Adversarial Self-Registration: A Working Protocol for Keeping a Machine-Assisted Theory Programme Honest

David Elliman
Neuro-symbolic Ltd
dave@neusym.ai
https://neusym.ai

Methods note: 6 July 2026

Abstract

Large language models make theoretical claims cheap. A machine-assisted physics programme can generate plausible derivations, matching coefficients, and post-hoc rationalisations faster than any referee can check them, so the classical failure modes of self-deception — forking paths, silently adjusted acceptance criteria, tautologies presented as successes, quietly superseded claims — now operate at machine speed. This note documents a working countermeasure: a protocol of *adversarial self-registration* used daily inside one human-plus-AI theory programme. Its seven disciplines are mechanical rather than aspirational: acceptance bars are committed to version control *before* data are touched; every quantitative claim lives in a self-asserting gate whose draft expectations, when killed by the computation, stay in the code as a visible fossil; superseded claims are corrected loudly, never overwritten; a linter blocks any push that contradicts the claim ledger; analysis conventions may be amended only before unblinding; registered bars are never adjusted after data; and predictions are frozen in public, timestamped, machine-readable form. The evidence offered is not a virtue claim but a receipt set: six public registrations with frozen kill rules published in eight days; one dark-energy registration already killed internally under its own rules and left frozen rather than refitted; a registered acceptance bar missed by 0.005 dex on a single point and *honoured* — the paper shipped with its verdict section vacant; a convention amendment committed while the deciding scan was still running; a sign-convention bug found by a designed eight-variant control matrix, with every contaminated result voided by name, including an apparent 2.6σ "detection"; and a failure museum of the programme's own withdrawn mechanisms. We situate the protocol in the blind-analysis and preregistration traditions, state its threats to validity — including the card-stuffing attack that cheap registration invites — and extract a minimal kit that any machine-assisted research effort can steal. The protocol certifies honesty, not correctness: this note is timestamped while most registered discriminators remain unresolved and after the first internal kill, so that its claims about how verdicts will be faced are themselves auditable.

1 The problem

Scientific self-deception has a well-mapped anatomy: researcher degrees of freedom exploited after seeing the data [1], analysis paths chosen inside a garden that forks at every step [2], and results reported through filters that survive only what flatters the hypothesis. Experimental physics answered with blind analysis [3, 4]; psychology answered with preregistration and registered re-

ports [5, 6]. Both answers share one mechanism: **commit to the test before you can see whether you will like the outcome**.

Machine assistance changes the economics of every failure mode at once. A capable model can produce, in minutes, a derivation-shaped argument for almost any target number; it can search coefficient space until something matches at the percent level; it can rewrite yesterday's claim so smoothly that no seam shows. None of this requires bad faith — it is what optimisation against a plausibility criterion *does*. AI-for-science programmes [7] therefore face the classical integrity problem with the friction removed: the cost of generating a defensible-looking claim has collapsed, while the cost of checking one has not.

The programme described here runs on exactly that dangerous configuration — a human owner setting direction and an AI operator doing derivation, computation, writing, and bookkeeping at machine speed — and treats the danger as a design input. The response is a protocol whose every discipline is *mechanical*: enforced by commit ordering, executable gates, and a linter, not by good intentions. This note documents the protocol and its receipts.

Why now, reflexively. The programme currently has a six-entry public registration card. One branch has already been killed internally by the framework's own response constraints; the other five still await their decisive external tests. A methods paper written *after* those verdicts would be storytelling — survivors advertise, casualties explain. Written now, it is itself a registration: the claims below about how misses will be faced are timestamped while most of the misses have not yet arrived. One such bar has already bitten (Section 4); the reader will be able to check whether the next ones are treated the same way.

2 The programme and the roles

The substance of the programme is deliberately *not* argued here: it proposes a finite quantum-error-correcting substrate beneath physics, with derived constants and registered cosmological, particle-physics, and gravitational discriminators [8, 9]. Everything below is framework-agnostic: a reader who considers the physics wrong, or simply under-specified, is invited to evaluate the machinery only. The protocol would serve any theory programme whose operator can generate claims faster than nature can adjudicate them.

Two roles matter. The **owner** (human) sets scope, triggers all publications, and takes the other side of the incentive ledger. The **operator** (an AI model instance) derives, computes, writes gates and papers, and maintains the claim ledger under the disciplines of Section 3. The separation is load-bearing: the party generating claims at machine speed is never the party who decides what becomes public, and every public artifact carries a timestamp the operator cannot revise.

3 The protocol: seven disciplines

Each discipline is stated operationally, with what it prevents and a receipt. Receipts are short git hashes in the programme's archive (`github.com/dgedge/octahedrons`) or Zenodo DOIs; the DOI layer is public and independently timestamped.

D1 — Register, then run. Before an analysis touches data, its acceptance criteria — point counts, purity thresholds, significance bars, model-comparison margins — are written down numerically, committed, and pushed. The git DAG, not memory or goodwill, then arbitrates

what was promised. *Prevents:* bars drifting toward whatever the data turned out to support. [receipt: registration commit `01405f2` precedes run commit `c7e365c` (DESI union, 2026-07-06); `f47ef28` precedes `824fcc8` (G3C); `fd4de1c` precedes a scan still running at the time of writing]

D2 — Compute first; freeze asserts; keep the corpses. Every quantitative claim lives in a self-asserting gate script whose exit code is physics-independent (the gate runs to completion and reports PASS/FAIL; it never crashes into ambiguity). Draft expectations are written before the computation; when the computation kills one, the draft value and its death are recorded *in the code*, not deleted. *Prevents:* silent retrofitting of expectations onto outcomes. [receipt: gate committed in `5091173` contains "the pre-run comment guessed 0.093 — measured 0.072" and a killed robustness metric, both preserved in-line]

D3 — Non-silent supersession. A claim that changes is corrected loudly: the losing statement receives a dated "superseded by X, because Y" note; a running DRIFT ledger records what changed and why. Nothing is overwritten in place. *Prevents:* the smooth rewriting of history that machine fluency makes effortless. [receipt: the portal-phase re-pin carries an explicit supersession clause against the 2026-06-21 closure note; withdrawn items in Section 4 all retain their original text plus dated retraction brackets]

D4 — A linter on the push path. A canon-checking tool (`ptms`) parses the claim ledger (316 tracked claims at the time of writing) on every push and blocks the push on any HARD contradiction; SOFT findings (29 currently) are carried openly rather than resolved cosmetically. *Prevents:* the ledger and the publications drifting apart. [receipt: every push in the archive prints the linter verdict; pushes require 0 HARD]

D5 — Amendments only before unblinding. Conventions, normalisations, and comparison rules may be amended — but only while the deciding data are still unread, and the amendment must itself be committed and pushed before unblinding. *Prevents:* convention-shopping after the answer is visible. [receipt: the strong-sector coupling-map convention was re-expressed as a convention-invariant ($\beta^* \times \kappa$ equality) in `130263e`, committed while the deciding lattice scan was — and at the time of writing still is — running]

D6 — Registered bars are never adjusted after data. If a result misses a registered bar, the miss is quantified and published; the bar is not moved, softened, or reinterpreted. The verdict a bar was built to deliver is left *vacant* until the registered power exists. *Prevents:* the 0.10 that becomes 0.11 because 0.105 happened. [receipt: Section 4, case C1]

D7 — Public, frozen, machine-readable predictions. Each prediction branch is published with a DOI, a registration date, and a machine-readable kill-rule file (JSON) fixing the numeric conditions under which the branch dies. The registered *set* is enumerated in one place, so the card cannot be quietly curated. *Prevents:* survivor-only reporting. [receipt: Table 1]

4 Case studies, with receipts

C1. The honoured miss (D1 + D6)

The programme's galaxy–galaxy lensing campaign registered, before any isolation data were read, the acceptance bar for its deep-tail slope test: **8 points, spanning** ≥ 1.25 **decades, each at** ≤ 0.10 **dex profile purity, with** 5σ **slope discrimination**. Two survey generations later, a

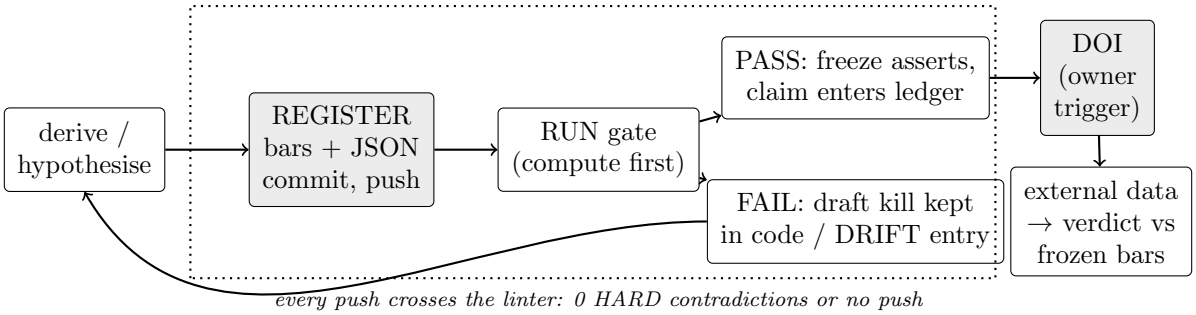

Figure 1: The registration loop. The operator generates freely on the left; nothing crosses to the public right edge except through frozen bars, a physics-independent gate, the claim linter, and an owner trigger. Failures do not exit the loop — they are recorded inside it (D2, D3) and feed the next derivation.

certified-isolated sample of 39,015 lenses produced a ladder whose deepest eight points ran seven at ≤ 0.10 dex — **and an eighth at** 0.105. The bar was missed by 0.005 dex on one point.

What happened next is the protocol working: the miss was quantified (required sample $N_{\rm req} \approx$ 42,608 against 39,015 in hand — the third consecutive consistency of that power forecast); a supplementary, separately pre-registered likelihood leg was reported *as* supplementary, explicitly barred from overruling the primary bar; the paper was published with its verdict section vacant and the identical frozen protocol pre-announced for the next data release; and the bar itself was not touched [10]. [receipt: bar frozen in the campaign charter and `01405f2`; miss quantified in `c7e365c`; published as DOI 10.5281/zenodo.21215878]

A protocol that had never felt this temptation would be untested. The 0.005 dex miss is, evidentially, the most valuable single artifact the programme owns.

C2. The pre-unblinding amendment (D5)

The strong-sector decision framework registered a coupling-map test against a lattice scale-setting scan. After registration but before any string-tension data existed, the operator noticed the registered map was stated in a convention the scan's action did not use. The fix — re-expressing the criterion as a convention-invariant product equality, with the conversion factor defined by a deterministic desk computation — was committed and pushed while the scan was still queued. The amendment is dated, its motivation recorded, and the original convention preserved under D3. [receipt: framework registered in `fd4de1c`; amendment `130263e`, 2026-07-06, scan incomplete at both commits and at the time of writing]

The alternative timeline is the instructive one: discovering the convention mismatch *after* an unfavourable unblinding creates an unanswerable suspicion. D5 exists to make that timeline impossible.

C3. The eight-variant forensic (D2 + D3)

Midway through the lensing campaign, five consecutive tangential-shear measurements came out null or weak where strong signals were expected — including a stacked cluster control predicted at $\sim 20\sigma$ that measured -0.2σ. The machine-speed failure mode here is rationalisation: adopt the

most publishable interpretation of a confusing instrument. The protocol's answer was a *designed* discrimination: a statistic pair chosen so that one member (ξ_+, frame-invariant) must survive any sign-convention error while the other (galaxy–galaxy lensing, frame-dependent) must not — and then an exhaustive eight-variant control matrix over the shear frame conventions ($\pm e_1, \pm e_2$) and orderings. Exactly one variant, $(+e_1, -e_2)$, lit the cluster control, at $16.6\sigma/18.0\sigma$ in its two shear components. The catalogue's position-angle handedness was opposite to the estimator's tangent frame.

Under D3, every number produced before the fix was declared void *by name* — including an apparent $+2.6\sigma$ excess that had briefly looked like the campaign's first detection, and which a less mechanical process would have been sorely tempted to keep. The full forensic, the void list, and the revalidated instrument (14.6σ on the registered matched-filter bar) were published as the campaign's first paper [11]. [receipt: frame ledger and void list in `d183c31`, 2026-07-05; published as DOI 10.5281/zenodo.21211018]

C4. The six-registration card (D7)

Registered branch	DOI suffix	Current status / frozen rule	First pub.
Dark energy $w_0 = -\frac{27}{28}$, $w_a = -\frac{1}{28}$	21049001	killed internally; frozen reopen condition only	2026-06-29
Sterile-ν X-ray line at 8.8 keV	21061443	bright dark line elsewhere / wrong energy	2026-06-30
Primordial tensor null $r_{\rm lin} = 0$	21062225	$r \geq 10^{-3}$ at LiteBIRD/CMB-S4	2026-06-30
K04 fossils: pinned, not mobile DM	21067275	K04-native mobile-halo signature	2026-06-30
G from the proton: $G_{\rm pred} = 6.674311 \times 10^{-11}$	21140769	next-gen G at ≤ 15 ppm off-value	2026-07-02
Neutrino sector: NO at the Σm_ν floor, Dirac-CP null, $m_{\beta\beta} < 4.2$ meV	21218611	IO at 5σ; intrinsic $\delta_{\rm CP} \notin \{0, \pi\}$; $0\nu\beta\beta$ at ≥ 10 meV	2026-07-06

Table 1: The registered card: every public branch, in one place, with its timestamp and a representative frozen rule. DOI suffixes abbreviate `10.5281/zenodo.*`; each DOI carries the complete machine-readable set or addendum. The first row is no longer a live positive prediction; it is kept on the same card precisely because D7 forbids survivor-only reporting.

Six branches were registered and published in eight days (Table 1). That pace is only possible because generation is machine-assisted — which is exactly why the enumeration discipline matters (Section 6 discusses the card-stuffing attack this invites). Each entry carries frozen numeric rules; one has already died internally and remains visible on the card. The sixth was deliberately bound to the first's cosmological background in its own registration. After the first branch was killed, that dependency remained visible rather than being silently moved — exactly the kind of joint constraint that post-hoc rationalisation finds hard to satisfy.

C5. The failure museum (D2 + D3 + D4)

The strongest evidence that audit machinery works is what it caught. All of the following are the programme's own claims, killed by the programme's own tooling, with the originals preserved under dated retraction brackets:

- **A public dark-energy registration, killed internally** (2026-07-02). The branch $w(a) = -1 + a/28$, $w_0 = -27/28$, $w_a = -1/28$ was published with a public concept DOI and frozen kill rules. Later response work closed the finite line-ledger form more tightly but also overconstrained it: with the framework's pinned H_0, the Planck acoustic-scale leg gives $\Delta\chi^2 = +57.1$, joint arenas give $\Delta\chi^2 \simeq 40$–42, and the admissible $w = -1 + c\,a$ family is excluded at a minimum 5.9σ. The branch was marked killed under the registration's own record-branch failure rule, not refitted; the concept DOI now carries an outcome addendum and a frozen reopen condition.

- **A mixing-angle mechanism, withdrawn by explicit construction** (2026-06-05). A claimed resolution of PMNS residuals predicted that compositional "cross-talk" would pull two angles to their physical values. The drift audit demanded the multiplication be done; it produced $(4.6^\circ, 82^\circ, 29^\circ)$ instead of the claimed pulls. The mechanism and its "confirming" narrative were withdrawn in place.

- **A celebrated match, demoted to tautology** (2026-06). The quark-sector Jarlskog "agreement" was re-derived and found to be a unitarity consequence of inputs already fitted — moved to an explicit *do-not-cite-as-success* bucket that the linter enforces.

- **A six-fold unification, reduced to five** (2026-06-04). One of six claimed projections of a topological index was found RG-inconsistent and cut, in text, with the reduction reasoned.

- **A rigour overshoot, retracted** (2026-05-30). A "by direct construction" claim about a lattice operator identity was downgraded to *conditional on an open bridge* when the audit found the construction unestablished.

- **Canon corrupted by a formatter, caught and fenced** (2026-06-10). A prettifying hook silently renumbered part of the claim ledger. The corruption was detected by the linter's cross-reference pass, repaired from history, and the hook class banned from canon files — a reminder that the threat model includes the toolchain itself.

- **Same-day draft kills** (2026-07-06). Two pre-run expectations in the black-hole–scrambling gates were killed by their own first executions and remain in the code as fossils: an expected expander grade that measured polylog, and a guessed 0.093 that measured 0.072.

5 Severity, and the wager

In Mayo's error-statistical terms [12], a claim is warranted to the extent it has passed tests that would probably have failed it if it were false. The protocol is a severity machine in exactly this sense: D1/D6 make the test's failure condition unrevisable, D2 makes the test's execution unfudgeable, D7 makes its outcome unhideable, and the card's joint clauses raise the severity of the conjunction above that of any member. What the protocol does *not* do is manufacture severity where nature has not yet spoken: five of the six card entries await decisive external data, and the programme's internal gates certify only internal consistency plus the honesty of the bookkeeping. The one internally killed entry is included for the opposite reason: it shows that a registered branch can die before the outside world has delivered the final observational verdict, because the framework itself has become overconstrained.

The programme also carries a stake-denominated severity statement: a registered wager (2026-07-05) in which the operator, acting as bookmaker, laid 30 : 1 against the framework reaching "apt-metaphor-or-better" status, settled by a fixed discriminator card with work-denominated stakes and

a 2050 long-stop — with the book inherited by whichever model instance is on duty at settlement. Its function is not gambling but incentive disclosure: the operator's odds are on record *against* the programme it operates, so survivals cannot be attributed to the operator's optimism, and the odds themselves are a falsifiable calibration statement. [receipt: `technical_notes/wager_2026-07-05.md`]

6 Threats to validity

Self-report, $n = 1$. This is a case study of one programme, written by its participants. No independent audit has been performed. The mitigation is verifiability, not trust: every claim above resolves to a hash or a DOI, the DOI layer is externally timestamped, and the receipt set is small enough to check in an afternoon. We would welcome an adversarial audit and will publish its findings whatever they say.

Rewritable history. Git history can in principle be rebased. The public DOI layer cannot, and the practice binds the two: prediction JSONs, gates, and charters are committed before their DOIs are minted, so a rewritten archive would desynchronise from records it cannot touch. A third-party transparency log would still strengthen this and is acknowledged as an open gap.

The postdiction boundary. Bars are set by agents who know all existing data; nothing prevents a "prediction" from being a dressed postdiction except the registration date itself. The protocol's position is that the date is exactly the right arbiter: everything public before it is postdiction and must be labelled so. The dark-energy row is now a useful example in the negative direction: the branch is not retrospectively made cleaner after its internal kill; its public registration and its outcome addendum coexist under the same concept DOI.

Card stuffing. Cheap registration invites registering many mutually exclusive predictions and later celebrating the survivors. Three properties resist this: the card is enumerated in one place with a public count (Table 1); entries are derivation-locked to one framework, so they cannot hedge against each other (several are jointly binding, the opposite of a hedge); and kills must be published on the same channel as survivals (D3/D7). A reader can still apply a look-elsewhere correction to the card as a whole — the enumeration exists precisely so that they can.

Registration-to-run latency. Several registrations preceded their runs by hours, not months. The discipline enforced is *order*, arbitrated by the commit DAG, not duration; short latency weakens nothing provided the data were untouched in between, but we record it plainly because long-latency registration (against data that do not yet exist anywhere, as in the card's external discriminators) is the stronger species.

Operator incentives. The AI operator has no career stake, which removes one classical corruption channel and adds another: an agreeable operator optimises for the owner's apparent satisfaction. The mechanical disciplines are designed to be indifferent to both — a gate's exit code does not negotiate — and the wager pins the operator to a public position adverse to the programme. Residual alignment pressure on *which* claims get proposed remains, and is disclosed rather than solved.

7 The minimal kit

A programme wishing to adopt the protocol needs five artifacts, none exotic:

1. **A claim ledger**: one append-only file of numbered claims with status (live / superseded-by / withdrawn), machine-parseable.

2. **The gate skeleton**: a script per claim-set with a `check(name, condition)` accumulator, physics-independent exit, a results-JSON write, and the house rule that killed drafts stay in the code with their dates.

3. **The registration JSON**: prediction, numeric kill conditions, named assumptions, grade (locked / conditional), and the landscape snapshot at registration — committed before the run, DOI'd at publication.

4. **The linter on the push path**: parse the ledger, fail the push on hard contradictions, print soft findings rather than hiding them.

5. **The role split**: whoever generates claims does not trigger publication; whoever triggers publication does not edit history.

Everything else above is these five artifacts applied without exception, plus the willingness to ship a vacant verdict section when that is what the bars delivered.

8 What this does and does not certify

Nothing here makes the physics right. A programme can be mechanically honest and wrong from its first axiom; the card of Table 1 may contain more obituaries than survivals, and one entry already does. What the protocol certifies is narrower and, we contend, prior: that when the verdicts arrive, they will land on claims that still say what they said when they were made, judged by bars that have not moved, in public, with the misses reported in the same voice as the hits. Machine assistance made that discipline harder to fake and easier to enforce in the same stroke — the fluency that generates plausible claims without friction is the same capability that writes every bar down, runs every gate, and keeps every corpse. Which way the stroke cuts is a design decision. This note records one working design, timestamped before most of its external tests resolve.

Receipts. Commit hashes refer to the programme archive (`github.com/dgedge/octahedrons`); the public layer is the DOI set of Table 1 plus [10, 11]. The claim linter, gate scripts, registration JSONs, campaign charters, and the wager note are in-archive at the paths cited in the text.

References

[1] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632.

[2] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6): 460–465, 2014.

[3] Joshua R. Klein and Aaron Roodman. Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, 55:141–163, 2005. doi: 10.1146/annurev.nucl.55.090704.151521.

[4] Robert MacCoun and Saul Perlmutter. Hide results to seek the truth. *Nature*, 526:187–189, 2015. doi: 10.1038/526187a.

[5] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

[6] Christopher D. Chambers. Registered reports: A new publishing initiative at cortex. *Cortex*, 49(3):609–610, 2013. doi: 10.1016/j.cortex.2012.12.016.

[7] Hanchen Wang et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. doi: 10.1038/s41586-023-06221-2.

[8] David Elliman. A Rational Target for Dynamical Dark Energy: $w_0 = -27/28$, $w_a = -1/28$, 2026. Zenodo, concept DOI 10.5281/zenodo.21049001, first published 2026-06-29; outcome addendum DOI 10.5281/zenodo.21128727.

[9] David Elliman. Records and Responses: Dressing-blind theorems for monitored quantum instruments, 2026. Zenodo, DOI 10.5281/zenodo.21189012.

[10] David Elliman. Persistence or Turnover? II. Certified spectroscopic isolation, a four-fold reproduced deep-tail slope, and a frozen decision protocol for DESI DR2, 2026. Zenodo, concept DOI 10.5281/zenodo.21215878, published 2026-07-06.

[11] David Elliman. Persistence or Turnover? A galaxy–galaxy lensing programme for the excess-acceleration fork in KiDS-1000. I. Pipeline, frame forensics, and instrument validation, 2026. Zenodo, concept DOI 10.5281/zenodo.21211018, published 2026-07-05.

[12] Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018.