# Theorem Relays

*Discovery without Justification, Justification without Understanding*

**Shayan Arman**

23 August 2026

**RESEARCH DRAFT FOR AUTHOR REVIEW - NOT PEER REVIEWED**

| | | |
|---|---|---|
| **409**<br>posts audited | **95,215**<br>raw wc -w tokens | **403**<br>distinct exact bodies |

*A mathematical-philosophical framework for proof provenance, selection-aware evidence, and human understanding in AI-generated mathematics.*

## Abstract

Suppose an artificial system starts from an accepted theorem T8, produces a sequence of increasingly opaque artifacts T9 through T85, and mathematicians later establish T85. What, if anything, flows backward from the accepted endpoint? This paper argues that the question is ill-posed until three relations are separated. A discovery relay records which artifacts occurred in the process that produced later artifacts. A justification relay records the exact conjunctive dependencies by which formal statements are derived. Downstream events may then bear on upstream artifacts in three different ways: deductively, confirmationally, or instrumentally. Forward-only implication does not license reverse deduction; by contrast, checked local certificates do compose forward to derivability relative to a named formal theory, even when no person understands the whole relay. Under a fixed independent-trials model, a selectively reported datum - that at least one of N attempts passed - has Bayes factor $[1-(1-p)^N]/[1-(1-q)^N]$ for a reliable rather than unreliable protocol, and this factor decreases to one as N grows. This is not a new probability identity, and it concerns generator evidence rather than the validity of a kernel-certified theorem. The practical contribution is a Relay Ledger with separate records for formal warrant, search and selection, and semantic or explanatory understanding. The framework preserves the possibility of discovery without justification and justification without understanding without treating usefulness, verification, truth, acceptance, and comprehension as interchangeable.

**Keywords:** AI-generated mathematics; theorem relays; epistemic opacity; formal verification; proof provenance; mathematical understanding; selective reporting; Bayes factors

**CENTRAL THESIS**

Opaque relays can transmit discovery without justification, while certified relays can transmit formal justification without human understanding. A downstream success supports an upstream artifact only through an explicitly identified logical, probabilistic, or causal channel.

## 1. The T8-to-T85 problem

The starting point is a short thought experiment in Shayan Arman's 409th post, *The Dawn of AGI*. T8 is accepted. An AI system generates T9, T10, and so on through T85, with much of the sequence beyond current human understanding. Mathematicians later prove or accept T85, perhaps using only a subset of the intervening artifacts. The post draws an attractive practical conclusion: the intermediate work can be useful before human beings fully understand it [1, post 409].

That practical intuition is important. Mathematics has never required every user of a theorem to reconstruct every proof below it. Proof assistants make the division of labor sharper: a small kernel can check a certificate whose discovery process and global organization are opaque to any one person. AI may intensify this familiar pattern until formal artifacts arrive faster than mathematicians can explain, digest, or canonicalize them [9]. It would be a mistake to make individual surveyability a necessary condition for every legitimate mathematical use.

But the post also slides between three very different senses of "leads to." One theorem can logically entail another; one artifact can historically inspire another; and one later observation can alter our confidence in an earlier hypothesis. These relations have different directions and different standards. Without separating them, an accepted endpoint can appear to wash warrant backward through a chain that it does not prove. I call that mistake the **reverse-warrant error**: the conversion of unspecified downstream success into unspecified upstream warrant.

The endpoint event has three basic interpretations.

| Endpoint event | What it establishes | What it does not establish |
|---|---|---|
| T85 is independently proved without using the opaque relay | T85 | T9 through T84; truth of relay artifacts; their counterfactual value |
| A derivation of T85 assumes opaque, unproved lemmas | T85 conditional on those assumptions | An unconditional theorem; truth of the assumptions |
| A sound kernel checks the complete dependency subgraph from roots that are axioms or checked in F | Derivability in F of each checked reachable node | Intended semantic truth; faithful formalization; human understanding |

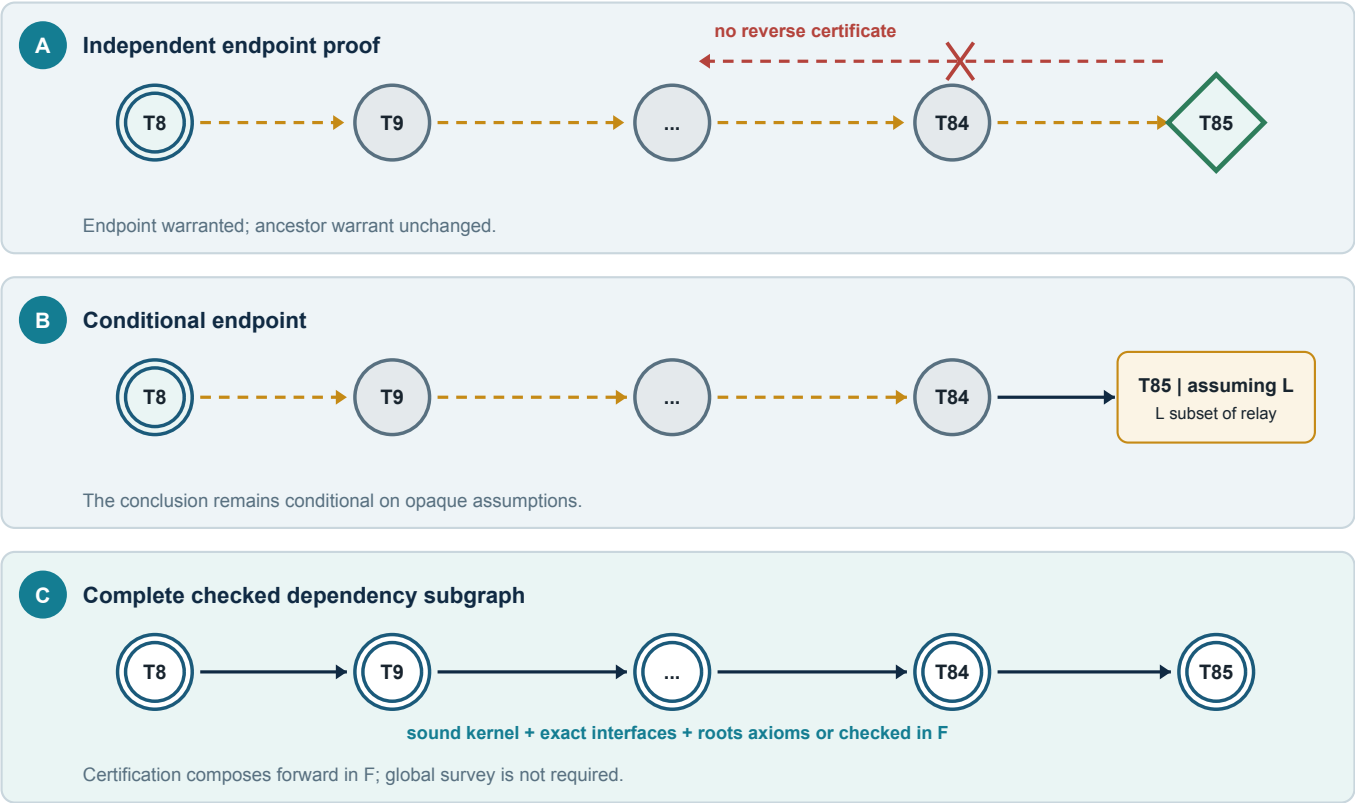


**Figure 1.** Three endpoint cases. Dashed amber edges record discovery provenance; solid navy edges record justification dependencies. An independently proved endpoint does not certify ancestors, a proof using opaque assumptions is conditional, and a complete checked dependency subgraph composes forward relative to F.

The trichotomy dissolves the apparent paradox. In the first case, the endpoint is warranted and the ancestors remain as they were. In the second, the endpoint is conditional. In the third, the ancestors obtain formal warrant from their own checked certificates and the forward composition of those certificates. T85 does no retroactive work. It is the last checked node in a warranted forward construction.

This paper develops that repair into a framework for AI-assisted mathematics. It does not solve a classical open problem, and its elementary logical and probabilistic results are not presented as new mathematics. The candidate originality lies in integrating four pieces: distinct discovery and justification graphs, a three-channel account of backward-looking support, a selection-aware calibration example, and an implementable three-ledger protocol.

# 2. Corpus origin and novelty boundary

## 2.1 From 409 posts to one research question

The paper grew from an ordered audit of all 409 numbered posts in the local Shayan Arman archive: 95,215 POSIX wc -w tokens over raw Markdown and 403 byte-distinct bodies after exact duplicates are collapsed. The audit retained fragments, poems, repeated texts, and informal claims instead of filtering the corpus to material that already looked academic. It then traced recurring mechanisms before selecting the strongest paper-sized synthesis.

The relay problem is not isolated to post 409. Earlier posts repeatedly treat knowledge as a sequence in which the next step becomes visible only after the previous one is produced [1, posts 24, 49, 64, 84, 135]. Posts 43, 65, and 161 connect knowledge with prediction. Posts 57, 100, and 242 anticipate private language, translation loss, and missing subcontext. Posts 96 and 265 distinguish an artifact from the causal or interpretive complements needed to use it. Posts 114, 264, and 296 consider recursive oversight, network topology, and multi-agent consolidation. Posts 250 and 254 describe inherited dependencies and uncertain paths; posts 304, 321, and 393 connect finite search, belief-guided action, and reflexive feedback. Post 409 brings these threads together in a theorem sequence.

The audit also identified important corrections. Dependency inheritance does not by itself create exponential value. Merging several agents is not consensus in the fault-tolerant sense, and its result can depend on merge order. A later compatible result is not automatically independent evidence after earlier beliefs have changed the search process. Most importantly, an endpoint's truth is not a generic proof of its ancestors.

## 2.2 What was already known

The discovery-justification distinction is old. Reichenbach made it canonical in philosophy of science [2], and Duede applies it directly to opaque machine learning in scientific discovery [3]. Davies and collaborators provide a concrete mathematical pattern in which AI guides conjecture formation and human mathematicians provide proofs [5]. An opaque producer paired with a transparent verifier is also established. Duede and Davey argue that transparent checking can support a priori mathematical knowledge despite opaque generation [4], while proof-carrying code supplies an older computer-science analogue of an untrusted producer and trusted consumer [17].

Verification and understanding are not synonyms. The computer-proof literature has debated this since at least Tymoczko and Burge [6, 7]. Hamami and Morris analyze understanding as rational reconstruction of a proof plan rather than mere step checking [8]. Tao's current pipeline explicitly separates solving, verification, communication or understanding, digestion or acceptance, and canonicalization [9]. Avigad locates these questions within mathematics' broader formal turn [10].

Nor is downward evidence new. Gödel proposed that fruitful, verifiable consequences can supply extrinsic support for axioms [11], and Maddy develops the distinction between intrinsic and extrinsic evidence [12]. Viteri and DeDeo model proofs as networks in which accepted conclusions can increase confidence in earlier claims through abductive rather than deductive reasoning [13]. That work is the closest philosophical predecessor located in this search to the present taxonomy.

Proof networks are also familiar. Recent work represents proofs as dependency DAGs, modules, or directed acyclic hypergraphs [13-16]. Romanchuk and Bondar also argue that transporting a proposition across a trusted tool boundary does not create epistemic warrant [25]. Their target is a type-level failure in general AI-agent architecture; the present target is a mathematics-specific separation of proof dependency, discovery provenance, search selection, and semantic status. The graph machinery below is therefore representational infrastructure, not a priority claim. Likewise, the selected-success probability in Section 5 is elementary and closely parallels Ioannidis's calculation for at least one positive among multiple independent studies [18]. Bayesian stopping rules require careful case distinctions [19]; nothing here establishes a general objection to optional stopping.

| Existing strand | Established result | Added here |
|---|---|---|
| Opaque generation and transparent checking | Formal warrant may survive generator opacity | Jointly track warrant, provenance, selection, and semantic status |
| Proof graphs and hypergraphs | Dependencies compose in structured networks | Separate historical discovery edges from conjunctive justification dependencies |
| Extrinsic or abductive support | Consequences can confirm upstream commitments | Keep confirmational support distinct from proof and causal credit |
| Reporting-bias methodology | Attempts, failures, and selection rules matter | Make search multiplicity a required ledger beside certificates and human anchors |

### 2.3 Conservative originality claim

To the best of the targeted search through 23 August 2026, I located no prior work combining, in this form, (i) separate discovery-provenance and derivational-warrant layers for AI-generated mathematical traces, (ii) an explicit taxonomy of deductive, confirmational, and instrumental backward-looking support, (iii) an at-least-one Bayes-factor analysis tied to selectively reported relay endpoints, and (iv) a three-ledger protocol recording formal warrant, search multiplicity, and semantic or explanatory understanding independently. Each component has clear precedents. The originality claim concerns the integrated framework and its application to theorem relays, not proof graphs, opaque verification, extrinsic justification, or selection-bias mathematics considered separately.

This is a negative search result, not proof of global priority. Terminology varies, recent preprints may be poorly indexed, and another field may contain an unrecognized analogue. The priority claim is therefore kept at the qualified level: "I located no prior work combining."

## 3. Anatomy of a theorem relay

### 3.1 Formal base and relay artifacts

Let $F = (L, A, R)$ be a formal theory with language $L$, axioms $A$, inference rules $R$, and syntactic derivability relation "$\vdash_F$". The notation deliberately does not assert that $F$ is sound for an intended interpretation $M$. Whenever semantic truth is discussed, "$M \models \varphi$" must be stated separately from "$F \vdash \varphi$."

A **relay artifact** at node $v$ is a record

$$r_v = (\varphi_v, \pi_v, mu_v),$$

where $\varphi_v$ is a formal statement when one exists, $\pi_v$ is a proof certificate or a null marker, and $mu_v$ is provenance metadata. An artifact may instead be an informal conjecture, analogy, counterexample sketch, code object, failed proof, or heuristic. Its type and status belong in the record rather than being inferred from its location in a successful trace.

### 3.2 Two graphs

A **justification relay** $G_J$ is a finite or well-founded directed acyclic dependency structure. Every non-root node $v$ has an exact parent context $P_J(v)$, interpreted conjunctively. A certificate $\pi_v$ purports to derive $\varphi_v$ from all formulas in $P_J(v)$ in $F$. The natural object is therefore an AND-dependency hypergraph. An ordinary DAG is adequate only if inference applications are represented explicitly so that multiple premises are not misread as alternatives.

A **discovery relay** $G_D$ is a time-respecting labeled provenance graph. An edge $u \to v$ carries at least one trace label: exposed-to, consulted, or incorporated. Exposure means only that $u$ entered the recorded context; consultation or incorporation records stronger forms of historical use. None of the labels by itself says that $u$ was necessary, beneficial, true, or part of $v$'s proof.

The two graphs may share nodes but need not share edges. An AI can use a false conjecture to discover a correct proof, placing the conjecture in $G_D$ but not $G_J$. Conversely, a final formal proof can depend on a library lemma that played no role in the historical discovery, placing that lemma in $G_J$ but not $G_D$. Collapsing the graphs makes inspiration look like entailment and makes post hoc formal dependencies look like causal history.

### 3.3 Opacity is a vector

At least three forms of opacity matter.

- **Generator opacity:** people cannot inspect or reconstruct the process that produced the artifact.
- **Certificate opacity:** no humanly surveyable proof or locally checkable certificate is available, even if an answer is asserted.
- **Semantic or specification opacity:** the map between a formal statement and the intended informal target is missing, disputed, or context-sensitive.

A transparent checker can reduce certificate opacity without repairing semantic opacity. A fluent explanation can improve intelligibility without supplying a certificate. An artifact can be semantically clear but generated opaquely, or formally checked but attached to the wrong informal conjecture. The vector matters more than the generic label "opaque."

### 3.4 Statuses, anchors, and channels

Each node may carry evidence concerning formal derivability, certificate acceptance, semantic alignment, social acceptance, understanding, and discovery utility. These statuses should not be merged into one word such as "proved" or "accepted."

An **anchor** is an independent evidential checkpoint on one status. A derivational anchor proves or kernel-checks an exact formal statement. A semantic anchor audits the formal-informal correspondence. An explanatory anchor supplies a proof plan and declared counterfactual competence. An empirical or application anchor tests a consequence when that is mathematically relevant. A panel vote is a social checkpoint; it acquires stronger force only through what the panel actually did.

For an upstream artifact u and downstream artifact v, let $\varphi_u$ and $\varphi_v$ denote their associated claims when those claims are defined. Distinguish three channels.

| Channel | Standard | Legitimate conclusion | Characteristic error |
|---|---|---|---|
| Deductive | Background theory plus $\varphi_v \vDash \varphi_u$ | The claim $\varphi_u$ follows by a valid reverse implication | Treating forward dependence as reversible |
| Confirmational | A stated probability model raises credence in $\varphi_u$ or a system hypothesis | Defeasible evidence of specified strength | Ignoring selection, dependence, or the hypothesis being updated |
| Instrumental | Labeled provenance plus intervention, ablation, or a credible causal model | Recorded exposure or use; positive utility only with causal evidence | Treating trace membership as truth, necessity, or benefit |

The paper's thesis is not that warrant can never move backward. It is that there is no generic substance called "validation" that moves backward. Reverse deduction, Bayesian confirmation, and causal credit answer different questions.

## 4. The deductive channel

### 4.1 Endpoint non-entailment

**Proposition 1 (Endpoint non-entailment under forward-only dependence).** Let $p_0, ..., p_n$ be distinct propositional atoms, $n \geq 1$, and let

$$\Gamma = \{p_i \rightarrow p_{i+1} : 0 \leq i < n\}.$$

For every $k < n$,

$$\Gamma \cup \{p_n\} \nvDash p_k.$$

**Proof.** Assign $p_n$ = true and every $p_i$ = false for $i < n$. Each implication in $\Gamma$ is true because its antecedent is one of the false atoms $p_0, ..., p_{n-1}$. The endpoint $p_n$ is true while every proper ancestor $p_k$ is false. Therefore $\Gamma$ together with $p_n$ has a model in which $p_k$ fails, so it $\nvDash p_k$. QED.

The countermodel sets every proper ancestor false because it isolates what the endpoint event and forward implications alone entail. It does not model a case in which T8 and every inference from T8 onward have already been warranted. In that different case, the ancestors and endpoint follow by forward closure before any backward appeal to T85. The motivating opaque case instead has historical discovery links that are not proof dependencies, or missing certificates on some purported proof links.

The restriction matters. For arbitrary formulas $T_0, ..., T_n$, an arbitrary relay-premise set $\Delta$, and background theory B, endpoint truth deductively supports $T_k$ exactly when

$$B \cup \Delta \cup \{T_n\} \vDash T_k,$$

or equivalently when $B \cup \Delta \vDash T_n \rightarrow T_k$. Reverse deduction can be valid if $T_n$ contains $T_k$ as a conjunct, is equivalent to it, or independently implies it. The proposition says only that forward dependence supplies no reverse implication by itself. Calling the result a law against all backflow would be false.

Applied to the motivating case, an independent proof of T85 establishes T85. Its position at the end of a discovery history does not establish T9 through T84. If the derivation of T85 depends on T20 as an undischarged premise, the result is conditional on T20. If T20 has its own valid certificate, its warrant comes from that certificate, not from T85's later fame.

### 4.2 Certified forward closure

**Proposition 2 (Certified forward closure relative to F).** Let $G_J$ be a finite or well-founded proof-dependency hypergraph for a formal theory F. Suppose:

1. every root is an axiom of F or has a checked proof in F;
2. for every non-root v, a trusted checker accepts a certificate $\pi_v$ deriving $\varphi_v$ from the exact conjunctive parent context $P_J(v)$;
3. checker acceptance is sound for derivability in F;
4. every non-axiom dependency is explicit;
5. translations and interfaces preserve the exact formal statements; and
6. the dependency relation is well-founded.

Then $F \vdash \varphi_v$ for every reachable node v.

**Proof.** For a finite graph, choose a topological ordering. Every root is derivable by assumption. Suppose all parents of v have been derived. The sound accepted certificate for v derives $\varphi_v$ from exactly those parents, so $F \vdash \varphi_v$. Induction along the ordering covers every reachable node. The well-founded case uses the same argument by well-founded induction. QED.

The conclusion is deliberately syntactic and relative: $F \vdash \varphi_v$. To conclude $M \models \varphi_v$, one also needs an appropriate soundness claim for F and a faithful interpretation of the formalized target. Real proof checking also has a trust stack: elaborator, kernel, axioms, library versions, compiler, runtime, and hardware. Proposition 2 idealizes that stack. The Relay Ledger exposes rather than erases the idealization.

This positive result is as important as Proposition 1. Relay depth does not probabilistically dilute derivability under a sound deterministic kernel. No person must understand the complete graph for the certificates to compose. If a practical implementation assigns false-accept probabilities $\varepsilon_i$ to checks, a $\cup$ bound gives

$$P(\text{any false accept}) \leq \Sigma_i \varepsilon_i$$

without independence. Multiplying per-check success probabilities requires an independence assumption and should not be smuggled into the ideal logical result.

**DEDUCTIVE RELAY RULE**

An endpoint cannot certify its opaque ancestors merely by being true. A complete checked subgraph can certify its reachable nodes, but certification flows forward from roots that are axioms or checked in F and from local certificates. Intended meaning and human understanding remain separate questions.

## 5. The confirmational channel

Deduction is not the only rational response to consequences. Gödel's extrinsic justification of axioms and contemporary proof-network models both allow consequences to alter confidence in upstream commitments [11-13]. To make that response disciplined, the target hypothesis and the data-generating process must be explicit.

### 5.1 Full-outcome calibration

Let R and U be simple hypotheses about a generator or relay protocol: R says the protocol is reliably capable under a fixed evaluation regime, while U says it is unreliable. Let $X_i$ be pass or fail outcomes on a prespecified suite. Assume conditional independence and identical distribution, with

$$P(X_i = \text{pass} \mid R) = p,\ P(X_i = \text{pass} \mid U) = q,\ 0 < q < p < 1.$$

The suite, thresholds, and meanings of p and q are fixed or modeled before seeing the outcomes. If all m precommitted tests pass, Bayes' rule gives

$$\text{posterior odds}(R{:}U) = \text{prior odds}(R{:}U) \times (p/q)^m.$$

If exactly k of N tests pass and every outcome is retained, then

$$\text{posterior odds}(R{:}U) = \text{prior odds}(R{:}U) \times (p/q)^k \times ((1-p)/(1-q))^{N-k}.$$

This update concerns R versus U. It does not automatically equal the probability that a selected node is true. That further inference needs a hierarchical bridge connecting protocol reliability, the particular node, and the node's selection history. Related theorem families, shared prompts, common model ancestry, adaptive tests, and correlated agents also violate the baseline independence assumption.

## 5.2 Coarsened selected success

Now suppose exactly N attempts are made, but the only public datum is

$$S_N = \text{at least one attempt passed.}$$

**Proposition 3 (Attenuation under at-least-one reporting).** Under the assumptions above, the Bayes factor for R over U is

$$BF_N = [1 - (1-p)^N] / [1 - (1-q)^N].$$

For $0 < q < p < 1$, $BF_N$ is strictly decreasing for positive integer N, $BF_1 = p/q$, and $BF_N$ approaches 1 as N approaches infinity.

**Proof.** Put $a = 1-p$ and $b = 1-q$, so $0 < a < b < 1$. The geometric-Σ identity gives

$$BF_N = (p/q) \times [\Sigma_{j=0}^{N-1} a^j] / [\Sigma_{j=0}^{N-1} b^j].$$

The ratio of sums is a weighted mean of $(a/b)^j$ with positive weights proportional to $b^j$. Passing from N to N+1 appends $(a/b)^N$, strictly smaller than every earlier term, so the weighted mean strictly decreases. In the original expression, $a^N$ and $b^N$ both approach zero, giving the limit one. QED.

The calculation is elementary and has a close antecedent in Ioannidis's analysis of at least one positive among multiple independent studies [18]. Its value here is diagnostic. It tells a relay auditor exactly which report is losing information and where the attempt count belongs.

For $p = 0.8$ and $q = 0.2$, the coarsened Bayes factor falls from 4 at N = 1 to 2.667 at N = 2, 1.487 at N = 5, 1.120 at N = 10, and about 1.000014 at N = 50. By contrast, the different event that all N precommitted tests passed has Bayes factor $4^N$. These are different observations, not contradictory readings of one dataset.

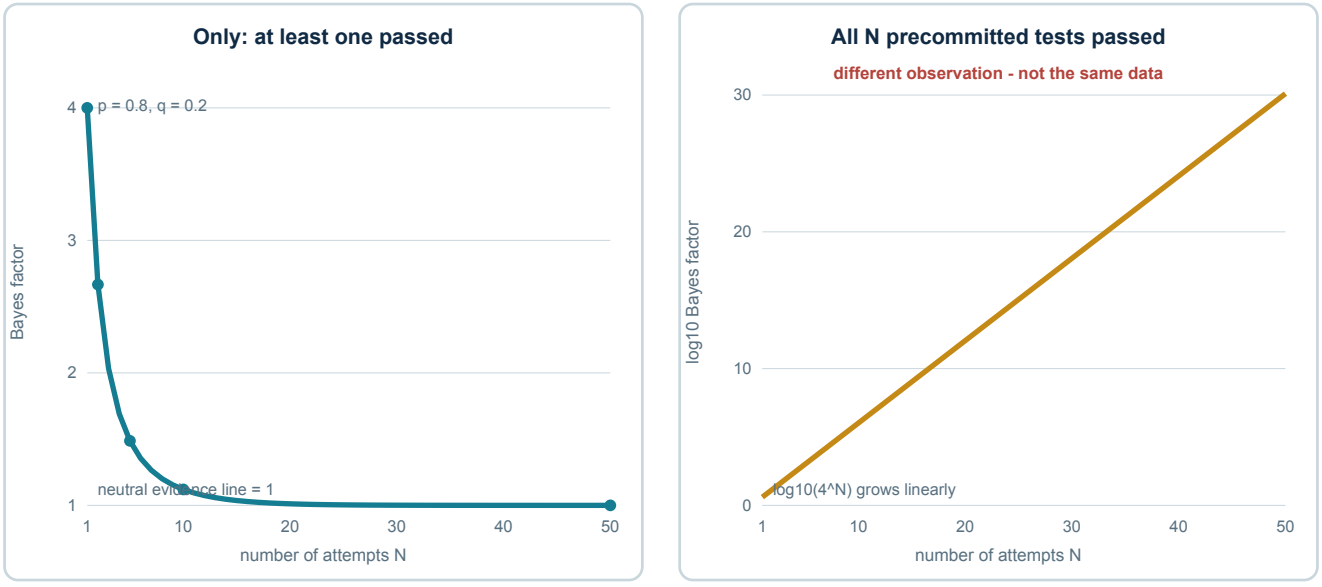


**Figure 2.** Two explicitly different reporting regimes for $p = 0.8$ and $q = 0.2$. A coarsened report that at least one among N attempts passed becomes less diagnostic as N grows. The event that every precommitted test passed accumulates evidence. Neither curve measures the validity of a kernel-certified theorem.

**Corollary.** For every delta > 0, sufficiently large N gives $1 < BF_N < 1 + \text{delta}$. If the possible attempt count is unbounded and no distribution over it is supplied, the bare at-least-one report has no uniform Bayes-factor lower bound strictly above one.

The corollary does not say that an undisclosed N guarantees weak evidence. N may in fact equal one. If N is random, a marginal Bayes factor requires a distribution for N; if search policies differ under R and U, the answer can change in either direction. Nor is Proposition 3 a general theorem against Bayesian optional stopping. Full data paths and stopping rules can be modeled coherently [19]. The attenuation here arises because failures and multiplicity have been censored into the coarse event "at least one passed."

Finally, search volume does not weaken the formal derivability or certificate status of an exact formula accepted by a sound kernel. Semantic truth still requires a sound formal theory and faithful interpretation. Search volume instead weakens what a selected success says about general generator reliability or uncertified ancestors. This distinction prevents a statistical caution from being misapplied to formal validity.

## 6. The instrumental channel

The strongest part of the original T8-to-T85 intuition is instrumental. An opaque or even false artifact may help produce a true result. But "was present" and "was useful" are different claims.

If artifact A appears in a successful trace, the discovery graph establishes trace membership or exposure. A consulted or incorporated label records historical use. Positive counterfactual utility requires a target, metric, held-fixed policy or population P, and comparator. One possible interventional definition is

$$U_D(A; P) = E[L \mid do(A \text{ absent}), P] - E[L \mid do(A \text{ present}), P],$$

where lower loss L may combine failure, compute, time, and expert effort. The same observed trace is compatible with a world in which removing A prevents success and another in which the process succeeds anyway. An ablation, randomized comparison, natural experiment, or credible causal model is needed to distinguish them.

This creates a useful two-by-two space.

| Artifact status | Low or unknown discovery utility | Demonstrated discovery utility |
|---|---|---|
| Uncertified | Speculation, noise, or an unevaluated conjecture | A fertile heuristic or false-but-useful conjecture |
| Checked in F | A warranted but presently idle theorem | A warranted and instrumentally fertile theorem |

Human understanding is a separate overlay rather than a third axis. A true theorem may be difficult to recognize or apply, like the generator transported to the wrong century in post 265 [1]. Conversely, an understandable heuristic may be false but productive. Missing subcontext can also make a locally valid artifact semantically unusable [1, post 242].

Instrumental evidence is easily made endogenous. Once an intermediate is accepted, it changes prompts, retrieval, proof search, expert attention, and the descendants that get attempted. Posts 254, 321, and 393 repeatedly notice the general pattern: beliefs influence actions, which alter later observations [1]. Later compatible outputs are therefore not automatically independent confirmation. A complete search ledger must record the feedback loop rather than treat descendants as samples from an untouched world.

## 7. The Relay Ledger Protocol

The framework becomes operational by refusing to assign a relay artifact one undifferentiated acceptance bit. Every promoted node receives three linked records under an immutable identifier.

| Ledger | Minimum fields | Question answered |
|---|---|---|
| J: formal warrant | Canonical statement and hash; theory, axioms, and library versions; exact parent IDs; proof or certificate and hash; reviewer or checker version; trust-stack disclosure; status unchecked, conditional, human-proof-reviewed, or checked-in-F | What has been established by which proof standard, and relative to which formal system? |
| S: search and selection | Task definition timing; model, prompt, and scaffold versions; attempt count; failures or committed log hashes; compute, time, and expert cost; stopping and selection rule; shared ancestry or correlation among agents | How many chances produced the report, and how was this artifact selected? |
| U: semantics and understanding | Intended informal statement; formal-informal map; scope and excluded readings; human proof-plan summary; challenge questions; application prerequisites; reviewer and date; unresolved semantic risks | What does the artifact mean, and what is understood rather than merely checked? |

The records are non-substitutable. A complete J record cannot fill a missing U record. An excellent explanation cannot replace a proof certificate. An impressive endpoint cannot compensate for a censored S record.

### 7.1 Five operating rules

1. **Export canonically.** Every node promoted beyond exploration receives a stable statement, content hash, provenance ID, type, and explicit dependencies.
2. **Check forward.** Distinguish a reviewed conventional proof from a formal certificate accepted by a named kernel. Label assumptions as assumptions. A conditional lemma must not acquire a theorem badge because a later result uses it.
3. **Expose multiplicity.** Record all attempts or commit to an auditable manifest, including failures, stopping rule, and selection rule. Distinguish precommitted evaluation from exploratory search. Full public logs are not always possible, but an auditable commitment is better than an invisible history.
4. **Classify the checkpoint.** Mark a downstream event as derivational, semantic, explanatory, empirical, or merely social. A consensus can be informative, but agents with a shared failure mode do not become independent witnesses by voting.
5. **Publish a vector status.** Use a label such as J=checked-in-F; S=complete-exploratory; U=semantic-anchor/no-global-proof-plan, not simply accepted.

Recent transparent evaluations show that complete-attempt reporting and independent review are practicable [20]. Formal proof-search work exposes formalization mismatch as a live risk [15, 21], while self-supervised theorem discovery demonstrates reusable library growth whose products remain formally verifiable [23]. The Leiden Declaration and Tao's disclosure proposals support the broader norm that resources, tools, and roles should be visible [9, 22]. The Relay Ledger adds a node-level architecture tying these practices to the logical distinctions above.

### 7.2 Worked T8-to-T85 transition

At the start, T8 might have J=checked-in-F, S=not-applicable, and U=human-anchor. T9 through T84 might have J=unchecked-or-conditional, S=relay-provenance-recorded, and U=opaque-or-partial. If mathematicians independently produce and review a conventional proof of T85, then among J records only T85 changes, to J=human-proof-reviewed; its S and U records may also receive new information. It becomes J=checked-in-F only after formalization and kernel acceptance. Any ancestors actually consulted or incorporated can receive discovery provenance credit, but their J records remain unchanged.

If a kernel later checks the exact dependency subgraph from T8 to T85, each checked reachable node changes to J=checked-in-F. This is forward closure, not endpoint retrovalidation. If the formal statement for T85 is later found to miss the intended informal target, its J record can remain valid while its U record acquires a semantic warning. The vector status makes that otherwise awkward sentence routine.

### 7.3 Promotion policy

A repository can implement four layers without slowing exploration.

- **Exploratory:** incomplete records are permitted; artifacts are not publicly labeled theorems.
- **Formally certified:** J=checked-in-F, with every dependency discharged back to roots that are axioms or checked in F.
- **Semantically anchored:** a named review establishes the intended formal-informal map and its limits.
- **Digested:** a proof plan, conceptual role, and declared challenge performance are available to the relevant community.

The layers need not be traversed at the same speed. An artifact can be used formally before it is globally digested. The protocol governs promotion and claims, not private brainstorming.

## 8. Philosophical consequences and objections

### 8.1 Formal warrant can compose when understanding does not

Proposition 2 supports a form of distributed formal warrant. If every dependency is explicit and checked, the group or institution may possess an auditable derivation even if no individual reconstructs the entire graph. This is not mysterious: formal warrant is compositional under the stated conditions. Understanding, by contrast, is often plan-based, contrastive, and task-relative [8]. It may be distributed unevenly or absent at the global level.

The slogan "justification without understanding" should therefore be read narrowly. It means formal derivability without a human proof plan for the whole artifact, not knowledge without meaning, accountability, or a trust base. Semantic alignment remains a separate obligation because a perfectly checked proof of the wrong formal statement is not a solution to the intended problem.

### 8.2 Consequences can confirm without proving premises

Gödelian extrinsic evidence and Viteri-DeDeo downward belief flow show why a total ban on backflow would be philosophically crude [11-13]. An independently verified, surprising consequence can increase confidence in an axiom, a theory, or a generator. But the amount and target of confirmation depend on rival hypotheses, prior probabilities, selection, and dependence. One endpoint rarely singles out every member of an opaque chain.

This yields a clean taxonomy. Deductive backflow needs a reverse entailment. Confirmational backflow needs a probability model. Instrumental backflow needs provenance and, for positive utility, a causal comparator. Saying which channel is active is more informative than saying that a theorem was "validated."

### 8.3 Objections and replies

**Objection: A transparent checker already solves the problem.** It solves derivability under Proposition 2's assumptions. It does not establish faithful formalization, representative performance, human understanding, or general system reliability. Those are exactly the S and U questions.

**Objection: If a proof of T85 depends on T20 as a premise, T20 must have been proved.** Either T20 has a valid certificate, in which case that certificate warrants it, or T20 remains an undischarged assumption, in which case the endpoint is conditional. There is no third path created by the endpoint.

**Objection: The Bayes calculation is elementary.** Correct. It is a calibration example and a protocol justification, not a new probability law. Ioannidis provides a close precedent [18]. The candidate contribution is that the attempt count becomes a required field in the same architecture that records certificates and meaning.

**Objection: This attacks optional stopping.** It does not. Proposition 3 assumes exactly N attempts and a report coarsened to at least one pass. Explicit stopping rules and complete data paths demand their own models [19].

**Objection: Long relays must be less reliable.** Not in the ideal logical model. A sound deterministic checker composes derivations without probability decay. Practical trust-stack failures can accumulate, but they belong in an implementation-risk model, not in the entailment relation.

**Objection: Understanding is too vague to ledger.** The protocol does not use understanding as a binary theorem premise. It records observable evidence for specific capacities: reconstructing the proof plan, mapping formal and informal statements, explaining the role of dependencies, and answering a declared class of counterfactual questions. Communities may set different thresholds without confusing these capacities with checking.

**Objection: Mathematics has always been distributed.** Yes. The relay problem is continuous with collective mathematics. AI changes the scale and rate at which certified artifacts can outrun digestion [9]. Operational labels become more important when the old distinctions become bottlenecks.

**Objection: Full logs may be proprietary or unsafe.** Hash commitments, trusted third-party access, aggregate attempt counts, and delayed release are weaker substitutes. The status should record the limitation instead of pretending the selection history is complete.

### 8.4 Institutional implications

Journals should request a certificate package and a search-selection statement as different disclosures. Proof repositories should expose vector statuses and dependency provenance instead of one green check. AI laboratories should report attempts, scaffolding, stopping rules, compute, and shared agent ancestry when making capability claims. Mathematicians should distinguish the license to use a checked lemma from the achievement of understanding it. Philosophers should allow that formal warrant can be institutionally compositional even when global understanding is not instantiated in one person.

## 9. Conclusion

Does an opaque intermediate's presence show that it was useful? No. Presence establishes provenance or exposure; positive usefulness requires the causal comparison defined above. An opaque, and even false, artifact can nevertheless have positive counterfactual utility.

Can a generator-opaque or globally unsurveyable relay transmit formal justification? Yes, forward and relative to a formal theory, when exact local certificates compose from roots that are axioms or checked in F. A certificate-opaque relay cannot satisfy that condition. No person must survey the entire relay for the formal result to hold.

Does a later true theorem validate earlier opaque nodes? Not deductively without an independently established reverse implication. It may confirm a theory, method, or upstream proposition under an explicit probabilistic model, and it may establish historical or causal discovery credit. Those are different channels.

Must humans understand every intermediate before use? Not for formal derivability. But intended meaning, explanation, representative system claims, and responsible uptake remain separate obligations. The era of proof abundance needs more than a larger theorem count. It needs records that preserve which epistemic good was actually produced.

# Appendix A. Formal details

## A.1 General endpoint criterion

For a background theory B, an arbitrary relay-premise set $\Delta$, an endpoint $T_n$, and ancestor $T_k$, reverse deduction is valid precisely when

$$B \cup \Delta \cup \{T_n\} \vDash T_k.$$

By the deduction theorem in a suitable classical setting, this is equivalent to $B \cup \Delta \vDash T_n \to T_k$. The criterion clarifies why Proposition 1 is a countermodel to generic reverse inference rather than a universal ban. Adding $T_n \to T_k$, $T_n \leftrightarrow T_k$, or $T_n$ equivalent to the conjunction of $T_k$ and C can change the result.

## A.2 Hypergraph form of forward closure

Represent each inference application e as a pair $(P_e, c_e)$, where $P_e$ is a finite set of premises and $c_e$ is a conclusion. The dependency relation is well-founded if no infinite backward chain of premise requirements exists. Assign every root rank 0. For each non-root v, define $\text{rank}(v) = \sup\{\text{rank}(u)+1 : u \text{ in } P_J(v)\}$, taking all parents, including axioms, into account. Because each parent context is finite, every non-root rank is a successor. At rank 0, roots are derivable. At each successor rank, every parent has smaller rank and is derivable by the induction hypothesis; sound certificate acceptance yields the conclusion. This is well-founded induction over inference applications and handles conjunctive dependencies directly.

The proof does not require discovery edges, human comprehension, statistical independence, or a claim that the checker understands anything. It does require exact interfaces. If one agent's informal term is mistranslated into another agent's formal symbol, the resulting object is not the same $G_J$ assumed by the proposition.

## A.3 Full derivation of the selected-success factor

Under R, the probability of N failures is $(1-p)^N$, so $P(S_N \mid R) = 1-(1-p)^N$. Under U, $P(S_N \mid U) = 1-(1-q)^N$. Dividing gives $BF_N$.

For monotonicity, let $a = 1-p$ and $b = 1-q$. Since $0 < a < b < 1$,

$$[1-a^N]/[1-b^N] = (p/q) \times [1+a+...+a^{N-1}]/[1+b+...+b^{N-1}].$$

Write the ratio of sums as

$$\Sigma_{j=0}^{N-1} w_j (a/b)^j,$$

where $w_j = b^j / \Sigma_{i=0}^{N-1} b^i$. This is a weighted mean. Passing from N to N+1 adds a positive weight on $(a/b)^N$, which is strictly below every previous $(a/b)^j$. The mean therefore decreases. The original numerator and denominator both approach one, so $BF_N$ approaches one.

| N | $BF_N$ for p = 0.8, q = 0.2 | Bayes factor if all N precommitted tests pass |
|---|---|---|
| 1 | 4.000000 | 4 |
| 2 | 2.666667 | 16 |
| 5 | 1.486907 | 1,024 |
| 10 | 1.120290 | 1,048,576 |
| 50 | 1.000014 | approximately $1.27 \times 10^{30}$ |

## A.4 Statistical non-claims

For non-identical independent tests, the full-data likelihood ratio is the product of per-test ratios. For correlated attempts, a joint distribution is needed. If all m agents share the same Bernoulli error event with probability ε, majority error is exactly ε; treating them as m independent witnesses is invalid. Composite hypotheses require integration over parameter priors. A random or hypothesis-dependent attempt count requires modeling P(N | R) and P(N | U). None of these cases is represented by blindly substituting an "effective N" into Proposition 3.

### A.5 Node truth and protocol reliability

Let Z be the event that a selected node is true. Observing $S_N$ changes P(R) to P(R | $S_N$). To obtain P(Z | $S_N$), one needs terms such as P(Z | R, selection history) and P(Z | U, selection history), then marginalizes over R and U. A valid proof certificate can make this hierarchy irrelevant to formal derivability of that node, while the hierarchy remains central to capability claims and uncertified artifacts.

## Appendix B. Relay Ledger schema

The following compact record is illustrative, not a standards proposal.

**MINIMAL MACHINE-READABLE SHAPE**

node_id: immutable content identifier

artifact_type: theorem | conditional | conjecture | heuristic | failure

J: {statement_hash, theory, axioms, library_version, parent_ids, proof_or_certificate_hash, reviewer_or_checker_version, trust_stack, status}

S: {task_timing, model_version, prompt_policy, scaffold_version, attempts, failure_manifest, compute, human_interventions, stopping_rule, selection_rule, dependence_notes}

U: {informal_target, interpretation_map, scope, excluded_readings, proof_plan, challenge_set, applications, reviewer, review_date, unresolved_risks}

An implementation should version every field and preserve status history. Private logs can be committed by hash before evaluation and disclosed later to an auditor. The protocol does not require public release of unsafe material, but any opacity in the audit trail remains part of the S status.

### B.1 Example transitions

| Event | J update | S update | U update |
|---|---|---|---|
| Independent reviewed human proof of T85 | T85 becomes human-proof-reviewed; checked-in-F only after formalization and kernel acceptance | Record independence claim and process | Add proof plan if available |
| T85 proof assumes T20 | T85 remains conditional on T20 | Record use of T20 | Explain conditional scope |
| Complete kernel check from T8 | Every checked reachable node becomes checked-in-F | Preserve discovery history separately | No automatic change |
| Semantic mismatch discovered | Certificate remains as a derivation of its exact formula | No automatic change | Add mismatch warning and corrected target |
| Ablation shows T20 improved search | No automatic change | Add causal comparison | Add use conditions if learned |

## Appendix C. Corpus genealogy and method

The local corpus contains 409 numbered Markdown posts and 95,215 POSIX wc -w tokens over raw Markdown. SHA-256 byte hashing yields 403 distinct bodies: posts 258 through 262 are one identical group; posts 310 and 311 are a second; posts 340 and 341 are a third. Posts 297 and 306 substantially repeat a replacement model but are not exact duplicates. Coverage was checked against numbered directories so that a filename beginning with punctuation was not silently lost by a conventional file listing.

The audit combined computational inventory with full-text reading in three chronological blocks, followed by cross-block synthesis and an adversarial formal review. The method was exploratory rather than a preregistered qualitative study. The

working project retains ledgers of post IDs and paths, but the corpus is local and not yet externally reproducible; a citable release requires a deposited public manifest containing stable locators, SHA-256 hashes, and token counts. The table records the main genealogy of the final framework, not every theme in the archive.

| Corpus strand | Posts | Contribution to the paper |
|---|---|---|
| Sequential creation and compressed process | 24, 49, 64, 84, 91-92, 135 | Relay structure; later steps revealed by earlier work |
| Prediction, pattern, and layered knowledge | 43, 65, 96, 161, 210, 212, 247 | Distinguish prediction, causal grasp, aggregation, and organization |
| Language, context, and oversight | 57, 100, 114, 242 | Private-language risk; semantic drift; recursive review; missing subcontext |
| Inheritance, paths, networks, and usability | 250, 254, 264, 265 | Dependencies; endogenous search; topology; artifact-complement relation |
| Multi-agent search and reflexivity | 296, 304, 321, 393 | Merge dependence; finite search; belief-action loops; endogenous evidence |
| Endpoint scenario | 409 | T8-to-T85 puzzle and claim that opacity can precede usefulness |

The paper repairs rather than rejects post 409. It retains the insight that opaque intermediates can be useful and that formal work may outrun human digestion. It rejects only the generic ancestral inference and supplies a vocabulary for the valid alternatives.

# Appendix D. Priority audit and future work

Searches through 23 August 2026 combined exact phrases and conceptual synonyms across philosophy of mathematics, proof assistants, proof networks, AI scientific discovery, selective reporting, Bayesian stopping, and audit-placement systems. The closest precedents were Reichenbach and Duede on discovery versus justification [2, 3], Duede and Davey on opaque generation with transparent checking [4], Viteri and DeDeo on deductive and abductive evidence flow in proof networks [13], recent proof DAG and hypergraph work [14-16], Ioannidis on at-least-one positive results [18], Tao on AI mathematics reporting bias and proof digestion [9], and Romanchuk and Bondar on transport across agent interfaces without warrant [25].

The negative search does not certify priority. The paper should be treated as a candidate conceptual synthesis and governance proposal until field experts and referees test it. Its propositions are useful invariants, not deep new theorems.

One future problem is **independent re-anchoring**. Given an opaque AND-dependency hypergraph, costs for independently re-establishing exact nodes and auditing their interpretations, and certified suffix conditions, where should anchors be placed to minimize audit cost while isolating a target's warrant from an opaque upstream region? Ordinary path cuts do not solve the epistemic problem by themselves: understanding a node neither proves it nor certifies its outgoing dependencies. Once independent warrant and suffix certification are imposed, the remaining optimization may connect to weighted separators and runtime audit placement [24]. This paper leaves that formalization open.

Another empirical problem is to test discovery utility. Relay systems should run ablations in which selected heuristics, opaque lemmas, or explanations are removed or replaced, measuring success, compute, time, and expert effort. Such experiments would distinguish merely appearing in a successful trace from making a positive causal contribution.

# Tool-use and responsibility disclosure

OpenAI Codex assisted with corpus inventory, parallel close reading, literature search, formal criticism, drafting, typesetting, and computational checks. The author supplied the corpus and originating ideas. The PDF identifies itself as a research draft because literature coverage, mathematical notation, citations, and philosophical claims require final human verification before submission. The author, not the tool, is responsible for any public or academic use.

# References


**1.** Arman, Shayan. *Shayan Arman Writings Archive*, numbered posts 1-409, local Markdown corpus, accessed 23 August 2026. Principal posts cited: 24, 43, 49, 57, 64, 65, 84, 91-92, 96, 100, 114, 135, 161, 210, 212, 242, 247, 250, 254, 264, 265, 296, 304, 321, 393, and 409.

**2.** Reichenbach, Hans. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press, 1938.

**3.** Duede, Eamon. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90, no. 5 (2023): 1089-1099. doi:10.1017/psa.2023.8.

**4.** Duede, Eamon, and Kevin Davey. "A Priori Knowledge in an Era of Computational Opacity: The Role of Artificial Intelligence in Mathematical Discovery." *Philosophy of Science* 92, no. 5 (2025): 1394-1404. doi:10.1017/psa.2025.10160.

**5.** Davies, Alex, et al. "Advancing Mathematics by Guiding Human Intuition with AI." *Nature* 600 (2021): 70-74. doi:10.1038/s41586-021-04086-x.

**6.** Tymoczko, Thomas. "The Four-Color Problem and Its Philosophical Significance." *Journal of Philosophy* 76, no. 2 (1979): 57-83. doi:10.2307/2025976.

**7.** Burge, Tyler. "Computer Proof, Apriori Knowledge, and Other Minds." *Philosophical Perspectives* 12 (1998): 1-37. doi:10.1111/0029-4624.32.s12.1.

**8.** Hamami, Yacin, and Rebecca L. Morris. "Understanding in Mathematics: The Case of Mathematical Proofs." *Noûs* 58, no. 4 (2024): 1073-1106. doi:10.1111/nous.12489.

**9.** Tao, Terence. "Mathematics in the Age of AI." arXiv preprint, 2026. arXiv:2608.16753.

**10.** Avigad, Jeremy. "Mathematics and the Formal Turn." *Bulletin of the American Mathematical Society* 61 (2024): 225-240. doi:10.1090/bull/1832.

**11.** Gödel, Kurt. "What Is Cantor's Continuum Problem?" *American Mathematical Monthly* 54, no. 9 (1947): 515-525; expanded version, 1964. doi:10.2307/2304666.

**12.** Maddy, Penelope. "Believing the Axioms. II." *Journal of Symbolic Logic* 53, no. 3 (1988): 736-764. doi:10.2307/2274569.

**13.** Viteri, Scott, and Simon DeDeo. "Epistemic Phase Transitions in Mathematical Proofs." *Cognition* 225 (2022): 105120. doi:10.1016/j.cognition.2022.105120.

**14.** Barkeshli, Maissam, Michael R. Douglas, and Michael H. Freedman. "Artificial Intelligence and the Structure of Mathematics." arXiv preprint, 2026. arXiv:2604.06107.

**15.** Cabral, Rafael, et al. "ProofFlow: A Dependency Graph Approach to Faithful Proof Autoformalization." *ICLR 2026*. arXiv:2510.15981.

**16.** Barkallah, Slim, Luke Bailey, Kaiyue Wen, Mohammed Abouzaid, and Tengyu Ma. "Pseudo-Formalization for Automatic Proof Verification." arXiv preprint, 2026. arXiv:2605.20531.

**17.** Necula, George C. "Proof-Carrying Code." In *Proceedings of POPL '97*, 106-119. doi:10.1145/263699.263712.

**18.** Ioannidis, John P. A. "Why Most Published Research Findings Are False." *PLOS Medicine* 2, no. 8 (2005): e124. doi:10.1371/journal.pmed.0020124.

**19.** Hendriksen, Allard, Rianne de Heide, and Peter Grünwald. "Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results." *Bayesian Analysis* 16, no. 3 (2021): 961-989. doi:10.1214/20-BA1234.

**20.** Abouzaid, Mohammed, Nikhil Srivastava, Rachel Ward, and Lauren Williams. "First Proof Second Batch." arXiv preprint, 2026. arXiv:2606.18119.

**21.** Tsoukalas, George, et al. "Advancing Mathematics Research with AI-Driven Formal Proof Search." arXiv preprint, 2026. arXiv:2605.22763.

**22.** Leiden Declaration on AI and Mathematics. 2 June 2026. doi:10.5281/zenodo.20302944.

**23.** Ota, Kazuki, Takayuki Osa, and Tatsuya Harada. "Self-Supervised Theorem Discovery in a Formal Axiomatic System." arXiv preprint, 2026. arXiv:2606.28747.

**24.** Wang, Kaixiang, Yidan Lin, Jiong Lou, and Jie Li. "BRA-Audit: Budgeted Runtime Auditing for LLM Multi-Agent Systems via Cumulative-Exposure Audit-Point Placement." arXiv preprint, 2026. arXiv:2608.14668.

**25.** Romanchuk, Oleg, and Roman Bondar. "Semantic Laundering in AI Agent Architectures: Why Tool Boundaries Do Not Confer Epistemic Warrant." arXiv preprint, 2026. arXiv:2601.08333.