Retro-Engineering Reality: Testing Retrocausal Selection by Future Devices in Today's Data

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 1, 2025

Suppose a future, ultra-capable information processor—say a post-selected quantum computer or fixed-point causal solver—can weight which past microhistories become our shared record, subject to ordinary physics and no-signaling. Then some puzzling features of today's world—quantum delayed-choice oddities, just-in-time discoveries, convergent evolution, improbable near-misses—could be statistical fingerprints of retrocausal selection rather than violations of known laws. This paper develops that hypothesis in a testable way. We formalize "future utility functions" as boundary conditions that bias otherwise lawful dynamics toward histories that score highly at a later time T. We then map twelve candidate phenomena to measurable predictions and propose preregistered protocols that deliberately lock future scoring functions before any data are taken. The aim is not metaphysics, but falsifiable signatures: conditional tilts that appear only when a future utility is well-defined, and that vanish in sham or de-scored arms. Alongside theoretical framing (two-state and fixed-point perspectives), we specify information-theoretic metrics (mutual information, MDL changes), quantum-compatible checks (within no-signaling), and governance guardrails for open, multi-site replication. Whether the outcome is positive or null, the program clarifies what "engineered reality" would mean in practice and offers a clean separation between story-like coincidences and statistically disciplined evidence.

Keywords: retrocausality, post-selection, two-state vector formalism, fixed-point causality, quantum eraser, entanglement swapping, no-signaling, algorithmic compressibility, mutual information, minimum description length, preregistration, utility function, convergent evolution, synchronicity, near-miss analysis, replication crisis, experimental design.

A collaboration with GPT-5. CC4.0. 57 pages.

Outline

1. **Motivation and Scope**
   1.1 Puzzles that invite a retrocausal reading
   1.2 From stories to tests: separating anecdotes from signatures
   1.3 Definitions: history, boundary condition, future utility

2. **Conceptual Framework**
   2.1 Retrocausal selection as boundary weighting, not law breaking
   2.2 Three mechanisms: post-selection, final-state constraints, fixed-point solvers
   2.3 Guardrails: conservation, locality, and no-signalling

3. **Statistical Signatures of Retro-Engineering**
   3.1 Conditional tilts vs unconditional baselines
   3.2 Information-theoretic fingerprints: MI and MDL deltas
   3.3 Negative controls: sham utilities and scrambled scoring

4. **Candidate Phenomena Mapped to Predictions**
   4.1 Bell/erasure and time-ordered entanglement swapping
   4.2 Weak values and post-selection anomalies
   4.3 Fine-tuning, low initial entropy, and design-like compressibility
   4.4 Entropy dips in mesoscopic systems (lawful improbabilities)
   4.5 Synchronicity at scale: semantic alignment beyond chance
   4.6 Just-in-time discovery and clustered success in replication
   4.7 Convergent evolution and digital evolution analogs
   4.8 Near-miss catastrophes and survival-compatible threading
   4.9 Algorithmic compressibility pockets in world data

5. **Experimental Program (Preregistered)**
   5.1 Utility-locked preregistration and cryptographic commitments
   5.2 Split generators and scorers; independence and timing
   5.3 Measurement axes: classical stats, MI/MDL, quantum outcome skews
   5.4 Multi-site protocols and stopping rules

6. **Pilot Studies**
   6.1 Utility-conditioned delayed-choice experiment
   6.2 Digital-evolution with blind future scoring
   6.3 Long-baseline calorimetry for entropy-dip surveillance
   6.4 World-data compressibility tracking under public targets

7. **Analysis and Interpretation**
   7.1 What constitutes evidence **for** retro-engineering?

1. Motivation and Scope

1.1 Puzzles that invite a retrocausal reading

Across domains, we repeatedly meet patterns that look wildly contingent up close but strangely "just-in-time" at larger scales. In quantum optics, delayed-choice and entanglement-swapping experiments make earlier records appear consistent with later measurement settings without enabling signalling. In complex systems, we see saltations and convergences (eyes, flight, camera-like organs) arising via distinct paths. In discovery, breakthroughs cluster at moments of maximal downstream leverage, as if the tape were spliced for utility rather than chronology. In risk, societies thread improbable needles—near-misses that preserve global capacity. None of these violate known laws; each can be modeled forward-causally with enough parameters. But taken together, they motivate a cleaner hypothesis: lawful microdynamics may be **weighted** by boundary conditions defined at a later time, T, such that histories with higher "future utility" are more likely to become our shared record. The scope of this paper is to turn that motivation into tests: derive signatures that would follow if such boundary weighting exists, and specify protocols that could detect (or bound) it.

1.2 From stories to tests: separating anecdotes from signatures

Anecdotes trade on pattern recognition after the fact; signatures demand preregistration before the fact. We therefore require three moves:

- **Lock the target**: Publicly commit (hash, timestamp) to a future scoring function U that will be computable only after data are collected (e.g., whether an unrelated paper is accepted by date D).
- **Split generation from scoring**: Ensure the present-time random sources and the future scoring process are physically and institutionally independent.
- **Define discriminators**: Look not for generic anomalies, but for **conditional tilts**: statistical, informational, or quantum-compatible skews that appear *only* in arms linked to U and vanish in sham or scrambled-U arms.

Concretely, we elevate evidence thresholds from "surprising coincidence" to **predeclared coupling** between present outcomes and later utilities. The result is a falsifiable wedge: either conditional tilts replicate across sites and utilities, or they do not. Either way, the narrative burden shifts from post hoc stories to measurable fingerprints.

1.3 Definitions: history, boundary condition, future utility

To keep the framework precise and copy-safe, we fix minimal definitions:

- **History (H)**: A full specification of micro/macro events over an interval [t0, T] that obey standard dynamics. In ordinary physics we estimate P(H) via forward evolution from initial data. Here we allow **weights** that may depend on boundary information at T.
- **Boundary condition (B_T)**: Information available at the terminal time T that constrains or reweights admissible histories. Examples include selected measurement outcomes, fixed-point constraints, or any external criterion that can be evaluated at T without violating locality or conservation laws.
- **Future utility (U)**: A scalar function U(H, T) in R (or bounded interval) evaluable at T using data generated by H. Intuitively, U quantifies the "goal satisfaction" of a history. Operationally, we preregister U's code and inputs in advance, then compute U only after the lock date.
- **Boundary-weighted measure**: Instead of bare P0(H) from forward dynamics, observed frequencies reflect P(H | B_T) proportional to P0(H) * W(U(H,T)), where W is a nonnegative weighting functional. We do **not** assume a form for W in advance; our tests look for **conditional excess coupling** between present outcomes and later U relative to baselines.
- **Signature**: Any statistically robust, preregistered deviation confined to U-linked arms—e.g., shifts in outcome frequencies, increases in mutual information I(present; U), or MDL reductions aligned with U—that disappear under sham utilities or timing permutations.

These definitions keep ordinary physics intact (no superluminal signalling, no broken conservation). They only allow that among all lawful histories, those scoring higher on preregistered U are more likely to be instantiated—precisely the claim our experiments are designed to confirm or bound.

2. Conceptual Framework

2.1 Retrocausal selection as boundary weighting, not law breaking

The proposal is minimal: keep standard microdynamics intact, but replace a single initial-value account with a two-boundary account in which terminal information at time T can reweight otherwise lawful histories.

Let H index admissible histories on [t0, T] that satisfy ordinary equations of motion and constraints (energy, momentum, charge). Let P0(H) be the forward probability (or path measure) assigned from initial data alone. Under retrocausal selection we posit an **observable** measure

$$P_{obs}(H) \propto P0(H) * W(U(H, T)),$$

where U(H, T) is a preregistered, well-defined utility evaluated at T, and $W(\cdot) \geq 0$ is an unknown weighting functional. Crucially:

- Dynamics remain unchanged; only **relative frequencies** of lawful histories differ.
- No new forces or signals are introduced; boundary information acts like a global Lagrange weight, not a local cause.
- Testability comes from **conditional tilts**: statistics shift only in arms tied to U, not in sham controls.

A path-integral analogy makes this concrete. In the usual formulation,

$$Amp(path) \propto \exp(i * S[path] / \hbar),$$

P0(H) comes from $|\Sigma_{paths}\ Amp|^2$.

Boundary weighting modifies only the ensemble weighting, not the phase evolution:

$$P_{obs}(H) \propto |\Sigma_{paths}\ Amp|^2 * W(U(H, T)).$$

Think of W as a soft constraint that prefers histories scoring higher on U, subject to all ordinary laws. When $W \equiv 1$ we recover standard statistics. Our experiments seek evidence that $W \neq 1$ **only when** U is locked and later evaluated.

2.2 Three mechanisms: post-selection, final-state constraints, fixed-point solvers

The same boundary-weighting idea can be realized (conceptually) via three canonical mechanisms. They differ in story but agree on the testable signature: conditional tilts tied to future utility.

(a) Post-selection (PSC)

- Idea: A future device retains only those outcomes that meet criterion C at T (e.g., a set of detector clicks), discarding all else. In thought experiments, perfect post-selection yields computational power up to PP.

- Mapping to weighting: W acts like an indicator/soft filter for U meeting C. In practice, perfect post-selection is unrealistic; our tests look for tiny, reproducible biases consistent with *partial* effective post-selection when U is predeclared.
- Lab cartoon: In a delayed-choice/eraser setup, the branch frequencies conditioned on a preregistered U (evaluated later) are fractionally skewed relative to baselines, yet remain no-signalling in any local partition.

(b) Final-state boundary conditions (two-state vector flavor)

- Idea: The world is constrained by both an initial state $|\psi_i\rangle$ at t0 and an effective final condition $\langle\psi_f|$ at T. Probabilities depend on both boundaries.
- Mapping to weighting: For a history H compatible with ($|\psi_i\rangle$, $\langle\psi_f|$), the Aharonov–Bergmann–Lebowitz-type rule implies reweighting that can be written as W(U(H, T)) once U is a function of the final boundary information.
- Lab cartoon: Weak-value anomalies arise not from law breaking but from the co-influence of initial and final constraints. Predeclared utilities correspond to "which final subspace" is selected.

(c) Fixed-point causal solvers (Deutsch-style CTC metaphor)

- Idea: A future computational fixed-point constraint forces self-consistent solutions across time (no paradoxes). Among the consistent solutions, some selection principle (e.g., minimal entropy production, maximal U) picks the realized fixed point.
- Mapping to weighting: W favors histories that sit at or near the chosen fixed point under the future solver's criterion U.
- Lab cartoon: When today's protocol is embedded in a feedback that will be closed only at T (e.g., a decision rule released later), present statistics align—slightly—with the fixed-point chosen by that later rule.

These mechanisms are **explanatory templates**; our paper does not assume their physical existence. What we test is the common prediction: if a future-evaluated U is locked now, then present-time statistics should exhibit preregistered, U-aligned tilts that disappear when U is removed or scrambled.

2.3 Guardrails: conservation, locality, and no-signalling

Retrocausal selection is often conflated with "magic." To avoid that, we state explicit guardrails that any acceptable model—and our empirical program—must respect.

(1) Conservation laws remain exact.
Energy, momentum, charge, angular momentum, and any other conserved quantities under the dynamics are preserved in every admissible history H. Boundary weighting never creates rule-breaking events; it only reweights which lawful histories are realized more frequently.

(2) Microdynamics are unmodified.
Schrödinger, Dirac, Maxwell, GR/NR mechanics—pick the regime—evolve as usual. No new interaction terms, hidden channels, or stochastic drives are added. If a standard model predicts a zero-probability event, boundary weighting cannot make it occur.

(3) No superluminal signalling.
For any spacelike partition A|B, the marginal statistics at A are independent of choices at B. Formally, if X_A are outcomes at A and S_B are settings at B,

$$P(X_A \mid S_B, U) = P(X_A \mid U)$$

for all S_B and preregistered U. Any observed dependence would falsify our guardrail and imply ordinary signalling (which we are not proposing).

(4) No "free energy" or perpetual motion.
Entropy bookkeeping must close on any bounded subsystem; apparent entropy dips (if any) occur only as lawful fluctuations whose *relative* frequency is boundary-weighted. Closed-system calorimetry and work-extraction controls are part of the experimental program.

(5) Transparency via preregistration and controls.
Because selection effects are subtle, we require: (i) cryptographic commitment of U before data collection, (ii) independence between present generators and the future scoring process, (iii) sham-U and timing-permutation controls, and (iv) multi-site replication with open materials.

(6) Nulls are decision-useful.
If U-conditioned tilts fail to appear under adequate power and clean controls, we set quantitative bounds on W's possible influence:

$$|E[g(X_{present}) \mid U] - E[g(X_{present})]| \leq \varepsilon$$

for predeclared functionals g and a small ε at 95% CI. This turns "no effect" into a scientific constraint on retrocausal selection.

Summary of the framework.

- Replace one-sided initial-value reasoning with a boundary-weighted ensemble on [t0, T].
- Realize this abstractly via post-selection, final-state constraints, or fixed-point solvers—without committing to any beyond their common prediction.
- Enforce strict guardrails so that any positive result cannot be reinterpreted as signalling, new forces, or energy creation.
- Make the hypothesis falsifiable by demanding U-locked, preregistered, conditional tilts that vanish under sham controls.

With this scaffold in place, the next sections derive measurable signatures and lay out protocols designed to distinguish "nice story" from **detectable boundary weighting**.

3. Statistical Signatures of Retro-Engineering

3.1 Conditional tilts vs unconditional baselines

Claim under test. Present-time outcomes X show small but reproducible skews only when they are prospectively tied to a future utility U that is computed at time T. Absent that tie, baselines hold.

Setup. Randomize units/runs into two arms before any data:

- Treatment arm A: outcomes X_A will later be scored by U (predeclared and hash-committed).
- Control arm B: identical protocol, but U is not applied (or is held blind until the end).

Let g be a predeclared statistic on X (e.g., success rate, a regression coefficient, a summary of counts). Define

Delta_g := E[g(X_A)] - E[g(X_B)].

Under the null (no boundary weighting), Delta_g = 0 up to noise. Under retro-engineering, Delta_g drifts away from 0 only when A is genuinely tied to U.

Binary outcomes. With counts n_A^+, n_A^- and n_B^+, n_B^-,

logOR := log((n_A^+ / n_A^-) / (n_B^+ / n_B^-)).

Estimate SE(logOR) in the usual way; preregister a minimal detectable effect (MDE) and power.

General outcomes. Use a preregistered analysis map g: X -> R^k and test

H0: E[g(X)|arm] equal across arms

H1: E[g(X)|arm] differs with sign predicted by U.

Prefer linear or GLM contrasts with sandwich SEs; avoid any post hoc feature mining.

Time-ordered designs. For streams X_1, X_2, … collected before T, fit a predeclared prequential model M_0, and test whether the one-step-ahead residuals in arm A tilt relative to B. Lock the residual functional (e.g., mean residual, log-score) ahead of time.

Quantum-compatible tilts. In Bell/eraser-style blocks, test arm-wise frequency shifts that remain no-signalling. For any spacelike partition A|B,

P(X_A | S_B, arm) = P(X_A | arm)

must still hold. We only test whether P(X_A|arm=A) differs from P(X_A|arm=B), not whether it depends on S_B.

3.2 Information-theoretic fingerprints: MI and MDL deltas

Tilts should also appear as **extra coupling** between present data and the future score that cannot be explained by chance.

Mutual information (MI). Let Z be a fixed, preregistered summary of X (e.g., binned outcomes or model-based scores). Estimate

I(Z; U) := E[log p(Z,U) - log p(Z) - log p(U)].

Procedure:

1. Specify discretization or a fixed density estimator *before* data.
2. Compute I_A := I(Z_A; U) in the treatment arm, and I_B := I(Z_B; U_sham) in control (see 3.3).
3. Test Delta_MI := I_A - I_B > 0 via block-preserving permutation that respects temporal order and all randomization constraints.

Bias control: use Miller–Madow or jackknife corrections for discrete MI; for continuous MI, use a fixed kNN estimator with k preregistered or a fixed neural estimator architecture frozen in advance.

Minimum Description Length (MDL). View “retro-engineering” as a pressure toward simpler descriptions aligned with U.

- Choose a fixed model class M and compressor C (both preregistered).

- For each run r, compute:
    - L0(r): codelength for X_r under baseline model M_0 (no reference to U).
    - LU(r): codelength for X_r under the same model class but with a *predeclared* side-information channel that may condition on the future U (no parameters learned from outcomes; only the wiring is allowed).
- Define per-run MDL gain:

Gamma(r) := L0(r) - LU(r).

Signature: Mean Gamma is larger in the U-tied arm than in control:

Delta_MDL := E[Gamma | arm=A] - E[Gamma | arm=B] > 0.

Prequential variant: use log-loss of a fixed online predictor with and without the side-information hook to U. All hyperparameters and the side-info wiring must be frozen before data.

Why MI/MDL? Both quantify coupling without relying on a single scalar outcome. MI captures dependence; MDL captures "explainability" improvement when U is available. Under the null, U is just noise and these deltas vanish (up to finite-sample bias).

3.3 Negative controls: sham utilities and scrambled scoring

To separate genuine boundary weighting from ordinary biases, we require controls that mimic all analysis steps but break the causal channel to the real U.

(a) Sham U. Predeclare U_sham that matches the type and distribution of U but is provably irrelevant to the experiment (e.g., a cryptographically scheduled coin flip published after T, or a random permutation of real U labels within strata).

- Compute the same statistics (Delta_g, Delta_MI, Delta_MDL) with U_sham.
- **Expectation:** No tilt. If tilts appear here, the pipeline leaks or overfits.

(b) Timing permutations. Preserve within-run structure but shift U forward/backward beyond any plausible influence window; recompute all metrics.

- **Expectation:** Effects disappear under misaligned timing.

(c) Arm masking and analyst blinding. Analysts receive only masked arm labels until the analysis code is frozen. Unmasking occurs once all scripts, thresholds, and reports are hash-committed.

(d) Overfitting guards. Fix all analysis choices in a *single* primary plan. If exploratory analyses are performed, route them into a "pilot" registry and demand a new, clean confirmatory run.

(e) Calibration checks. In binary tasks, verify that predicted probabilities stay calibrated within each arm (e.g., reliability plots, Brier decomposition). Retro-tilts should not be explainable by miscalibration alone.

(f) Multiple testing control. If multiple g's or multiple blocks are predeclared, control FWER/FDR (e.g., Holm-Bonferroni) or, better, use a single primary endpoint plus secondary endpoints labeled for corroboration.

(g) Power and bounds. Precompute power for your primary metric (e.g., logOR or Delta_MI). If null, report quantitative bounds:

|Delta_g| <= eps_g, Delta_MI <= eps_MI, Delta_MDL <= eps_MDL

with 95% CIs. Nulls then tighten constraints on any weighting W, rather than merely "failing to find an effect."

Summary. A positive result is a **pattern of converging evidence**:

- A prespecified scalar tilt (Delta_g) in the U-tied arm,
- An increase in dependence (Delta_MI > 0),
- A compression gain aligned with U (Delta_MDL > 0),
- All of which vanish under sham U and timing permutations,
- While respecting no-signalling and conservation guardrails.

This trio (tilt, MI, MDL) gives orthogonal lenses on the same hypothesis and makes "just a coincidence" increasingly difficult to maintain under rigorous preregistration and controls.

4. Candidate Phenomena Mapped to Predictions

4.1 Bell/erasure and time-ordered entanglement swapping

Phenomenon. Delayed-choice erasers and time-ordered swapping make earlier records consistent with later settings without signaling.
Retro-prediction. If a future utility U is preregistered to prefer interference-like (or particle-like) patterns, then arm=A (U-tied) shows a small, reproducible frequency skew in the congruent direction; arm=B (no U) does not.
Protocol. Two arms, identical optics. Arm=A is later scored by U (e.g., "interference clarity" computed at T from a locked algorithm). Arm=B is scored by a sham U_sham. The choice to insert/remove the eraser (or perform the Bell measurement in swapping) is governed by independent cosmic RNGs.

Signature. Delta_g > 0 for a preregistered g (e.g., contrast, fringe visibility) only in arm=A; Delta_MI > 0 between Z (binned counts) and U; no change in marginal statistics across spacelike partitions (no-signaling preserved).
Falsifier. Same tilts in arm=B or under U_sham; dependence of A's marginal on B's setting.

4.2 Weak values and post-selection anomalies

Phenomenon. Extreme weak values appear under pre- and post-selection.
Retro-prediction. When the future post-selection is tied to U (locked now, computed at T), the distribution of weak values exhibits a U-aligned skew beyond standard TSVF expectations, while sham filters do not.
Protocol. Fix pre- and post-selection states; preregister U (e.g., a numerical score from a future, independent event). Assign runs to A (real U) vs B (sham).
Signature. A preregistered tail metric (e.g., top-quantile mass q_95) increases only in arm=A; Delta_MDL > 0 when a side-channel wired to U is allowed in the compressor; disappears under timing permutations.
Falsifier. Identical tail behavior in B or under scrambled U labels.

4.3 Fine-tuning, low initial entropy, and design-like compressibility

Phenomenon. Cosmological parameters and initial low entropy appear finely tuned.
Retro-prediction. Joint codes of parameter tuples show an MDL below chance when compressed with codebooks aligned to observer-supporting utilities U (e.g., computational capacity proxies), compared with codebooks aligned to sham utilities.
Protocol. Construct model-agnostic encodings of parameter sets across credible cosmologies; preregister multiple U-aligned codebooks vs sham codebooks.
Signature. Groupwise MDL gain Gamma_U > Gamma_sham with predeclared thresholds; Delta_MI between parameter encodings and U > 0.
Falsifier. No differential compressibility beyond sampling noise; gains vanish after controlling for known astrophysical correlations.

4.4 Entropy dips in mesoscopic systems (lawful improbabilities)

Phenomenon. Rare entropy-reducing fluctuations are lawful but vanishingly unlikely.
Retro-prediction. When runs are tied to a future U (e.g., impact score at T), the frequency of small, lawful entropy dips is modestly elevated in arm=A relative to B, without violating the second law in expectation.
Protocol. Closed calorimetry on mesoscopic devices (e.g., colloidal particles, superconducting circuits) with high-resolution heat/work tracking. Predeclare thresholds for "dip" events; assign runs to A (U-tied) vs B (sham).
Signature. Elevated rate ratio RR_dip > 1 with CI excluding 1 only in arm=A; prequential log-loss

improves when side-info to U is allowed (MDL gain).
Falsifier. Similar elevations in B; energy non-closure; dependence on analyst choices post hoc.

4.5 Synchronicity at scale: semantic alignment beyond chance

Phenomenon. Meaningful coincidences cluster subjectively.
Retro-prediction. With preregistered semantic targets (U) computed at T by an independent panel/algorithm, streams produced now (texts, images) show excess mutual information with U only in U-tied arms.
Protocol. Generate independent data streams (crowd text, IoT telemetry). Lock a rubric to score "alignment" with future events (e.g., topic labels released at T). Arm=A binds to U; B binds to U_sham.
Signature. Delta_MI > 0 between Z (fixed embeddings) and U in A but not B; fixed classifier trained pre-experiment yields higher AUC in A; MDL gain with U side-channel in A only.
Falsifier. Gains persist with timing permutations or sham U; effects vanish after preregistered deconfounding (e.g., calendar, news baselines).

4.6 Just-in-time discovery and clustered success in replication

Phenomenon. Breakthroughs arrive "just in time"; replications succeed in clusters.
Retro-prediction. When labs preregister future utilities (acceptance, real-world impact at T), the present outcomes show conditional tilts: success probabilities modestly higher in U-linked arms than matched baselines.
Protocol. Multi-site preregistrations with split arms: A tied to U, B to U_sham. Outcomes are graded by independent committees after T.
Signature. Positive logOR for success in A vs B; MI between preregistered features and U increases in A; MDL gains on lab outcome logs with U side-info.
Falsifier. No tilt after stratification by known confounders (lab quality, sample size); mirror tilts under sham U.

4.7 Convergent evolution and digital evolution analogs

Phenomenon. Independent lineages converge on similar solutions (eyes, flight).
Retro-prediction. In digital evolution with blind, delayed scoring U (evaluated at T), lineages show excess convergence (measured by functional distance in phenotype space) in U-tied conditions relative to identical baselines.
Protocol. Evolve many populations with identical randomness budgets. Only after T, compute U on held-out tasks; compare convergence metrics for A (U-tied) vs B (no U or U_sham).
Signature. Lower pairwise phenotype distances and higher topological homology overlap in A; increased hit-rate of specific motifs preregistered as "useful" under U; MDL gains for genotype logs with U side-info.

Falsifier. Convergence equally strong in B; effects explained by selection pressure already encoded in forward fitness.

4.8 Near-miss catastrophes and survival-compatible threading

Phenomenon. Societies "thread the needle" through tail risks.
Retro-prediction. Cataloged near-miss events show an overrepresentation aligned with future utility indices (e.g., innovation capacity at T) versus hazard-model expectations, but only for U-linked strata.
Protocol. Predefine hazard baselines (weather, seismic, epidemiological), preregister U indices, and stratify geographies/periods into A (U-linked) vs B (controls).
Signature. Positive residuals (observed − expected) for near-miss counts in A but not B; MI between near-miss indicators and U; robustness across alternative hazard baselines.
Falsifier. Residuals vanish under alternative but preregistered hazard models; effects present in B or for sham U.

4.9 Algorithmic compressibility pockets in world data

Phenomenon. Datasets sometimes admit sudden, simple descriptions ("keys").
Retro-prediction. When a future target U is preregistered (e.g., a contest score at T), streams collected now exhibit **target-aligned compressibility drops** (shorter codes) only in U-tied arms.
Protocol. Fix compressors and model classes. Collect parallel streams for A (U-tied) and B (sham). At T, compute U and the side-information wiring.
Signature. Delta_MDL > 0 in A, stable across compressors; predeclared checkpoints show stepwise code-length drops temporally aligned to U milestones; Delta_MI(Z; U) > 0 only in A.
Falsifier. Equivalent MDL gains in B or under scrambled U; gains attributable to mundane covariates (time-of-day, platform shifts) after preregistered adjustment.

Synthesis across 4.1–4.9.
Each phenomenon yields the same triad of signatures—(i) scalar tilt in a preregistered outcome, (ii) MI coupling to U, (iii) MDL gain with a U side-channel—that:

- appear only when the arm is truly tied to the future utility U,
- vanish under sham utilities and timing permutations, and
- respect conservation and no-signalling guardrails.

A consistent, cross-domain appearance of this triad would count as evidence for boundary weighting; systematic nulls would bound any such effect to below predeclared epsilon thresholds.

5. Experimental Program (Preregistered)

5.1 Utility-locked preregistration and cryptographic commitments

Goal. Make “the thing scored in the future” immutable *before* any data exist.

5.1.1 What gets locked

- Full protocol text (design, arms, randomization, primary/secondary endpoints).
- Exact code to compute the future utility U (inputs, transforms, thresholds).
- Analysis plan (primary metric, estimator, SE/CI method, permutation rules).
- Blinding plan, data retention, and release schedule.

5.1.2 How to lock

- Produce a single, versioned bundle (e.g., retro_exp_v001.tar.gz).
- Compute SHA-256: H = sha256(bundle). Record H and UTC timestamp.
- Optional: Merkle-tree the bundle’s files to allow granular verification; publish tree root in the prereg record.
- Post H to at least two public, time-stamped channels (e.g., OSF/Zenodo prereg + a public blockchain note).
- For human-legible redundancy, include a short “commit string” hash (e.g., Base32-encoded first 80 bits).

5.1.3 Utility computation delayed by design

- U must depend on information unavailable until time T (e.g., an external panel’s decision, a third-party contest score, or a beacon value released after T).
- U code reads an input file U_inputs.json that will only be created at T by an independent party.

5.1.4 Example prereg header (minimal)

study_id: RETRO-001

hash_sha256: 6f5c...a913

arms: A=U_tied, B=U_sham

primary_endpoint: Delta_g (predeclared)

secondary_endpoints: Delta_MI, Delta_MDL

U_definition: ./code/U_eval.py # frozen; takes U_inputs.json at T

U_sham_definition: ./code/U_sham.py

alpha: 0.025 (one-sided) on primary_endpoint

power_target: 0.8 for MDE=logOR 0.10

stop_rule: group-seq OBrian-Fleming, K=3 looks, Pocock secondary

5.2 Split generators and scorers; independence and timing

Goal. Ensure "present generators" and "future scoring" cannot leak information or influence each other by ordinary means.

5.2.1 Physical and institutional separation

- *Generators* (devices, experiments, data streams) operate at Sites S1..Sk under Lab PIs.
- *Scorers* (who compute U at T) are an independent committee C or third party; different org, different network.
- Data firewalls: generators cannot read scorer networks; scorers cannot read any unanalyzed raw data prior to T.

5.2.2 Timing and clocks

- Synchronize all sites to authenticated UTC (e.g., GPS-disciplined NTP; log offsets).
- Define T_start, T_stop for collection; define T_score (strictly after T_stop).
- All logs include monotonic clock and UTC stamps; any drift beyond preregistered epsilon flags the run.

5.2.3 Randomness sources and settings

- Use preapproved RNG sources. Example:
    - Cosmic RNG for setting choices (astronomical photon arrival time; or trusted hardware TRNG mirrored at all sites).
    - PRNGs only for simulation arms; never for allocation or measurement settings.
- Store RNG seeds and raw draws; publish post hoc.

5.2.4 Arm assignment

- Randomize units/runs to A vs B with blocked randomization (site, day).
- Arm labels are masked (A_mask, B_mask) for analysts until the analysis code is frozen and hashed.

5.2.5 U-sham construction

- U_sham must match the marginal distribution and datatype of U but be independent by design (e.g., cryptographic beacon values time-shifted outside the study window, or label permutations within strict strata). U_sham is fully defined in the prereg bundle.

5.3 Measurement axes: classical stats, MI/MDL, quantum outcome skews

Goal. Triangulate with three orthogonal families of metrics, all frozen in advance.

5.3.1 Classical scalar tilt (Delta_g)

- Define a single primary scalar statistic g(X) (e.g., success rate, contrast score, log-loss improvement).
- Estimate arm difference:

Delta_g = E[g(X) | arm=A] - E[g(X) | arm=B].

- Predeclare estimator and SE:
    - Binary: log odds ratio with Wald or exact CI; stratified CMH if blocking used.
    - Continuous: difference-in-means with HC3 robust SE; or prereg GLM with fixed link and covariates.
- Primary hypothesis: H0: Delta_g <= 0 vs H1: Delta_g > 0 (one-sided, if direction is predeclared).

5.3.2 Information-theoretic coupling (Delta_MI)

- Fix a summary map Z = f(X) (e.g., binned counts, fixed embedding, pretrained model score). Freeze f.
- Compute mutual information using a prereg estimator (discrete plug-in with Miller-Madow, or kNN with fixed k):

I(Z; U) and I(Z; U_sham).

Delta_MI = I_A(Z; U) - I_B(Z; U_sham).

- Significance by block-preserving permutation consistent with randomization and time-order; report exact p or 10k-permutation p.

5.3.3 Compression gain (Delta_MDL)

- Freeze compressor C and model class M_0 (e.g., arithmetic coder over a fixed HMM, or a frozen lightweight neural predictor).
- Compute per-run code lengths:

L0(r) = codelen under M_0 without side info

LU(r) = codelen under same class with a fixed side-info wiring to U

Gamma(r) = L0(r) - LU(r)

Delta_MDL = E[Gamma | A] - E[Gamma | B].

- Predeclare checkpoints (time-slices) to test for stepwise code reductions aligned with U milestones.

5.3.4 Quantum-compatible outcome skews

- In Bell/eraser blocks, verify no-signalling constraints as guardrails:

For any spacelike A|B: P(X_A | S_B, arm) = P(X_A | arm).

- The only allowed effect is an armwise shift:

P(X_A | arm=A) vs P(X_A | arm=B).

- Predeclare which marginal(s) and contrast(s) are primary; use exact tests with multiplicity control.

5.3.5 Multiplicity and modeling

- One primary endpoint (Delta_g) at alpha_primary.
- Two secondary endpoints (Delta_MI, Delta_MDL), labeled corroborative; control FDR or report as descriptive unless preregistered for confirmatory use.
- All model hyperparameters fixed in the prereg; any exploratory analysis is quarantined to an “E-appendix” with new IDs.

5.4 Multi-site protocols and stopping rules

Goal. External validity and clean frequentist inference with transparent interim looks.

5.4.1 Multi-site orchestration

- Sites S1..Sk run identical hardware/software stacks validated by a shared "dry-run harness."
- Site-level stratification: randomize within site-by-day blocks; record environment covariates (temperature, power, operator).
- Data escrow: sites write append-only logs to a neutral escrow (e.g., S3-like bucket with object lock), plus local WORM media.

5.4.2 Blinding and unmasking

- Analysts receive masked arms and site labels; they finalize code, produce analysis_v001.tar.gz, and hash-commit.
- Unmask only after the analysis hash is posted and a checksum meeting is recorded.

5.4.3 Interim looks and stopping

- Use a group-sequential design with K looks (e.g., K=3 equally spaced information fractions).
- Primary endpoint only participates in alpha spending (O'Brien–Fleming or Pocock, preregister one).
    - Example OBF nominal one-sided alphas for K=3: 0.005, 0.014, 0.022 (final).
- Secondary endpoints are monitored but do not trigger stopping unless preregistered otherwise.
- Stop for efficacy if the Z-stat crosses the spending boundary; stop for futility via a fixed conditional power rule (e.g., stop if cond. power < 0.2 given current effect).
- If stopped early, all MI/MDL analyses must still be executed and reported with the current information fraction.

5.4.4 Replication and roll-up

- Require at least one independent replication site to confirm a positive before declaring "evidence."
- Predefine a fixed-effects meta-analytic roll-up of site-level Delta_g (inverse-variance weights); report heterogeneity (Q, I^2).
- For MI/MDL, use the same estimator and permutation scheme per site; combine p-values by Fisher's method (preregistered).

5.4.5 Data release and auditability

- Release: raw anonymized data, RNG draws, code, hashes, and environment logs within N days of lock.
- Provide a “1-click” reproducibility script that re-creates all figures and tables from frozen containers.
- Third-party audit: invite an external team to re-run the pipeline; preregister acceptance criteria for audit success.

5.4.6 Quantitative null reporting

- If null, publish epsilon-bounds on W’s possible influence via the predeclared metrics:

|Delta_g| <= eps_g, Delta_MI <= eps_MI, Delta_MDL <= eps_MDL (95% CI)

- Convert to interpretable caps (e.g., “any retro-weighting changes success odds by <= 1.02x at 95% CI under this design”).

Outcome. This preregistered program turns retrocausal “engineering” into a falsifiable target: only U-locked, time-delayed, independence-respecting designs can show the predicted conditional tilts. Clean positives would warrant scaled tests; disciplined nulls tighten quantitative bounds on any boundary-weighting functional W.

6. Pilot Studies

6.1 Utility-conditioned delayed-choice experiment

Objective. Detect a small, preregistered frequency tilt in an optical delayed-choice/eraser that appears only when runs are tied to a future utility U and vanishes under sham U.

Setup.

- Mach–Zehnder interferometer with fast electro-optic modulator (EOM) to insert/remove which-path marking.
- Single-photon source (attenuated laser or heralded SPDC), two single-photon detectors per output arm.
- Timing: GPS-disciplined NTP; event timers with 100 ps resolution.

- Randomness: cosmic RNG (astronomical photon arrivals) drives EOM setting; mirrored hardware TRNG as backup.
- Arms: A (U-tied), B (U_sham). Allocation blocked by site and hour.

Future utility U (locked before data).

- U := indicator that an independent committee, at date T, accepts a pre-written short paper derived from the A-arm dataset (binary: accepted vs not). Code U_eval.py reads accept_list.json released at T.

Procedure.

1. Preregister bundle: optics schematic; contrast score function g; primary endpoint; U and U_sham code; analysis script hash.
2. Collect N_A and N_B blocks of trials with identical hardware; EOM setting from cosmic RNG each trial.
3. Write append-only logs: (timestamp, arm, EOM_state, detector_counts, environment sensors).
4. At T, independent committee renders decision; U_eval.py outputs U in {0,1}.
5. Unmask arms; run frozen analysis.

Primary metric.

- g := fringe visibility estimate per block: g = (C_max − C_min) / (C_max + C_min).
- Delta_g := E[g | arm=A] − E[g | arm=B]. One-sided test H1: Delta_g > 0.

Secondary metrics.

- Delta_MI between Z (binned counts) and U vs Z and U_sham.
- Delta_MDL using a fixed arithmetic coder with/without a preregistered side-info bit that toggles when U=1.

Power sketch.

- Assume baseline visibility 0.20, SD 0.04 across blocks, MDE 0.01. With alpha=0.025 (one-sided), power 0.8 requires ~ (2 * (1.96+0.84)^2 * SD^2) / MDE^2 ≈ 2 * (7.84 * 0.0016) / 1e-4 ≈ 251 blocks total (balanced A/B).

Guardrails.

- Verify no-signalling: P(X_A | S_B, arm) = P(X_A | arm) via exact tests for each spacelike partition.
- Photonic loss and detector afterpulsing modeled identically in both arms; calibration interleaved.

Failure modes & mitigations.

- Drift: alternate A/B blocks frequently and include dummy blocks.
- RNG leakage: log raw cosmic events; post public seeds; mirror hardware TRNG for audit.
- p-hacking: single primary endpoint; others labeled corroborative.

Deliverables.

- Frozen container to reproduce all tables/plots, raw logs, RNG traces, committee records, hash receipts.

6.2 Digital-evolution with blind future scoring

Objective. Test whether lineages show excess convergence toward late-revealed tasks when future utility U is locked now but computed only at T.

Setup.

- Evolution engine: simple grid or population-based EA with fixed mutation/crossover rates; fitness defined only on *current* proxy tasks (no peek at U).
- Genotype: fixed-length bitstring or small DSL programs; phenotype: functional behavior trace on a task bank.
- Two arms: A (U-tied), B (U_sham). Identical randomness budgets per population.

Future utility U (locked).

- U := score on a *hidden* task set H evaluated at T (e.g., generalization to withheld environments, or performance on a puzzle dataset released by a third party). U_sham := hash-shuffled H mapping with same marginal difficulty.

Procedure.

1. Preregister engine, seeds, mutation schedule, phenotype distance metric d(.,.).
2. Run P populations per arm, each for G generations; store checkpoints every K generations (genotypes, phenotypes, proxy fitness).
3. At T, release H and compute U for final phenotypes (and optionally along trajectories) with U_eval.py.
4. Unmask arms; run frozen analysis.

Primary metrics.

- Convergence: mean pairwise phenotype distance at generation G, d_bar(G). Test Delta_d := d_bar_B − d_bar_A > 0 (A more convergent).
- Hit-rate of preregistered motifs M that are known to help on H (but are not beneficial on proxy tasks).

Secondary metrics.

- Delta_MI between Z (fixed phenotype summary) and U across populations.
- Delta_MDL: code length of genotype logs under a compressor with/without a side-info hook to U.

Power sketch.

- Suppose d_bar SD ≈ 0.08 across populations; MDE 0.02. With alpha=0.025, power 0.8 requires ~ 2 * (1.96+0.84)^2 * 0.08^2 / 0.02^2 ≈ 2 * 7.84 * 0.0064 / 4e-4 ≈ 251 populations total.

Controls & leakage prevention.

- Strict separation: the engine container does not include H; checksum at start and end; no network egress.
- Timing permutation: recompute metrics with H temporally shifted or label-permuted.

Failure modes & mitigations.

- Over-regular fitness shaping accidentally aligns with H: preregister ablation runs with alternate proxy tasks.
- Mode collapse: monitor diversity; enforce niching or novelty search equally in both arms.

Deliverables.

- Engine source, seeds, checkpoints, H release artifact, and a one-click repro.sh to rebuild figures.

6.3 Long-baseline calorimetry for entropy-dip surveillance

Objective. Detect a modest elevation in the lawful frequency of small entropy-reducing fluctuations ("dips") only when runs are tied to U.

Setup.

- Mesoscopic platform with precise heat/work readout: e.g., optically trapped colloidal particle with calibrated potential; or superconducting circuit with calibrated Q-meter.
- Closed-system approximation with continuous monitoring of bath temperature and drift sensors.
- Two arms: A (U-tied), B (U_sham); identical operation schedules.

Future utility U (locked).

- U := a delayed external score (e.g., whether a pre-submitted data descriptor passes peer review by T). Binary or scalar; code U_eval.py provided in prereg.

Procedure.

1. Define dip event: a predeclared statistic S_t over a sliding window (e.g., stochastic entropy production sigma_t) falls below a fixed threshold for at least w samples.
2. Calibrate energy accounting with reversible protocols; estimate noise floors.
3. Record long baselines (weeks) alternating A/B epochs; store raw traces with UTC stamps.
4. At T, compute U; unmask and analyze with frozen scripts.

Primary metric.

- RR_dip := rate_A / rate_B, where rate_x is dip count per unit time. H1: RR_dip > 1 with CI excluding 1.
- Alternative: Delta_g := E[sigma_t | A] − E[sigma_t | B] < 0 (more negative excursions in A).

Secondary metrics.

- Prequential log-loss improvement when a fixed predictor is allowed one side-info bit tied to U (Delta_MDL).
- Delta_MI between dip indicators and U vs U_sham.

Power sketch.

- If baseline dip rate is ~1 per 6 hours, target RR_dip=1.15, alpha=0.025. Approximate events needed via Poisson rate test: ~ ((1.96+0.84)^2 * (1+1/1.15)) / (log 1.15)^2 ≈ 900–1,100 dip events total; thus several weeks of acquisition across sites.

Guardrails.

- Energy closure tests daily; exclude intervals with failed closure by preregistered rules.
- Environmental covariates logged and included as fixed strata; exact tests within strata.

Failure modes & mitigations.

- Hidden drift: randomized epoch scheduling; interleave calibration pulses.
- Definition hacking: single preregistered S_t and threshold; any alternatives in E-appendix only.

Deliverables.

- Raw traces, calibration files, environment logs, scripts, and energy-closure reports.

6.4 World-data compressibility tracking under public targets

Objective. Observe target-aligned compressibility drops (shorter codes) in real-world streams only when a future target U is preregistered and public.

Setup.

- Streams: choose at least two independent, stationary-ish feeds (e.g., public sensor telemetry, financial tick subsets, social text with strict sampling rules).
- Compressor suite: fixed set C = {LZMA with frozen dict, arithmetic coder over fixed n-gram tables, lightweight frozen RNN}.
- Arms: A (U-tied) vs B (U_sham) as parallel streams matched by time of day, source, and rate.

Future utility U (locked & public).

- U := score of a model on a challenge released at T (e.g., MAPE on a public forecasting leaderboard or accuracy on a benchmark split announced in advance). U_sham := time-shifted beacon with same marginal distribution. All definitions and URLs are in the prereg.

Procedure.

1. Publish prereg bundle and a public statement of the target and collection window.
2. Collect streams for the window [t0, T]; checkpoint compressed sizes at fixed intervals (e.g., hourly).
3. At T, compute U; wire the side-info channel per prereg (e.g., a single bit or a small codebook index derived from U).
4. Run frozen analysis; release all raw streams and compressed artifacts.

Primary metric.

- MDL gain per interval i: Gamma_i := L0_i − LU_i where L0_i is code length without side-info and LU_i with side-info wired to U.
- Delta_MDL := E[Gamma | A] − E[Gamma | B] > 0.

Secondary metrics.

- Delta_MI between Z_i (fixed summaries of each interval, e.g., low-order moments or fixed embeddings) and U; permutation p-values within time-matched blocks.
- Step-change detection: preregistered CUSUM on Gamma_i to test for alignment with target milestones (e.g., leaderboard freeze).

Power sketch.

- Suppose per-interval Gamma SD ≈ 0.8% of L0, MDE 0.2%. With alpha=0.025 (one-sided), power 0.8 needs ~ 2 * (1.96+0.84)^2 * 0.008^2 / 0.002^2 ≈ 250 intervals total (balanced A/B), e.g., ~10 days at hourly checkpoints.

Controls and deconfounding.

- Match A/B streams by time and source; include calendar and exogenous event covariates in stratified analyses.
- Timing permutations: recompute Gamma with U shifted by ±k intervals; expect loss of effect.

- Compressor ablations: effect should persist across at least two distinct compressors in C.

Failure modes & mitigations.

- Regime shifts: preregister change-point exclusion rules; report sensitivity with and without flagged intervals.
- Hidden leaks: ensure U cannot be inferred from streams prior to T (e.g., when U is a hidden benchmark score).

Deliverables.

- All raw intervals, compressed files, compressor binaries, prereg hashes, and a dashboard notebook that recomputes Delta_MDL, Delta_MI, and step-changes end-to-end.

Across pilots 6.1–6.4.

- Each design enforces the same triad: a preregistered scalar tilt, MI coupling to U, and MDL gain with a U side-channel—appearing only in U-tied arms, disappearing under sham U and timing permutations, and respecting conservation/no-signalling guardrails. Positive results justify scaled follow-ups; disciplined nulls yield explicit epsilon-bounds on any boundary-weighting effect.

7. Analysis and Interpretation

7.1 What constitutes evidence for retro-engineering?

Evidence is **convergent, preregistered, and conditional on the future utility U**—never a single surprising p-value.

Minimal evidentiary package (confirmatory run):

1. **Primary scalar tilt:** the preregistered arm contrast is positive at the prespecified one-sided α (e.g., $\Delta g > 0$ with adjusted CI excluding 0).
2. **Independence-respecting corroboration:** no-signalling checks pass; conservation/closure checks pass; environment covariates behave as balanced noise.
3. **Dual information signatures:** $\Delta MI > 0$ **and** $\Delta MDL > 0$ using frozen estimators/compressors.
4. **Negative controls are null:** the same analyses with U_sham and with timing permutations yield no effects beyond noise.

5. **Replication:** at least one independent site reproduces (1–4) with compatible effect sizes; fixed-effects meta-analysis reports a pooled positive with low heterogeneity (e.g., $I^2 \leq 25\%$).

Strength tiers (for the abstract):

- **Tier S (strong):** All five criteria met, plus a preregistered Bayesian check (e.g., $BF_{10} \geq 10$ for the primary endpoint under a weakly informative prior on Δg).
- **Tier M (moderate):** Criteria 1–4 met at one site, with a second site showing directional agreement but underpowered.
- **Tier B (bounds only):** Primary null with tight CIs → report quantitative ε-bounds on $|\Delta g|$, ΔMI, and ΔMDL.

Reporting units (assay-agnostic, comparable across pilots):

- Δg in domain-natural units (e.g., logOR, contrast points).
- ΔMI in bits (per run or per block).
- ΔMDL in bits or % of baseline codelength per interval.

7.2 Distinguishing biases, p-hacking, and ordinary selection effects

Retro-claims are fragile unless we rule out mundane paths to "effects." We therefore bake in diagnostics aimed at the usual suspects.

(a) Researcher degrees of freedom (RDoF).

- Single primary endpoint; all preprocessing, estimators, and hyperparameters frozen and hash-committed.
- Masked arms until analysis code is finalized; containerized pipeline with deterministic seeds.
- Any exploratory variants (alt thresholds, alt features) go to a labeled E-appendix and **cannot** alter the confirmatory call.

(b) Specification curve & multiverse transparency.

- Precompute a *preregistered* specification curve (finite menu of reasonable alternatives known in advance). The primary must win under its own α; the curve is descriptive context, not a license to cherry-pick.

(c) Ordinary selection / survivorship.

- **Forward-only baselines:** simulate data under standard dynamics and site-measured noise to estimate how big Δg, ΔMI, ΔMDL could look **without** any boundary weighting. Real effects must exceed the 95th percentile of these null distributions.
- **Negative-outcome controls:** include outcomes believed a priori to be orthogonal to U; they should stay null (placebo outcomes).
- **Sham U and timing permutations:** identical pipelines that break the U link must yield nulls; if not, the pipeline leaks or overfits.

(d) Confounding and leakage.

- **DAG discipline:** preregistered causal diagrams for each pilot; any back-door that could couple arm with U by mundane channels must be blocked (physical separation, escrow, air-gaps).
- **Leak checks:** information-flow audits (e.g., hash-timelines for file creation, network logs, RNG provenance).
- **Sensitivity analysis:** E-value/R-value style bounds—how strong an unmeasured confounder would have to be to explain Δg; report alongside results.

(e) Calibration & model misfit.

- Reliability plots and Brier decompositions per arm for any probabilistic predictions used in g or Z.
- Posterior predictive checks for any GLM components; refuse “wins” that ride on miscalibration.

(f) P-hacking tripwires.

- Independent data clerk re-derives all primary numbers from raw logs.
- All analysis timestamps post-date prereg hash; any edits after unmasking invalidate confirmatory status.

7.3 Power analysis, effect sizing, and robustness checks

Because predicted tilts are small, power and robustness must be explicit—not aspirational.

Effect sizes (report all three):

- **Scalar tilt:** logOR (binary) or standardized mean difference (continuous).
- **Information coupling:** ΔMI in bits with small-sample bias correction (e.g., Miller–Madow).
- **Compression gain:** ΔMDL as $(L_o - L\_U) / L_o \times 100\%$ per unit time or per run.

Back-of-envelope planning (illustrative):

- For a mean-difference endpoint with SD = σ and MDE = δ at one-sided α, power ≈ 0.8 needs per-arm $n \approx 2(z\_\{\alpha\}+z\_\{0.8\})^2 \sigma^2 / \delta^2$.
- For logOR with baseline p and target OR, use standard formulas or a simulation power curve preregistered in the bundle.
- For MI/MDL, simulate under the site-specific null to get the finite-sample variance and set MDEs (e.g., ΔMI ≈ 0.01–0.05 bits/block; ΔMDL ≈ 0.1–0.5% per interval).

Sequential designs & α-spending.

- Group-sequential K-look plan preregistered (e.g., O'Brien–Fleming). Only the primary endpoint spends α; MI/MDL are monitored descriptively unless declared co-primary with appropriate multiplicity control.

Robustness checks (must pass to claim Tier S/M):

1. **Site heterogeneity:** pooled fixed-effects estimate positive with low I^2; a random-effects sensitivity reported.
2. **Estimator stability:** primary Δg holds under HC3 vs classical SEs; ΔMI holds for the prereg NN-free estimator and for the discretized plug-in; ΔMDL holds across at least two compressors in the frozen suite.
3. **Time-slice stability:** effect persists across preregistered temporal strata (early/mid/late blocks).
4. **Control invariance:** placebo outcomes and U_sham analyses remain null; timing permutations kill the signal.
5. **No-signalling & closure:** all guardrail tests pass at the prereg α (or stricter), with plots included.

Quantitative nulls (decision-useful).
If primary is null under adequate information, publish bounds in the abstract:

$|\Delta g| \le \varepsilon_g$, $\Delta MI \le \varepsilon_{MI}$ bits, $\Delta MDL \le \varepsilon_{MDL}$ %

(95% CI; pooled across sites)

Translate to plain language (e.g., "any boundary weighting would change success odds by ≤ 1.02× under this design").

Put together:
A credible positive is a *triangle*—(i) a small, directional Δg that clears a prespecified, well-powered bar; (ii) independent information-theoretic coupling (ΔMI); and (iii) algorithmic compressibility gain (ΔMDL)—all confined to U-tied arms, absent in sham/perm tests, replicated across at least one independent site, and housed within intact guardrails. Anything less is either exploratory or a bound, and we label it as such.

8. Risks, Ethics, and Governance

8.1 Open science constraints and adversarial preregistration

Why: Small effects + exotic claims create a high risk of motivated reasoning. Open, adversarially robust preregistration keeps us honest.

Core constraints

- **Full-lock prereg:** One tarball for protocol, code, estimators, compressors, RNG sources, and utility definitions; SHA-256 of the bundle posted to two independent, time-stamped venues (e.g., registry + public ledger). Include a Merkle tree for file-level verification.
- **Immutable analysis path:** One primary endpoint; frozen estimators/hyperparameters; masked arms until analysis hash is posted. Any deviations move results to an exploratory appendix.
- **Tamper-evident provenance:** Content-addressed data objects (hash-named files), append-only run logs, and signed environment manifests (container digests, library versions).
- **No-surprises clause:** All transformations of raw data to analysis tables must be defined in the prereg. If an unexpected cleaning step is required (e.g., a sensor fault), it is documented, hashed, and treated as a protocol amendment.

Adversarial prereg (treat your future self as an adversary)

- **Spec curve precommitment:** Publish a finite set of "reasonable alternatives" *before* data (e.g., 3 binning schemes, 2 MI estimators). The primary call cannot depend on choosing among them; the curve is descriptive context only.
- **Time-lock inputs to U:** U must depend on inputs created *after* data collection (e.g., third-party beacon, independent panel decisions). Store U_inputs.json under an external custodian; labs have no access before T.
- **Sham symmetry:** Define U_sham to match U's type and marginal distribution. Every test run on U is mirrored on U_sham.
- **Red-team review window:** A pre-data "attack period" (e.g., 10 days) invites external critiques; responses (or non-responses) are appended to the prereg bundle.
- **Threat model & leak tests:** Diagram potential back-doors that could couple arms to U (networks, people, scheduling, filenames). Add concrete leak probes (e.g., embedded canary tokens, fake U variants that must remain null).

Transparency deliverables

- 1-click container to rebuild figures/tables.
- Public hashes for all artifacts; human-readable "how to verify me" checklist.
- A negative-result template with prewritten language for ε-bounds.

8.2 Dual-use concerns in utility design

Risk: If U encodes harmful goals, the whole framework becomes a mechanism for covert optimization toward undesirable outcomes.

Disallowed utility classes (bright lines)

- **Physical harm:** Any U that rewards injury, property damage, environmental degradation, or unsafe lab behavior.
- **Illegal manipulation:** U tied to election outcomes, market manipulation, targeted harassment, surveillance, or privacy violations.
- **Discrimination:** U that treats protected classes differently or encodes biased outcomes.
- **Sensitive inference:** U constructed from personal data without explicit consent and minimization.

Discouraged utility classes (require extra review)

- **High-stakes societal levers:** U tied to policy shifts or critical infrastructure KPIs.
- **Opaque third-party scores:** U from proprietary black-box systems without audit rights.

Preferred utility classes (low-risk)

- **Scientific benchmarks:** Preannounced leaderboards, blinded peer-review outcomes, public forecasting challenges.
- **Synthetic beacons:** Cryptographic randomness beacons, time-locked puzzles, or delayed coin-flips with verifiable entropy.
- **Benign performance metrics:** Accuracy on public datasets, compression on non-personal corpora, pass/fail of preregistered checklists.

Design principles

- **Minimize coupling power:** The side-information channel to U is tightly bounded (e.g., 1 bit, fixed mapping), reducing any real-world leverage.
- **Data minimization:** Only features strictly needed to compute U are retained; apply k-anonymity or differential privacy where appropriate.
- **Independent custody:** U_inputs.json is prepared and released by an external custodian under an MoU; labs cannot influence its contents.
- **Ethics checklist (pre-data):**
    - Does U create incentives for unsafe shortcuts?
    - Could U be reverse-engineered from available signals before T?
    - Does any stakeholder face material downside if $\Delta g > 0$?
    - Are there non-contactable or vulnerable populations implicated?

Red-team/blue-team pattern

- **Blue team:** Authors propose U, argue minimal harm and necessity.
- **Red team:** Independent reviewers attempt to identify misuse channels or back-doors and propose safer U' with equivalent scientific value.
- **Resolve:** Adopt U' or document why U is essential; both positions are public in the prereg.

8.3 Community review, data escrow, and replication incentives

Goal: Make results legible, auditable, and worth replicating—even (especially) when null.

Community review

- **Registered Reports:** Submit as Stage 1 (methods accepted in principle before data). This bakes in editorial commitment independent of outcome.
- **Open protocols forum:** Host a pre-data Q&A thread; archive criticisms with timestamps. Encourage competing designs under the same theme.
- **Replication playbooks:** Alongside the methods section, provide a "Replicate in 7 steps" playbook with calibrated bill of materials, power targets, and expected wall-clock.

Data escrow & custody

- **Tri-party escrow:** Raw data and RNG traces mirrored to (i) an institutional repository, (ii) a commercial cloud bucket with object lock, and (iii) a community mirror (e.g., IPFS snapshot) with published CIDs.
- **Key-split governance:** Decryption keys (if any) split among two or more custodians; opening requires quorum.
- **WORM logging:** Write-once storage for run logs; daily hash manifests signed by site PIs.
- **Delayed PII release policy:** If any human data exist (generally avoid), release only de-identified derivatives, with a separate, IRB-approved access path for restricted raw.

Replication incentives

- **Bounties:** A dedicated fund pays small fixed bounties for (a) faithful replications that confirm, (b) serious, adequately powered failures to replicate, and (c) pre-registered methodological critiques that reveal a leak.
- **Badges & credit:** Issue citable DOIs for "Replication Packages" and award visible badges in the paper metadata (Confirmatory Replication, Robust Null, Leak Discovered). Prominently list replication teams on the project page.
- **Negative-result prestige:** Maintain a public leaderboard of ε-bounds across sites; highlight the tightest nulls as field-advancing artifacts.
- **Shared utilities repository:** A curated "Registered Utility Repository" (RUR) with audited U definitions, sham counterparts, and reusable code. Replicators can adopt existing U to reduce setup friction.

- **Rolling meta-analysis:** Predeclare how new replications will be pooled (effect estimator, heterogeneity model); maintain a living meta-analysis with versioned DOIs.

Governance tiers

- **Bronze (minimum):** Preregistration + dual escrow + masked analysis + public container.
- **Silver (recommended):** Add Registered Report, red-team review, bounties, and RUR use.
- **Gold (flagship):** Multi-site from day one, key-split escrow, third-party audit, and rolling meta-analysis governance.

Exit ramps

- **Safety stop:** If any guardrail fails (no-signalling violation, energy non-closure, leak detected), the study auto-pauses; a preappointed oversight panel decides remediation before resumption.
- **Communications protocol:** Prewritten language for positive, null, or ambiguous outcomes; avoid speculative over-reach. All press materials are linked to prereg artifacts.

Bottom line. The combination of adversarial preregistration, careful utility design, escrowed transparency, and tangible rewards for replication turns a high-risk/high-weirdness program into something the community can evaluate on ordinary scientific terms. Whether the results are positive or tight nulls, the governance makes them durable, auditable, and useful.

9. Implications

9.1 If signals are found: updates to causality narratives

Causality with two boundaries. Positive, multi-site evidence (scalar tilt + MI + MDL, all U-conditioned and guardrail-clean) would support a **boundary-weighted** picture: ordinary dynamics propagate forward, but realized frequencies obey
$P_{obs}(H) \propto P0(H) * W(U(H,T))$.
Causality remains local-in-the-small (no signalling), yet **explanations** become two-sided: some regularities are best read as consequences of both initial data and future boundary information.

What changes—and what doesn't.

- **Unchanged:** Equations of motion, conservation, and no-signalling constraints; lab interventions still work exactly as before.

- **Updated:** The *explanatory arrow* allows lawful, utility-conditioned reweighting. Narratives about "just-in-time" discovery, convergent evolution, or fine-tuned coincidences can be reframed as the visible effect of W(U) rather than as brute luck or hidden variables.

Thermo & information. Entropy bookkeeping stays intact, but rare, lawful fluctuations may appear **slightly enriched** when they advance preregistered U. Information-theoretically, the world exhibits a tiny, reproducible I(present; U_future) > 0 that cannot be removed by conditioning on measured confounders.

Computation and physics. Evidence favors models where post-selection or fixed-point constraints **exist in nature** only as *tiny, aggregate biases*—far below the power of idealized PP/CTC thought experiments, yet nonzero. This motivates effective theories of **bounded retro-selection** and new complexity-theoretic classes for "U-weighted sampling."

Method and practice. Where utilities are public (benchmarks, societal milestones), one should expect **detectable micro-tilts** if designs are sensitive enough; in purely private or undefined-U settings, standard baselines should prevail. That split itself becomes actionable design guidance.

Ethics & governance. If U can bias frequencies (even slightly), utility choice becomes a first-class experimental parameter with dual-use risk; community oversight and a "Registered Utility Repository" gain lasting importance.

9.2 If null: strengthened bounds on retrocausal influence

Quantitative ceilings. Well-powered, well-controlled nulls translate into **bounds on W**. For preregistered functionals g, MI, and MDL:

$|\Delta g| \le \varepsilon_g, \quad \Delta \mathrm{MI} \le \varepsilon_{\mathrm{MI}}$ bits, $\quad \Delta \mathrm{MDL} \le \varepsilon_{\mathrm{MDL}}$ %

(95% pooled CI)

These caps rule out any boundary weighting larger than ε under the tested designs, time windows, and utility classes.

Implications for theories.

- **Two-state / final-boundary models:** Constrained to regimes where effective reweighting is operationally negligible in lab-scale statistics.
- **Post-selection power:** Practical nature cannot approximate PP-like selection beyond ε under open, adversarial designs.

- **CTC/fixed-point metaphors:** If they exist, their macroscopic influence must be **screened** so that present-future coupling is below detection at tested scales.

Where next after nulls.

- **Tighten the priors:** Adopt these ε-bounds as informative priors for future experiments; design for 2× tighter CIs or orthogonal endpoints.
- **Alter regimes:** Move to longer horizons T, different U classes (e.g., cryptographic beacons vs human adjudication), or higher-sensitivity platforms (optical counts, superconducting calorimetry).
- **Meta-lesson:** Absence of U-conditioned tilts across domains suggests that most "engineered-looking" coincidences are adequately explained by forward selection, publication filters, and garden-of-forking-paths—now with data-driven justification.

9.3 Broader lessons for methodology and philosophy of science

Methodology

- **Adversarial preregistration works.** Treating your future self as an adversary—locking utilities, estimators, compressors, and negative controls—turns high-weirdness into ordinary, auditable science.
- **Information metrics belong in the core toolkit.** Reporting Delta_MI (bits) and Delta_MDL (% codelength gain) alongside effect sizes provides geometry-free evidence of coupling, less gameable than a single p-value.
- **Outcome design is theory.** Choosing U is not clerical; it instantiates what the hypothesis claims can couple across time. Publishing audited, reusable U's (and matched shams) should become standard.

Philosophy of science

- **From causes to constraints.** Even if null, the program clarifies a distinction: local *causes* vs global *constraints*. Our default stance remains initial-value causality; boundary-weighted views are meaningful only if they clear disciplined empirical bars.
- **Teleology without mysticism.** If positives emerge, they operationalize a carefully fenced "final cause": utilities act as boundary weights, not as forces. If null, the result still sharpens our language for excluding teleological gloss from lawful dynamics.

- **Demarcation by design.** The field gains a template for testing claims often dismissed as unfalsifiable. Preregistered utilities, sham symmetry, timing permutations, and no-signalling guardrails make the difference between story and science legible.

Culture

- **Prestige for robust nulls.** Tight ε-bounds are genuine contributions; they close degrees of freedom for speculation and guide future power analyses.
- **Replication as first-class output.** Bounties, Stage-1 acceptance, and rolling meta-analysis realign incentives so that confirmations and clean failures both advance knowledge.

Bottom line.

- **If signals:** We modestly enlarge causality—keeping dynamics and locality—by adding empirically warranted boundary weighting tied to preregistered utilities.
- **If null:** We tighten the fence: any retro-influence, if real at all, is smaller than ε under open conditions, and standard forward-causal narratives suffice.
  Either way, the program upgrades how we study coincidence, constraint, and explanation—moving debates about "engineered reality" onto testable, sharable, community-governed ground.

10. Conclusion

10.1 What we have tested, what remains open

What we tested.
We translated a provocative idea—**boundary-weighted histories conditioned on future utilities**—into falsifiable science. Across Sections 3–6 we specified (i) a **primary scalar tilt** (Δg) that must appear only in arms tied to a preregistered future utility U, (ii) **information-theoretic coupling** (ΔMI in bits) and **compression gains** (ΔMDL as % shorter codes) that corroborate the tilt, and (iii) **guardrails** (no-signalling, conservation, energy closure) and **negative controls** (sham U, timing permutations) that must stay null. We proposed four **pilot classes**—utility-conditioned delayed-choice, blind-scored digital evolution, long-baseline calorimetry for entropy dips, and world-data compressibility tracking—each with power sketches, leakage defenses, and one-click reproducibility.

What this closes.

- It **demarcates** retrocausal claims from anecdotes: only preregistered, U-locked, arm-specific tilts count.
- It **rules out** many ordinary artifacts via mirrored sham utilities, timing misalignments, site stratification, and escrowed provenance.
- It **quantifies** outcomes in common units (effect size, bits, % code length), making cross-study synthesis possible.

What remains open.

- The **form of W**: Even with positives or nulls, the weighting functional W(U(H,T)) remains only bounded, not identified.
- **Scaling laws**: How Δg, ΔMI, ΔMDL vary with horizon length (T−t0), utility complexity, and system size is unknown.
- **Mechanistic attribution**: Post-selection, final-state constraints, and fixed-point solvers are phenomenological templates; discriminating among them would require new probes (e.g., utility geometry, cross-utility interference tests).
- **External validity**: Pilots are deliberately "toy-ish." Whether any effect—positive or ε-bounded null—transfers to other domains (biology, economics, geophysics) needs targeted replications.

10.2 A roadmap for larger, cleaner, cheaper next-generation tests

Design upgrades (cleaner).

1. **Stage-1 Registered Reports + red-team prereg** by default: methods accepted in principle, adversarial critique before data.
2. **Stronger independence**: hardware data diodes, key-split custody of U_inputs.json, and third-party beaconing to guarantee post-T inputs.
3. **Stricter guardrails**: automatic no-signalling dashboards and daily energy-closure attestations baked into the run harness.
4. **Unified primary endpoint**: one global Δg across labs, with predeclared site-level nuisance adjustments; MI/MDL as co-primary only if jointly α-controlled.

Statistical power (larger).

1. **Meta-power planning**: simulate multi-site power curves under site heterogeneity (I^2 targets ≤ 25%).
2. **Longer horizons T** where feasible (e.g., committee decisions, leaderboard freezes) to maximize any boundary weighting—paired with more units to keep variance in check.
3. **Adaptive allocation** (ethically benign): response-adaptive randomization between A and B guided by blinded conditional power to reach information thresholds sooner without inflating α.

Cost control (cheaper).

1. **Commodity rigs & open stacks**: off-the-shelf optics kits; Raspberry-Pi-class controllers; frozen containers; reuse of community compressors and MI estimators.
2. **Shared Utility Repository (RUR)**: audited U/U_sham bundles reduce setup time and eliminate bespoke coding; bounties for adding well-documented utilities.
3. **Cloud escrow + IPFS mirrors**: one-time setup yields durable, low-touch data custody for many studies.
4. **Replication micro-grants**: small, standardized payouts for confirmatory runs (positive or null) drive economies of scale.

Sensitivity & robustness (sounder).

1. **Orthogonal endpoints**: pair Δg with a predeclared placebo outcome and an instrumented negative-control stream every run.
2. **Compressor/estimator diversity**: require effects to persist across ≥2 compressors and ≥2 MI estimators fixed in prereg.
3. **Timing triage**: preregistered CUSUM to auto-flag regime shifts; prewritten logic either trims or stratifies affected intervals.
4. **Leak audits**: scheduled "canary" drops (fake U variants) that must remain null; any non-null auto-invalidates confirmatory status.

Exploration to mechanism (smarter).

1. **Utility geometry**: run factorial sets of U with known overlaps to test superposition/competition (does coupling add, saturate, or interfere?).

2. **Horizon scaling**: systematic scans over T to map decay/accumulation of ΔMI and ΔMDL with delay.
3. **Cross-domain triangulation**: pair a quantum pilot with a classical information pilot in the same window to see if tilts co-vary with the same U.

Reporting & culture (stickier).

1. **Living meta-analysis**: rolling DOIs that update pooled ε-bounds or pooled positives with each replication.
2. **Prestige for nulls**: abstract includes ε-bounds; badges for "Robust Null" and "Leak Discovered" earn equal placement to "Confirmatory Replication."
3. **Public dashboards**: per-study pages with prereg hashes, run clocks, guardrail lights (green/amber/red), and 1-click repro links.

Closing thought.
Whatever the outcome, the program upgrades our **standards for testing constraint-based explanations**. If signals appear, we gain a calibrated, laboratory handle on boundary weighting without breaking dynamics. If nulls dominate, we secure sharp ε-bounds that retire a class of speculative narratives and redirect attention to forward-causal accounts. In both cases, the payoff is the same: better demarcation, better methods, and a reusable blueprint for turning high-weirdness into plain, auditable science.

Appendices

A. Formalization of boundary-weighted histories (copy-safe math)

A.1 Objects and measures

- Time window: [t0, T].
- History space: H = set of admissible micro/macro histories h obeying standard dynamics and constraints.
- Forward (initial-value) reference measure: P0(h), induced by ordinary equations of motion plus noise.
- Future utility: U: H -> R, computable only at time T from h's terminal records.
- Weighting functional: W: R -> R_{\ge 0}, unknown a priori.

Boundary-weighted observational law

P_obs(h) = [P0(h) * W(U(h))] / Z, Z = sum_{h in H} P0(h) * W(U(h)).

Special cases:

- No retro-weighting: W(u) = 1 => P_obs = P0.
- Hard post-selection on threshold u*: W(u) = 1{u >= u*}.
- Soft preference: W(u) = exp(lambda * u) with small |lambda|.

A.2 Conditional tilts and arm structure

Two arms A,B differ only in whether their histories are scored by U (A) or U_sham (B). Let g: H -> R be a predeclared statistic.

Delta_g := E_obs[g(h) | arm=A] - E_obs[g(h) | arm=B].

Under W=1 for both arms (null), Delta_g = 0 up to noise. Under boundary weighting tied to U in A only:

E_obs[g | arm=A] = sum_h g(h) * [P0(h) * W(U(h))] / Z_A,

E_obs[g | arm=B] = sum_h g(h) * [P0(h) * W(U_sham(h))] / Z_B ~ sum_h g(h) * P0(h).

A.3 No-signalling guardrail (copy-safe)

Let X_A be outcomes in spacelike region A, S_B settings in region B. For any arm:

P(X_A | S_B, arm) = P(X_A | arm). (no-signalling)

Positive evidence is an **armwise** shift:

P(X_A | arm=A) != P(X_A | arm=B),

never a dependence on S_B.

A.4 Information signatures

Let Z=f(h) be a fixed summary. Mutual information (discrete plug-in):

I(Z;U) = sum_{z,u} p(z,u) * log(p(z,u) / (p(z)*p(u))).

MDL gain for a fixed model class M and compressor C:

Gamma(h) = L0(h) - LU(h),

L0 = codelength under baseline model M without U,

LU = codelength under same class with a predeclared side-info channel from U.

Treatment signature: Delta_MDL := E[Gamma | arm=A] - E[Gamma | arm=B] > 0.

A.5 Bounding W from nulls

For small effects and smooth W, first-order expansion:

E_obs[g] - E0[g] ~ Cov_0(g(h), log W(U(h))).

If |Delta_g| <= eps_g (95% CI), then for any g in a predeclared class G,

sup_{g in G} | Cov_0(g, log W(U)) | <= eps_g.

This turns nulls into quantitative constraints on log W's effective variation under P0.

B. MDL/MI estimation recipes and code stubs

B.1 Discrete MI (plug-in with Miller–Madow correction)

- Preregister bins for Z and quantization for U (if scalar).
- Compute counts n_{z,u}, marginals n_z, n_u over N samples.
- Plug-in MI:

I_hat = sum_{z,u} (n_{z,u}/N) * log((n_{z,u} * N) / (n_z * n_u)).

- Miller–Madow bias correction (approx):

bias ~ ((|Z|-1) * (|U|-1)) / (2N * ln 2),

I_mm = I_hat - bias.

B.2 kNN MI (continuous or mixed)

- Fix k (e.g., k=5) and distance metric; use Kraskov-type estimator KSG-1.
- Freeze k and tie-breaking in prereg.
- Use block-preserving permutations for p-values.

B.3 MDL with fixed compressors

Two safe choices that avoid learning from outcomes:

1. **Arithmetic coder over fixed tables**
    - Freeze symbol alphabet and order-n context tables learned **before** the study from public corpora.
    - L0: code Z (or raw X) with fixed tables.
    - LU: same, but with a **predeclared** side-channel bit or index determined by U (e.g., choose 1 of 2 frozen tables).
2. **Frozen lightweight predictor**
    - A tiny RNN/HMM with weights frozen from pre-study data; emit probability for next symbol; arithmetic code under that probability.
    - Side-channel: a fixed gate that selects one of K frozen heads based on U.

B.4 Code stubs (Python)

Discrete MI with plug-in + Miller–Madow:

```
import numpy as np


def mi_discrete_mm(z, u):
    # z, u are 1D integer arrays of same length
    assert len(z) == len(u)
    N = len(z)
    z_vals, z_idx = np.unique(z, return_inverse=True)
    u_vals, u_idx = np.unique(u, return_inverse=True)
    Z, U = len(z_vals), len(u_vals)


    counts = np.zeros((Z, U), dtype=np.int64)
    for i in range(N):
        counts[z_idx[i], u_idx[i]] += 1
    nz = counts.sum(axis=1, keepdims=True)
    nu = counts.sum(axis=0, keepdims=True)
    with np.errstate(divide='ignore', invalid='ignore'):
        term = counts * (np.log((counts * N) / (nz @ nu)) )
    term[np.isnan(term)] = 0.0
    I_hat = term.sum() / N / np.log(2)  # bits
    bias = ((Z - 1) * (U - 1)) / (2.0 * N * np.log(2))
    return I_hat - bias
```

Block-preserving permutation test:

```
def perm_test_mi(z, u, blocks, B=10000, rng=None):
    # blocks: list of index arrays, permutations must stay within each block
    rng = np.random.default_rng() if rng is None else rng
    obs = mi_discrete_mm(z, u)
```

```
    cnt = 0
    for _ in range(B):
        u_perm = u.copy()
        for idx in blocks:
            u_perm[idx] = rng.permutation(u_perm[idx])
        val = mi_discrete_mm(z, u_perm)
        if val >= obs - 1e-12:
            cnt += 1
    pval = (cnt + 1) / (B + 1)
    return obs, pval
```

Minimal MDL (two frozen models, side-channel selects model A or B):

```
def code_length(seq, model_probs):
    # seq: list of symbols in {0..S-1}
    # model_probs[t, s] = P(symbol=s | history up to t) from frozen model
    import math
    L = 0.0
    for t, s in enumerate(seq):
        p = max(model_probs[t, s], 1e-12)
        L -= math.log2(p)
    return L


def mdl_gain(seq, probs_A, probs_B, U_bit):
    L0 = code_length(seq, probs_A)  # baseline uses A
    LU = code_length(seq, probs_A if U_bit == 0 else probs_B)
    return L0 - LU  # positive = gain
```

B.5 Prequential metrics (log-score)

Fix an online predictor M0. For each t:

logloss0_t = -log p0(x_t | past),

lossU_t = -log pU(x_t | past, U_sideinfo),

Gamma_t = logloss0_t - lossU_t.

Aggregate Gamma_t by block; compare arms (Delta_MDL).

B.6 Power via site-specific null simulation

- Freeze the estimator(s) and block structure.
- Fit a forward-only generator to pilot baselines.
- Simulate many datasets under W=1; record the empirical SDs of Delta_g, Delta_MI, Delta_MDL; set MDEs accordingly and select N.

C. Protocol templates, preregistration checklists, and audit artifacts

C.1 Minimal preregistration template (YAML)

```
study_id: RETRO-XYZ
version: v001
authors:
  - name: ...
    affiliation: ...
primary_question: >
  Do U-tied runs exhibit a positive Delta_g and corroborating increases in MI and MDL
  that vanish under sham utilities and timing permutations?

arms:
  A: U_tied
  B: U_sham

design:
  units: blocks
  allocation: blocked_randomization(site, day)
  horizon:
    t_start: 2025-10-05T00:00:00Z
    t_stop:  2025-10-20T00:00:00Z
    t_score: 2025-10-27T00:00:00Z

utility:
  U_eval_path: code/U_eval.py
  U_inputs_spec: docs/U_inputs_spec.md
```

```
  U_sham_path: code/U_sham.py
  blinding: external_custodian_provides_U_inputs_json_at_t_score

endpoints:
  primary:
    name: Delta_g
    definition: docs/g_definition.md
    alpha_one_sided: 0.025
    estimator: GLM_identity_HC3
    mde: 0.01
  secondary:
    - name: Delta_MI
      estimator: discrete_plugin_mm
      binspec: docs/Z_binning.json
    - name: Delta_MDL
      compressor_suite: ["arith_ctx3", "frozen_rnn_small"]
      sideinfo_channel: "1bit_from_U"

guards:
  no_signalling: true
  energy_closure: true
  leakage_diagram: docs/DAG.png

randomness:
  setting_rng: cosmic_rng + hw_trng_mirror
  seeds_release: "post-lock"
```

```
  logs: "raw_rng_events_worm_log"

analysis:
  container: ghcr.io/org/retro:v001
  script: scripts/run_primary_and_secondaries.sh
  hash_sha256: TBD_AFTER_BUILD

registration:
  bundle_hash_sha256: TBD
  merkle_root: TBD
  registries:
    - osf: osf.io/xxxx
    - ledger: txid:0xabc...

stopping:
  group_sequential: obrien_fleming
  looks: 3

replication:
  sites: ["S1", "S2"]
  meta: fixed_effects_IVW
```

C.2 Pre-data checklist (adversarial)

1. Single tarball contains: protocol, code, frozen models, compressors, RNG specs, U/U_sham.
2. SHA-256 and Merkle root posted to 2+ venues.
3. Arm labels masked; analysts sign off that analysis container builds from the tarball only.

4. U_inputs.json held by external custodian; labs cannot access or infer contents.
5. Sham symmetry verified: U_sham matches type and marginal distribution.
6. Leak threat model documented; canary tokens embedded.
7. Negative controls (placebo outcomes, timing permutations) fully scripted.
8. Power curve produced from site-specific null simulation.
9. No-signalling and energy-closure tests scripted and unit-tested.
10. Red-team window completed; responses appended to prereg.

C.3 Run-time checklist

- UTC/GPS sync within tolerance; offsets logged.
- RNG health checks passed; raw RNG streams mirrored to escrow.
- Block randomization executed; A/B balance by site/day verified online.
- WORM logs rolling; daily hash manifests signed.
- Guardrail dashboards green (no-signalling prelims, closure prelims).
- Any deviations documented as amendments (hash + timestamp).

C.4 Post-lock analysis checklist

- Freeze analysis container; record its image digest.
- Post analysis hash, then unmask arms.
- Execute primary endpoint first; record Z, CI, p.
- Run MI/MDL with prereg estimators; run sham and timing-perm variants.
- Run guardrail tests; compile a single “traffic-light” summary.
- Produce epsilon-bounds if null.
- Export reproducibility package; verify 1-click rebuild passes on clean machine.

C.5 Audit artifacts (deliverables)

- prereg_bundle.tar.gz with SHA-256 and Merkle root.
- Raw data (or de-identified derivatives), RNG traces, environment logs.
- Containers: acquisition, preprocessing, analysis (image digests).

- DAG and leak-test reports (including canary results).
- Guardrail reports: no-signalling tables, energy-closure plots.
- Primary and secondary endpoint notebooks with exact seeds.
- Replication playbook (bill of materials, power target, SOPs).
- Signed attestations: site PIs, external custodian, independent analyst.
- DOI'd replication packages and a living meta-analysis entry.

C.6 Minimal "how to verify me" (human-readable)

1. Check the prereg hash posted before data collection matches the bundle.
2. Build the analysis container; confirm its digest matches the recorded one.
3. Recompute primary Delta_g and its CI from raw data using scripts/run_primary.sh.
4. Run scripts/run_controls.sh to see sham/timing-perm nulls.
5. Verify no-signalling and closure summaries are green.
6. Confirm that re-running on a clean machine reproduces all numbers.

References

Foundations of retrocausality, two-boundary pictures, and weak values

1. Aharonov, Y., Bergmann, P. G., & Lebowitz, J. L. (1964). Time symmetry in the quantum process of measurement. *Physical Review*, 134(6B), B1410–B1416.
2. Aharonov, Y., Albert, D. Z., & Vaidman, L. (1988). How the result of a measurement of a component of the spin of a spin-½ particle can turn out to be 100. *Physical Review Letters*, 60(14), 1351–1354.
3. Aharonov, Y., & Vaidman, L. (1991). Complete description of a quantum system at a given time. *Journal of Physics A: Mathematical and General*, 24(10), 2315–2328.
4. Ma, X.-S., Kofler, J., Zeilinger, A. (2016). Delayed-choice gedanken experiments and their realizations. *Reviews of Modern Physics*, 88(1), 015005.
5. Price, H. (1996). *Time's Arrow and Archimedes' Point*. Oxford University Press.
6. Wharton, K. (2010). Time-symmetry as a solution to the measurement problem. *Symmetry*, 2(1), 272–283.

Bell tests, delayed-choice/eraser, entanglement swapping

7. Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49(25), 1804–1807.
8. Hensen, B., et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 km. *Nature*, 526, 682–686.
9. Shalm, L. K., et al. (2015). Strong loophole-free test of local realism. *Physical Review Letters*, 115(25), 250402.
10. Kim, Y.-H., Yu, R., Kulik, S. P., Shih, Y., & Scully, M. O. (2000). Delayed "choice" quantum eraser. *Physical Review Letters*, 84(1), 1–5.
11. Walborn, S. P., Terra Cunha, M. O., Pádua, S., & Monken, C. H. (2002). Double-slit quantum eraser. *Physical Review A*, 65(3), 033818.
12. Pan, J.-W., Bouwmeester, D., Weinfurter, H., & Zeilinger, A. (1998). Experimental entanglement swapping. *Physical Review Letters*, 80(18), 3891–3894.
13. Ma, X.-S., Zotter, S., Kofler, J., Ursin, R., Jennewein, T., & Zeilinger, A. (2012). Experimental delayed-choice entanglement swapping. *Nature Physics*, 8, 479–484.
14. Handsteiner, J., et al. (2017). Cosmic Bell test: Measurement settings from high-redshift quasars. *Physical Review Letters*, 118(6), 060401.

Post-selection, final-state, and fixed-point/CTC models

15. Aaronson, S. (2005). Quantum computing, postselection, and probabilistic polynomial-time. *Proceedings of the Royal Society A*, 461(2063), 3473–3482.

16. Deutsch, D. (1991). Quantum mechanics near closed timelike lines. *Physical Review D*, 44(10), 3197–3217.
17. Brun, T. A., Harrington, J., & Wilde, M. M. (2009). Localized closed timelike curves can perfectly distinguish quantum states. *Physical Review Letters*, 102(21), 210402.
18. Horowitz, G. T., & Maldacena, J. (2004). The black hole final state. *Journal of High Energy Physics*, 02, 008.

No-signalling, causal constraints, and analysis

19. Wood, C. J., & Spekkens, R. W. (2015). The lesson of causal discovery algorithms for quantum correlations. *New Journal of Physics*, 17(3), 033002.
20. Cavalcanti, E. G., & Lal, R. (2014). On modifications of Reichenbach's principle of common cause in light of Bell's theorem. *Journal of Physics A*, 47(42), 424018.

Weak measurements and related experiments

21. Dressel, J., Malik, M., Miatto, F. M., Jordan, A. N., & Boyd, R. W. (2014). Colloquium: Understanding quantum weak values. *Reviews of Modern Physics*, 86(1), 307–316.
22. Lundeen, J. S., et al. (2011). Direct measurement of the quantum wavefunction. *Nature*, 474, 188–191.

Statistical foundations, MI/MDL, prequential analysis

23. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.
24. Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3rd ed.). Springer.
25. Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.
26. Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6), 066138.
27. Miller, G. A. (1955). Note on the bias of information estimates. *Information Theory in Psychology*, 95–100.
28. Dawid, A. P. (1984). Present position and potential developments: Some personal views—Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147(2), 278–292.

Sequential testing, error spending, and replication design

29. O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3), 549–556.
30. Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2), 191–199.
31. Lakens, D. (2021). Sample size justification. *Collabra: Psychology*, 7(1), 33267.

32. Van Calster, B., et al. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, 230.

Thermodynamics of fluctuations and entropy production

33. Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14), 2690–2693.
34. Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation. *Physical Review E*, 60(3), 2721–2726.
35. Seifert, U. (2005). Entropy production along a stochastic trajectory. *Physical Review Letters*, 95(4), 040602.
36. Ciliberto, S., Joubaud, S., & Petrosyan, A. (2010). Fluctuations in out-of-equilibrium systems: From theory to experiment. *Journal of Statistical Mechanics*, 2010(12), P12003.
37. Parrondo, J. M. R., Horowitz, J. M., & Sagawa, T. (2015). Thermodynamics of information. *Nature Physics*, 11, 131–139.

Evolution, convergence, and digital evolution

38. Lenski, R. E. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME Journal*, 11, 2181–2194.
39. Wilke, C. O., & Adami, C. (2002). The biology of digital organisms. *Trends in Ecology & Evolution*, 17(11), 528–532.
40. Clune, J., Mouret, J.-B., & Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society B*, 280(1755), 20122863.

Open science, preregistration, and registered reports

41. Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.
42. Chambers, C. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49(3), 609–610.
43. Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Mapping the universe of Registered Reports. *Nature Human Behaviour*, 2, 793–796.

Algorithmic compressibility and practical compression

44. Mahoney, M. V. (2012). *Data Compression Explained*. (Self-published monograph; widely cited for practical MDL/compression methods.)
45. Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4), 629–636.

Causal discovery & methodology guardrails

46. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

47. Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Miscellaneous tools cited in protocols

48. NIST. (2011–). Randomness Beacon. (For cryptographic, time-stamped entropy beacons referenced as safe utilities.)

49. ISO 8601 / IETF RFC 3339. (Timestamp standards for UTC logging used in protocol checklists.)

Note: The above list groups core sources by theme (quantum retrocausality, information theory/MDL, fluctuation thermodynamics, experimental/replication design, and open-science governance) to support all sections of the paper (conceptual framework, signatures, pilots, and analysis). Where multiple canonical demonstrations exist (e.g., Bell tests, quantum erasers), representative high-quality exemplars are cited.