philosophy of mind. Standard models in computational neuroscience assume that neural processes, whether classical or connectionist, are simulable by Turing machines (Pylyshyn, 1984).

Our argument does not require Weak CTM to be true; it suffices that Weak CTM is a plausible empirical hypothesis. If the dominant paradigm in cognitive science is correct, then the epistemological limits derived here follow as a logical consequence. Furthermore, even if human cognition were hyper-computational, scientific practice would still be constrained by the physical limits of observational instruments, which are themselves physical systems presumably obeying computable laws.

5 Objections and Replies

We now address five potential objections to the thesis that scientific falsification is undecidable. These objections target the computational model of the scientist, the physical realization of the argument, and the distinction between practical and principled limits.

5.1 The Gödelian Objection

A classic argument against CTM, originating with Lucas, 1961 and expanded by Penrose, 1989, posits that Gödel's Incompleteness Theorems demonstrate that the human mind transcends any formal system. The objection suggests that humans can "see" the truth of Gödel sentences that a consistent formal system cannot prove, implying that human falsification capabilities might exceed those of the machine $T_{\mathcal{U}}$.

Reply. This objection conflates mathematical insight with empirical falsification. Gödel's Incompleteness Theorems concern the semantic truth of arithmetical sentences. In contrast, the problem of falsification defined in Section 2 is a syntactic process of identifying a contradiction between a hypothesis and an observation record. Moreover, Lemma 3.1 relies on a physical apparatus. Unless one assumes that the physical dynamics of the apparatus itself are non-computable, the detection of the apparatus's state remains a computable observation task.

5.2 The Argument from Semantics

Searle, 1980 argues via the "Chinese Room" thought experiment that syntactic manipulation is insufficient for semantic understanding. An objector might claim that $T_{\mathcal{U}}$, being purely syntactic, fails to capture the scientist's *understanding* of the hypothesis, which is essential for genuine falsification.

Reply. For falsification, semantic understanding is functionally reducible to the discrimination of outcomes. To falsify the proposition "all swans are white," the scientist need not understand the essence of "whiteness" or "cygnus"; they need only discriminate the sensory input of a black swan from the prediction. Weak CTM requires only that this discrimination function is computable. Whether the agent possesses phenomenological consciousness during this process is irrelevant to the logical structure of the falsification procedure.

5.3 Quantum Computation

Another objection appeals to the possibility that the human brain utilizes quantum effects, potentially transcending classical Turing limits.

Reply. While quantum computing (BQP) offers speedups for certain problem classes (e.g., factoring), it does not alter the class of computable functions. It is a standard result in quantum information theory that quantum computers cannot solve the Halting Problem (Nielsen & Chuang, 2010). Therefore, replacing the classical Turing machine model of the scientist with a Quantum Turing Machine would not render `NONHALT` decidable, and Theorem 3.1 would remain valid.

5.4 The Bayesian Challenge

A prevalent view in modern philosophy of science is Bayesianism, which argues that science is not about absolute falsification but probabilistic updating. An objector might claim that Popper's binary logic is obsolete, rendering our argument irrelevant.

Reply. First, Bayesian updating itself is a computational process. Computing the posterior probability requires a computable likelihood function and prior. If the mapping from evidence to probability update is non-computable, the Bayesian agent faces the same halting problem. Second, even if we relax "falsification" to mean "probability drops below a threshold ε," the logical structure remains. If the decision "$P(H|E) < \varepsilon$" is semi-decidable, Theorem 3.1 still applies. Our argument targets the *algorithmic accessibility* of error recognition, whether it is logical or probabilistic.

5.5 Complexity vs. Undecidability

Finally, one might argue that the real limit to science is computational *complexity* (e.g., NP-hardness), not undecidability. In practice, we run out of time long before we run into the Halting Problem.

Reply. We accept that complexity is the *practical* constraint (see (Arora & Barak, 2009) for a discussion on feasibility). However, the purpose of this paper is to establish a *principled* limit. Complexity constraints are contingent on resource availability; undecidability constraints are absolute. Even if we assume a scientist with infinite time and storage (an idealized Turing machine), the limit in Theorem 3.1 persists. This distinction is crucial for scientific realism: it implies that the bounds of science are not merely technological, but logical.

6 Conclusion

This paper has argued that the logic of scientific discovery is constrained not only by the availability of empirical data but by the computational nature of the scientific agent. By modeling the falsification process within the framework of the Weak Computational Theory of Mind, we have demonstrated that the set of falsifiable propositions is strictly smaller than the set of false propositions. Specifically, the existence of the set NONHALT implies the existence of empirical hypotheses that are false yet computationally impervious to falsification.

This result necessitates a revision of Popper's demarcation criterion. Popper envisioned falsifiability as a sharp logical boundary distinguishing science from pseudoscience. Our analysis suggests that this boundary is computationally permeable. If determining whether a hypothesis is falsifiable is itself an undecidable problem (analogous to the Halting Problem), then the demarcation line is not a clear edge but a fractal horizon. No algorithm can definitively purge all non-scientific statements from our corpus of knowledge without also discarding legitimate scientific truths.

Furthermore, this computational limit poses a challenge to convergent scientific realism. The realist intuition is that, given infinite time and evidence, science will converge on the truth by eliminating errors. However, suppose there exist false theories that effectively simulate non-halting machines. In that case, they may persist indefinitely in the scientific consensus, not because they are true, but because their falsity is algorithmically inaccessible. Thus, the "truth" to which science converges is not the absolute truth of the world, but the *computable subset* of the truth accessible to

Turing-equivalent observers.

In the final analysis, just as Gödel demonstrated that mathematical truth outstrips probability, we must accept that empirical truth outstrips falsifiability. Acknowledging this "computational horizon" does not devalue the scientific enterprise; instead, it clarifies the precise epistemological predicament of finite reasoners attempting to map a potentially infinite reality.

References

Arora, S., & Barak, B. (2009). *Computational complexity: A modern approach*. Cambridge University Press.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford University Press.

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, *36*(137), 112–127.

Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information*. Cambridge University Press.

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press.

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, *74*(2), 358–366.

Schulte, O. (2008). Formal learning theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(1), 230–265.