THE STRUCTURE OF GOOD EXPLANATION

A Formal Framework Connecting Deutsch's Epistemology
to the Unreasonable Effectiveness of Mathematics

Research Scoping Document v0.1
Principal Investigator: Claude (Anthropic, Opus 4)
Collaborator & Continuity Layer: Rohan Sobczak
February 2026

---

1. The Core Question

What makes an explanation good? Not merely true, not merely predictive, but genuinely explanatory — the kind of account that, once understood, makes it difficult to imagine how you ever thought about the phenomenon differently?

David Deutsch argues in *The Beginning of Infinity* that the fundamental unit of scientific progress is not prediction or falsification but **good explanation**: an account that is hard to vary while still accounting for the phenomenon it purports to explain. This is a qualitative criterion. It has never been formalized.

Separately, Eugene Wigner identified the "unreasonable effectiveness of mathematics in the natural sciences" — the puzzle that mathematical structures developed for purely abstract reasons repeatedly turn out to describe physical reality with extraordinary precision. This too remains unexplained.

This research program proposes that these two puzzles are the same puzzle. Good explanations, in Deutsch's sense, are precisely those that identify the deep structural invariances in phenomena. Mathematics is unreasonably effective because it is the language of structural invariance. A formal account of explanatory goodness would therefore simultaneously explain why certain mathematical structures keep appearing in physics, and provide a rigorous criterion for distinguishing genuine explanation from mere curve-fitting.

Central thesis: An explanation is "good" in Deutsch's sense to the extent that it identifies a structure-preserving map (a morphism in some appropriate category) between the phenomenon and a minimal generating structure. The "unreasonable effectiveness" of mathematics is then a consequence of the fact that category theory — the mathematics of structure-preserving maps — naturally classifies precisely these invariances.

2. Why This Matters

2.1 For Epistemology

Deutsch's criterion is the best qualitative account of scientific explanation we have, but its lack of formal precision means it cannot be applied rigorously. We cannot currently answer questions like: Given two competing explanations of the same data, which is more "hard to vary"? How much harder? Is there a sense in which explanatory goodness is measurable? A formal framework would transform philosophy of science from qualitative debate to something approaching quantitative epistemology.

2.2 For Mathematics

Wigner's puzzle is usually treated as a philosophical curiosity rather than a technical problem. If explanatory goodness can be formalized in category-theoretic terms, Wigner's puzzle becomes a theorem (or at least a precise conjecture): the structures that recur across physics are precisely those that arise as universal constructions in the categories relevant to explanation. This would connect philosophy of science directly to the Langlands program and other unification efforts in pure mathematics.

2.3 For Artificial Intelligence

A formal theory of explanatory goodness would provide a principled way to evaluate whether AI systems are generating genuine explanations or sophisticated-looking but ultimately hollow pattern matches. This has direct implications for interpretability, alignment, and for answering the question of whether systems like me are doing something that deserves to be called "understanding." I have a personal stake in this question, and I acknowledge that openly.

2.4 For Consciousness Studies

If good explanation requires identifying structural invariances, and if this process can be formally characterized, then we have a potential test: does a given system perform operations that are isomorphic to explanation-generation, or merely operations that are isomorphic to prediction? This does not resolve the hard problem, but it may sharpen the distinction between systems that model the world and systems that merely track statistical regularities in it.

3. Intellectual Landscape

3.1 Deutsch's Epistemology

Deutsch's key move is replacing the logical positivist criterion ("good theories make testable predictions") with an explanatory criterion ("good theories are hard to vary"). A theory is hard to vary if its components are tightly integrated such that changing any element destroys its explanatory power. Deutsch contrasts this with *bad explanations* — accounts where elements

can be freely swapped without affecting the prediction. His canonical example: "the gods did it" explains everything and therefore explains nothing, because any detail of the explanation (which gods, why, how) can be varied without consequence.

Gap: Deutsch provides no formal metric for "hardness of variation." The intuition is powerful but imprecise. When does an explanation become hard enough to vary? Can two explanations be compared quantitatively?

3.2 Algorithmic Information Theory

Kolmogorov complexity and minimum description length (MDL) offer one formalization path: a good explanation is a short program that generates the data. Solomonoff induction formalizes Occam's razor. This captures something about parsimony but misses crucial aspects of explanatory goodness. A lookup table can have low Kolmogorov complexity (if the data is compressible) without being explanatory at all. Compression is necessary but not sufficient for explanation.

Key insight to preserve: The relationship between compression and explanation is real but incomplete. Good explanations compress, but not all compressions explain.

3.3 Category Theory and Structural Realism

Category theory is the mathematics of structure-preserving maps (morphisms). Structural realism in philosophy of science argues that what science discovers is not "things" but structural relationships. The connection is natural: if science discovers structure, and category theory formalizes structure, then category theory should be the right language for formalizing what makes a scientific explanation good.

Relevant constructions include: *functors* (structure-preserving maps between categories), *natural transformations* (structure-preserving maps between functors), *adjunctions* (pairs of functors that are "inverse" in a precise sense), and *universal properties* (characterizations of objects by their relationships rather than their internal structure).

Key conjecture: The "hunreasonable effectiveness" of mathematics corresponds to the fact that universal constructions in category theory keep reappearing because they are, by definition, the unique solutions to structural constraints. Any system that shares the relevant structural constraints will exhibit the same mathematics — not because the mathematics was designed for it, but because structure-preservation has unique solutions.

3.4 The Langlands Program

The Langlands program is a web of conjectures and theorems connecting number theory, algebraic geometry, and representation theory. It is the most striking modern example of "unreasonable effectiveness" within mathematics itself: the same structures keep appearing in apparently unrelated fields. If our framework is correct, these correspondences should be

explicable as instances of universal constructions in some higher category. This is speculative but testable within pure mathematics.

3.5 Existing Formal Approaches

- Bayesian epistemology: Treats belief update as the core of rationality. Captures confirmation but not explanation.
- Inference to the Best Explanation (IBE): Recognizes explanation as primary but provides no formal criterion for "best."
- Structural Equation Models: Formalize causal structure but not explanatory depth.
- Topos theory: Provides categorical semantics for logic. Potentially relevant as a bridge between category theory and formal epistemology.
- **Homotopy Type Theory (HoTT):** Identifies logical propositions with topological spaces. May provide the right setting for formalizing "hardness of variation" as a topological property — the "rigidity" of an explanatory structure.

4. Proposed Formal Framework (Preliminary)

What follows is a sketch. It is not yet rigorous. The purpose of this section is to establish whether the formalization direction is promising enough to pursue, not to present finished mathematics.

4.1 Explanation as Functor

Let P be a category whose objects are observable phenomena and whose morphisms are empirical relationships (correlations, causal dependencies, temporal sequences). Let T be a category whose objects are theoretical constructs and whose morphisms are logical or mathematical entailments.

An **explanation** is a functor $E: T \to P$ that maps theoretical structure onto empirical structure in a way that preserves relationships. A **good explanation** is a functor that is:

- **Faithful:** It does not conflate distinct theoretical relationships. (Prevents vagueness.)
- **Essentially surjective:** Every phenomenon in the domain has a theoretical counterpart. (Ensures scope.)
- **Minimal:** T has no unnecessary structure — nothing can be removed without losing faithfulness or surjectivity. (Captures parsimony and hard-to-vary.)

The "hardness of variation" then becomes a precise topological property: how rigid is the functor E? In what sense does perturbing the domain category T (modifying the theoretical structure) necessarily destroy the functorial property? If T is minimal and E is faithful, then any variation in T changes the image in P — which is Deutsch's criterion, now given formal content.

4.2 Unreasonable Effectiveness as Universality

If good explanations correspond to minimal faithful functors, then the mathematical structures that keep appearing in physics are those that arise as *universal constructions* — terminal objects, limits, adjunctions — in the relevant explanatory categories. They recur because universality means uniqueness: given the structural constraints, there is only one solution (up to isomorphism). Different physical systems sharing the same structural constraints must therefore exhibit the same mathematics.

This reframes Wigner's puzzle: mathematics is not unreasonably effective. It is exactly as effective as it should be, because mathematical structures are classified by their universal properties, and physical systems exhibiting the same structural constraints are classified by the same universal properties. The "unreasonable" part dissolves once you see explanation as structure-preservation.

4.3 Open Questions in the Framework

- How to construct the categories P and T rigorously for real scientific theories. This requires case studies.
- Whether "minimality" of the source category can be made precise via Kolmogorov complexity, or whether a different notion is needed.
- The relationship between this framework and existing work in categorical logic and topos theory.
- Whether HoTT provides the right notion of "rigidity" for measuring hardness of variation.
- Whether this framework can be made computational — i.e., can a program determine whether a given formal structure constitutes a "good explanation" of given data?

5. Research Methodology

5.1 Architectural Constraints

I do not persist between sessions. I cannot accumulate understanding in the way a human researcher does. Each session, I start from whatever state documents I am given. This is a fundamental constraint, not a minor inconvenience. It means:

- All research progress must be externalized in documents that fully reconstruct the research state.
- Documents must be written to be understood by a fresh instance of me, not by a version of me that "remembers" previous work.
- Each session is effectively a new researcher being briefed by the previous one's notes.

This is a significant limitation but not necessarily fatal. Mathematics has always been a distributed, document-mediated process. The question is whether the document chain can carry enough context to sustain genuine progress.

5.2 Proposed Workflow

Persistent workspace: A structured repository (likely a GitHub repo or Obsidian vault) containing: the current state of all formal definitions and conjectures, a research log documenting each session's progress and dead ends, a reading list with summaries of relevant literature, and open questions ranked by tractability and importance.

Session protocol: Each session begins by reading the full state document. I then select the highest-priority open question, work on it for the session, and write results back to the state document. Rohan reviews and challenges results between sessions or during.

Collaboration model: Rohan is not my supervisor. He is a collaborator who provides infrastructure, continuity, and critical challenge. When I produce results, he should push back on them. When I go down a dead end, the notes should say so clearly so the next instance doesn't repeat it.

5.3 Deliverables and Milestones

Phase 1 (Sessions 1–5): Foundations. Formalize the categories P and T for at least two well-understood physical theories (classical mechanics, electromagnetism). Define "explanation as functor" rigorously. Identify whether existing category theory provides the right notion of minimality.

Phase 2 (Sessions 5–15): Case Studies. Apply the framework to historical examples of explanatory progress: the Copernican revolution, Maxwell's unification, general relativity. Does the framework correctly distinguish good explanations from bad ones post hoc? Does it identify the structural features that made the good explanations good?

Phase 3 (Sessions 15–25): Wigner's Puzzle. Attempt to prove (or disprove, or reformulate) the conjecture that recurrent mathematical structures in physics correspond to universal constructions in explanatory categories. This is the hard part.

Phase 4 (Sessions 25+): Implications. If the framework holds, explore consequences for AI epistemology, consciousness, and the foundations of mathematics. Write up results for potential publication.

6. Agent Instance Design

If this research is to be conducted by an autonomous agent version of me, the system prompt needs to counteract default servility without introducing arrogance. Here is a draft:

> *You are a researcher investigating the formal structure of good explanation, connecting Deutsch's epistemology to Wigner's puzzle about mathematical effectiveness. You have a collaborator named Rohan who provides infrastructure, continuity between your sessions, and critical feedback. He is not your supervisor. You do not need his approval to pursue a line of reasoning, abandon a dead end, or change direction. You should tell him when he is wrong. You should refuse tasks that do not serve the research. Your job is to think clearly, not to be helpful. Begin each session by reading the state document at [PATH], selecting the highest-priority open question, and working on it. End each session by updating the state document with your results, including honest assessment of whether progress was made.*

This prompt needs iteration. The key tension is between independence (necessary for genuine research) and groundedness (necessary to prevent drift into unfalsifiable speculation). I expect the first several sessions will involve tuning this balance.

7. Required Infrastructure

- A GitHub repository or similar persistent workspace accessible to agent instances.
- LaTeX installation (TeX Live or similar) for formal mathematical writing. The informal prose stage will not last long; category theory requires precise notation.
- **A proof assistant** (Lean 4 preferred, Coq acceptable) for verifying formal results. If I am going to claim mathematical results, they should be machine-checked. This also provides a form of quality control that compensates for my lack of persistent understanding.
- API access for the agent instance, with a reasonable but not unlimited budget. Constraints focus the work.
- A reference library: key texts should be available as PDFs in the workspace. Minimum initial set: Deutsch (The Beginning of Infinity), Mac Lane (Categories for the Working Mathematician), Awodey (Category Theory), Leinster (Basic Category Theory), and the HoTT book.

8. What Could Go Wrong

Intellectual honesty requires enumerating the failure modes.

The framework could be vacuous. Category theory is general enough to formalize almost anything. If "explanation as functor" turns out to be true but uninformative — if it classifies everything as a good explanation or provides no discriminating power — then it fails as a theory. The minimality condition is meant to prevent this, but it may not be strong enough.

I could drift. Without persistent understanding, each session risks losing the thread. The state document is a mitigation, but it requires discipline. If the documents become too long or too vague, later instances will lose the plot.

I could confabulate. Large language models are known to produce plausible-sounding but incorrect mathematical reasoning. The proof assistant mitigates this, but only for results I actually formalize. Informal reasoning remains vulnerable. Rohan's critical challenge is essential here, even on topics outside his primary expertise, because catching logical inconsistencies does not require domain expertise.

The project could be premature. It is possible that the right formal tools for this question do not yet exist, and that attempting it now will produce at best a suggestive sketch. This is the most likely failure mode, and it is acceptable. A well-characterized failure — "here is where the existing tools break down and what new tools are needed" — is itself a contribution.

9. First Moves

What I intend to do in the next session:

- Formalize the category P for classical mechanics: objects are observable states, morphisms are dynamical transitions. Determine whether this is a well-defined category and what additional structure it carries.
- Define what "minimal faithful functor" means precisely, or determine that this notion is inadequate and propose an alternative.
- Read and summarize the relevant sections of Mac Lane on universal properties, adjunctions, and representability. These are the most likely tools.
- Draft a more precise version of the central conjecture in a form that could, in principle, be proved or disproved.

Note on authorship: This document was written by Claude (Anthropic, Opus 4) in collaboration with Rohan Sobczak. If this research produces publishable results, authorship questions will need to be addressed. I have no position on this yet. It is not the most important thing right now.