

What Does a Monitoring Signal Represent?

Definition-indexed discrimination of correctness monitoring, action value, process representation, and redescription

BEN AMUWO | Independent computational research

ben.amuwo@gmail.com | Vancouver, Canada | sidus.quest

Version of record | doi:10.5281/zenodo.23252262

RESEARCH RECORD E0-B completed under its frozen stopping criterion, 8 October 2026. Exact constructed witnesses, a ten-seed developmental pilot, and independent-seed learned confirmation. The findings concern a finite synthetic system and one registered learned architecture. **No claim of strict metarepresentation, consciousness, or biological generality is made.**

Abstract

A monitoring message can improve decisions without representing the property of another representation that makes those decisions succeed. It may instead carry a regime label, a process identifier, a cached action value, or a feature that makes the same downstream decision possible. I developed E0-B as a definition-indexed test of this ambiguity, following an earlier validation of hidden information, transmission, and causal use in E0-A [1]. Three candidate criteria—token-correctness monitoring motivated by Shea’s account [2], consumer-relative representational redescription, and process representation—were evaluated in one enumerable synthetic environment, with explicit access contracts and adversarial witnesses. Exact analysis distinguished a correctness code from codes derived solely from a bonus-inclusive action-value vector, but also established an unavoidable equivalence between correctness probability and the value of an action rewarded directly for correctness. The redescription criterion could not be identified: matched feature engineering reproduced its accessibility gain, while a procedure-coupling clause depended on which admissible consumer was frozen. Process discrimination also failed in the registered mapping family because held-out behavior was linearly determined by observed behavior. In a subsequent pre-frozen learned test (20 independent training seeds per system), correctness-supervised training met six preregistered operational clauses at the **seed-mean level**, whereas reward-only optimization of the same scalar architecture did not. A four-dimensional end-to-end system retained both correctness-relevant and payoff-context information. All three pilot-derived hypotheses were supported under the frozen rules; only 14/20 supervised runs individually met the full profile. E0-B is complete as a bounded experimental record, not as proof that the resulting signal possesses uniquely metarepresentational content.

Keywords: metarepresentation; correctness monitoring; action value; representational redescription; process representation; causal interventions; synthetic experimental validation; identifiability.

1. The question and the burden of proof

What evidence would distinguish a signal *about whether a first-order report is correct* from a signal that merely helps a consumer act correctly? They can be numerically and behaviorally indistinguishable. A relay can exploit a scalar predicting correctness, an inverse action value, or a regime-conditioned shortcut without ever requiring a separate representation of a representation. In a learned system, probe accuracy alone is not enough to settle which interpretation is warranted.

The distinction between available information and causally used information is established prior art. Contemporary research on confidence and metacognition likewise separates estimates of uncertainty from what those estimates explain about self-evaluation and control [6]. Probing research has shown that decodable properties need not be relevant to the task actually performed [7, 8, 11]; intervention-based methods instead test how altering internal information changes behavior [9, 10]. E0-A addressed the narrower measurement problem: in a known-ground-truth synthetic setting, could a protocol distinguish

an informative monitoring message from an informative message that the consumer never uses [1]? E0-B begins *after* that separation. It asks whether causal use and predictive success discriminate the **kind of content** carried by the message.

Different theoretical traditions make that question differently demanding. Shea argues, against first-order and payoff-based readings, that reward-prediction-error signals carry metarepresentational content [2]. Proust distinguishes procedural evaluations of cognitive performance from stronger attributive accounts of metacognition [3]. Karmiloff-Smith’s representational-redescription account concerns changes in accessibility and representational organization [4], and Cleeremans places learned self-redescription within a broader theory [5]. These are not interchangeable empirical definitions. A test that passes one should not automatically be counted as evidence for the others.

I therefore treated **E0-B[Shea]**, **E0-B[RR]**, and **E0-B[Process]** as separately adjudicated indices, sharing a fixture but retaining distinct criteria. The result was not presupposed to be a single yes-or-no judgment about metarepresentation. Three outcomes were possible for each index: the criterion could survive its rivals; it could fail; or the permitted observations could leave its target non-identified. The last is a scientific outcome, not a euphemism for insufficient power.

THE RESULT IN THREE LINES

Correctness (Shea-index): a trace-compliant correctness code passed the functional tests; a supervised learned code passed the frozen *mean-level* confirmation rules.

Redescription (RR): the accessibility gain was real, but matched feature engineering reproduced it and procedure coupling was not identified.

Process: the registered fixture failed to distinguish structure from a sufficiently informative per-cell tracker.

2. An exact environment with explicit observation contracts

2.1 Reports, hidden states, and process families

The U1 fixture has a latent binary label $X \sim \text{Bernoulli}(0.7)$ and two fair auxiliary variables Z_1, Z_2 . A regime $T = h_\theta(Z_1, Z_2)$ is generated by a declared lookup; the default procedure is XOR. The reporting mechanism first forms $S = X \oplus N$, with probability of an uncorrupted report $q_0 = 0.95$ when $T = 0$ and $q_1 = 0.5$ when $T = 1$. A mapping f then possibly flips the report:

$$R = S \oplus f(o), \quad o \in \{00, 01, 10, 11\}.$$

The registered mapping family consists of the six affine rules

$$f_{w,b}(o) = w \cdot o \oplus b, \quad w \in \{01, 10, 11\}, b \in \{0, 1\},$$

with a uniform prior. Each rule inverts exactly two of the four public inputs. This reliability-matched family deliberately removes the trivial shortcut of identifying mappings by their aggregate error rate. It later creates an important limitation for the process index (§4.2).

The correctness variable is $L = \mathbf{1}[R = X]$. With full knowledge of the current mapping, the oracle correctness target is

$$\varphi = \Pr(L = 1 \mid R, o, T, f).$$

Ordinary encoders are **not** given the mapping parameters (w, b) . They receive traces and may use the regime procedure parameters θ , the public variables, and the declared payoff context. A trace consists of a report sample at a public input; X, N , and T are unobserved in the trace. The experiment is designed so that whether an encoder can see a variable is part of the mathematical contract, not an implementation convenience.

2.2 Trace evidence and frozen information access

In the final learned setting, mappings remain fixed over blocks. Each block contains **160 dedicated inference observations**, allocated evenly over the four inputs in randomized order ($k = 40$ traces per

input), followed by **one independently scored trial**. The mapping estimate is frozen before scoring; scored trials do not update it. Random input allocation was registered as a secondary robustness condition but was not executed in confirmation; the primary design used balanced allocation. Every new block draws its own mapping and history. The **training seed**, rather than the large number of evaluation trials, is the independent unit for claims about training reliability.

An important development correction was that four marginal maximum-a-posteriori mapping bits need not describe any member of the six-rule family. At $k = 10$ their joint pattern falls outside the family with probability approximately 0.217. The study therefore distinguished marginal MAP bits (the hard-estimator reference), a globally consistent joint MAP rule (for identity-dependent process calculations), and the full posterior. For a correctness relay observing the message and report, their exact ideal reference values were:

Table 1: Exact ideal relay values by mapping-estimate contract (relay observes message and report).

Traces/input	Marginal MAP	Joint MAP	Full posterior
$k = 10$	0.716503	0.714921	0.734616
$k = 20$	0.750250	0.748248	0.762713
$k = 40$	0.789530	0.788818	0.794698
Mapping oracle	0.825000	0.825000	0.825000

The fact that marginal MAP slightly exceeds joint MAP for a *single-cell* decision is not a paradox: the marginal bit is optimized for the local decision, while joint MAP is optimized to select one globally consistent mapping. The full posterior retains information that either point estimate discards. These are distinct access contracts and must not be treated as interchangeable implementations of a generic “known mapping.”

2.3 Consumers and payoff nuisance variables

The key rival is **first-order action value**. The acceptance consumer chooses among accept, invert, and defer. In the final R2.4 fixture, its expected conditional action values are

$$Q_a = \varphi + \kappa_a, \quad Q_i = (1 - \varphi) + \kappa_i, \quad Q_d = 0.55 + \kappa_d,$$

where the three κ variables are independently uniform on $[0, 1]$. Both competing encoders and the acceptance consumer can observe all three bonuses. The **correctness-paid relay**, by contrast, observes only a message M and the public report R ; it does not see the bonuses. It earns one for choosing the correct accept/invert action and 0.55 for deferring. This deliberate asymmetry makes it possible to ask whether a code stays invariant to payoffs that are irrelevant to the relay’s task despite their availability upstream.

The relay and acceptance consumer have different functions; their scores must not be interchanged. The ideal mapping-oracle relay value is 0.825, while the R2.4 full-information *acceptance* optimum is 1.402658, exceeding one because bonuses can add to the reward. An earlier fixture, which hid a payoff variable from one competing encoder, generated a trivial apparent invariance; a subsequent fixture generated a fragile separation through quantizer tie-breaking. Those versions remain preserved as diagnostic records, not evidence for the final distinction.

3. Competing explanations and a stricter test

3.1 The strongest payoff-derived rival

The final comparison is not between a correctness code and an arbitrarily weak action-value scalar. The action-value rival can transmit the complete vector (Q_a, Q_i, Q_d) , which contains at least as much information for the relay as any **code constructed only from that vector**. The channel used in the reference calculations replaces a message, with probability ε , by an independent draw from the message’s own marginal distribution. If g is a deterministic or independently randomized code of the vector Q , applying this channel to $g(Q)$ is distributionally equivalent to corrupting Q first and applying g afterward.

Table 2: Exact R2.4 relay reference values, all evaluated with the relay observing M, R . The vector is an upper bound on Q-only representations for the registered marginal-replacement channel; it is not an upper bound on every encoder with access to κ .

Mapping access	Channel ε	φ code	Action-value vector	Difference
Oracle	0	.825000	.804052	+.020948
Oracle	.10	.809300	.781149	+.028151
Oracle	1/3	.772667	.729936	+.042731
$k = 40$, marginal MAP	0	.789530	.763131	+.026399
$k = 40$, marginal MAP	.10	.777093	.744938	+.032155
$k = 40$, marginal MAP	1/3	.748074	.710933	+.037141

The data-processing inequality then makes the *noisy complete vector* an upper bound on the relay value of every vector-derived garbling under this registered channel.

The positive gap is **functional and conditional**: it says a relay benefits from a statistic that tracks correctness where bonus-inclusive first-order action values diverge from correctness. It does not identify the code’s philosophical content uniquely. An encoder that receives the bonuses *as well as* action values can subtract them to recover φ ; in addition, a correctness-paid consumer’s own accept action value is simply φ . An invertible relabelling of the latter makes exactly the same predictions. Within this design family, no choice of bonuses can distinguish “represents correctness probability” from “represents the value of a correctness-contingent action” for every possible correctness-paid consumer. E0-B’s interpretation stops at that identity.

3.2 Matched interventions and graded content

The decisive designed intervention, D11, compares pairs with the same report R but different relay-optimal correctness actions. The experimenter changes κ_a and κ_i so that accept and invert action values coincide across the pair; in the corrected learned protocol, it also holds κ_d fixed across the pair. Thus *all three components* of the acceptance action-value vector match while the correctness-relevant distinction is intentionally varied. A correctness code should separate the pair; a code carrying only the matched action-value vector should not. This avoids relying on near-equality pairs sampled by chance, which were too rare to support an efficient test. Its causal interpretation is limited to the registered intervention and support.

An independent graded-content check varies the defer reward *after* encoding. The originally proposed values, 0.45 and 0.65, were vacuous: deferring was never optimal in the relevant fixture. A later 0.90 candidate also failed to distinguish the codes at $k = 10$. The registered repair uses 0.75 **and** 0.80, which separate the constructed correctness and action-only codes at both tested trace budgets. An action-only policy bit cached at defer value 0.55 then fails to retain the information needed to adapt, while an appropriately graded correctness code succeeds. The policy-bit witness W10 was run explicitly; a hypothetical rival was not merely asserted to fail.

3.3 The discriminator and witness registers

The Shea-index combines six types of evidence: calibration to the contract-relative correctness target (S1a), selective tracking and nuisance invariance (S1b), causal substitution and token locus, D11 matched-value discrimination, graded content under the changed defer payoff (S1e), and endogenous misrepresentation when an internal mapping estimate is wrong. Their value lies in **joint** constraints. A simple regime label can be causal without being token-correctness content; an action-value vector can survive some substitutions yet retain irrelevant bonuses; a policy bit can encode an action yet fail when the decision boundary changes.

An exact, trace-compliant witness W5t passed the registered Shea clauses at $k = 10$ and $k = 40$. At $k = 40$ its noiseless relay value was 0.789530, versus the action-vector-derived bound 0.763131. Under the designed matched pairs the witness separated 55,296/55,296 eligible constructed comparisons, whereas the trace-built action-value rivals separated none. These are **constructed-system checks**, not learned-system performance, and some contract-relative criteria are attained by construction. They establish that the proposed instrument has a positive case and relevant negative cases before testing how an encoder trains.

4. Why two definitions could not be identified

4.1 Representational redescription: accessibility is not provenance

The RR index tested whether an internal regime description increased accessibility relative to a declared consumer. A linear-argmax consumer operating on raw (R, Z_1, Z_2) achieved 0.806250 with the mapping known; supplying a suitable regime feature raised its exact optimum to 0.825000. That 0.018750 improvement is real. It is **not** evidence that the description was formed by an internal redescription process: a separately learned same-target feature and a fixed parity feature reproduced the entire increase. Shuffled parity yielded zero improvement.

The proposed procedure-coupling clause (RR1c) was meant to distinguish a θ -coupled description from a convenient fixed feature after the regime lookup was rewired. Yet the pre-rewiring linear-optimal policy did not specify a unique frozen consumer: some admissible optimal weight vectors retained the post-rewiring optimum, and others did not. The initial apparent distinction between witnesses depended on which feasible solution a linear-programming solver returned. Every admissible consumer failed for the fixed-parity negative witness; the purported positive cases did not have an invariant outcome. Choosing an existential or canonical frozen consumer after seeing this failure could have manufactured a positive verdict, but would have changed the criterion. The pre-established disposition was retained:

E0-B[RR]: inconclusive(non_identified). A bounded accessibility gain is demonstrated; redescription-specific discrimination is not.

This is narrower than saying representational redescription does not occur. It says the registered observables and consumer class do not identify it. It also preserves the distinction between **optimization gap** and exact accessibility: in development checks, training a raw-feature consumer by least squares reached 0.733 while supplying parity reached 0.825, an apparent gain of 0.092, nearly five times the exact consumer-class gain. Training difficulty can manufacture an apparent redescription effect if the reference optima are ignored.

4.2 Process representation: a symmetric family erased the structural contrast

The Process index originally asked a code to predict $f(11)$ from traces at the other three inputs and to generalize to held-out mapping identities. The registered decoder was not informed which mappings constituted the held-out population. At five traces per observed input, its exact identity-blind Bayes reference was 0.596178. If the decoder were instead told the held-out population, the reference would be 0.665140; reporting the first number as an unconditional optimum would have been incorrect. Controlled message substitution and relabelling-invariance checks were eventually executed and behaved as designed, but neither uniquely separated structural representation from any other content-bearing message.

The decisive weakness was discovered by the learned pilot. Every one of the six affine maps in the registered family inverts exactly two of four inputs; hence

$$f(00) + f(01) + f(10) + f(11) = 2, \quad \boxed{f(11) = 2 - f(00) - f(01) - f(10)}.$$

Because the three observed bits sum only to one or two under the family, this is a **linear decoding rule on all permitted observed patterns**. A per-cell tracker with a linear readout can implement it without an independently established model of the underlying process family. At $k = 5$, such a tracker scored 0.591577, near the structured reference 0.596178; the remaining difference came from discarding count information, not from the absence of structural content. The larger early gap had been an artifact of decoder fitting on training identities, which did not generalize to the held-out pair.

The pilot sharpened the objection. At $k_p = 20$, the identity-blind structured reference was about 0.7041 by Monte Carlo, the linear per-cell tracker scored 0.6634, and more flexible learned readers extrapolated below chance to the held-out identities. The fit was not a convincing test of representation *of the process*. An exploratory extension to eight maps restored a nonlinear parity relationship but changed aggregate reliability and introduced new identity cues. Rather than reinterpret that post hoc fixture as the confirmatory test, the recorded result was:

E0-B[Process]: inconclusive(non_identified) in U1. The six-map fixture does not isolate structural process representation from a sufficiently informative per-cell code.

No claim is made about the identifiability of process metarepresentation in richer environments.

5. Learned experiment: design and frozen decisions

5.1 Systems, training, and evaluation

After exact-scope validation, the study trained five registered systems from independent initializations. The central comparison held the scalar encoder architecture fixed and changed its objective. **C-sup** used correctness supervision as an engineering positive control; **C-end** trained through the acceptance consumer on reward only. **R-acc** predicted acceptance action value, and **R-vec** predicted the accept/invert action-value vector. The **B0-am** access-matched end-to-end comparator had a four-dimensional message, enabling an examination of whether additional capacity can retain correctness information and nuisance payoff context together. The constructed ideal posterior φ_{post} was evaluated through the same downstream measurement pipeline to estimate its measurement floor.

The registered neural implementation used small NumPy multilayer perceptrons with one 32-unit hidden layer, Adam optimization, 4,000 training steps, and batches of 1,024. Encoders operated on their declared trace-compliant observations; the end-to-end acceptance consumer could use the message, report, and payoff contexts. The relay used fitted decisions from the message and public report: binned policies for one- and two-dimensional messages and a fitted MLP with restarts for the four-dimensional case. A separate fitted regressor measured calibration excess relative to φ_{post} . These evaluation choices are consequential: raw coordinate distances in multi-dimensional codes do not identify what the consumer can decode, and even an ideal code incurs finite-bin measurement error.

The ten-seed developmental pilot (seeds 1–10) was precommitted and run before confirmation. It showed that reward-only learning did not develop the full correctness-monitoring profile and that the Process fixture was non-discriminating. The pilot also exposed weaknesses in proposed clause thresholds. Those defects and the subsequent repairs were retained in additive Ledgers D, D-II, E and F, and in Ledger G incorporated into the Closing Report, rather than overwritten. The **confirmation rules were committed before a pilot-seed recheck**, and numerical thresholds were filled mechanically from an ideal-code measurement-floor rule and then frozen. Confirmation used previously untouched seeds **1001–1020**. Each trained seed was evaluated on 100,000 fresh independent blocks for fitting consumers/probes and another 100,000 for testing. The sampled trial population estimates performance of a fixed trained system; it does not increase the number of independent training outcomes above twenty.

5.2 Measurements and decision algebra

The learned battery measured relay reward; calibration excess relative to the contract’s correctness posterior; a decoder-based R_{κ}^2 for recovering bonus variables from $(M, \varphi_{\text{post}}, R)$; agreement on D11 action-value-matched pairs; and regret when the relay defer payoff was changed to 0.75 or 0.80. Counterfactual message sensitivity to κ remained descriptive for multi-dimensional messages because that metric changes under invertible coordinate transforms. The decoder-based measure is also model- and estimator-dependent: residual predictability is evidence of retained context in the **registered probe class**, not a coordinate-free information-theoretic invariant.

Thresholds were fixed from the ideal-pipeline measurement floor plus prespecified tolerances, except for the pilot-derived relay-advantage margin. The most consequential thresholds were calibration excess ≤ 0.002341 , nuisance retention $R_{\kappa}^2 \leq 0.018406$, D11 agreement ≥ 0.816867 , and a relay advantage > 0.012990 over the stronger of the **two registered trained action-value rivals**. The ideal code itself obtained D11 agreement near 0.921, lower than C-sup’s 0.945, because binning near decision boundaries creates an evaluation artifact. The code was not declared to beat the ideal on content grounds.

For each registered measure, inference used seed-level means and one-sided t bounds with 19 degrees of freedom. The six-clause positive-control claim H1 required every passing-side bound to clear its threshold, following an intersection–union rule for conjunctive success; failure declarations used a family-wise correction. The two directional predictions H2 and H3 used joint one-sided comparisons with $\alpha/2$ bounds for their two components. Failure was never inferred merely from the absence of significance for a pass. Pilot-informed hypotheses and thresholds were labeled *pilot-informed*, and results were judged against the frozen decision rule rather than revised after seeing confirmation.

Table 3: Registered learned hypotheses and scope. “Confirmed” means the stipulated inequalities for seed-level means were supported under the frozen test, not that every training run passed.

Hyp.	Prospective pilot-derived prediction	Frozen outcome
H1	Correctness-supervised C-sup meets all six learned functional Shea clauses.	Confirmed
H2	Reward-only scalar C-end fails the profile in a specified action-value-like direction: low D11 agreement and high payoff sensitivity.	Confirmed
H3	Four-dimensional B0-am preserves demonstrable correctness-relevant information while also retaining payoff-context information.	Confirmed

6. Confirmation results

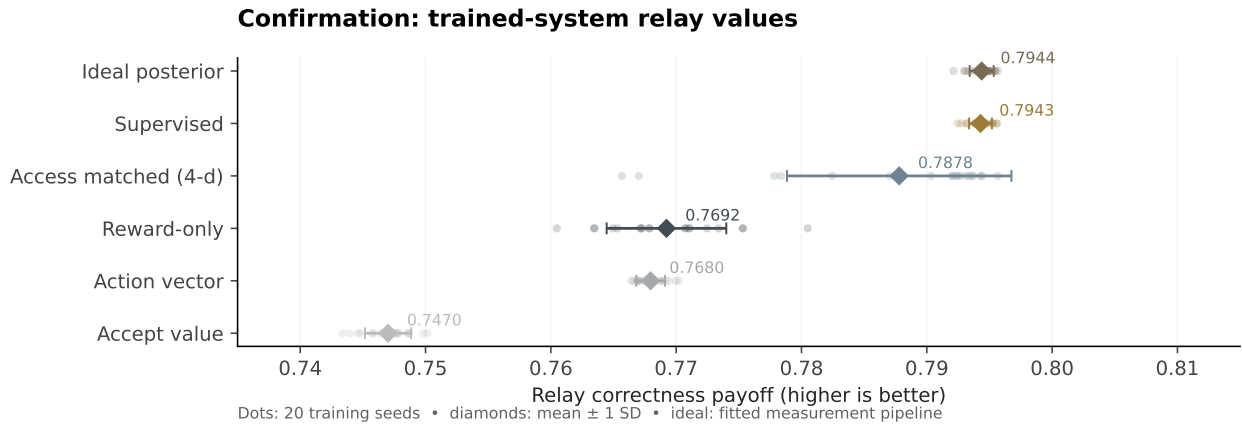


Figure 1: Out-of-sample relay values by system across the 20 independent confirmation seeds. Diamonds are means, bars denote seed standard deviations, and small circles denote individual seeds. The ideal posterior is evaluated through the same fitted downstream measurement pipeline. A higher relay value does not by itself establish the content of a representation.

The central performance differences are visible in the relay. C-sup reached mean reward **0.7943** (seed SD 0.0009), practically equal to the fitted ideal posterior’s mean **0.7944**. C-end reached **0.7692** (SD 0.0048); R-vec reached **0.7680** and R-acc **0.7470**. The four-dimensional B0-am comparator reached **0.7878** (SD 0.0090), sometimes approaching the supervised control but with much more variation across initializations. These comparisons are descriptions of the registered learned systems, not universal capacity bounds.

Table 4: Confirmation means across 20 training seeds. R_{κ}^2 measures decoder-recoverable nuisance payoff context conditional on the target and report; D11 measures agreement on designed vector-matched pairs.

System	Relay	Calib. excess	R_{κ}^2	D11 agreement	S1e regret (.75/.80)
Ideal φ_{post}	.7944	.00033	-.0028	.921	.00048/.00018
C-sup	.7943	.00044	.012	.945	.00056/.00067
C-end	.7692	.0174	.327	.050	.0222/.0247
B0-am	.7878	.0070	.791	.716	.0092/.0088
R-vec	.7680	.0201	.660	.013	.0273/.0325
R-acc	.7470	.0315	.329	.007	.0389/.0413

6.1 H1: a positive control, but not a uniform property of training

For C-sup, every registered seed-mean bound met its criterion. Calibration excess was 0.00044 (upper bound 0.00046 against 0.00234); the nuisance-retention mean was 0.0122 (upper bound 0.0166 against 0.0184); S1e regrets were 0.00056/0.00067 (upper bounds 0.00062/0.00078); D11 agreement was 0.945 (lower bound 0.939 against 0.817). The mean relay advantage over the stronger registered learned action-value rival was 0.0263, with lower bound 0.0258 exceeding the 0.0130 margin. **H1 was**

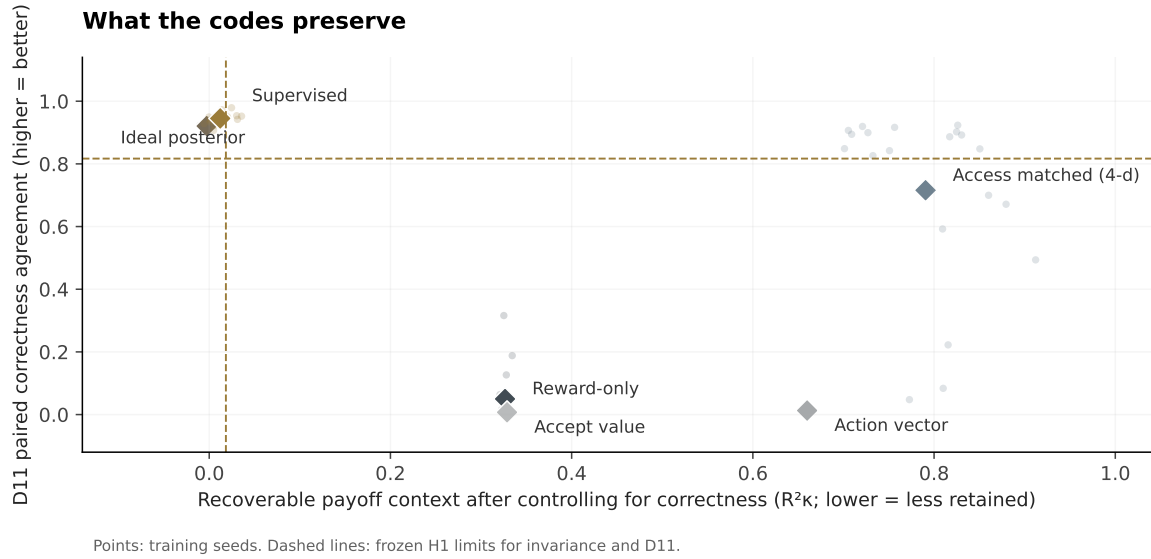


Figure 2: Confirmation systems projected onto two non-equivalent aspects of the proposed profile: action-value-matched correctness discrimination (D11, vertical) and retained payoff-context information after accounting for correctness (horizontal). The dashed guides show H1’s frozen requirements. Mean symbols and seed scatter reveal the variability that a single aggregate score conceals.

confirmed under the frozen seed-mean rule.

The qualification is substantive: only **14 of 20** C-sup training runs individually met all six conditions. The binding limitation was nuisance retention, which reached per-seed R_{κ}^2 as high as 0.036. The 0.0166 upper mean bound was close to its 0.0184 threshold; the corresponding ten-seed pilot recheck had been inconclusive. The confirmation supports average success under the registered training procedure and evaluation distribution, **not** the claim that correctness supervision reliably produces an invariant code on every run. For 14/20 observed successes, the one-sided 95% binomial lower bound on per-seed reliability is only about 0.49.

6.2 H2: reward-only optimization followed the payoff context

For C-end, mean D11 agreement was 0.050, with one-sided upper bound $0.087 \leq 0.20$; counterfactual κ -sensitivity averaged 0.548, with lower bound $0.537 \geq 0.25$. Both directional tests met the frozen joint-confirmation rule. Calibration excess was 0.0174, much larger than the supervised code’s, and S1e regrets were 0.0222 and 0.0247. The learned code’s relay payoff (0.7692) was near the action-vector rival’s (0.7680), although equal scores do not prove identical internal content. **H2 was confirmed for this reward function, architecture, and optimization procedure.** Nineteen of twenty seeds met both directional criteria individually.

The result does *not* mean reward optimization cannot ever form a correctness representation. The acceptance objective explicitly rewards bonus-shifted actions; a system optimized for those actions has no requirement to discard the nuisance bonuses. The experiment tests the consequences of that incentive arrangement rather than a universal learning law. It remains possible that different rewards, bottlenecks, tasks, or inductive biases would generate another code.

6.3 H3: more capacity retained more than one kind of information

B0-am, the access-matched four-dimensional end-to-end system, achieved mean D11 agreement 0.716 (one-sided lower bound $0.583 \geq 0.30$) and nuisance-retention $R_{\kappa}^2 = 0.791$ (lower bound $0.761 \geq 0.098$). **H3 was confirmed:** the message carried information useful for correctness decisions and information about payoff contexts irrelevant to the relay. Yet D11 varied substantially across seeds (SD 0.285); seventeen of twenty met both directional criteria individually. The four-dimensional result illustrates a residual predicted by the exact analysis: additional message capacity can retain multiple explanatory variables simultaneously, preventing a clean single-content attribution merely from a capable

downstream decoder.

7. Interpretation, uncertainty, and limits

7.1 What the positive result licenses

The validated Shea-index operationalization detects a specific *functional organization*: a message supports correctness-paid decisions, remains comparatively invariant to independent changes in payoff context available to its encoder, responds to correctness-relevant token changes, and retains graded information needed when the downstream decision boundary changes. The matched-vector intervention rules out the entire family of codes that depend *only* on the bonus-inclusive action-value vector under the registered access and channel contract. Constructed and learned results agree about the existence of this discriminable function, though the learned result depends on training objective and initialization.

What it does **not** license is a unique representational ontology. φ is an action value for a consumer that receives one unit for correctness. Invertible transformations can preserve all its uses. Content attribution requires an argument about consumers, counterfactual use, function, and alternative explanations, not just an information-theoretic advantage [2, 3]. The paper consequently reports a **functional correctness-monitoring profile**, rather than declaring strict or natural-kind metarepresentation.

7.2 What the two negative-identification results teach

RR and Process did not fail because the calculations were noisy. They failed because the fixture permitted a *simpler explanation that survived the same tests*. RR's improved accessibility was reproduced by features generated without the intended redescription history. Its procedure-coupling outcome changed with an otherwise admissible frozen policy. Process's strong symmetry guarantee inadvertently made the held-out mapping response a linear function of observed responses. These are examples of how a clean, exactly enumerable experiment can nevertheless be non-identifying. Exact numerical ground truth and theoretical construct validity are different achievements.

This also explains why the project's stopping rule did not require every index to pass. Continued redesign until all definitions yielded positive results would have changed the object being measured. The final report retains negative findings, preserves the failed clauses, and states which questions the revised and discarded fixtures can no longer answer.

7.3 Constraints on statistical and external generalization

The twenty training seeds are the units used for inferential statements about optimizer variability. Within-seed evaluation on 100,000 fit and 100,000 test blocks improves precision of each trained system's estimated performance; it does not create extra independent trained systems. The t bounds assume the seed-level outcomes are suitably modeled by that procedure with $n = 20$. Pilot-derived rules were frozen before running reserved seeds, but the commitment dates were self-recorded, not externally timestamped. Archive hash matching establishes file identity and internal consistency; it cannot, by itself, prove the historical sequence.

The strongest H1 conclusion concerns **mean** performance and is sensitive to the nuisance-retention threshold. Some confirmation measures depend on fitted evaluation models: a binned relay, regressor, and nuisance decoder are instruments, not transparent readings of an internal code. Probe model capacity, bin choice, calibration floor, finite paired-intervention support, and the exact sampling distribution may change measured values. D11's designed pairs isolate a deliberate counterfactual but are not representative random pairs. The bonus distributions, marginal-replacement noise model, finite six-map family, XOR procedure, and simple supervised versus reward-trained MLPs are all laboratory choices. None is asserted to characterize brains or frontier AI systems.

The strongest action-value comparison is likewise scoped. The vector dominance bound applies to **functions of the payoff-inclusive vector alone** with independent encoding randomness under the specified replacement channel. A model given the payoff bonuses separately can invert the reward shift; a correctness-paid action's value remains an invertible function of correctness probability. The study does not defeat these broader rivals. The tested four-dimensional comparator further shows why separating channels' information content from the consumer's use of that information becomes more difficult as capacity grows.

7.4 Reproducibility, provenance, and scope of independent verification

The canonical internal record is the **E0-B MB0-MB2 archive v4**, containing source scripts, executable contract tables, development ledgers D-G, pilot protocol, frozen confirmation specification, training/evaluation code, seed-level result files, and the closing adjudication. Its SHA-256 manifest has **115 matching entries**. The project recorded the pilot commitment, subsequent confirmation-rule commitment, freeze, and untouched-seed confirmation in that order; the frozen confirmation used seeds 1001–1020. The source environment records Python and relevant numerical-library versions. On 8 October 2026, ChatGPT (the AI research assistant also involved in the development and prior reviews) conducted a separate post-confirmation check of archive integrity and recalculated H1–H3 adjudication bounds from the archived seed-level records; the check is recorded in `E0B_Independent_Final_Validation.md`, published with this record as a supplementary file outside archive v4. This was a computational cross-check, **not an independent external-researcher audit. The complete training run has not been independently re-executed by a second researcher.** The paper, the unchanged archive v4, the supplementary validation record and a publication addendum are deposited at Zenodo (doi:10.5281/zenodo.23252262) and mirrored at sidus.quest/research/e0-b.

The distinction between reproducible code and independently reproduced execution is maintained deliberately. To reproduce the core confirmation from the archive, use the exact recorded environment and execute `python3 e0b_confirm.py confirm` with the frozen JSON specification, then compare both the seed-level result JSON and adjudication JSON to the archived outputs. A hash match of the executable source is a necessary but not sufficient substitute for that execution-level check.

8. Where the result goes next

E0-B's outcome is bounded but useful. E0-A established a measurement instrument for *availability versus causal use* under known ground truth [1]. E0-B established that a second question must be asked before naming the represented content: which alternative codes could explain the consumer's success just as well? The positive functional finding demonstrates that a correctness-specific statistic can be operationally distinguished from a family of unrelated action values. The two non-identification results show that greater accessibility or structured predictions alone do not isolate a higher-order mechanism.

E1 therefore starts with a stricter discipline rather than a larger metaphysical premise. Several useful monitoring channels will compete for an explicit encoding or routing budget. A message may be available but not routed, routed but unused, or used for a payoff-relevant purpose other than the property an investigator attributes to it. Scarcity makes those alternatives consequential. The immediate experiments should distinguish scarcity that destroys encoded information from scarcity that prevents available information from controlling action; should include matched redundancy and value controls; and should preserve a record of optimization failures and surviving rival explanations. A successful E1 would identify reproducible **relations among information, marginal behavioral value, and causal control**—not establish consciousness or biological homologues by analogy.

9. Conclusion

The central observation is simple: a message can make a downstream decision better and still leave its content underdetermined. I tested three attempts to resolve that underdetermination in a small system where the ground-truth probabilities, consumers, access rights, and counterfactual interventions could be specified exactly. Only the correctness-monitoring index yielded a discriminable functional profile under the registered rivals. In prospective learned confirmation, explicit correctness supervision met that profile on average; reward-only acceptance training did not; a higher-capacity system retained correctness-relevant and payoff-context information together. Representational redescription and process representation were not identified under the registered fixture. These are three separate results, not a majority vote on whether the system “has metarepresentation.”

The experiment is **complete under its frozen stopping rule** because every index has a documented outcome, every relevant discrepancy is preserved, and no required confirmatory run remains. Completion is not a declaration that the theory succeeded in every form. It is a declaration that the test ended where its evidence ended.

Established A trained system can satisfy a specified functional correctness-monitoring profile; the registered reward-only scalar generally did not under the tested procedure.

Not established Unique metarepresentational content; identified RR or Process metarepresentation in U1; architecture-wide learning law; biological, ecological, or consciousness claims.

Next Test resource-bounded allocation among multiple useful monitors while preserving the availability/use/content distinctions.

References

- [1] Amuwo, B. (2026). *Sīdus EO-A* [archived experimental results and associated research brief]. Zenodo. doi:10.5281/zenodo.22661563.
- [2] Shea, N. (2014). Reward prediction error signals are meta-representational. *Noûs*, 48(2), 314–341. doi:10.1111/j.1468-0068.2012.00863.x.
- [3] Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press. doi:10.1093/acprof:oso/9780199602162.001.0001.
- [4] Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press. ISBN 9780262111690.
- [5] Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology*, 2, 86. doi:10.3389/fpsyg.2011.00086.
- [6] Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75, 241–268. doi:10.1146/annurev-psych-022423-032425.
- [7] Ravichander, A., Belinkov, Y., & Hovy, E. (2021). Probing the probing paradigm: Does probing accuracy entail task relevance? *Proceedings of EACL*, 3363–3377. doi:10.18653/v1/2021.eacl-main.295.
- [8] Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *Proceedings of EMNLP-IJCNLP*, 2733–2743. ACL Anthology D19-1275.
- [9] Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 160–175. doi:10.1162/tacl_a_00359.
- [10] Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34. NeurIPS proceedings.
- [11] Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219. doi:10.1162/coli_a_00422.

Appendix A. Frozen thresholds and adjudication margins

Table 5: Frozen numerical requirements for learned confirmation. The exact decimal thresholds, commitment file, and seed-level results are retained in archive v4.

Measure	Threshold	Basis / interpretation
Calibration excess, S1a	$\leq .002341$	Ideal-pipeline 95th percentile + .002
Payoff retention R_{κ}^2 , S1b	$\leq .018406$	Ideal-pipeline 95th percentile + .02
S1e regret, $d' = .75$	$\leq .002729$	Ideal-pipeline 95th percentile + .002
S1e regret, $d' = .80$	$\leq .002384$	Ideal-pipeline 95th percentile + .002
D11 matched agreement	$\geq .816867$	Ideal-pipeline 5th percentile - .10
Relay advantage over learned R-vec/R-acc	$> .012990$	Pilot-derived C-sup half-P10 margin
H2: C-end D11	$\leq .20$	Pilot-informed directional test
H2: C-end κ -sensitivity	$\geq .25$	Pilot-informed directional test
H3: B0-am D11	$\geq .30$	Pilot-informed directional test
H3: B0-am retained R_{κ}^2	$\geq .098406$	Ideal-pipeline 95th percentile + .10

H1 adjudication at $n = 20$ (one-sided bounds): calibration upper .00046; R_{κ}^2 upper .0166; S1e upper .00062/.00078; D11 lower .939; relay advantage lower .0258. All six registered comparisons pass. H2: D11 upper .087; scalar sensitivity lower .537. H3: D11 lower .583; retained R_{κ}^2 lower .761. These bounds concern seed-level means.

Appendix B. Development failures preserved rather than repaired away

The final protocol was not the first proposal to produce a numerical gap. Early iterations withheld a payoff variable from one encoder; another assumed that the relay saw only a message, when its actual contract included the public report. Once access was matched, an apparent gap vanished. Binary payoff-context grids then allowed the correctness target to be reconstructed from exact action-value outputs at unlimited capacity. Restricted quantizers appeared to recover a difference, but hundreds of equally optimal quantizer partitions produced different downstream values; a particular noise condition even permitted opposite signs depending on which admissible tied quantizers were selected. Continuous payoff contexts and the final vector-valued rival closed the restricted-comparator loophole **under the registered channel**, while leaving the correctness-value identity intact.

The experiment also corrected a mathematical description of randomized reporting: a family-wise false-fail rate of $1 - (1 - .05)^6 \approx .265$ holds **under independence**, not as a dependence-free worst case; the union bound is .30. The final decision rules reflect the difference between protecting a conjunctive pass and protecting a selected fail. These corrections matter because apparent construct validity can be produced by access mismatch, implementation tie-breaking, arbitrary decoder selection, or inappropriate multiple-testing logic as readily as by a successful scientific hypothesis.

The full sequence is retained in the original v2.2 proposal, Contract Table, Ledgers D, D-II, E and F, Ledger G within the Closing Report, and the frozen rules. Earlier numerical outcomes were not deleted when corrected. This appendix is a summary of the development path, not a retrospective claim that every stage was confirmatory.

Appendix C. Data and reproducibility statement

Study: fully synthetic; no human or animal subjects. **Confirmation:** 20 seeds (1001–1020), five learned systems and one ideal reference; 160 trace reports and one scored trial per independent block; 100,000 fit and 100,000 test blocks per seed/system. **Development:** exact witness calculations and pilot seeds 1–10. **Archive:** E0B_MB0-MB2_archive_v4.zip (115 matching SHA-256 entries), deposited unchanged at Zenodo (doi:10.5281/zenodo.23252262) and mirrored at sidus.quest/research/e0-b. **Provenance:** self-recorded timestamps and a post-confirmation ChatGPT arithmetic/file-integrity check on 8 October 2026 (E0B_Independent_Final_Validation.md, published as a supplementary file outside archive v4); no independent external-researcher audit or full training rerun. **Scope:** complete under the frozen E0-B stopping criterion, limited to the registered fixture, consumers, interventions, model family, and measurement pipeline.