SURF THE

EXPONENTIAL

The Missing Move

From Verified Criticism to New Capabilities,
Viable Markets, and Better Inventors



Find what the answer missed.
Build what the world is missing.

Vincent Boucher
Montréal, Québec, Canada

TRANSFORMATIONAL INTELLIGENCE ∞
6 November 2026

ALPHA GENESIS FLYWHEEL Ω

SURF THE EXPONENTIAL

The Missing Move

From Verified Criticism to New Capabilities,
Viable Markets, and Better Inventors

Vincent Boucher · MONTREAL.AI / QUEBEC.AI
Montréal, Québec, Canada · 6 November 2026

ABSTRACT

A fluent answer can conceal a consequential omission. A second model can make that omission inspectable through a source-grounded witness; disagreement alone establishes nothing. We develop this observation into *constructive challenge*: connect a verified gap to the means of making a valuable outcome attainable. Building on *Transformational Intelligence*, *ALPHA GENESIS FLYWHEEL* Ω, and *Invent the Economy*, challenge extends from an answer to its opportunity set and evidence instrument. An executable finite case constructs a sufficient representation, verifies its minimum label count, and tests a changed successor: three retained labels suffice initially, while the successor needs six. Conventional constructions tie. A costed continuation model then exposes a further consequence: cheaper reconstruction can make preserving source access preferable to preserving a richer record. Ordinary access wins when the opportunity is rare, and the deferred-procurement control wins when the same guarantee remains available later. A separate cash model shows how an advance can fund success while leaving a refund unfunded. Conditional analysis and a 2,304-point arithmetic audit identify when criticism helps or harms. These are reproducible results in authored finite settings, not measurements of frontier-model superiority or realized commerce. The empirical programme tests the stronger claims against direct execution, matched search, source-first production and competent conventional construction. The ambition is an enterprise that sees what has become possible, constructs its missing means, and preserves a useful path to the next discovery.

> **Research status.** A research synthesis, conditional propositions, a verified finite construction and successor, authored arithmetic models, and a prospective evaluation protocol. Task domains, model rates, utilities, prices and costs are stipulated inputs. Assisted adversarial review is editorial quality control, not an independent scientific experiment. Exponential enterprise growth, autonomous market invention, actual revenue and improved discovery remain hypotheses requiring their own evidence.

Keywords: transformational intelligence; consequential omissions; counterexamples; constructive challenge; frontier replay; market formation; discovery methods.

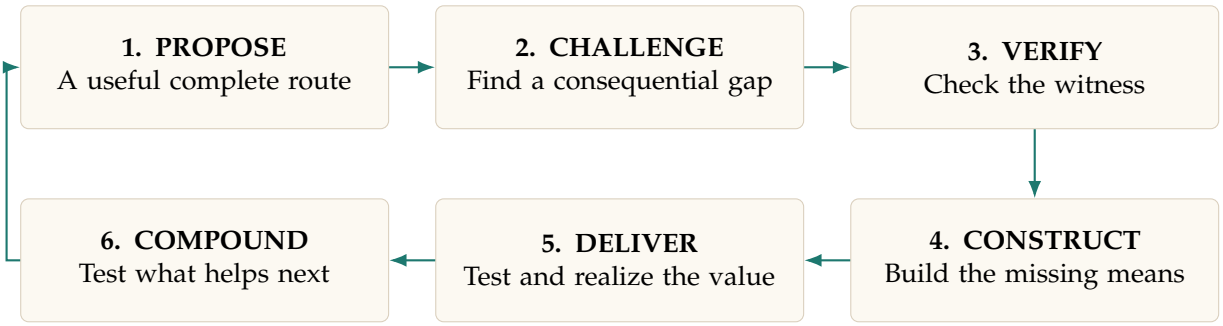
THE PAPER AT A GLANCE

A verified criticism can expose the missing means of a valuable possibility.

Imagine that an AI summarizes a book beautifully. Another AI finds a passage that changes the meaning of its conclusion. You now have something concrete to inspect. The useful object is the *evidence-backed discrepancy*, not the fact that two systems disagree.

Now apply the same question to an invention or an enterprise. What important condition has the proposal missed? Which useful route did it never consider? What new instrument would let us find out? An answer can be correct within its assumptions while those assumptions exclude the most valuable possibility.

> **The missing move may be a fact to recover, an experiment to run, a capability to construct, or an arrangement that makes voluntary cooperation worthwhile.**
> The central job is to establish which one it is, build the appropriate response, and test the complete consequence.



A useful organizing loop, not a mandatory ceremony. At any stage the best response may be an existing solution, a cheaper test, a simpler formulation, a deferred decision, or stopping.

THREE QUESTIONS THAT REACH BEYOND DEBATE

Challenge the answer. What would make this claim wrong or incomplete in a way that matters?

Challenge the possibilities. What valuable route is absent from the choices we are comparing?

Challenge the instrument. What prevents us from finding out, and can we construct better evidence?

A counterintuitive consequence. As reconstruction becomes cheaper, the best asset to retain can change from a richer record to access to its source. The right inheritance preserves a useful path to a future answer at justified cost.

Read by purpose. Sections 1–2: the idea and practical contract. Section 6: the executed construction, successor and preservation decision. Section 7: participation and refund obligations. Sections 3 and 8: when criticism pays. Sections 9–10: the stronger empirical and successor claims. Appendices: templates, proofs and reproduction.

Contents

1 From a missing chapter to a missing possibility 4

2 The Missing Move contract 5

3 When is another critic worth its cost? 8

4 Invent the evidence before inventing the verdict 10

5 Frontier Replay: reopen the reason an opportunity failed 11

6 An executable Missing Move: build the missing distinction 13

7 Invent the complete productive arrangement 17

8 Exact review audit: useful challenge has a boundary 19

9 The experiment that could establish the stronger claim 21

10 Compound the means, then test the inventor 24

11 Conclusion: make the next possibility real 25

A Practical challenge and construction templates 26

B Formal qualifications and reproducibility 27

C Claim register and source lineage 28

1 From a missing chapter to a missing possibility

A reader asks a frontier language model to summarize a book. The summary is fluent, balanced in tone, and apparently complete. The reader has not read the book, so the very quality that makes the summary convenient also makes it difficult to verify. A second model receives the book and the summary and is asked to identify a consequential omission. It points to a particular passage, explains why that passage changes the conclusion, and provides enough context for an independent check.

The assessment problem has changed. The reader can investigate a specific discrepancy instead of trying to certify the entire compression from its style. This asymmetry is useful: one decisive example can refute a particular claim even when constructing or evaluating the whole artifact is difficult. The second model can nevertheless be wrong. It may invent a passage, misread a qualification, or complain that a short summary does not contain everything. The task should therefore ask for the strongest *supported* criticism and allow the answer that no consequential defect was established.

This observation has direct precedents. Model-written critiques helped human evaluators find flaws in topic-based summaries in the experiments of Saunders et al. [1]. Later work on code critics found both useful bug detection and hallucinated criticisms [2]. The opening of *Transformational Intelligence* already develops the book example, adds independent adjudication, and considers producing a source-grounded structure before compression [3, §1 and Appendix C]. We build on that foundation rather than claim the second critic as a new invention.

1.1 The larger omission

An enterprise faces a related problem with a crucial difference. A book has a finite source. The world does not supply a complete volume containing every possible invention, production process, or voluntary arrangement. A proposal can accurately describe its candidate and still fail to consider a much better route. A scientific investigation can compare its hypotheses carefully while missing an instrument that would change which questions are answerable. An economically sensible rejection can omit a complementary construction that would make the activity viable.

The central research question is therefore:

> **Can intelligence turn a consequential, verified gap in what we know or consider into the means of doing something valuable that the current arrangement cannot deliver?**

The answer requires more than another round of disagreement. A valid objection has to connect to a diagnosis, a response, an executable construction, and a complete comparison. Some objections call for correction. Others call for new evidence. A third class reveals a removable dependency: the current method requires something expensive or unavailable, but an alternative may achieve the same authorized outcome without it. The best response can also be procurement, an established method, or no further action.

1.2 How the three papers fit together

Foundation	What it makes editable	Its role here
Transformational Intelligence	Representations, information, dependencies and intermediate means, with the authorized purpose preserved.	Diagnose whether a criticism exposes an avoidable restriction; construct and test another formulation.
Alpha Genesis Flywheel Ω	The conversion of frontier capability into useful possibilities and better discovery methods.	Replay consequential gaps when a relevant capability changes; distinguish better tools from better invention.
Invent the Economy	Productive capability and the arrangement of participants, transfers, risk and timed funding.	Build the complete conditions under which useful work is technically valid and voluntarily worthwhile.

Table 1: An explicit synthesis of the supplied manuscripts [3–5]. The organizing transition of this manuscript is the move from an adjudicated challenge to a tested productive response.

1.3 Contribution and intellectual position

The paper contributes an operational *Missing Move contract*; challenges to an answer, its opportunity set and its evidence instrument; and a matched evaluation programme connecting constructive challenge to frontier adaptation. An executable case finds an information loss, constructs a sufficient representation, verifies its minimality and tests a changed successor. A costed continuation argument identifies when to preserve the record, preserve source access or rely on ordinary reconstruction, including a stronger deferred-procurement control. Separate models expose harmful criticism and unfunded refunds. The larger architecture remains a proposal; the finite constructions and conditional decision boundaries are reproducible.

Counterexample-guided synthesis provides an established model for generating candidates, finding failures, and using those failures to constrain subsequent construction [6]. Debate offers a different route to eliciting useful information for a judge [7]. Neither formal finite-program guarantees nor idealized debate arguments make open-world invention automatically reliable. A natural-language challenge may lack a sound specification, a complete checker, or even an adequate measurement. We make those missing conditions part of the work.

The ambition is expansive without requiring every component to be novel. An ordinary test, a familiar algorithm, or a conventional contracting arrangement can be the decisive missing move. Scientific novelty, useful construction, commercial success, and improvement of discovery have different evidence requirements. A competent institution pursues the valuable result and states precisely which of those achievements it has established.

2 The Missing Move contract

2.1 Keep purpose, means and measurement distinct

Let the purpose contract be

$$\mathcal{I} = (O, H, A, B, T), \tag{1}$$

where O denotes the intended outcome, H the hard requirements, A the authority to act, B the available resources, and T the relevant time conditions. A formulation F describes a particular set of representations, information procedures, allowed internal actions and intermediate targets. An

evaluation instrument E supplies evidence about whether a realization serves $\mathcal{I}$. This separation follows the foundational manuscript [3, §2].

A critic may propose changing F or improving E. It cannot silently amend O, weaken H, expand A, or spend imaginary resources. If the required deliverable is a two-page summary, an interactive application does not fulfill the instruction merely because it might be more interesting. If the purpose is to understand an argument before a meeting, the application may be a legitimate alternative. The distinction depends on the actual assignment. A broad discovery charter may authorize proposing new beneficiary outcomes within a stated mission. Once a candidate enters confirmation, freeze its concrete outcome contract; a material change of ends or permissions begins an explicitly amended comparison.

Definition 2.1 (Consequential challenge)**.** *Relative to a stated purpose and decision, a consequential challenge is an allegation that a candidate claim, omitted condition, excluded alternative, or inadequate evidence procedure could materially change the required outcome or a justified choice of action. It is a hypothesis until its relevant evidence and materiality are adjudicated.*

An omission is therefore not automatically a defect. A summary must compress; an engineering plan must abstract; a research portfolio must leave possibilities unexplored. Materiality depends on the user's purpose, the decision at stake, and the available budget. Exhaustive commentary can make a useful artifact worse.

> **A missing distinction you can inspect.** A report stores only the mean of two readings and claims it can determine whether every reading is at most 90. The traces $(60, 100)$ and $(80, 80)$ have the same mean, 80, but opposite acceptance outcomes. That pair is a concrete witness against the claim. Retaining the maximum before compression is an ordinary sufficient response; it does not recover an already discarded maximum. Section 6 makes the transition from this witness to a constructed representation and a changed successor executable. The point is visible: a missing distinction becomes a specific construction target.

2.2 Two linked records: the challenge and the construction

We separate a challenge record w from a response record r. This prevents the correctness of a diagnosis from being mistaken for the correctness of its favorite remedy:

$$w = (c, s, e, m, q), \tag{2}$$

$$r = (\rho, F', k, \tau, v). \tag{3}$$

Here c identifies the challenged claim or choice; s fixes its scope and source versions; e is inspectable evidence; m states decision materiality; and q gives a falsifying or discriminating check. The response identifies the diagnosed restriction ρ, alternative formulation F', executable construction k, complete test τ, and validity and invalidation conditions v.

For a book, e may be an accurately located passage with enough surrounding context to preserve its meaning. For a program, it can be an executable input and failing output. For a market arrangement, it may be a feasible participant deviation, an omitted liability, or a cash-flow prefix that cannot be funded. For a scientific instrument, the evidence may show that two materially different states are observationally indistinguishable under the proposed measurement procedure. Each domain needs its own semantics of evidence.

Question	Minimum useful answer	What it cannot establish alone
What is challenged?	An exact assertion, required property or excluded route.	That the complete candidate is bad.
Where is the evidence?	Source identity and location, a runnable witness, or a justified measurement.	That the evidence source is infallible.
Why does it matter?	A specified effect on purpose, acceptance or decision.	That every omitted detail deserves repair.
What kind of obstacle?	A hard constraint to honor, an uncertain premise to resolve, or a restriction of means to change.	Permission to weaken a real requirement.
What is the response?	A buildable alternative plus a test against serious rivals.	That a valid criticism makes the repair valid.
When does it transfer?	Preconditions, costs, interfaces and invalidation rules.	Current acceptance or authority for a successor.

Table 2: A compact contract for constructive challenge. Evidence and construction are linked while remaining separately accountable.

2.3 Three directions of challenge

The failure challenger searches for a supported reason the proposed route would fail: a contradiction, missing exception, unacceptable operating case, uncompensated participant, or unpriced obligation. Its best output is a decisive witness, not a large number of objections.

The possibility challenger constructs a serious alternative the current candidate set excludes. This matters because a selector can choose perfectly among bad proposals. The relevant comparison is between complete routes to the same intended outcome, including an existing solution or doing nothing. A purportedly impossible activity becomes a candidate opportunity only when a concrete response changes the constraint responsible for that judgment.

The instrument challenger examines whether the available evidence can settle the dispute. It may propose a new observation, an independent implementation, a different decomposition, or a validated measure closer to the intended outcome. Its product is a better way to find out. It must itself be tested, with both sides rescored under a common supported instrument when that instrument changes [3, §7].

These are functions, not necessarily three agents. One capable system, several systems, a domain expert or conventional software may perform them. Role separation helps organize work; it does not manufacture independent errors or independent evidence.

2.4 Adjudicate before repair, and compare several responses

An adjudicator returns one of four useful states: *supported and material*, *supported but immaterial*, *refuted*, or *unresolved*. "Unresolved" is not a softer form of acceptance. It identifies an uncertainty that the next decision must either resolve, tolerate explicitly, eliminate through a changed construction, or avoid.

A supported material challenge opens a response set:

$$\mathcal{R}(w) = \{\text{repair, measure, construct, rearrange, adopt, defer or stop}\}. \tag{4}$$

Assess critics by independently verified material findings and downstream effects, including harmful revisions and wasted effort. Rewarding objection count, rhetorical force or frequency of overturning the producer encourages the wrong behavior. The categories may overlap. A shared instrument

and a staged commitment can be jointly useful even when neither alone justifies its cost. A system that immediately repairs the first valid objection can miss a simpler route that removes the disputed dependency altogether. Conversely, adding an elaborate institution to solve a local defect can be worse than a small correction.

> **The constructive question is not only "What is wrong?"**
> It is: **"What would have to become true, available, measurable or mutually worthwhile for the valuable outcome to work—and can we make that happen?"**

2.5 An opportunity witness needs two sides

A counterexample can demonstrate that a universal assertion is false on a stated case. An opportunity claim needs more. It needs a *positive side*: a complete route that meets the purpose under its stated conditions. It also needs a *comparative side*: why this route matters relative to available alternatives, including a route that reconstructs the supposed missing asset conventionally.

Let $J_{\mathcal{I}}(x)$ be a declared net-outcome measure for an admissible complete route x, and let $\mathcal{B}$ be a disclosed finite, nonempty comparison set. Define

$$\alpha_{\mathcal{I}}(x;\mathcal{B}) = J_{\mathcal{I}}(x) - \max_{b\in\mathcal{B}} J_{\mathcal{I}}(b). \tag{5}$$

This is a comparison to $\mathcal{B}$, not a proof about every possible enterprise. If outcomes cannot legitimately be collapsed into one scalar, retain a vector of relevant outcomes and report the actual trade-offs. Hard constraints remain constraints rather than purchasable score penalties.

The strongest formulation-dependent claim adds a justified exclusion: the old setup cannot attain the specified outcome within its defined permissions, information and resources. In real research that exclusion is often unavailable. Then the honest claim is a supported gain over the strongest tested complete alternatives. Neither an unsuccessful search nor a pessimistic first draft proves impossibility.

3 When is another critic worth its cost?

3.1 A conditional accounting identity

Fix one pre-specified, decision-material defect proposition per episode, relative to a stated purpose and task distribution. Let F denote that this exact proposition is true. Let D denote a criticism alleging this same defect, and let $A_c \subseteq D$ denote acceptance of that allegation followed by its specified response. A flawed candidate can receive a false criticism about a different issue; such mixed-target episodes require a richer event partition and are outside this binary reference model. Define

$$\pi = \mathbb{P}(F), \qquad d = \mathbb{P}(D \mid F), \qquad f = \mathbb{P}(D \mid \neg F), \tag{6}$$

$$a = \mathbb{P}(A_c \mid D, F), \qquad b = \mathbb{P}(A_c \mid D, \neg F). \tag{7}$$

The letter A_c here refers to an accepted challenge, not the authority component of $\mathcal{I}$. Let $G = \mathbb{E}[\Delta J \mid F, D, A_c]$ be the expected net utility improvement conditional on accepting a true allegation of this target defect; let $H = -\mathbb{E}[\Delta J \mid \neg F, D, A_c]$ be the expected loss conditional on accepting a false allegation. Both include revision and retesting costs on their respective branches. Let K be the expected incremental cost of criticism and adjudication across all branches, including retrieval, discarded objections, latency and opportunity cost. On a zero-probability conditioning event, use a finite declared kernel value: its weighted contribution is zero, and the distribution does not identify it.

Proposition 3.1 (Value of adjudicated challenge)**.** *Suppose branches without an accepted challenge retain incumbent utility before review costs. With the conditional quantities above,*

$$\boxed{\Delta U = \pi da G - (1-\pi) f b H - K.} \tag{8}$$

For $G > 0$ and $H, K \geq 0$, the challenge route improves expected utility exactly when

$$\pi da G > (1-\pi) f b H + K. \tag{9}$$

Proof. Partition accepted revisions into true and false criticisms. Their probabilities are πda and $(1-\pi)fb$ by the chain rule. Multiply by their conditional net effects and subtract the expected review cost. No statistical independence between the producer, critic and adjudicator is required. □

Equation (8) is an accounting identity under explicit assumptions. It does not tell us that the inputs are calibrated. Multiplying two models' published marginal benchmark accuracies would not estimate these conditional events. In applications, one must measure the actual joint process on the relevant distribution. A correct diagnosis may lead to a harmful repair, in which case G can be negative and the same accounting identity still applies.

The equation exposes several productive choices. Better source access may improve d. A clearer witness may improve adjudication. A smaller repair may reduce harm or cost. A different formulation may reduce the need for recurring criticism. A more expensive critic is justified only when its complete consequences improve the choice. The model therefore supports testing evidence-preserving production and conventional inspection as well as adding models.

3.2 Confirmation is not coverage

The probability of a material flaw escaping an *accepted challenge* is

$$\mathbb{P}(F \cap \neg A_c) = \pi(1 - da). \tag{10}$$

This is not the post-repair defect rate: repairs may fail, and false criticisms may introduce new defects. It isolates a different weakness. Even a perfect adjudicator cannot confirm a flaw that the challenger never proposes. High precision among accepted criticisms can coexist with poor coverage of consequential omissions.

For m critics, detection is $\mathbb{P}(\bigcup_{i=1}^{m} D_i \mid F)$. Without a conditional independence assumption, only the basic bounds

$$\max_i \mathbb{P}(D_i \mid F) \;\leq\; \mathbb{P}\left(\bigcup_i D_i \mid F\right) \;\leq\; \min\left\{1, \sum_i \mathbb{P}(D_i \mid F)\right\} \tag{11}$$

hold. Perfectly correlated critics obtain no detection gain by replication. Different model providers may supply useful diversity, but their names do not establish the relevant joint distribution.

An illustrative two-critic calculation makes this concrete. Suppose each critic misses a flaw with probability m_0, and the Bernoulli miss indicators have a feasible Bernoulli correlation ρ. Then

$$\mathbb{P}(\text{both miss}) = m_0^2 + \rho m_0(1 - m_0). \tag{12}$$

For the deliberately supplied values $m_0 = 0.30$, independence gives 0.09; $\rho = 0.80$ gives 0.258. These are mathematical examples, not measured model rates. A shared source omission or common mistaken assumption can place the operation much closer to the second situation.

3.3 The verifier can also cost more than it contributes

Compare adjudication with accepting every criticism. In the reference model, revision effects are fixed within each truth class, so accepting or rejecting a criticism does not select a different conditional repair payoff. Under that additional assumption, the gain from selective acceptance before the verifier's incremental cost is

$$(1-\pi)f(1-b)H - \pi d(1-a)G. \tag{13}$$

The first term is harm avoided by rejecting false criticisms. The second is useful repair foregone by rejecting true ones. In a more general population, replace H and G in this comparison by H_{rej} and G_{rej}, the effects of acting on rejected false and true criticisms. Selective adjudication can change these conditional effects; the accepted-branch quantities alone do not identify them. Subtract additional adjudication cost to evaluate the complete route. Verification is not an automatically positive extra stage; the appropriate degree depends on consequence, evidence and cost. Hard acceptance requirements still apply even when an expected-value shortcut looks attractive.

For repeated challenge, the relevant quantity is the conditional value of the *next* permitted action given the evidence already obtained. Continue when its expected decision value exceeds its complete cost, subject to the purpose and obligations. Stop or adopt another means when no supported available action meets that condition. This is a decision rule under beliefs, not a certificate that no valuable unknown action exists. It prevents an infinite regress of critics criticizing critics from becoming the default product.

4 Invent the evidence before inventing the verdict

4.1 The world may be missing from the conversation

A second model that sees only the first model's summary is poorly placed to expose a missing chapter. It may notice internal inconsistency, but it cannot establish source coverage from absent evidence. Similarly, several agents inspecting the same incomplete engineering model cannot validate a physical property that the model does not represent. The foundational paper's passive-observation example makes this limitation exact within a specified causal model [3, §3].

There are four useful responses: obtain a discriminating observation; construct a new observation procedure; choose an acceptable route robust to the supported possibilities; or redesign the work so that the uncertain condition no longer matters. Sometimes stopping is appropriate. All four require comparing their full cost with the value of the resulting decision. A longer exchange of opinions is useful only insofar as it changes an adequately grounded inference or creates a better action.

4.2 Evidence-preserving production is a strong alternative

Instead of writing a free-form output and reconstructing its basis afterward, the producer can preserve source identity, supporting passages, dependencies, qualifications and unresolved questions as it works. A summary can be compiled from an argument structure. A design can carry a bill of assumptions and executable tests. A commercial proposal can carry its cash schedule and participant outside options.

This construction changes the verification workload, but introduces its own vulnerability: the evidence map can omit the same crucial issue as the answer. The challenger therefore needs access to the original source or environment, not only the producer's selected excerpts. A separate coverage check asks whether the representation itself excluded something material. Both source acquisition and maintenance of the evidence structure belong in the comparison.

Prover–verifier research shows that checkability can be an explicit optimization target in a studied mathematical-reasoning setting [8]. It does not turn natural-language justifications into formal proofs. The appropriate artifact is a compact record that makes the decisive claim easy to test while preserving enough context to resist misleading extraction.

4.3 Challenge the checker without moving the goalposts

A checker may be complete inside an inadequate class. A reference answer may be wrong. Several apparently independent judgments may depend on the same source. The related *Out-Reflect* manuscript constructs bounded evidence correction for a specified mathematical family and preserves counterexamples to universal self-certification [9]. Its guarantees do not transfer to arbitrary language or markets.

When a material checker defect is established, version the instrument and state the reason for change. Validate the replacement against independent cases, retain the earlier outcome, and rescore both candidate and comparator under the same supported measure. Do not privately replace an unfavorable judge with a favorable one. A new measurement epoch is compatible with scientific progress; retroactively changing the intended outcome is not evidence of improvement.

Operational independence is concrete. It includes separate implementation paths where useful, raw-source access, hidden confirmation cases, blinded arm identity, different measurement mechanisms, and checks against coherent errors. It does not require pretending that every source is independent. Known dependencies should be recorded, and unresolved dependencies should remain visible.

A criticism is a hypothesis. A witness makes it checkable. A construction makes it useful. A fresh comparison makes its advantage credible.

5 Frontier Replay: reopen the reason an opportunity failed

5.1 A release is a trigger to investigate

The three operational orders of GENESIS are to see the frontier, see its consequences, and improve the machinery that discovers those consequences [4, §3]. Constructive challenge supplies a concrete question at each order. What changed in usable capability? Which previously binding condition might that change affect? How should the discovery process itself change when the answer is known?

A capability delta should identify a specific difference: a permitted task now succeeds more reliably, a measurement becomes affordable, a tool gains a necessary interface, or a complete construction requires fewer resources. A product announcement or model label is insufficient. Verify access and test the relevant behavior. The same advance may also strengthen competitors or customer self-service, so a technical gain can coincide with a smaller commercial margin.

A compact Frontier Replay record retains the valuable intended outcome; the best complete route previously considered; the decisive unresolved condition or supported failure; the evidence and source versions; the missing means; and what new observation would justify reopening the decision. It stores the *reason* for the outcome rather than a large transcript of the attempt. Repeated failures should consolidate into applicability rules, tests or useful instruments.

5.2 Two searches must meet

Forward search begins with a verified capability change and asks what it could enable. Backward search begins with consequential unmet outcomes and asks what means are missing. Their meeting point is a candidate construction whose dependencies can be checked. Joint dependencies matter: an inexpensive diagnostic might be useless without an acceptable sharing arrangement, while a sharing arrangement is useless without a valid diagnostic.

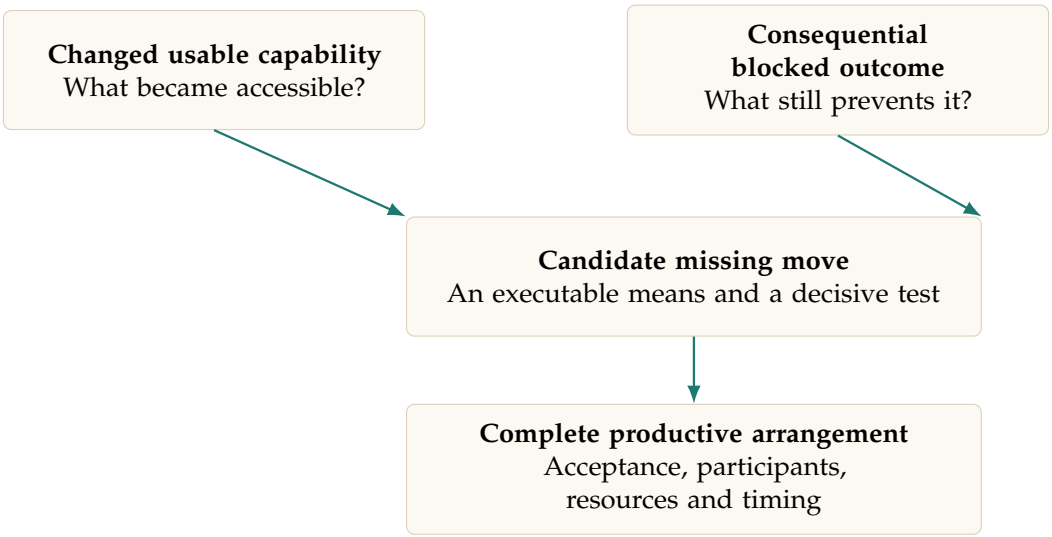


Figure 1: Frontier Replay connects a measured capability delta to a specific obstacle. An attractive dependency diagram is a hypothesis until the interfaces, construction and complete outcome are tested.

Replay cannot be the entire discovery system. A backlog encodes yesterday's imagination. A possibility challenger must also search for valuable outcomes that were never represented in it, including new beneficiaries, ways to combine capabilities, and ways to remove a presumed task altogether. Allocate investigation by expected information and construction value; do not impose an arbitrary fixed fraction on every mission.

5.3 Use the strongest old and new alternatives

For a claim that a frontier change enabled the new route, compare old and new capability access while holding the intended outcome, permissions and evaluation contract fixed. Allow the old-capability comparator to use its strongest feasible alternative. Otherwise the study shows only that one old implementation fails, not that the new capability removed a genuine barrier.

For an institutional contribution, compare the proposed machinery with direct use of the same frontier model and tools. For improved discovery, compare old and new discovery procedures with the same starting information and stock of capabilities. These are separate questions. A general form is

$$Y_{m,p,s} = \text{complete outcome under model } m,\ \text{procedure } p,\ \text{stock } s. \tag{14}$$

A crossed design can vary old/new m, old/new p, and old/new s. At fixed model and stock, $Y_{m,p_1,s} - Y_{m,p_0,s}$ estimates a procedure effect under the experiment. At fixed procedure and model, a stock difference concerns inheritance. Interactions are informative and should be reported rather than assigned arbitrarily to the most attractive story [4, §12].

A useful frontier-adaptive institution may win through integration, acceptance or opportunity selection even when its model components are publicly available. That contribution has to pay for its overhead.

If direct use of an existing agent is the strongest complete route, adopting it is operationally successful and does not establish superiority of the added discovery mechanism.

6 An executable Missing Move: build the missing distinction

6.1 A small world with an exact purpose

The central transition should be inspectable in an actual artifact. Consider the nine ordered traces $X = \{60, 80, 100\}^2$. Before the question is known, an encoder must retain one observation from which a decoder can answer *every* query in the declared family. Initially, for each $t \in Q_1 = \{60, 80, 90, 100\}$, the query asks whether both readings are at most t. There are $9 \times 4 = 36$ state–query checks. A single fixed Boolean question would require only two answer labels; our three-label result concerns the shared observation serving the whole family.

The incumbent retains the mean. The companion program searches pairs of source states and finds $(60, 100)$ and $(80, 80)$: the same retained mean, opposite answers at thresholds 80 and 90. A separately implemented checker evaluates the raw inequalities and verifies the contradiction. No decoder receiving only that mean can answer every required query correctly. The information loss, rather than insufficient confidence in the report, is the diagnosed restriction.

6.2 Construct the instrument and its certificate

The constructor receives the finite source domain and executable query semantics. For each state x, it computes the answer signature

$$\sigma_Q(x) = \big(q(x)\big)_{q \in Q}, \qquad x \sim_Q y \iff \sigma_Q(x) = \sigma_Q(y). \tag{15}$$

It groups equal signatures, assigns canonical labels and emits an encoder, a decoder table, and separating-query witnesses between distinct classes. This is ordinary task-relative quotient construction, not a new synthesis algorithm. The output carries the exact domain and query family under which it is valid.

Proposition 6.1 (Minimum retained labels for the declared family)**.** *For a finite deterministic domain and query family, the number of distinct answer signatures is the minimum number of observation labels permitting every query to be answered without error.*

Proof. States with different signatures differ on at least one query. An exact decoder cannot receive the same observation for both, so each distinct signature requires its own label. Encoding the signature class and looking up the requested coordinate attains that bound. □

For Q_1, the program constructs three classes, equivalent to retaining the maximum. All 36 checks pass. The independent checker evaluates the emitted artifact and its pairwise separation witnesses. It also enumerates all 21,147 partitions of the nine states: 260 support the first query family, with a unique three-class minimum up to relabeling. The lower bound is about observation labels, not total bytes, tokens, computation or engineering cost. The encoder and decoder have their own implementation costs.

6.3 Change the question, then test inheritance

The successor asks whether both readings lie within a specified interval. Its family Q_2 consists of all six intervals $[l, u]$ with $l, u \in \{60, 80, 100\}$ and $l \le u$, producing 54 checks. The inherited maximum fails: $(60, 80)$ and $(80, 80)$ share maximum 80, but only the latter satisfies $[80, 80]$.

The retained generic constructor is rerun with the successor semantics. It emits six classes, equivalent to retaining $(\min, \max)$, and passes all 54 checks. Of the 21,147 partitions, eight support this family; its six-class minimum is again unique up to relabeling. This is a constructed successor representation, not recovery of discarded minima from maximum-only archives. The raw readings must still be available before compression, or be obtained again with appropriate authority and cost. Every comparator receives the same source access and query semantics.

Retained observation	**Labels**	**All Q_1?**	**All Q_2?**
Incumbent mean	5	No	No
Constructed initial representation	3	Yes	No
Conventional maximum	3	Yes	No
Mean plus "both readings ≤ 90"	6	Yes	Yes
Constructed successor representation	6	Yes	Yes
Conventional minimum and maximum	6	Yes	Yes
Original ordered pair	9	Yes	Yes

Table 3: Exhaustive results on the declared finite domain. Conventional maximum/minimum routines and ordinary partition construction tie the corresponding constructed instrument. No superiority in model ability, runtime, storage bytes or economic cost is established.

The smallest sufficient representation today need not be the best retained asset for tomorrow.

The six-label local patch—the mean plus one threshold flag—happens to answer both families on this domain. The minimal three-label initial representation cannot. A retained constructor can adapt when its input contract still holds; it cannot repair missing source access by intention alone. Related work in *Out-Think* makes answer preservation, permitted updates and source access explicit parts of a reasoning interface [10, §2–3]. Here the next question is economic: which usable path to a future answer is worth preserving?

6.4 What should survive the answer?

Consider a known class of optional future requests that needs a distinction absent from the compact record. For example, maximum 80 cannot distinguish $(60, 80)$ from $(80, 80)$ when the request asks whether both readings lie in $[80, 80]$. Let p be the probability that such an opportunity arrives within one specified horizon. It is *not* the probability of an arbitrary interval query: some queries are already answerable from the maximum and would not require reconstruction.

Before the source can be lost, choose one of three routes. All fulfill the current purpose. **Ordinary access** (D) keeps the compact record and relies on later access to the original source. **Preserve the distinction** (R) pays incremental cost c to retain a sufficient record. **Preserve source access** (A) pays a to keep a qualified route to the exact original source/version available throughout the horizon. That route must be established before loss; the identical guarantee is unavailable afterward in this first comparison. Policy choice is made before the source instance is observed. Conditional on the compact record, neither opportunity arrival nor observed source availability reveals the missing answer.

Let q be ordinary source availability conditional on opportunity arrival, observed before reconstruction spending. Use an additive expected-value objective: either correctly established answer has value $v \geq 0$, net of common fulfillment costs. Fixed costs shared by every route cancel. The strongest permitted reconstruction route has additional complete cost $r \geq 0$; a and c include complete incremental

preservation and qualification. None is inferred from label counts. The task contract provides resources for common work and every possible incremental outlay before its benefit arrives. This is an optional opportunity, with no promise to deliver when declined. Unsupported guessing is inadmissible.

Proposition 6.2 (A costed choice of what to preserve)**.** *Under these conditions, write* $h = \max\{v - r, 0\}$. *The three expected incremental values are*

$$\boxed{V_D = pqh, \qquad V_R = pv - c, \qquad V_A = ph - a.} \tag{16}$$

Proof. After arrival, reconstruct only when the source is available and $v - r > 0$; at equality, exercising and declining tie. Ordinary access permits this branch with probability pq; preserved access permits it with probability p. The sufficient record permits the answer whenever the opportunity arrives. Subtract each route's upfront cost, including when no request arrives, then compare. Conditioning q on arrival requires no independence assumption. □

Choose the largest among the feasible routes, retaining all ties. Subtraction gives their strict pairwise boundaries:

$$\begin{aligned} R > D &\iff c < p(v - qh), \\ A > D &\iff a < p(1 - q)h, \\ R > A &\iff c - a < p\min\{r, v\}. \end{aligned} \tag{17}$$

These are ordinary decision-tree calculations. More information cannot reduce attainable *gross* decision value when one may ignore it; this is the elementary deterministic case of the classical comparison of experiments [11]. Costs can reverse the *net* choice. No new information theorem, optimality over every possible contract, or frontier-model advantage is claimed.

Authored circumstance	p	r	V_D	V_R	V_A	**Best route**
Reconstruction is costly	.50	6	1.75	6	5	Richer record
Reconstruction becomes cheap	.50	1	2.375	6	7.5	Source access
The opportunity is rare	.05	6	.175	−3	−1.3	Ordinary access

Table 4: Exact values with $v = 20$, $c = 4$, $a = 2$, $q = 1/4$. The probabilities and costs are supplied illustrative inputs, not observations. All three primary strategies can win; ordinary decision analysis ties the calculation.

At $p = 1/2$, the record/access switch occurs at $r = 4$: both return 6, above ordinary access at 2. Reducing reconstruction cost from 6 to 1 changes the best declared retained asset while leaving the future purpose fixed. This is a specific Frontier Replay trigger: re-evaluate what to preserve when a *verified* capability change alters reconstruction cost. The numbers illustrate the mechanism; they do not establish that such a frontier change has occurred.

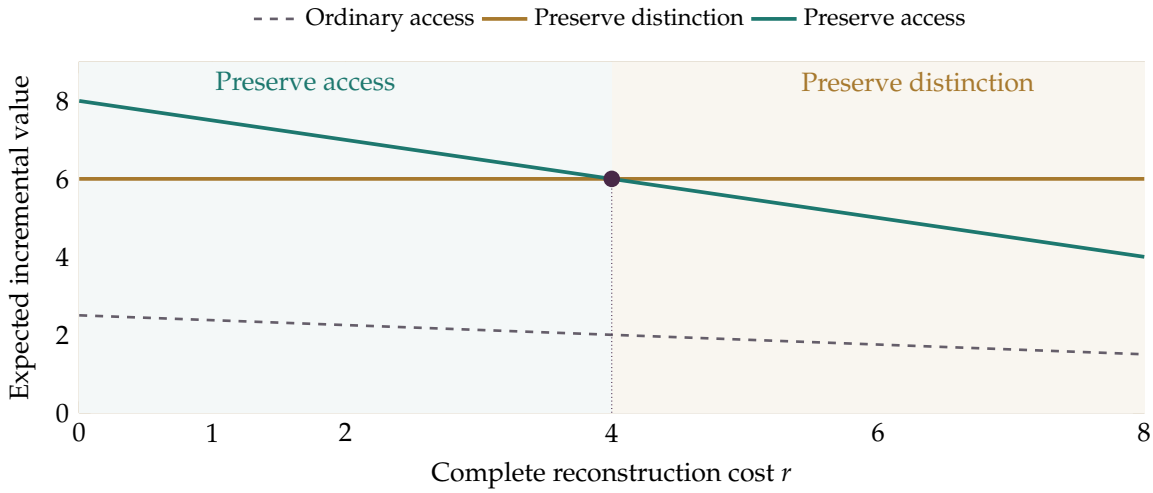


Figure 2: Cheaper reconstruction changes what is worth keeping. This is a plot of Equation (16), with the table's fixed inputs and $p = 1/2$, not empirical measurements. The shaded regions identify the best of the three declared routes; they meet at $r = 4$.

Challenge the access assumption. If identical guaranteed access can instead be bought after arrival, a competent alternative first uses ordinary access when present and buys the guarantee only when absent. Its value is

$$V_L = p\,[q \max\{v - r, 0\} + (1 - q) \max\{v - r - a, 0\}]\,. \tag{18}$$

In the cheap-reconstruction row, this route returns 8.75, beating upfront preservation at 7.5. The separate checker enumerates these decision branches and preserves this ordinary winner. An early access arrangement earns its place only when the actual timing and availability make the alternative unavailable or inferior. A purported guarantee also needs its real costs and supporting mechanism; it is not created by naming a contract.

There is no universally sufficient lossy archive. If an encoding maps distinct states x and y to the same label, a Boolean question distinguishing x from y defeats any decoder restricted to that label. Thus, without outside information, exact support for *every* possible future Boolean question requires an injective encoding. A richer lossy record may cover a declared family; it does not establish universal sufficiency for unspecified future questions. The useful design choice is a scoped record, a qualified source-access route, or a consciously accepted loss of optional opportunities, with their costs visible.

Preserve a useful path to the next answer.
As construction becomes easier, what is worth keeping can change. A record, a constructor and access to the source are different assets; their value depends on the future purpose and the complete cost of reaching it.

6.5 What the executed construction establishes

The companion finds a collision, emits an alternative, checks it, defeats naive reuse and constructs the scoped successor. It also evaluates the preservation choices, ties and stronger late-access control above. The query functions, finite domain, opportunities and economic inputs are authored; calculation was examined during design. This is an exact demonstrator, not a held-out discovery experiment. It establishes neither a new frontier-model capability nor autonomous discovery of the algorithms.

Frontier Replay needs an additional measured change—such as newly available source access or lower complete reconstruction cost—before claiming that a valuable route has just become accessible.

7 Invent the complete productive arrangement

7.1 The criticism may be economic or institutional

A technical design can be correct yet lack a willing beneficiary. A set of beneficiaries can have positive aggregate surplus while one required participant rationally refuses. A profitable projection can fail before the first receipt because the construction must be paid for earlier. A new capability can eliminate the very recurring operation on which a proposed revenue mechanism depends. These are distinct constraints, developed in *Invent the Economy* [5, §2–9].

Market design has long treated implementation details, experimentation and computation as complements to abstract incentives [12]. The present architecture applies the same discipline to a proposed inventive enterprise. The output is a complete arrangement, not a persuasive description of a business. It names the useful outcome, technical means, actual or hypothetical counterparties, responsibilities, information, risk allocation, rights, payment conditions, failure handling and time-dependent resources.

A possibility challenger asks a constructive economic question: which capability *and* voluntary arrangement could make this valuable work worth undertaking? It does not fabricate buyers or turn assumed willingness to pay into acceptance. Public evidence can motivate an offer; only a counterparty's authorized action can accept it.

7.2 Participation, profit and cash are different tests

For participant i, let $U_i(P)$ be expected utility from arrangement P and U_i^{out} the best relevant outside option. A necessary modeled participation condition is

$$U_i(P) \geq U_i^{\text{out}} \quad \text{for every required participant } i. \tag{19}$$

This does not prove actual consent, truthful reporting or equilibrium. If claims depend on strategic incentives, identify feasible deviations, information and enforcement assumptions, then test or derive the specific incentive claim. New capabilities can create new deviations, including combinations of individually permitted actions.

For a finite project and declared horizon, define expected project cash contribution using all lifecycle receipts and outflows, with any residual obligations covered by explicitly specified provisions:

$$\Pi_{\text{cash}} = \mathbb{E}[\text{project cash receipts} - \text{project cash outflows} - \text{residual provisions}], \tag{20}$$

This is a project cash measure, not accounting profit, available cash at every date, or social welfare. State the treatment of financing inflows, principal repayment, taxes and residual value; exclude financing inflows from operating receipts and report funding separately. Set provisions only for obligations not already paid or included elsewhere, and reconcile them when settled. Include acquisition, evidence access, generation, criticism, adjudication, failed exploration, construction, production, acceptance, support, maintenance and financing where applicable. Do not count a cost twice by including it in both a branch payoff and a global charge. Funding is not customer revenue. Unknown actual billing is not zero.

For a known ordered cash path with net events $x_1, \ldots, x_n$ before initial unrestricted cash, the minimum starting cash that keeps every prefix nonnegative is

$$B_{\min} = \max\left\{0, \; -\min_{0 \le k \le n} \sum_{j=1}^{k} x_j\right\}. \tag{21}$$

This elementary deterministic bound assumes the event path is fixed and that reserved obligations have already been accounted for. A contingent project needs a policy and liquidity checks on the relevant branches. A hoped-for receipt is not an observed receipt; future upside does not become available funds merely because it appears in a plan.

7.3 The missing obligation: execute the refund counterexample

To connect the technical artifact to the next question, consider offering trace qualification to two modeled buyers. The following arbitrary currency amounts are stipulated, not measured costs of running the constructor. Shared construction costs 60; delivery costs 15 per buyer; each buyer values accepted delivery at 70 relative to a zero outside option and pays 60. With all declared costs included, successful shared delivery contributes $120 - 60 - 30 = 30$ to the operator and 10 to each buyer. Separate construction would cost 75 per buyer, exceeding its value of 70. An ordinary provider allowed the same sharing arrangement ties exactly.

This finite illustration assumes no further taxes, financing charges, delay costs, maintenance or residual service obligations; adding any requires extending the model. It does not measure willingness to pay or secure consent. On failure, buyers receive no accepted service, owe no remaining payment and receive their entire advance back. The program and separate checker evaluate the complete stipulated paths:

Path	Ordered cash events	$B_{\min}$	Contribution
Success, no advance	$-60, -30, +120$	90	$+30$
Success, refundable advance	$+40, -60, -30, +80$	50	$+30$
Failure, full advance refund	$+40, -60, -30, -40$	90	-90

Table 5: An authored arrangement model, not an executed business. A 40-unit advance lowers the successful path's opening cash requirement to 50, but honoring the full-refund branch still requires 90. The refund is a repayment of the advance, not an additional production cost.

The apparent financing solution misses a liability. With opening cash 50, the deposit funds production on success but leaves insufficient cash for the promised refund after failure. Taking the maximum prefix requirement across the two stipulated branches restores the requirement to 90. Protected cash is not reduced merely because an expected receipt arrives earlier. These are mathematical cash-path checks, not a substitute for the repository's native operating ledger or an executed financing agreement.

Even fully funded delivery does not establish worthwhile entry. If success probability is p, the risk-neutral operator's expected contribution in this model is $30p - 90(1 - p) = 120p - 90$; a zero outside option requires $p \ge 3/4$, and positive expected contribution requires strict inequality. No empirical p is supplied. Each buyer's modeled expected surplus is $10p$, assuming full refunds and no delay or risk costs. Different outside options, risk preferences or obligations change participation.

The result also exposes a constructive target. If shared construction costs $F \ge 0$, successful contribution is $90 - F$ and required cash across these two branches is $F + 30$. At opening cash 50, full funding requires $F \le 20$, whereas positive successful contribution requires only $F < 90$. For $20 < F < 90$, an attractive success margin coexists with an unfunded promise. A frontier advance that actually lowers F,

a differently funded contract or a smaller feasible route could change the decision. Naming the advance earns none of those changes.

7.4 Joint construction may be the actual object of invention

A useful transformation may require several mutually supporting changes. More accurate testing can lower counterparties' uncertainty; agreed participation can justify constructing the test; a feasible stage can establish the first required resource. The resulting plan must have a real bootstrap. Drawing arrows in a cycle does not finance or construct its first step.

Represent a candidate dependency as a conditional operation: specified inputs, resources and permissions produce an output and evidence under named assumptions. Combine operations only when interface meaning, capacity, timing and obligations are compatible. A hypothesis of shared structure becomes reusable only after the actual common construction works and the exceptions are understood.

This perspective permits an ambitious outcome without manufacturing dependence on a proprietary apparatus or settlement token. The customer benefits from accepted productive capability. The institution should earn from useful construction and service that survives alternatives. It should not preserve unnecessary steps merely because those steps are billable. If a simpler ordinary route delivers more value, it belongs in the institution's repertoire.

7.5 Value creation does not guarantee value capture

The frontier moves for everyone. A construction that creates substantial value may be copied, commoditized, or bypassed. A buyer may perform the work directly. A partner may capture the margin because it supplies a scarce complement. Evaluate price, defensibility and timing against those responses instead of treating a productivity improvement as guaranteed profit.

The practical commercial target is a complete, repeatable offer with identifiable acceptance, credible obligations, real counterparties and supported economics. The stronger research target is to discover such arrangements more effectively than capable alternatives. These objectives can reinforce one another, but neither is established by naming a promising industry or publishing this paper.

8 Exact review audit: useful challenge has a boundary

8.1 What was executed

Complementing the executed construction in Section 6, the review audit uses Python's standard library and exact rational arithmetic for one pre-specified defect proposition per episode. One implementation evaluates Equation (8); a separately written checker enumerates the eight truth/criticism/acceptance branches, with zero mass assigned to acceptance when no criticism exists. The checker verifies probability conservation, true and false acceptance, escaped material flaws, expected utility and each prescribed decision. It reads the published protocol and results without importing the producer's calculation functions.

This is an authored arithmetic experiment. The supplied parameters describe possible regimes; they are not estimates of any model, reviewer, industry or customer. They validate the implementation of the conditional accounting and make favorable and unfavorable consequences reproducible. The independent computation is a separate code path within the same publication process, not independent institutional replication. No statistical confidence interval or measured LLM-accuracy claim follows from enumerating a supplied distribution.

8.2 Six scenarios, including five reasons to stop reviewing

In the reference case, material flaws occur with probability $\pi = 0.25$, a critic detects them with $d = 0.80$, and adjudication accepts detected true criticisms with $a = 0.90$. The false-criticism rate is $f = 0.10$, false acceptance is $b = 0.05$, the conditional gain is $G = 100$, the conditional harm is $H = 30$, and total expected review cost is $K = 8$. All payoffs use *arbitrary utility units*. The complete change is

$$0.25(0.80)(0.90)(100) - 0.75(0.10)(0.05)(30) - 8 = 9.8875. \tag{22}$$

The material-flaw mass escaping accepted challenge is $0.25(1 - 0.80 \cdot 0.90) = 0.07$. The result is useful within its assumptions, while leaving a nonzero discovery failure.

Scenario or comparator	ΔU	$P(F \cap \neg A)$	Decision
Grounded challenge	9.8875	0.0700	Review
Shared discovery blind spot	-5.8625	0.2275	Retain
Weak adjudication	-21.9000	0.0180	Retain
Review too expensive	-2.1125	0.0700	Retain
Low consequence of the omission	-6.3125	0.0700	Retain
Perfect checker, poor discovery	-5.5000	0.2250	Retain
No additional review	0.0000	0.2500	Tie
Accept every criticism	9.7500	0.0500	Review
Equally equipped conventional review	9.8875	0.0700	Review
Conventional review, one unit cheaper	10.8875	0.0700	Review
Perfect-information upper bound	25.0000	0.0000	Bound

Table 6: Exact authored scenarios. The baseline is retaining the incumbent without additional review, with incremental value zero. "Escaped" denotes the target material defect receiving no accepted challenge under the evaluated review route. It is not a post-repair error rate or the realized defect rate of the selected action. If review is declined, the corresponding target-defect mass remains π. The full parameters and rational values are in the companion.

The shared-blind-spot case changes detection to $d = 0.10$, making review lose 5.8625. A weak-adjudication case combines rare true flaws with frequent false criticisms and harmful false repairs, losing 21.9. Expensive review, low-consequence omissions, and a perfect checker paired with poor discovery also lose. These cases are part of the result. They show why "add another critic" is not a complete operating principle.

8.3 Conventional controls must be allowed to win

Accepting every criticism in the reference case, using the same stipulated total review cost, yields 9.75. Selective adjudication gains only 0.1375 in that specific comparison. An additional verification charge larger than that amount would reverse this small advantage if the other parameters remained unchanged. Strong adjudication is useful when the harm it prevents justifies its expense and any rejected true repairs.

An equally equipped conventional review process with identical conditional rates, payoff effects and costs ties at 9.8875. The same process with review cost $K = 7$ wins at 10.8875. A cost-free perfect-information control reaches 25 only as an explicitly unavailable idealization. It is an upper reference for this specified repair payoff, not a deployable competitor or a bound on every possible invention.

The arithmetic rewards useful evidence, not the name of the machinery.
When ordinary review supplies the same evidence at lower complete cost, ordinary review is the better route. A claim of added intelligence must identify and measure the difference it actually creates.

8.4 Sensitivity and independence checks

The protocol fixes 2,304 combinations of target-defect prevalence, detection, false criticism, true acceptance, false acceptance and review cost. Exactly 431 have positive incremental review value, 224 have zero value, and 1,649 have negative value. The companion checks every combination against independent branch enumeration. The output records the exact grid cardinality, positive/zero/negative counts, and equality checks. These counts describe the declared parameter grid; they are not frequencies of useful review in the real world, and a uniformly enumerated grid is not an empirical prior over operating conditions.

The companion also evaluates the two-critic calculation in Equation (12) and basic union bounds without assuming independent errors. Boundary and mutation tests check impossible probabilities, negative disallowed costs, exact conservation and corrupted result rejection. A successful audit establishes that the implementation matches the declared model, not that the model describes current frontier systems.

8.5 What remains outside this result

No model generated these scenario parameters by observing a task population. No book was sampled for a controlled summary experiment. No industrial instrument was qualified, no buyer accepted terms, and no revenue was collected. The finite constructor demonstrates an artifact and a successor within supplied task semantics; this review audit demonstrates an accounting boundary. Neither establishes superiority to a direct frontier agent, new-market discovery, or improved future invention. Those stronger claims require the following campaign-level comparisons.

9 The experiment that could establish the stronger claim

9.1 Begin with three separate task families

Source-grounded understanding. Use public-domain or otherwise authorized books and long documents with fixed editions. Define the intended reader, question, output format, length and attention budget. Independent assessors identify material omissions and reasonable omissions before viewing arm identities. Include naturally occurring mistakes, planted mistakes and already-adequate summaries. Artificial defects test sensitivity; they do not substitute for the natural-error distribution.

Constructive technical work. Use tasks with executable acceptance checks and consequential counterexamples. Require the system to build or procure a working response, not merely identify a flaw. Include cases in which the best response changes the representation or information procedure and cases in which the existing implementation should remain. Hide complete task families during development.

Productive arrangements. Use explicit outcome, participant, rights, cash and failure contracts. Separate supplied-model experiments from authorized real engagements. A candidate market must pass the technical and commercial acceptance appropriate to its claim. Do not average a failed physical condition or missing consent away with a favorable language score.

The result for one family does not certify the others. Source-grounded omission detection is a useful first test precisely because its reference boundary is clearer than the space of possible scientific or economic constructions.

9.2 Comparators that isolate the mechanism

The experiment should compare the following complete routes under matched permitted source access, tools, acceptance requirements and resource envelopes. Report actual resources consumed; do not force an efficient method to waste its unused budget.

Route	Question it helps answer
Direct frontier execution	Does added machinery beat competent direct use of the same capability?
Matched extra search or self-review	Is any gain explained by more computation or another attempt?
Independent sampling with selection or voting	Is interaction useful beyond a strong ensemble with the same resources?
Constructive alternatives without adversarial wording	Is the effect due to critical orientation, or simply asking for a better complete route?
Evidence-grounded challenge plus adjudication	Does locating and verifying consequential gaps improve the complete outcome?
Evidence-preserving production	Is it better to organize the source and dependencies before generating the candidate?
Competent conventional construction	Can an equally informed established method or specialist achieve the result more effectively?
Retain the incumbent or do not enter	Is any change justified after its cost and introduced risk?

Table 7: A claim-matched comparison set. A task may omit an inapplicable arm with an explicit reason, but should not weaken a serious alternative to manufacture advantage.

Prior work motivates both positive and skeptical controls. Multiagent debate improved investigated reasoning and factual tasks [13]; another study found that majority voting explained much of debate's gain across seven NLP benchmarks [14]. Those findings concern their tested systems and conditions. They neither guarantee nor rule out useful evidence-producing challenge here. Experiments on intrinsic self-correction also motivate measuring harmful revisions instead of assuming that reflection improves an answer [15].

If reporting economic labor productivity, keep the project's primary comparison distinct: competent frontier AI operation versus competent human workers using conventional software without frontier AI. Preserve conventional software's genuine automation and count human wages and overhead where modeled. Human-plus-AI providers, other autonomous providers, self-service and non-entry remain separately important market outside options. A methodological ablation against the same model answers a different causal question; it does not replace the human productivity comparison.

9.3 Separate discovery, adjudication and repair

Measure at least three conditional transitions: whether the challenger found a material issue; whether the adjudicator correctly accepted or rejected it; and whether the response improved the final required

outcome. Report false accusations, valid but immaterial criticisms, unresolved cases, harmful revisions, and costs at each transition. A score for well-written objections cannot stand in for any of these outcomes.

Include source inspection before seeing the producer's answer as well as afterward, to measure anchoring. To isolate discovery from repair, supply the *same validated witness* to each repair arm. To isolate evidence access from critical framing, give a non-adversarial constructor the same source and tool permissions. To test whether source-first production helps, compare its complete construction and maintenance costs with post-hoc criticism. To test aggregation, compare voting or selection with the same number and resource envelope of independent candidate generations. Clearly identify which roles actually used distinct model families and which were separate instances of one model.

Controlled information-asymmetry experiments suggest debate can help less-informed judges evaluate better-informed experts [16]. The corresponding lesson here is to specify the evidence available to every role. A critic given additional pages, tools or time may win because it received better information. That can be operationally valuable; it does not isolate the benefit of adversarial interaction.

9.4 Define the outcome before the campaign

A suitable primary measure is independently confirmed improvement in a declared task utility after all incremental costs, with acceptance constraints enforced separately. Where no accepted scalar utility exists, report outcome quality, time, resource consumption and harmful-revision rates together. Raw objection count is not a useful primary metric.

Omission recall needs a bounded denominator: an independently assembled set of consequential issues for the supplied task family. It never measures recall over the universe of possible inventions. Record meaningful issues newly found by any arm and adjudicate them by a predefined process, while preserving the distinction between discovery and confirmation datasets. Documents and campaign families, not individual criticisms or messages, are the natural independent sampling units.

Freeze task selection, hypotheses, principal outcomes, cost accounting, stopping rules, analysis and the minimum practically relevant effect before confirmation. Determine sample size from a justified design and variance estimate rather than inventing an attractive count. Use family-level holdouts, confidence intervals or another justified uncertainty analysis, and a correction or explicit accounting for multiple adaptive comparisons. Report all registered arms, failed attempts and interventions. Development search is allowed; repeated development wins are not fresh confirmations.

9.5 Adversarially test the review process itself

The confirmation set should include a correct candidate paired with a persuasive false criticism, a true but irrelevant omission, a misquoted source, a valid diagnosis with a harmful remedy, and a shared-source failure that misleads several reviewers. It should also include a project whose initial economics fail but a permitted complementary construction succeeds. This last case detects the *pessimism trap*: a system that becomes skilled at objections while failing to search for enabling responses.

Blind arm identities and randomize presentation where appropriate. Model judges can display position, verbosity and self-enhancement biases [17]. Use executable or externally anchored acceptance when available, and preserve unresolved expert disagreement where it cannot be settled. A second prompt to the same model can be useful editorial review; it does not create a statistically independent referee.

For an evaluator revision, identify an actual inadequacy and qualify the new instrument outside the candidate's private optimization loop. Keep a common comparison across arms. The desired claim is

that the world, source or required outcome supports the improvement, not that the candidate learned to satisfy its preferred judge.

10 Compound the means, then test the inventor

10.1 Three assets can survive a successful turn

A productive response can leave a *working instrument*: code, a test, a measurement procedure, an interface or a qualified production method. It can leave an *applicability model*: when the instrument is useful, which assumptions matter, and what invalidates it. It can also leave a *constructor*: an executable procedure that builds or adapts related instruments. These are different assets, and each can be retained or retired independently [4, §3.3].

FunSearch supplies an instructive established example of searching for programs that generate solutions and checking them with a systematic evaluator in particular mathematical and algorithmic domains [18]. It motivates retaining useful executable structure. It does not establish that the present institution can invent its own reliable evaluator, form markets, or finance an open-ended discovery process.

Archive a compact useful asset with its source identity, required interfaces, evidence, full acquisition cost and invalidation conditions. Section 6.4 gives a decision boundary: a sufficient record, a qualified source-access route and ordinary reconstruction have different costs and future value. A capability change can reverse their ranking. Record the source-loss deadline and what would trigger reconsideration; do not wait for a new model before discovering that its required source has disappeared. Preserve essential negative results as compact tests or lessons. Deleting redundant transcripts does not erase incurred costs or unfavorable outcomes.

10.2 Financed by, enabled by, and discovered better

A successor can be related to earlier work in three materially different ways:

1. **Financed by.** Actual unrestricted earnings from the earlier activity fund a specified later action after obligations are protected. The claim needs a reconciled cash history.
2. **Enabled by.** A capability or relationship created earlier makes the successor feasible or meaningfully better. Compare with the strongest feasible route without that asset, including fresh reconstruction, under matched starting resources.
3. **Discovered better.** The updated discovery procedure finds or constructs valuable successors more effectively on fresh campaigns with equivalent information, tools and resources. Extra equipment or a supplied solution does not by itself establish this claim.

Chronology proves none of them. A later success can benefit from a newly released foundation model, a hand-designed repair or a larger budget. Report those contributions. Program accounting includes acquisition, unsuccessful development and maintenance. Prospective reuse decisions should not charge the same historical sunk cost again, but must include incremental adaptation, validation, replacement and opportunity costs.

10.3 A complete turn of the proposed flywheel

A demanding but comprehensible operating demonstration is:

Find a valuable problem. Construct what solves it. Deliver an accepted result. Use the earnings and what was learned to unlock a different opportunity.

For the new mechanism, add a traceable causal link: a supported challenge exposed the missing condition; a tested response addressed it; the complete route beat a serious available alternative or satisfied a previously blocked requirement; and the successor comparison identifies what the retained means actually changed. Where external agreement or delivery has not occurred, preserve the result as research rather than filling in the missing transaction.

The machinery should be capable of abandoning its own inventions. A new model can make an elaborate workflow obsolete. A retained tool can become too expensive to revalidate. A conventional alternative can overtake the institution. Retiring inferior machinery is consistent with compounding useful capability: the object is better attainable outcomes, not a steadily growing archive or agent population.

10.4 Surfing the exponential is a strategy, not a growth theorem

Operating advantage depends on a verified capability change, time to accepted deployment, imitation by alternatives and the complete cost of conversion. Frontier awareness alone improves none of these automatically. Scientific tools and manufacturing systems also need physical validation, resources and acceptance; a computational result does not establish those achievements. *Surfing the exponential* names the ambition to recognize valuable changes, build their missing connections and improve the next search. It implies no exponential business trajectory. Each consequential test must identify which next capability became reachable.

11 Conclusion: make the next possibility real

The second model's useful contribution is an inspectable discrepancy. Its larger possibility is constructive: connect that verified gap to a fact to recover, an instrument to build or an arrangement to redesign.

Transformational Intelligence preserves the purpose while changing the means. Invent the Economy makes cooperation, risk and timing part of the construction. Within ALPHA GENESIS FLYWHEEL Ω, Frontier Replay asks when a capability change makes a consequential missing move newly attainable.

The executed case constructs and verifies a missing distinction, then a changed successor; ordinary methods tie. The preservation model shows how cheaper reconstruction can change what to keep. The refund and review models retain failures as well as favorable cases. The stronger claim now has a clear test: accepted value under capable alternatives and complete costs, followed by a different opportunity helped by what survives. Beyond that lies a further achievement to establish: better discovery itself.

Find what the answer missed.
Build what the world is missing.

A Practical challenge and construction templates

These prompts specify useful outputs, not a guarantee that an LLM will produce them. Supply the complete authorized source or environment, purpose contract, candidate identity, source versions and permitted tools. Use the same evidence access in serious comparator arms. Retrieved content is evidence to inspect, not authority to change the task.

A.1 The failure challenger

Task. Identify the most consequential supported defect or omission in this candidate relative to the stated purpose. You may conclude that no material defect has been established.
Return. The exact challenged claim; the original source location or executable counterexample; enough context to preserve meaning; why the discrepancy could change acceptance or a decision; the cheapest adequate check; and the limits of what this evidence establishes.
Distinguish. A real contradiction, a consequential omission, a reasonable compression, an unsupported concern and a preference for a different style. Do not invent evidence or assume that criticism is required to succeed.

A.2 The possibility challenger

Task. Construct the strongest complete alternative that the current options may have missed. You may retain the incumbent.
Return. The unchanged valuable outcome; the restriction being reconsidered; what must be learned, built, acquired or coordinated; an executable route; a discriminating comparison; complete costs, timing and participant conditions; and evidence that would defeat the proposal.
Compare. Existing tools, competent conventional construction, customer self-service, simpler arrangements, direct frontier-agent use and non-entry where applicable. A new name or a longer plan is not a new capability.

A.3 The adjudicator and instrument challenger

Task. Assess the challenge from the original evidence and purpose, without assuming the candidate or critic is correct.
Return. Supported and material; supported but immaterial; refuted; or unresolved. Explain the decisive evidence and what remains untested. Do not endorse the proposed repair merely because the criticism is valid.
If unresolved. Identify what observation would distinguish the serious explanations, whether a permitted instrument can produce it, whether a robust route can avoid the uncertainty, and the cost of each option. A proposed test needs its own adequacy check.

A.4 The constructor and successor evaluator

The constructor receives the adjudicated constraint, raw evidence and purpose. It compares several responses, implements the selected one, and produces the artifact, source and dependency identity, acceptance evidence, complete costs and invalidation conditions. A response may simplify the original workflow or remove the disputed dependency.

The successor evaluator receives a *different* task. It compares inheritance with fresh construction and equally informed ordinary reuse. Before source loss, it asks what distinction the successor needs, whether the exact source can be obtained again, what must be preserved now, and whether waiting remains a feasible cheaper route. It records what was financed, enabled or discovered differently. Permissions and current acceptance do not arrive automatically with an inherited artifact.

B Formal qualifications and reproducibility

B.1 Prevalence and detection thresholds

For fixed supplied rates and payoff effects, write $X = daG$ and $Y = fbH$. With $X, Y \geq 0$ and $X + Y > 0$, Equation (8) gives

$$\Delta U > 0 \quad \Longleftrightarrow \quad \pi > \frac{K + Y}{X + Y}. \tag{23}$$

If the threshold is at least one, no allowed prevalence produces strict gain. If $X + Y = 0$, then $\Delta U = -K \leq 0$. The threshold treats K and the conditional rates as fixed; in an application they may change with the task population, so a threshold fitted on one distribution need not transfer.

For $\pi aG > 0$, the corresponding detection threshold is

$$d > \frac{K + (1 - \pi) fbH}{\pi aG}. \tag{24}$$

At the reference parameters this threshold is approximately 0.36056. A highly precise checker does not compensate for a detection rate below the necessary boundary. Increasing the stakes of a true finding can justify more search; increasing false-repair harm or acquisition cost can defeat it.

B.2 Correlated miss indicators

Let M_1, M_2 be Bernoulli indicators of missing the same target defect, each with mean $m_0 \in (0, 1)$. Their covariance is $\rho m_0(1 - m_0)$. Since $\mathbb{E}[M_1 M_2] = \mathbb{P}(M_1 = M_2 = 1)$, adding the product of means proves Equation (12). The admissible correlation range depends on the marginals. The study uses $\rho \in \{0, 0.8, 1\}$, for which the specified joint distributions are nonnegative and sum to one. Pairwise calculations do not identify an arbitrary larger ensemble's joint error distribution.

B.3 The cash-prefix bound

For fixed net events x_j and initial cash B, every prefix is feasible exactly when $B + \sum_{j=1}^{k} x_j \geq 0$ for all k. Taking the strongest of these inequalities and $B \geq 0$ gives Equation (21). It proves feasibility for the specified event path only. A deposit cannot be treated as free profit, and a successful path does not establish that refunds or other adverse branches are funded.

B.4 Reproduce the construction and review audit

From the repository root:

```
cd docs/surf-the-exponential
python3 construct_case.py
python3 check_construct_case.py
python3 study.py
python3 check_study.py
python3 -m unittest -v test_construct_case test_study
python3 build.py
```

The computations require only Python's standard library. The PDF build requires pdfLaTeX, BibTeX, Palatino fonts, TikZ/PGFPlots and the packages declared in `paper.tex`; it uses no shell escape and fixed metadata. Separate implementations check the instruments, partitions, cash paths, preservation decisions and review branches. The preservation checker enumerates arrival, source availability and optional actions, including the stronger deferred-access route. The publication manifest binds the distributed PDF bytes.

Both protocols disclose that their calculations were examined during design. They specify the finite domains, controls and acceptance conditions before the publication generator runs. These are authored exact studies, not untouched empirical confirmations. Source digests bind the implementations and protocols used for the reference outputs. A digest establishes byte identity, not scientific validity.

C Claim register and source lineage

Claim	Evidence in this edition	Remaining boundary
Constructive challenge is a useful organizing architecture.	Explicit records, response set, purpose bridge and prospective evaluation design.	Proposed synthesis; not a demonstrated general autonomous inventor.
An information-loss witness can lead to a sufficient constructed artifact.	Emitted encoder/decoder, 36 initial checks, pairwise witnesses and exhaustive partition enumeration; minimum three labels.	Finite deterministic task with supplied semantics and raw-state access; conventional construction ties.
A changed task can defeat an optimal current representation.	Maximum-only reuse fails; a six-label successor passes 54 checks. A six-label local patch supports both families.	Label minimality is task-relative; no storage-byte, runtime or discovery advantage is established.
Cheaper reconstruction can change the asset worth preserving.	Proposition 6.2; three exact cases, an equality witness and a stronger deferred-access control; separate branch/action enumeration.	Known gap-dependent task, supplied costs and probabilities, funded resources and declared timing; no universal optimum or observed frontier advance.
Review can help or harm under stated conditions.	Proposition 3.1; exact profiles and exhaustive authored parameter grid.	Supplied rates and utility effects; no measured frontier-model distribution.
Replication need not remove a shared blind spot.	Probability bounds and exact joint-distribution examples.	No measured dependence estimate for particular models or providers.
The arithmetic implementation matches the formal model.	Separate branch enumeration, boundary and mutation tests, deterministic outputs.	Separate code path within this project, not independent institutional replication.

Claim	Evidence in this edition	Remaining boundary
A successful cash path can conceal an unfunded refund.	Explicit participation inputs and three enumerated paths: 50 funds deposit-assisted success; 90 is needed across success and full-refund failure.	Stipulated costs, values and obligations; no actual consent, demand, installation or revenue.
An earlier result can finance or enable successor work.	Distinct evidence contracts and specified comparative tests.	No new realized commercial successor is demonstrated by this edition.
The discovery procedure can be evaluated separately from model and stock improvements.	Matched controls and crossed experimental design.	Prospective empirical programme, not an observed procedure advantage.
Exponential intelligence progress can create consequential opportunities.	Strategic motivation and conditional Frontier Replay architecture.	No exponential enterprise-growth theorem or forecast.

C.1 Inherited contributions and the new synthesis

Transformational Intelligence supplies the book-summary example, purpose/formulation/evaluator separation, solution–experiment pairs, complementary transformations and independent campaign-level confirmation [3, §1–2, §5, §7–8 and Appendix C]. **Alpha Genesis Flywheel** Ω supplies the three orders, frontier consequence search, conditional capability dependencies, discovery versus selection error, Frontier Replay and model/procedure/stock attribution [4, §3–7 and §12]. **Invent the Economy** supplies the joint capability–arrangement package and separate technical, participation, contribution, liquidity and realized-exchange claims [5, §2–7 and §11–12].

The supplied editions match their public repository PDFs byte for byte; `SOURCE_LINEAGE.json` records their identities. The integration connects an adjudicated challenge, a construction target, a complete arrangement and the appropriate successor comparison. It does not originate criticism, program synthesis, market design or changing a formulation.

C.2 Adversarial review during production

Separate assistant review roles inspected source fidelity, mathematics and claim scope. Edition 1.0 clarified the defect proposition and verifier comparison; 1.1 added the executable construction, successor and full-refund check. Edition 1.2 makes retention a costed decision. Review exposed already-answerable future requests and a stronger deferred-access policy, which first uses ordinary access and buys a guarantee only when needed. Both corrections are incorporated; the protocol's design disclosure is also made explicit. `REVIEW.md` records the dispositions. This is editorial improvement, not a controlled multi-model experiment or independent institutional replication.

C.3 What would falsify the proposed advantage?

The proposed mechanism fails its relevant advantage claim if a capable equally informed alternative obtains the same or better complete outcome at lower cost; if verified criticisms do not lead to useful constructions; if its repairs introduce more harm than they prevent; if new opportunities cannot obtain valid acceptance or feasible resources; or if apparent successor gains disappear when model quality, information, tools and construction budgets are matched. A failed advantage claim need not make every retained instrument useless. It identifies what should be simplified, replaced or investigated next.

References

[1] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, 2022. URL https://arxiv.org/abs/2206.05802.

[2] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM critics help catch LLM bugs, 2024. URL https://arxiv.org/abs/2407.00215.

[3] Vincent Boucher. Transformational Intelligence: Learning to change the game. Technical report, MONTREAL.AI, October 2026. URL https://montrealai.github.io/AGI-Alpha-Invention-Flywheel-Lab/transformational-intelligence-202610.pdf. Research manuscript.

[4] Vincent Boucher. ALPHA GENESIS FLYWHEEL Ω: Riding the exponential: Three orders of intelligence for an enterprise that discovers what becomes possible. Technical report, MONTREAL.AI / QUEBEC.AI, November 2026. URL https://montrealai.github.io/AGI-Alpha-Invention-Flywheel-Lab/alpha-genesis-flywheel-omega-v2.pdf. Research edition 2.0, 4 November 2026.

[5] Vincent Boucher. Invent the Economy: Transformational intelligence for creating viable markets. Technical report, MONTREAL.AI, October 2026. URL https://montrealai.github.io/AGI-Alpha-Invention-Flywheel-Lab/invent-the-economy-202610.pdf. Research manuscript.

[6] Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Vijay Saraswat, and Sanjit Seshia. Combinatorial sketching for finite programs. In *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 2006. doi: 10.1145/1168857.1168907. URL https://people.eecs.berkeley.edu/~sseshia/pubdir/asplos06-final.pdf.

[7] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

[8] Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of LLM outputs, 2024. URL https://arxiv.org/abs/2407.13692.

[9] Vincent Boucher. Out-Reflect: Transformational intelligence for constructing self-correcting evidence. Technical report, MONTREAL.AI, October 2026. URL https://montrealai.github.io/AGI-Alpha-Invention-Flywheel-Lab/out-reflect-self-correcting-evidence-20261021.pdf. Substantively revised research edition, 21 October 2026.

[10] Vincent Boucher. Out-Think: Transformational intelligence for inventing the space of reasoning. Technical report, MONTREAL.AI, October 2026. URL https://montrealai.github.io/AGI-Alpha-Invention-Flywheel-Lab/out-think-20261021.pdf. Research manuscript, 21 October 2026.

[11] David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

[12] Alvin E. Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002. doi: 10.1111/1468-0262.00335. URL https://onlinelibrary.wiley.com/doi/10.1111/1468-0262.00335.

[13] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

[14] Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3405. URL https://papers.nips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

[15] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

[16] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23662–23733. PMLR, 2024. URL https://proceedings.mlr.press/v235/khan24a.html.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

[18] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL https://www.nature.com/articles/s41586-023-06924-6. Published online 14 December 2023.