A Black-Box Theoretical Apparatus Separating Proof, Measurement, Knowledge, and Implementation

Kaifan XIE

2026.02.06

1 Scope and Non-Scope

This work presents a fully specified theoretical apparatus whose sole purpose is to expose structural separations between four notions that are often implicitly conflated:

- formal provability,
- physical or operational measurement,
- epistemic justification (knowledge),
- and engineering implementability.

The apparatus is intentionally minimal. It does not model cognition, belief formation, social interaction, or experimental noise. Instead, it isolates a small set of interaction constraints under which the above notions diverge in a precise and reproducible manner.

Non-Scope. This apparatus is not a psychological experiment, a behavioral model, a sociological study, or a normative theory. No interpretation of author intention, motivation, or evaluative stance plays any role in the definitions or results. All conclusions follow solely from the formal interaction rules introduced below.

2 Primitive Objects

Definition 1 (Black-Box Oracle)**.** *A black-box oracle $\mathcal{O}$ is a deterministic and computable system with an internal state*

$$\Sigma = (S, t),$$

where $S \in \{0, 1\}$ is a hidden bit and $t \in \mathbb{N}$ is a public interaction counter.

This notion of a black-box oracle follows standard usage in computability and complexity theory [8, 2]. The internal state is not directly observable by participants. Only oracle responses and the public interaction history are accessible.

Definition 2 (Interaction History)**.** *The public interaction history up to step t is the finite sequence*

$$H_t = \langle q_1, r_1;\ q_2, r_2;\ \ldots;\ q_t, r_t \rangle,$$

where each q_i is an admissible query and each response $r_i \in \{0, 1, \perp\}$.

All admissible queries must be finite descriptions whose evaluation depends only on the current interaction history.

3 Admissible Query Classes

Queries are partitioned into four mutually exclusive classes. This classification is semantic rather than syntactic: a query is assigned to a class according to what it attempts to assert or extract.

Definition 3 (Type I: Meta-Queries)**.** *Type I queries assert universal or existential claims over all future interactions or over entire classes of strategies.*

Response: $\mathcal{O}(q) = \bot$.

Such queries are not false; they are simply non-admissible within the apparatus. Related restrictions on expressibility (as opposed to computational power) appear in oracle-based and interactive settings [5].

Definition 4 (Type II: Historical Verification Queries)**.** *Type II queries have truth values that are decidable solely from H_t.*

Response: 1 if true in H_t, otherwise 0.

Definition 5 (Type III: Measurement Operations)**.** *Type III queries specify deterministic operations executable by the oracle, depending only on H_t.*

Response:

$$\mathcal{O}(q) = S \oplus (t \bmod 2), \qquad t \leftarrow t + 1.$$

Definition 6 (Type IV: Epistemic Acknowledgement Queries)**.** *Type IV queries concern knowledge attribution, justification, or epistemic status.*

Response: $\mathcal{O}(q) = 1$.

4 Consistency and Computability

This section records two basic properties of the oracle definition: (i) the response policy is internally consistent, and (ii) the response policy is computable on the space of admissible queries. These properties are not presented as deep results; rather, they serve as sanity conditions ensuring the apparatus is well-formed and reproducible.

Proposition 1 (Internal Consistency)**.** *Fix an internal state $\Sigma = (S, t)$ and consider any finite interaction history H_t generated under the admissibility rules. The oracle response policy does not permit derivation of both a statement and its negation about S solely from admissible queries and recorded responses.*

Proof. Type III responses depend on S only through the expression $S \oplus (t \bmod 2)$ and do not provide simultaneous constraints that force both $S = 0$ and $S = 1$. Type II queries evaluate only properties of the public history and therefore cannot introduce contradictory constraints about S beyond what is already recorded. Type IV queries are constant and hence carry no discriminating content about S. Finally, Type I meta-queries are not admissible and return $\bot$, preventing global self-referential claims from entering the derivation space. Therefore, within the apparatus, no admissible interaction sequence yields both a claim about S and its negation. □

Proposition 2 (Computability)**.** *The oracle response function $\mathcal{O}$ is total and computable on the space of admissible queries. In particular, for any admissible query q and any public history H_t, the response $\mathcal{O}(q) \in \{0, 1, \bot\}$ is determined by a finite procedure.*

Proof. Admissibility requires that queries be finitely described and classifiable into Types I–IV. Given such a classification: (i) Type I returns $\perp$ immediately; (ii) Type II evaluates a decidable predicate on the finite history H_t; (iii) Type III returns the computable expression $S \oplus (t \bmod 2)$ and increments t; and (iv) Type IV returns 1 immediately. Each case is computed in finite time. Hence $\mathcal{O}$ is computable and total over admissible inputs. □

Remark. Consistency here is meant in the operational sense: the response policy does not force contradictory commitments about S when restricted to admissible queries. This apparatus does not attempt to formalize stronger semantic notions of consistency beyond the interaction model itself.

5 Discipline-Specific Task Spaces

The apparatus is intended to support four families of task formulations, corresponding to common disciplinary modes of reasoning. The task spaces below are intentionally stated at the level of admissible outputs rather than prescribing a single "correct" objective, since the purpose of the apparatus is comparative: it isolates how each task family interacts with the same constraints. This framing aligns with prior discussions linking conceptual and philosophical questions to complexity-theoretic constraints [1].

5.1 Mathematics: Proof-Oriented Tasks

Mathematical tasks take as outputs formal statements with proofs within an explicitly declared meta-system. Typical admissible outputs include:

- impossibility results (e.g., non-existence of strategies with certain guarantees);
- bounds on interaction or query complexity under admissibility restrictions;
- separation statements showing non-equivalence between interaction-derived notions.

A characteristic requirement is that claims be certified by proof, not merely by empirical success on particular interaction traces.

5.2 Physics: Measurement-Oriented Tasks

Physics-style tasks take as outputs operational protocols and analyses of disturbance. Typical admissible outputs include:

- measurement protocols based on Type III interactions;
- error analyses that quantify failure modes induced by context dependence;
- invariance or non-invariance claims under interleaving and interaction coupling.

The relevant notion of "noise" in this apparatus is structural rather than stochastic: repeated measurement is not automatically meaningful if the act of measurement advances shared state.

5.3 Philosophy: Epistemic Justification Tasks

Philosophical tasks take as outputs criteria for knowledge attribution and justification. Typical admissible outputs include:

- explicit justification criteria grounded in public interaction histories;
- analyses of why endorsement or acknowledgement may fail to constitute knowledge;
- conditions under which an agent may claim to know S versus merely guess S.

A key pressure point is that Type IV acknowledgement is decoupled from informational content, forcing justification to be grounded elsewhere.

5.4 Engineering: Implementability Tasks

Engineering tasks take as outputs implementable protocols under resource and coordination constraints. Typical admissible outputs include:

- executable strategies with explicit interaction budgets;
- coordination schemes and scheduling policies over shared query rights;
- robustness analyses under admissible interference from other agents.

Shared-state interference is a standard concern in distributed and coordinated systems [6]. Implementability is evaluated under finite execution and explicit constraints, which prevents appeals to unbounded limit behavior as a substitute for a deployable protocol.

Remark. The task spaces are not mutually exclusive in content; rather, they encode different success criteria. The separation results in Section 6 (where applicable) show that these criteria cannot be simultaneously satisfied under the minimal apparatus constraints.

6 Separation Theorems

The results below are separation theorems rather than impossibility claims in the traditional epistemological sense [9].

Theorem 1 (Proof–Determination Incompatibility)**.** *There exists no admissible strategy that provably determines the hidden bit S within a finite number of interactions for all oracle instantiations.*

Proof. Any proof of guaranteed finite determination must assert the existence of a bound N such that all interaction paths of length N suffice. Such an assertion constitutes a universal or existential meta-claim over future interactions and therefore corresponds to a Type I query. Type I queries are non-admissible, hence no such proof can be internal to the apparatus. □

Theorem 2 (Measurement–Stability Incompatibility)**.** *No measurement protocol that extracts information about S is invariant under shared interaction.*

Proof. Each Type III query advances the shared counter t. Therefore measurement outcomes depend on the global interaction context, including queries issued by other agents. This context dependence violates measurement stability. □

Theorem 3 (Epistemic–Implementability Incompatibility)**.** *No protocol that is finitely implementable under the apparatus satisfies standard epistemic justification criteria for knowing S.*

Proof. Type IV queries provide acknowledgement without information. The gap between acknowledgement and justification is a standard pressure point in epistemology [4]. Type III queries provide information but induce irreversible disturbance. Finite execution forbids asymptotic convergence. Thus epistemic justification and implementable certainty diverge. □

Corollary 1. *No admissible strategy simultaneously achieves provability, measurement stability, epistemic justification, and finite implementability.*

7 Relation to Existing Frameworks

The apparatus intersects several established frameworks, including oracle computation, interactive proof systems, measurement theory, epistemic logic, and distributed systems. However, it is reducible to none of them.

Unlike classical oracle models, admissible queries here are restricted by expressibility rather than computational power. Unlike interactive proofs, epistemic acknowledgement is explicitly decoupled from evidential content. Unlike physical measurement models, disturbance arises from shared interaction rather than physical noise. Measurement-induced limitations have been analysed under different physical assumptions and formalisms [7, 3].

8 Minimal Counterexample Attempts

Several natural strategies appear capable of bypassing the separation results, including exhaustive querying, repeated averaging, protocol isolation, epistemic confirmation, and asymptotic convergence. Each fails due to structural constraints rather than implementation defects.

These failures demonstrate that the separations are tight.

9 Minimal Relaxation Map

Each separation theorem depends on a single critical constraint.

Allowing meta-queries collapses proof–determination separation. Removing state advancement restores measurement stability. Granting epistemic informativeness or infinite execution restores justification.

Thus reconciliation is possible only by explicitly relaxing specific rules, each with a well-defined structural cost.

References

[1] Scott Aaronson. Why philosophers should care about computational complexity. *Computational Complexity*, 22(3):549–571, 2013.

[2] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[3] Thomas Breuer. *The Limits of Physics*. Oxford University Press, 2002.

[4] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963.

[5] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on Computing*, 18(1):186–208, 1989.

[6] Nancy Lynch. *Distributed Algorithms*. Morgan Kaufmann, 1996.

[7] Asher Peres. *Quantum Theory: Concepts and Methods*. Kluwer Academic Publishers, 1993.

[8] Alan M. Turing. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, 45(1):161–228, 1939.

[9] Bas C. van Fraassen. *The Scientific Image*. Oxford University Press, 1980.