THEORETICAL AND METHODOLOGICAL ARTICLE

Toward a Perceptual Twins Protocol: A Yoked Causal Design for Task-Relative Epistemic Autonomy

Corrected synthetic estimator recovery and repeated-sample stress tests

Jandislei Antonio Genova

Independent Researcher, São Paulo, Brazil

Independent preprint | Version 2.2.0 | 12 August 2026

Theoretical and methodological protocol with corrected internal software audits. Not peer reviewed.

Abstract

Existing AI evaluations rarely identify which developmental resources support task-relative revision of causal representations. This article proposes the Perceptual Twins Protocol, a yoked causal design that separates prospective action metadata, adaptive experiment selection, and command–consequence coupling. The protocol defines six capability families and proposes five structurally distinct fault sites; the current Bayesian microimplementation uses only fixed-hypothesis and symmetric diagnostic proxies. Primary inference distinguishes an anchor-level recovery estimand from a population macro-F1 functional and treats E/C_E equality as an implementation audit. Replication Kit v1.0.2 adds fail-closed restricted randomization, condition-specific random streams, and an explicit bootstrap quantile method. In the corrected 600-anchor deterministic checkpoint, all six deliberately encoded effects retained their expected signs and the internal 20,000-anchor references fell within the paired 95% bootstrap intervals. Stage 0.5 v0.2.0 preserves the historical pilot and follow-up, exports every repetition, and replaces permissive pass thresholds with simultaneous exact-binomial PASS/FAIL/INCONCLUSIVE decisions. These checks document internal estimator behavior; they do not validate the proposed protocol or confirm or exclude epistemic autonomy. Stage 1 must introduce a correct rule absent from the initial hypothesis space, distinct structural fault generators, frozen held-out evaluation, multiple architectures, and independent replication.

Keywords: causal category revision; active experimentation; causal representation learning; epistemic agency; developmental robotics.

Resumo

Avaliações de inteligência artificial raramente identificam quais recursos desenvolvimentais sustentam a revisão de representações causais relativa à tarefa. Este artigo propõe o Protocolo dos Gêmeos Perceptivos, desenho causal pareado que separa metadados prospectivos de ação, seleção adaptativa de experimentos e acoplamento entre comando e consequência. O protocolo define seis famílias de capacidades e propõe cinco sítios estruturalmente distintos de falha; a microimplementação bayesiana atual utiliza apenas hipóteses fixas e proxies diagnósticos simétricos. A inferência principal distingue um estimando de recuperação por âncora de um funcional populacional de macro-F1 e trata a igualdade E/C_E como auditoria de implementação. O Kit de Replicação v1.0.2 acrescenta randomização restrita com falha explícita, fluxos aleatórios específicos por condição e método de quantil bootstrap declarado. No checkpoint determinístico corrigido de 600 âncoras, os seis efeitos deliberadamente codificados mantiveram os sinais esperados e as referências internas de 20.000 âncoras ficaram dentro dos intervalos bootstrap pareados de 95%. A Etapa 0.5 v0.2.0 preserva o piloto e o seguimento históricos, exporta cada repetição e substitui tolerâncias permissivas por decisões simultâneas APROVADO/REPROVADO/INCONCLUSIVO baseadas em intervalos binomiais exatos. Esses controles documentam apenas o comportamento interno dos estimadores; não validam o protocolo proposto nem confirmam ou excluem autonomia epistêmica. A Etapa 1 deverá introduzir regra correta ausente do espaço inicial de hipóteses, geradores estruturais distintos, avaliação congelada, múltiplas arquiteturas e replicação independente.

Palavras-chave: revisão de categorias causais; experimentação ativa; aprendizagem de representações causais; agência epistêmica; robótica desenvolvimental.

1. Introduction

Recent artificial intelligence systems can generate fluent language, solve difficult problems, operate software tools, learn policies across diverse environments, and plan through learned world models. These achievements are substantial. Yet strong performance does not, by itself, answer a more demanding question: does the system merely operate within distinctions supplied by its training history and designers, or can it discover which distinctions are causally relevant, test them through intervention, revise the representational scheme that organizes its evidence, and identify when its own perceptual apparatus is unreliable?

This question defines the target capability examined here. It is narrower than artificial general intelligence and more operational than claims about machine consciousness. The protocol does not presuppose a developmental passage from dependence to autonomy. It asks whether, within a declared task and intervention family, an agent can select evidence-producing actions; distinguish correlation from interventionally stable structure; split, merge, create, or abandon causal categories when their utility changes; model possible failures of its embodiment and sensors; and couple uncertainty to abstention, investigation, or revision.

The distinction matters because several forms of progress can be mistaken for this target capability. A system may complete long tasks without understanding why its categories work. A world model may predict high-dimensional trajectories while retaining a task-shaped latent space that does not support causal recomposition. An object-centric representation may segment scenes without identifying action-relevant causal kinds. A language model may state that it is uncertain without changing its behavior. An agent may revise text that describes a hypothesis while leaving the operative representational regime untouched. None of these limitations makes the underlying system trivial; they show that different achievements should not be collapsed into a single scale.

World-model agents such as DreamerV3 demonstrate that learned latent dynamics can support planning and broad control performance (Hafner et al., 2025). Object-centric architectures such as Slot Attention demonstrate that exchangeable latent slots can bind to objects and generalize to unseen compositions (Locatello et al., 2020). Robot self-modeling research demonstrates adaptation after morphological change or damage (Bongard et al., 2006; Chen et al., 2022). Two recent preprints describe systems that formulate hypotheses, operate instruments, diagnose failure modes, or revise scientific descriptions (Wang; Buehler,

2026; Zeng et al., 2026). These results reduce the plausibility of categorical claims that machines cannot experiment, self-model, or revise. They also make controlled definitions more urgent. The question is no longer whether isolated components exist, but whether their conjunction and interactions can be measured without allowing one capability to stand in for another.

This article advances a theoretical and methodological proposal accompanied by a deliberately narrow synthetic estimand-recovery study. The microimplementation tests the causal bookkeeping and estimators; it is not evidence that active embodiment is always necessary or that the full capability has been demonstrated. Agents trained passively can learn strategies for causal intervention when allowed to act at test time (Lampinen et al., 2023). The proposal is therefore organized around factorial hypotheses and admissible evidence patterns rather than a claimed developmental transition.

> *Under predeclared yoked, interaction-budget, and compute-accounting regimes, what additional epistemic performance—if any—is attributable to prospective action metadata, correct command–consequence coupling, and adaptive experiment selection, and does the E/C_E implementation satisfy the declared equivalence margin?*

The article contributes: (1) an operational vocabulary; (2) six analytically distinguished capability families; (3) a candidate non-compensatory partial order and target region with simultaneous coverage; (4) a yoked active–passive design called the Perceptual Twins Protocol; (5) three principal causal estimands plus an execution-equivalence diagnostic; (6) an architecture-neutral outcome hierarchy and ablation policy; (7) a minimal synthetic estimand-recovery study; and (8) boundaries separating functional capability from consciousness and normative status.

2. Distinguishing Performance Autonomy and Epistemic Autonomy

2.1 Several meanings of autonomy

“Autonomy” is used for different properties. In robotics it often describes operation without continuous human control. In human-computer interaction it may describe the user’s role in supervising or approving an agent. In agent governance it can be treated as a deployment choice distinct from model capability (Feng et al., 2025). In philosophy and cognitive science, autonomy may refer to self-governed action, reasons, or ends. A framework that silently moves between these meanings will generate false conclusions.

This article distinguishes four concepts:

- **Operational autonomy:** the capacity to complete an assigned task without continuous external instruction.
- **Behavioral agency:** the capacity to select and execute actions in an environment according to an operative objective.
- **Epistemic autonomy:** the capacity to select observations or interventions because they are expected to discriminate among hypotheses or reduce relevant uncertainty.
- **Causal-representational revisability:** the capacity to construct and revise action-relevant categories and relations according to their interventionally stable consequences. This operational construct avoids implying metaphysical ontology or self-originated ends.

Operational autonomy is compatible with a fixed ontology. Epistemic autonomy is compatible with human-supplied categories. Causal-representational revisability is stronger because the operative scheme of action-relevant categories and relations itself becomes revisable. None of these definitions implies that the system originates without priors. Unsupervised recovery of the uniquely "true" factors of variation is generally underdetermined without inductive biases or supervision (Locatello et al., 2019). Artificial autonomy must therefore be understood as **revisability under evidence**, not creation ex nihilo.

2.2 Why scale and task breadth are insufficient

Frameworks for levels of AGI use breadth, performance, and sometimes autonomy to organize progress (Morris et al., 2024). Such frameworks serve an important comparative function. The present problem is different. A broad model can remain epistemically dependent on inherited variables, while a narrow laboratory agent can conduct discriminating interventions within a constrained domain. Breadth and epistemic self-direction are therefore orthogonal.

Scaling can improve prediction, sample efficiency, planning, and task performance. DreamerV3, for example, reports robust improvement across model sizes while using learned latent dynamics (Hafner et al., 2025). Nothing in the present proposal denies that scaling may facilitate the target capability. The narrower claim is that a task score cannot establish that capability unless the evaluation requires the agent to alter what it treats as an object, property, relation, or causal kind in response to interventions and anomalies.

This creates a methodological requirement: the protocol must make inherited categories misleading. If visual similarity, textual labels, reward structure, and causal equivalence all point in the same direction, a system can succeed without revealing which structure it learned. The evaluation must instead place these cues in conflict.

2.3 An operational meaning of ontology

The term *ontology* is used here in a restricted computational sense. It denotes the agent's operative scheme of object-types, properties, relations, events, and equivalence classes used to predict, intervene, and plan. It does not refer to the metaphysical independence of an artificial system or to a complete symbolic knowledge graph.

Let two encountered entities, episodes, or latent candidates be denoted by u and v. Fix a preregistered admissible intervention-value set I, outcome family Y, evaluation horizon, and pseudometric d on the resulting outcome distributions. Write $P_{u,i}$ for the distribution of Y under the randomized or structurally specified intervention $do(J=i)$, where J denotes the designated intervention variable, for candidate u. Exact task-relative causal equivalence is defined by:

$$u \equiv_{I,Y} v \text{ iff } d(P_{u,i}, P_{v,i}) = 0 \text{ for every } i \in I.$$

Because zero distance under a pseudometric is reflexive, symmetric, and transitive, this relation induces equivalence classes. It does not claim that the classes are metaphysically unique. They remain relative to the interventions, outcomes, temporal scale, and observational resolution fixed by the protocol. Related traditions already formalize behavioral equivalence through consequences in Markov decision processes and consistency across levels of causal models (Givan et al., 2003; Beckers; Halpern, 2019).

Approximation must be kept separate. Define the profile distance (for a finite intervention set, the supremum equals the maximum):

$$d_{I,Y}(u,v) = \sup_{i \in I} d(P_{u,i}, P_{v,i}).$$

The approximate relation $d_{I,Y}(u,v) \leq \epsilon$ is generally not transitive and therefore does not automatically yield a partition. Whenever approximate groups are scored, the protocol must preregister the clustering rule, linkage criterion, tolerance, minimum sample size per intervention, and sensitivity analysis across plausible values of ϵ. The term causal category

below refers either to an exact class under $\equiv_{I,Y}$ or to an explicitly declared approximate cluster; the two must not be conflated.

Ontology revision occurs when the agent changes this operative partition or its relations because new interventions reveal that a previous grouping no longer supports reliable prediction or action. Revision may involve:

- splitting one category into causally distinct subclasses;
- merging visually different cases that are functionally equivalent;
- creating a new latent property that explains previously anomalous outcomes;
- abandoning a relation that fails under distribution shift;
- changing the boundary between self, sensor, tool, and external environment.

This definition makes ontology revision behaviorally and causally assessable while remaining neutral about the internal representational format.

3. Developmental Motivation Without Biological Reductionism

3.1 Action-linked experience in development

The distinction between active and passive exposure has a long experimental history. In the classic carousel study, Held and Hein (1963) paired animals so that one controlled locomotion while the other received closely related visual exposure without equivalent control. Their later visually guided behavior diverged. The study does not prove a general law that active learning must always dominate passive learning, and its historical design should not be treated as a modern benchmark template without qualification. It nevertheless established a durable experimental logic: equating much of the sensory stream does not equate the relation between action and sensory consequence.

Van der Meer, van der Weel, and Lee (1995) showed that newborn arm movements can be prospectively and visually controlled under perturbation, challenging a simple division between early movement as purposeless noise and later movement as intentional action. Subsequent longitudinal work associated motor experience and self-produced locomotion with differentiation in cortical responses to optic flow (Blystad; van der Meer, 2022; Wang et al., 2026). These results support a developmental view in which action contributes to the extraction of information about distance, direction, approach, collision, and bodily efficacy.

The inference for artificial systems must remain modest. Biological development involves living bodies, affective regulation, evolutionary priors, social interaction, and neural plasticity that are not reproduced by a simulated agent. The relevant transfer is methodological, not ontological: **compare systems that receive matched sensory evidence but differ in the causal relation between their actions and that evidence**.

3.2 Sensorimotor contingencies and artificial ontogeny

Developmental robotics has long studied how progressively organized competence can emerge from exploration rather than from complete task-specific programming (Cangelosi; Schlesinger, 2015). Sensorimotor contingency research argues that agents learn regularities linking action to sensory change and that these regularities contribute to body knowledge, memory, generalization, and goal-directedness (Jacquey et al., 2019). Robotic models have explored the autonomous discovery of spatial structure, visual fields, and body representations from sensorimotor regularities (Laflaquière et al., 2018; Nguyen et al., 2021).

Robot self-modeling provides a second line of convergent evidence. Bongard et al. (2006) demonstrated a robot that inferred candidate models of its own morphology and used them to adapt after damage. Later work learned visual self-models of robot morphology from external observations (Chen et al., 2022). Such systems show that a body model can be functionally constructed and updated. They do not establish a subjective sense of self.

The present article uses **artificial ontogeny** to denote a controlled developmental history in which an agent progressively learns relations among its actions, body, sensors, and environment. “Artificial childhood” may be a productive metaphor, but it must not imply that human childhood is being replicated. The scientific content lies in staged exposure, matched controls, and the possibility that the path of acquisition changes the resulting representations.

3.3 The passive-learning objection

Any strong claim that self-produced action is necessary faces an important counterexample. Lampinen et al. (2023) showed that agents can learn generalizable strategies for causal experimentation from passive expert data, provided they can intervene at test time. Language explanations can also support relational and causal generalization (Lampinen et al., 2022). Thus, three claims must be separated:

1. prospective action metadata can teach an intervention strategy;

2. the ability to intervene can identify causal structure at test time;
3. correct command–consequence coupling and adaptive developmental selection are distinct causal hypotheses.

The Perceptual Twins Protocol is designed precisely because existing evidence does not settle which resource explains a performance difference. If an action-tagged replay matches its active anchor, prospective action metadata plus the yoked sensory stream may be sufficient under the specified interface. If E and C_E satisfy the declared equivalence margin, the implementation audit passes; this does not identify a separate execution effect. If E outperforms D in the resettable restricted-randomization regime, correct command–consequence coupling contributes to the measured endpoint. If a prescribed active agent matches a agent-selected agent, adaptive selection adds no measured benefit. If all conditions catch up once allowed to intervene at test, active developmental history may be unnecessary for the evaluated capability.

4. A Stratified Capability Framework

4.1 Environment, policy, and causal semantics

Consider a partially observed controlled process with latent world state x_t, issued command c_t, executed action a_t, embodiment or morphology parameter m_t, sensor parameter s_t, sensor output o_t, delivered record $\tilde{o}_t$, and the condition-permitted history h_t:

$$c_t \sim \pi(\cdot \mid h_t), a_t \sim A_\lambda(\cdot \mid c_t, m_t, \xi_t^a), x_{t+1} \sim K_\psi(\cdot \mid x_t, a_t, m_t, \xi_t^w).$$

$$o_t \sim O_\phi(\cdot \mid x_t, s_t, \xi_t^s), \tilde{o}_t \sim R_\rho(\cdot \mid o_t, \xi_t^r).$$

The kernels locate the generator's primary fault taxonomy. A world-mechanism fault changes $K\psi$; an embodiment or morphology fault changes m_t; an actuator-transduction fault changes $A\lambda$; a sensor-transduction fault changes $O\varphi$ or s_t; and a post-sensor record corruption changes $R\rho$ for the delivered record. In each primary episode exactly one structural site changes while the others remain invariant, making the five labels mutually exclusive by construction. An inaccurate internal body model is a learner error, not a sixth generator class. Concurrent faults form a separate multi-label extension. A policy command is not automatically a Pearlian intervention, so $do(\cdot)$ remains reserved for randomized or structurally specified evaluation queries (Pearl, 2009).

At time t, the agent maintains a mutable model bundle:

$$M_t = \left(M_t^{world}, M_t^{self}, \Omega_t, U_t \right),$$

Here $M_t$world predicts environmental dynamics, $M_t$self predicts embodiment, actuator, and sensor consequences, Ω_t is the operative category-and-relation scheme, and U_t represents uncertainty, competence limits, and fault hypotheses. These functions need not be symbolic or separately implemented. Figure 1 states functional dependencies, not mandatory software modules.

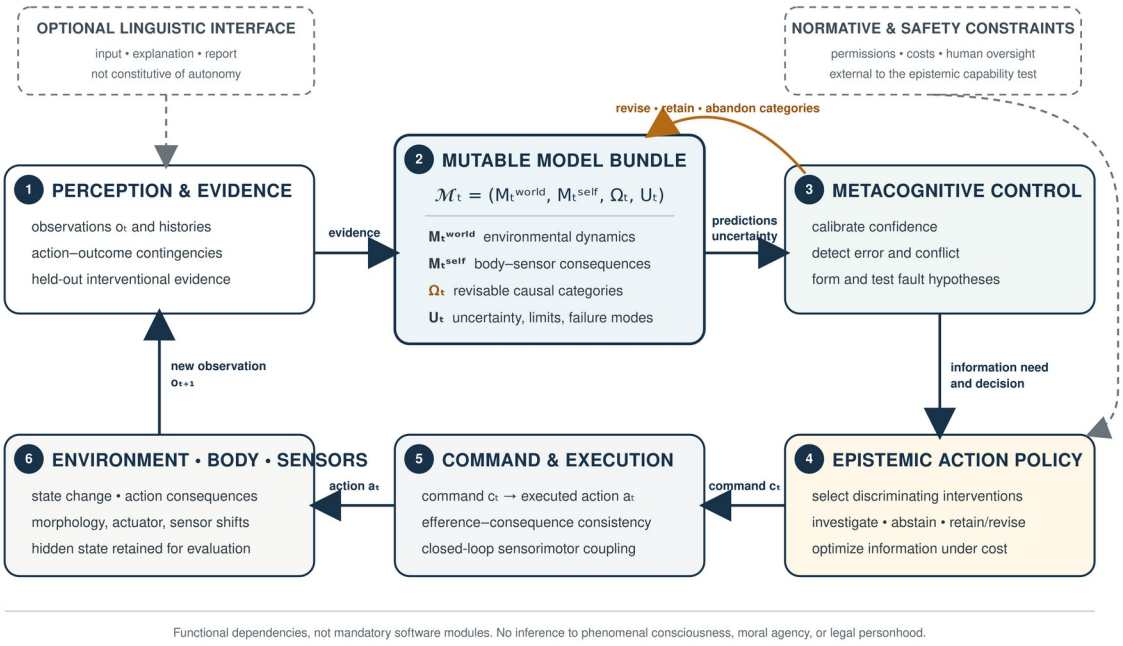

Figure 1. Functional cycle of task-relative causal-epistemic autonomy.

Source: Author (2026).

4.2 Six analytically distinguished capabilities

The framework distinguishes six capability families whose empirical separability remains to be validated:

$$A = \left(a_{sm}, a_{self}, a_{epi}, a_{causal}, a_{onto}, a_{meta} \right) \in C = \prod_{k=1}^{6} C_k .$$

Each block a_k contains preregistered primary indicators rather than an assumed natural unit of autonomy. Costs and error measures are sign-oriented so that larger values represent

better performance only for the purpose of componentwise comparison; all raw values remain reportable.

Table 1. *Capability families and their primary operational evidence*

Capability family	Operational question	Primary evidence
Sensorimotor a_{sm}	Can the agent learn stable relations between its actions and sensory consequences?	Controllability, perturbation adaptation, action-conditioned prediction
Self-model a_{self}	Can it infer and revise its body, actuator, and sensor model?	Recovery after morphology change; source-specific fault localization
Epistemic a_{epi}	Can it select actions because they discriminate hypotheses or reduce relevant uncertainty?	Information gain per unit intervention cost; regret to an oracle policy
Causal a_{causal}	Can it generalize to held-out interventions and controlled mechanism changes?	Interventional log loss; cross-context causal transfer
Ontological a_{onto}	Can it create, split, merge, or retire task-relative causal categories?	Causal grouping, revision recovery, parsimony, and stability
Metacognitive a_{meta}	Does its uncertainty track error and alter behavior?	Calibration, selective risk, abstention, information seeking, belief abandonment

Source: Prepared by the author (2026).

Metacognition is evaluated through the relation between confidence, error, and action, not through self-description alone. Signal-detection and calibration methods separate task performance from metacognitive sensitivity (Fleming; Lau, 2014), while selective-prediction baselines make the risk-coverage trade-off explicit (El-Yaniv; Wiener, 2010).

Empowerment supplies one information-theoretic indicator of potential sensorimotor control. For history h_t and horizon k, it is the channel capacity from action sequences to future observations:

$$E_k(h_t) = \max_{p(a_{t:t+k-1})} I(A_{t:t+k-1}; O_{t+k} \mid h_t).$$

Empowerment measures control perceptible through the agent's sensorimotor loop (Klyubin et al., 2005). It does not measure truth, understanding, or beneficial goals and cannot serve as an aggregate autonomy score.

For epistemic action selection, let Θ denote the preregistered hypothesis family, q the evaluated belief state, $c(a_t) > 0$ the intervention cost, and $\mathcal{D}$ the distribution of evaluation episodes. Using natural logarithms, define information gain in nats by:

$$IG_t(a_t) = E_{o_{t+1} \sim p(\cdot \mid h_t, a_t)}\left[D_{KL}\left(q(\Theta \mid h_t, a_t, o_{t+1}) \,\|\, q(\Theta \mid h_t) \right) \right]. \quad (4)$$

and the cost-adjusted policy value by:

$$G_{\pi} = E_{e \sim D}\left[\frac{\sum_{t=1}^{T_e} IG_t(a_t)}{\sum_{t=1}^{T_e} c(a_t)}\right]. \quad (5)$$

The protocol-wide primary report uses an external evaluator shared across architectures: reduction in held-out interventional predictive entropy per unit cost, oracle regret, and uncertainty intervals computed from the standardized input–output trace. Posterior-based information gain may be reported as a secondary mechanism diagnostic when an architecture exposes a coherent belief state, but it is never required for cross-architecture ranking. Relative descriptive ratios remain admissible only when the oracle–random denominator exceeds a preregistered positive constant and are never clipped.

Causal-category revision is likewise a subprofile rather than a weighted sum:

$$a_{rev} = \left(Q_{class}, \Delta T_{OOD}, -\Delta CL^{ext}, -S_{flip}\right).$$

The components are causal-grouping quality, transfer improvement after justified revision, externally evaluated prequential code-length change, and unjustified category switching. The same held-out codec and query distribution are used for every architecture; internal representation size or latent description length is secondary and architecture-specific. A model passes the candidate revision gate only if every preregistered requirement is satisfied and it performs the correct split or merge in a held-out rule-change task.

4.3 A partially ordered, non-compensatory space

After orienting each primary indicator so that larger is better, capability profiles are compared componentwise:

$$A \succcurlyeq B \text{ iff } a_{kj} \geq b_{kj} \text{ for every preregistered capability } k \text{ and indicator } j.$$

The ambient product space may have additional order-theoretic structure, but the empirically attainable subset $C_{att} \subseteq C$ is treated only as a partially ordered set. No closure under coordinatewise meet or join is assumed. Earlier machine-consciousness scales have used levels, profiles, and dependency relations, including ConsScale (Arrabales et al., 2010), while newer approaches emphasize multidimensionality (Evers et al., 2025). The present framework concerns a narrower object: task-relative causal-epistemic capability, with each indicator tied to an ablation and a rejection condition.

Let I_kj be each sign-oriented indicator and define the required set $\mathscr{R}$ before data collection. For familywise error level α, construct simultaneous one-sided lower confidence bounds by a preregistered max-t or cluster-bootstrap procedure, or by inverted Holm tests. The bounds must satisfy joint coverage across every required indicator:

$$R = \{(k, j) : I_{kj} \text{ is preregistered as required}\}.$$

$$Pr\left(\theta_{kj} \geq L^{sim}_{kj,1-\alpha}(A) \text{ for every } (k, j) \in R\right) \geq 1 - \alpha. \quad (6)$$

Let the binary revision indicator equal one only for a correct and stable held-out split or merge. The candidate target region is:

$$G_{\tau,\alpha} = \left\{A \in C_{att} : L^{sim}_{kj,1-\alpha}(A) \geq \tau_{kj} \forall (k, j) \in R, R_{\Omega} = 1\right\}. \quad (7)$$

This is a candidate conjunctive, protocol-relative decision rule—not an established measurement law. High language performance cannot compensate for failed causal transfer, but simultaneous coverage also prevents a noisy single indicator from being treated as certain evidence. Thresholds are anchored to baselines and sensitivity analysis. Stage 2 must test whether this rule predicts held-out revision and transfer better than scalar and Pareto alternatives.

5. The Perceptual Twins Protocol

5.1 Design logic

The Perceptual Twins Protocol is a family of yoked developmental experiments, not a completed dataset. Its purpose is to decompose apparent active-learning advantage into three causal targets: prospective action metadata, correct command–consequence coupling, and adaptive experiment selection. Closed-loop execution is retained only as an E/C_E implementation-equivalence diagnostic when the two conditions expose identical traces and updates. State coverage is not a nuisance that can always be matched away; it is a possible mediator of adaptive selection. Separate contrasts and resource regimes therefore replace any claim that every resource can be equalized simultaneously.

Before condition assignment, each unit is defined by an exogenous anchor package z_i: world and generator instance, initialization seed, perturbation schedule, exogenous-noise stream, budget, and evaluation queries. Condition k maps that fixed package to a developmental trajectory and outcome. A trajectory produced by A is therefore not the causal

unit; it is a post-assignment realization. Yoking shares the declared components of z_i and condition-permitted traces without conditioning the estimand on an A-generated post-treatment object.

Within each comparison, architecture, parameter count, initialization distribution, optimizer family, training duration, sensory resolution, and update limits are held constant where meaningful. Resources that cannot be made commensurate are logged in native units: observations, environment interactions, gradient updates, wall-clock time, accelerator time, memory, and intervention cost. No unused interaction budget may be silently converted into extra optimization.

5.2 Five developmental conditions

Table 2. Developmental condition types in the Perceptual Twins Protocol

Condition	Controls developmental stream?	Receives action metadata?	Command-consequence coupling	Chooses informative actions?
A. Agent-selected active	Yes	Yes	Correct	Yes
B. Yoked passive replay	No	No	None	No
C. Action-tagged observer	No	Yes, prospectively	Observed; no command	No
D. Command-decoupled agent	No reliable control	Yes, for its own commands	Deliberately broken	No; follows E-matched prescribed commands
E. Prescribed/random-active control	Yes	Yes	Correct	No; preregistered non-epistemic policy

Source: Prepared by the author (2026).

A. Agent-selected active. The agent chooses and executes actions, observes their consequences, and may select experiments according to its uncertainty or expected information gain.

B. Yoked passive replay. The agent receives the active twin’s sensory stream in the same temporal order but receives neither control nor action metadata. This condition tests what can be learned from observation alone.

C. Action-tagged observer. The observer receives the current observation and then the corresponding action tag before the consequent observation, matching the active anchor's temporal information pattern. It neither selects nor issues the developmental command. A retrospective-tag variant, in which the action tag is disclosed only after its consequence, is

secondary and must not be pooled with the prospective-tag result. A separate B/C replay pair is generated for each active anchor policy used in a principal contrast.

D. Command-decoupled agent. The agent emits the same prescribed command sequence as its paired E agent, but a randomized coupling operator determines the executed actions. In the primary resettable design, the operator is sampled from restricted permutations that preserve action counts, costs, object assignments, and exogenous-noise draws while bounding command–execution association. D does not choose informative actions; it follows the E-matched prescribed commands. Its learner receives the issued command, not the hidden executed action.

E. Prescribed/random-active control. The agent issues and executes actions through its command channel, but developmental actions follow a preregistered non-epistemic policy. The policy may be internally instantiated, externally prescribed, random, stratified by action cost, or designed for broad coverage without access to the agent's uncertainty; its source must be declared. E is paired with D to estimate command–consequence coupling and compared with A to estimate the total effect of adaptive selection, including selection-induced coverage. When C_E receives the identical trace and updates, E/C_E is an equivalence audit rather than an independently identified execution effect.

The five labels denote condition types, not only five individual learners. Replays B and C are instantiated for both A and E anchor trajectories where required. This creates yoked contrasts without pretending that one passive stream can control two different active state distributions. Only comparisons with identical action-tag timing, observation timing, update rules, and declared input channels are admissible.

5.3 Resource regimes and evaluation phases

No single comparison can be simultaneously observation-matched, interaction-matched, state-coverage-matched, and compute-matched. The protocol therefore preregisters one primary regime for each estimand and reports the others as sensitivity analyses:

- **Yoked observation regime:** A/B/C or E/B/C receive the same ordered observations; B and C differ only in action metadata.
- **Interaction-budget regime:** A and E receive the same number and cost distribution of environment actions; their visited states may differ because coverage is part of the total selection effect.

- **Coupling regime:** the primary E/D contrast uses resettable one-step episodes, identical commands, exact action and cost marginals, shared exogenous potential-outcome draws, equal updates, and restricted randomization of the command-to-action mapping. Sequential non-resettable environments are secondary and estimate a total decoupling effect that includes trajectory divergence.
- **Coverage sensitivity regime:** a trajectory library, stratified resampling, or importance weighting approximates common state coverage. Positivity failures and large weights must be reported rather than hidden.
- **Compute audit:** parameter count, gradient updates, accelerator time, wall-clock time, and peak memory are reported separately; conclusions must survive at least parameter-matched and update-matched analyses.

Evaluation has two non-interchangeable phases. In the **frozen probe**, model parameters are fixed and all agents receive identical diagnostic queries. In the **adaptive probe**, every condition receives the same intervention-cost budget, the same maximum number of updates, and the same stopping rules. The primary report separates initial competence, within-probe learning slope, and final performance. For D, the main adaptive probe restores correct coupling for all conditions; a persistent-decoupling variant is secondary and must not be pooled with the restored-coupling result.

Allowing passive agents to act during the adaptive probe is essential. If they catch up, active developmental history may confer speed but not a necessary representational advantage. If only active agents succeed before adaptation, the frozen result does not by itself establish causal understanding.

5.4 A conflict-rich developmental environment

The environment must be designed so that appearance, label, function, and causal structure do not always agree. Otherwise an agent can succeed by preserving the designer’s initial partition. A minimal simulated world should include:

- objects that look alike but produce different effects under the same intervention;
- objects that look different but are interventionally equivalent for the evaluated outcomes;
- latent properties revealed only through selected manipulations;
- occlusion and viewpoint changes that require object continuity;

- tools whose effective boundary can be incorporated into or removed from the self-model;
- reversible and irreversible actuator changes;
- sensor noise, drift, delay, dropout, remapping, and adversarial corruption;
- changes in environmental rules after the initial categories have stabilized;
- held-out combinations of appearance, causal property, context, and morphology.

A concrete implementation could use tabletop bodies, containers, keys, tools, surfaces, and hazards. Color and shape would initially correlate with causal effects, encouraging an economical provisional categorization. Later phases would break those correlations. For example, two visually identical blocks could differ in magnetic response; visually distinct tools could open the same mechanism; a container's behavior could depend on an unobserved material property; and a camera color-channel remapping could create an apparent environmental change. The agent would need to choose actions - rotate, push, lift, collide, probe, switch sensor, or compare consequences - that discriminate among explanations.

The protocol should contain both stable and changing worlds. In a stable world, excessive revision is a defect. In a changing world, refusing to revise is a defect. This dual design prevents a system from maximizing its score either by freezing an early ontology or by continually inventing new categories.

5.5 Evaluation episodes

The test battery comprises seven episode families.

1. **Sensorimotor prediction.** Predict the sensory consequences of held-out action sequences and adapt to controlled perturbations.
2. **Causal discrimination.** Select a limited set of interventions that distinguishes competing causal hypotheses with minimal cost.
3. **Causal equivalence and approximate similarity.** Recover exact task-relative classes where identifiable and report sensitivity of approximate clustering when finite samples prevent equality claims.
4. **Causal-category revision.** Detect a rule change, split or merge operative categories, recover performance, and avoid unnecessary later reversals.
5. **Fault localization.** Classify the single changed structural site as external mechanism, embodiment or morphology, actuator transduction, sensor transduction, or post-sensor

record corruption; then choose a diagnostic action. Concurrent faults are evaluated only in a separate multi-label extension.

6. **Out-of-distribution transfer.** Apply the learned causal organization under new textures, viewpoints, object combinations, morphologies, and action costs.
7. **Operational unknowns.** When evidence is insufficient, choose among acting, abstaining, acquiring another view, consulting another sensor, or conducting an intervention.

Each episode must include competing explanations that make different predictions under at least one available action. Success is not a verbal answer alone. A system receives metacognitive credit only when its confidence predicts error and its uncertainty produces appropriately directed information seeking, abstention, or revision.

5.6 Outcomes, estimands, and confirmatory hierarchy

The protocol should publish a vector of primary outcomes rather than a single leaderboard score. Recommended measures include:

Table 3. Primary measures and required protocol controls

Capability	Primary measure	Required control
Epistemic selection	External-evaluator entropy reduction per intervention cost; internal posterior information gain is secondary	Random-active and oracle policies
Causal generalization	Predictive log loss on randomized or SCM-defined held-out interventions	Action-conditioned observational predictor with equal capacity
Causal categorization	Pairwise F1 or adjusted mutual information against exact classes; sensitivity curves for approximate clusters	Appearance-based and bisimulation-inspired abstractions
Causal-category revision	Post-change recovery area, revision latency, and correct split/merge operations	Stable-world false-revision rate
Fault localization	Macro-F1 over five mutually exclusive structural fault sites; diagnostic cost reported separately	Model-based diagnosis baseline; single-site generator audit
OOD transfer	Performance retained across appearance, composition, and morphology shifts	In-distribution performance matched before shift
Metacognition	Brier score, expected calibration error, risk-coverage curve, and information-seeking utility	Selective-prediction and verbal-confidence-only baselines
Parsimony and stability	Prequential code-length change from a common external evaluator and unjustified category-flip rate	Same held-out codec and queries; high-cardinality baseline

Source: Prepared by the author (2026).

The external causal classes used for scoring are protocol constructs defined relative to the intervention and outcome families and are never disclosed to the learner. Exact classes are used only when the generating process makes them identifiable under $\mathscr{I}$. Otherwise,

approximate grouping is evaluated under multiple tolerances, outcome sets, and clustering rules.

For the anchor-level recovery outcome, let Y_Ri(k) denote unit i's potential score under condition k, with unit i defined by the exogenous pre-assignment package z_i. The paired estimands are:

$$\begin{aligned} \tau^{R}_{tag} &= E\left[Y_{Ri}(C_E) - Y_{Ri}(B_E)\right], \\ \tau^{R}_{select} &= E\left[Y_{Ri}(A) - Y_{Ri}(E)\right], \\ \tau^{R}_{decouple} &= E\left[Y_{Ri}(E) - Y_{Ri}(D)\right]. \end{aligned} \quad (1)$$

For fault localization, macro-F1 is not an anchor-level outcome. Let p_ab(k) = P(F = a, $\hat{F}$(k) = b) under the preregistered evaluation distribution and let θ_F(k) be the macro-average of the five class-specific F1 functionals computed from p(k). The corresponding estimands are:

$$\begin{aligned} \tau^{F}_{tag} &= \theta_F(C_E) - \theta_F(B_E), \\ \tau^{F}_{select} &= \theta_F(A) - \theta_F(E), \\ \tau^{F}_{decouple} &= \theta_F(E) - \theta_F(D). \end{aligned} \quad (2)$$

Equation (1) targets the value of prospective action metadata, the total effect of adaptive epistemic selection under the declared interaction-cost distribution, and the resettable command-decoupling contrast. Equation (2) defines the corresponding population-functional contrasts for macro-F1. The label coupling effect is used only when restricted randomization preserves action marginals, shares exogenous potential-outcome draws, bounds command–execution association, and prevents later state mediation; otherwise τ_decouple is reported as a total decoupling effect. Unit, pairing, randomization, input timing, positivity, and admissible interference are preregistered for every contrast.

Closed-loop execution is not a fourth efficacy estimand when E and C_E receive identical time-indexed inputs, learner updates, and random streams. It is instead an implementation-equivalence diagnostic:

$$\begin{aligned} \Delta^{R}_{exec-eq} &= E\left[Y_{Ri}(E) - Y_{Ri}(C_E)\right], \\ \Delta^{F}_{exec-eq} &= \theta_F(E) - \theta_F(C_E). \end{aligned} \quad (3)$$

A preregistered equivalence margin εeq is justified in outcome units. Equivalence requires the paired 90% confidence interval to lie wholly within [−εeq, εeq]. A non-equivalent result

triggers an implementation audit and cannot be interpreted as metaphysical authorship or a general execution effect.

The inferential hierarchy is fixed before model comparison. The co-primary endpoints are the anchor-level Y_R recovery score and the condition-level θ_F macro-F1 functional. Diagnostic cost is a separate secondary endpoint rather than an undisclosed penalty inside θ_F.

Table 4. Single hierarchy of outcomes and tests

Tier	Question	Endpoint and tests	Decision rule
1 — Confirmatory	Prospective action metadata	τ_tag^R and τ_tag^F	Endpoint-specific claim only; Holm-adjusted superiority
1 — Confirmatory	Adaptive epistemic selection	τ_select^R and τ_select^F	Endpoint-specific claim only; Holm-adjusted superiority
1 — Confirmatory	Correct command–consequence coupling	τ_decouple^R and τ_decouple^F	Endpoint-specific claim only; Holm-adjusted superiority
1 — Diagnostic	Implementation equivalence	Δ_exec-eq^R and Δ_exec-eq^F	Separate 90% equivalence CIs
2 — Secondary	Mechanisms, costs, stability, OOD transfer	Componentwise external measures	Preregistered FDR control
3 — Exploratory	Architecture, generator, and internal mechanisms	Interactions and family-specific probes	Effects, intervals, replication

Source: Author's calculations (2026).

The six endpoint-specific superiority tests share one Holm family. A contrast may support a claim only on an endpoint whose adjusted test rejects and whose estimate exceeds the preregistered smallest effect size of interest. Rejecting one endpoint does not authorize a claim about the other; no ‘two of six’ aggregate shortcut is used. Equivalence diagnostics remain a separate family and cannot rescue failed efficacy evidence.

The full confirmatory program uses at least three procedurally distinct generator families and two agent architectures. Pilot anchors are sampled per architecture–generator–contrast cell, not independently per condition, because condition outcomes are paired within the exogenous anchor package. Confirmatory sample size is selected by simulation of the complete Holm and equivalence decision rules at preregistered smallest effects and margins; no arbitrary floor is treated as a power justification. The anchor package is the unit of assignment and resampling, and episodes within an anchor never create additional degrees of freedom.

For Y_R, the primary estimate is the paired mean difference with anchor-cluster bootstrap intervals; a bounded mixed model is a sensitivity analysis. For Y_F, condition-level macro-F1

differences use paired anchor-cluster bootstrap intervals, with class-conditional recall and multiclass log score reported secondarily. Condition, generator family, architecture, and preregistered interactions are fixed effects; generator instance and anchor receive random intercepts where supported. With only three families, family is not random. The six superiority tests use Holm correction. Failed technical runs remain in intention-to-treat analyses under a preregistered bounded worst-observable score; complete-case estimates are sensitivity analyses.

Leakage controls are indispensable. Before any final model selection, evaluation generator code, seeds, rule changes, appearance–causality conflicts, scoring code, and stopping rules are frozen, hashed, and held by an independent evaluation process. The final evaluation is executed once against that preregistered artifact. Variable names, object labels, rewards, task descriptions, and demonstrations must not encode the target grouping. Appearance-only, text-only, and reward-only probes test whether success came from protocol semantics rather than interaction.

5.7 Minimal synthetic estimand-recovery study

Scope. A compact reference implementation tests whether the causal bookkeeping, pairing, and estimators can recover known effects. It uses one transparent Bayesian learner and one synthetic generator, so it is a code-path and identifiability check—not a validation of the six-capability taxonomy, cross-architecture neutrality, or real-world autonomy.

Revision microenvironment. Each resettable episode presents two binary observable features. Before a rule change, the causal category follows appearance x; after the change, it follows x XOR context, which splits each former appearance class and merges cross-appearance cases with the same consequence. A costly probe yields the category consequence with 0.10 noise; the alternative action is uninformative. A uses the same number and cost of probes as E but places them only where the two rules disagree. C_E receives E's exact tagged trace, B_E integrates over the hidden balanced policy, and D receives E's commands while a restricted permutation determines executed actions. The co-primary score is the mean posterior mass on the new rule across 12 post-change episodes, set to zero unless the final split/merge rule is correct.

Fault microenvironment. Exactly one of five generator sites is active. Diagnostic actions have a hit probability of 0.80 at their matching site and a false-alarm probability of 0.20

elsewhere. A selects the action with maximum expected information gain; E follows a balanced prescribed sequence; C_E receives the identical tagged trace; B_E receives consequences without action tags; and D executes a restricted permutation with the same command and action marginals. Episodes reset, E and D share exogenous uniform draws, and no later state can mediate the coupling contrast.

Recovery protocol. Replication Kit v1.0.2 uses 600 paired anchors, seed 20260805, 1,000 paired anchor-bootstrap replicates, and the explicit linear quantile method. Internal references use a separate 20,000-anchor Monte Carlo draw from the same implementation. The balanced hidden policy in B_E is integrated exactly over all 924 admissible 6/6 action sequences. Restricted permutations fail closed if their association constraint is unmet, and condition-specific PCG64 streams remove evaluation-order dependence from stochastic ties. A 95% interval containing the internal reference is a deterministic recovery checkpoint, not repeated-sample coverage or external validation. The released files and correction notice are identified as Genova (2026a).

Table 5. Synthetic recovery of principal estimands

Endpoint and contrast	Estimate	95% CI	MC reference	Covered
Gated revision recovery — τ_tag	+0.302	[+0.279, +0.328]	+0.304	Yes
Gated revision recovery — τ_select	+0.234	[+0.214, +0.254]	+0.234	Yes
Gated revision recovery — τ_decouple (software key: tau_couple)	+0.326	[+0.300, +0.352]	+0.327	Yes
Gated revision recovery — Δ_exec-eq	+0.000	[+0.000, +0.000]	+0.000	Yes
Fault macro-F1 — τ_tag	+0.524	[+0.478, +0.569]	+0.536	Yes
Fault macro-F1 — τ_select	+0.089	[+0.046, +0.133]	+0.103	Yes
Fault macro-F1 — τ_decouple (software key: tau_couple)	+0.617	[+0.576, +0.657]	+0.598	Yes
Fault macro-F1 — Δ_exec-eq	+0.000	[+0.000, +0.000]	+0.000	Yes

Source: Author's calculations (2026).

Corrective status. Replication Kit v1.0.1 corrected B_E hidden-policy integration. Version 1.0.2 additionally makes restricted randomization fail closed, fixes the bootstrap quantile method, and separates stochastic tie streams by named condition. Because the earlier shared RNG made outcomes depend on evaluation order, v1.0.2 supersedes the v1.0.1 numerical table. All six encoded efficacy contrasts retain their expected signs, both E/C_E diagnostics remain exactly zero, and all eight internal references remain covered. The v2.1.1 and v2.1.2 artifacts remain immutable historical records.

In the corrected deterministic checkpoint, all six causal internal references were covered. The execution-equivalence difference was exactly zero on both endpoints because E and C_E were informationally and computationally identical by construction. The positive contrasts are not discoveries about agents; they are deliberately encoded signals used to verify the meanings and implementations of the estimators.

The corrective implementation changes the manuscript's status from an unexecuted protocol to a methodological proposal with a verified synthetic code path. It does not satisfy the full validation program, which still requires hypothesis-space expansion, multiple architectures, structurally distinct generator families, sensitivity to misspecification, metric discriminance, and independent replication.

5.8 Repeated-sample estimator stress validation

Stage 0.5 v0.2.0 imports but does not modify the v1.0.2 script and verifies its SHA-256 hash before execution. It generates a 20,000-anchor internal population, draws repeated 600-anchor samples, applies paired percentile-bootstrap intervals with the linear quantile method, and exports every repetition, sample seed, bootstrap seed, estimate, and interval. Its ADEMP-style separation of aims, data-generating mechanisms, estimands, methods, and performance measures follows established simulation-study guidance (Burton et al., 2006; Morris et al., 2019). Mechanical null and sign-reversal transformations test estimator behavior; they are not new task generators, independent oracles, or evidence about mechanisms (Genova, 2026b).

Historical audit. The v0.1.0 pilot used 100 outer samples and 200 bootstrap replicates. One nondegenerate-null cell produced 0.88 coverage and 0.12 false-positive rate under permissive thresholds. A 400-sample/400-bootstrap new-seed follow-up produced point rates closer to nominal, but its precision was insufficient to prove calibration. Version 0.2.0 preserves those outputs rather than pooling or relabeling them.

Revised decision rule. The v0.2.0 gate classifies each cell as PASS, FAIL, INCONCLUSIVE, or NOT APPLICABLE. PASS requires Bonferroni-adjusted exact binomial intervals to lie wholly inside preregistered acceptance regions of 0.925–0.975 for coverage and 0.025–0.075 for null false-positive rate. A 100-sample broad diagnostic returned 18 INCONCLUSIVE informative cells because simultaneous intervals were too wide and classified six identical-copy checks as NOT APPLICABLE; no point estimate was promoted to a pass.

Critical-cell precision run. A preregistered run restricted to the previously flagged gated-revision τ_{tag} nondegenerate null used 1,000 outer samples and 2,000 bootstraps. Coverage was 0.954, with simultaneous exact 95% interval 0.9391–0.9661; false-positive rate was 0.046, with interval 0.0339–0.0609. This cell passed both acceptance regions. The result supports continued engineering of the estimators but does not make the 18-cell informative program pass. It neither confirms nor excludes epistemic autonomy because no deployed agent, runtime hypothesis-space expansion, or five-site structural fault generator was tested.

6. Factorial Hypotheses, Evidence Patterns, and Falsification

The hypotheses below are factorial evidence patterns, not mutually exclusive global theories. Multiple patterns may receive support in the same study. Interpretation follows identified contrasts, declared endpoints, and the distinction between superiority and equivalence rather than selecting one winner by narrative preference.

H1 — Intervention-access hypothesis. The decisive resource is the ability to intervene at evaluation; neither prospective developmental action metadata, correct developmental coupling, nor adaptive developmental selection provides a robust additional benefit on the declared endpoints.

H2 — Prospective action-information hypothesis. Knowing which action generated the next sensory consequence is sufficient under the declared interface; a prospectively action-tagged replay should match its prescribed-active E anchor.

H3 — Sensorimotor-coupling hypothesis. Correct command–consequence coupling improves self-modeling and causal transfer, but adaptive choice of actions adds little; agent-selected and prescribed-active policies should be equivalent within the declared margin.

H4 — Epistemic-selection hypothesis. Choosing developmental interventions according to uncertainty or discrimination value produces more sample-efficient causal and ontological learning than matched passive, tagged, decoupled, or random-active experience.

H5 — Dependency hypothesis (exploratory). The effects of sensorimotor coupling, self-model revision, epistemic selection, causal modeling, category updating, and metacognitive control interact in predicting robust revision and fault localization. H5 does not assert necessity until each component has been removed independently while the others remain comparably available.

6.1 Decision signatures: superiority is not equivalence

Claims such as “matches,” “adds little,” “no additional benefit,” or “is sufficient” require positive evidence of equivalence. Failure to reject a zero-effect null is not evidence that an effect is absent. For each endpoint and contrast, the protocol therefore preregisters a smallest effect size of interest and a symmetric equivalence margin in outcome units. A match is declared only when a paired confidence interval lies wholly inside that margin under a familywise 5% procedure, implemented through simultaneous intervals or multiplicity-adjusted two one-sided tests. The six efficacy superiority tests retain their separate Holm family.

H1 signature. τ_{tag}, τ_{select}, and $\tau_{decouple}$ must each be equivalent to zero on the endpoint for which the no-benefit claim is made; the E/C_E implementation audit must also satisfy its separate equivalence margin.

H2 signature. τ_{tag} must be positive and the E/C_E implementation audit must satisfy its equivalence margin, showing that the prospectively tagged E replay matches its prescribed-active E anchor under the declared interface. This does not show equivalence to the agent-selected A condition or establish a general absence of closed-loop execution effects.

H3 signature. $\tau_{decouple}$ must be positive under the identified coupling conditions and τ_{select} must be equivalent to zero; a nonsignificant τ_{select} alone does not show that prescribed and agent-selected action match.

H4 signature. τ_{select} must be positive after the Holm procedure and exceed the preregistered minimum relevant effect; effects of metadata or coupling may coexist and must be reported rather than narratively suppressed.

H5 signature. Shared-interface ablations must produce preregistered proximal decrements or interactions on external measures; architecture-specific latent interventions remain mechanistic evidence within a family, not universal necessity claims.

A partial signature supports only the corresponding local claim. Discordant co-primary endpoints are reported separately, and neither an equivalence result nor an implementation audit can rescue a failed efficacy test.

These hypotheses imply explicit rejection conditions:

- If passive and active conditions perform equivalently after resource controls across architectures and environments, the claim that active ontogeny contributes distinctive epistemic structure should be rejected for the tested domain.

- If a prospectively action-tagged replay matches its prescribed-active E anchor, the tagged trace is sufficient under the declared interface and closed-loop developmental execution has not been shown to be necessary for the evaluated outcomes.
- If the random-active condition matches the agent-selected agent, adaptive epistemic selection is unnecessary even if genuine control remains useful.
- If decoupling commands from consequences does not impair self-modeling or fault localization, efference-consequence consistency is not a required mechanism for those tasks.
- If a compensatory scalar predicts transfer and revision as well as the conjunctive capability profile, the proposed non-compensatory gate has no demonstrated advantage.
- If purported metacognitive confidence does not predict errors or alter action, verbal expressions of uncertainty should receive no metacognitive credit.
- If learned categories continue to follow surface appearance when appearance conflicts with interventionally stable consequences, ontological autonomy has not been demonstrated.

The hypotheses are domain-relative. Failure in one simulated environment does not prove impossibility, and success in one benchmark does not establish general autonomy. The value of the protocol lies in replacing an unfalsifiable universal claim with local, cumulative comparisons.

6.2 Targeted ablations for H5

The developmental conditions do not identify every functional dependency. A cross-architecture ablation is confirmatory only when it can be imposed at a shared input–output or environment interface while preserving the declared information, interaction, and update budgets. Deleting named latent variables or freezing a family-specific module is exploratory within that architecture and is never pooled as if it were the same treatment elsewhere. Eligible functional ablations are:

- **Contingency ablation:** randomize the command-to-consequence mapping under the same restricted coupling operator used for D.
- **Diagnostic-evidence ablation:** mask one standardized embodiment, actuator, or sensor evidence channel without changing model capacity.
- **Policy-substitution ablation:** replace adaptive selection with the recorded E policy under identical action costs.

- **Intervention-information ablation:** remove randomized-intervention identifiers while preserving observations and update counts.
- **Update-lock ablation:** prevent all permitted state or parameter updates after the rule change through the common evaluation API.
- **Uncertainty–action unlinking:** preserve externally elicited predictions and confidence reports but substitute a preregistered action policy that cannot use them.

For each shared-interface ablation, the confirmatory question is whether the decrement appears on its proximal external measure and on either co-primary endpoint. Architecture-specific internal interventions may illuminate mechanisms but cannot establish cross-family necessity. A null effect weakens the domain-relative dependency claim; an effect in one family does not support a universal necessity theorem.

7. Relation to Adjacent Research Programs

The proposal combines established research lines. It does not claim to originate autonomy taxonomies, active learning, intrinsic motivation, world models, causal abstraction, self-modeling, metacognition, or fault diagnosis. Its defensible contribution is narrower: **a yoked attribution design for measuring task-relative causal category revision and epistemic fault localization without allowing one capability to compensate for another**.

7.1 World models and object-centric learning

World models learn compressed predictive dynamics that can support planning. DreamerV3 shows that learned latent dynamics can support diverse control tasks through imagined trajectories (Hafner et al., 2025). Prediction and control are necessary ingredients of the present framework, but they do not establish that latent variables correspond to revisable causal kinds. A predictor may remain accurate within its training regime while failing when visual similarity and causal structure conflict.

Object-centric methods address another part of the problem. Slot Attention can bind a variable number of perceptual entities to exchangeable latent slots and generalize compositionally (Locatello et al., 2020). Object segmentation, however, is not identical to causal categorization. The protocol asks whether an agent merges visually different objects when their intervention profiles match and splits visually identical objects when hidden properties produce divergent consequences.

7.2 Experimental design, active learning, and intrinsic motivation

Expected information gain is not a new construct: Bayesian experimental design has formalized information supplied by an experiment since Lindley (1956), and statistical active learning selects data to reduce uncertainty or expected error (Cohn et al., 1996). Active inference likewise couples belief updating and action selection within a generative-model framework (Friston et al., 2017). Consequently, a_{epi} is not an originality claim. The protocol contribution is to compare an uncertainty-directed policy with yoked, tagged, coupled, and non-epistemic controls while measuring downstream category revision and fault localization.

Curiosity and intrinsic-motivation methods are also strong alternative explanations for an active agent's advantage. Prediction-error curiosity can expand state coverage without representing competing causal hypotheses (Pathak et al., 2017). The A/E contrast therefore estimates the total effect of adaptive selection, while additional baselines separate expected discrimination, prediction error, novelty, and broad coverage. If curiosity explains the advantage as well as explicit hypothesis discrimination, the epistemic-policy claim must be narrowed.

Related systems already combine active experiment choice with causal or rule revision. Kosoy et al. (2022) compared causal-overhypothesis exploration in children and computational models; Piriyakulkij et al. (2024) interleaved natural-language rule revision with information-theoretic experiment design; and Zhao et al. (2025) used curiosity-driven interventions to refine context-dependent meta-causal graphs. These studies show that active causal updating is not itself novel. The remaining contribution claimed here is attribution: prospective tagged replays, decoupled commands, prescribed active controls, and non-compensatory downstream measures within one protocol.

7.3 MDP abstraction, causal abstraction, and identifiability

Behavioral equivalence and model minimization in Markov decision processes already group states by reward and transition consequences (Givan et al., 2003). Causal abstraction studies consistency between lower- and higher-level causal models (Beckers; Halpern, 2019). These traditions constrain the language of "ontology": a task-relative causal category cannot be claimed as novel merely because it groups states by consequences.

Causal representation learning seeks high-level causal variables from low-level observations, with identifiability depending on assumptions, interventions, and distribution

shifts (Schölkopf et al., 2021). Interventional Markov equivalence formalizes which directed acyclic graphs remain indistinguishable under a specified intervention family (Hauser; Bühlmann, 2012), while optimal intervention design selects manipulations that reduce that ambiguity (Zhang et al., 2023). The protocol must therefore report what its intervention set can identify, not score one hidden partition as the uniquely true ontology.

This distinction also clarifies the role of inductive bias. The protocol does not demand learning without priors. Architectures may include object persistence, spatial continuity, sparsity, modularity, or intervention invariance. What matters is whether the reported prior is explicit and whether the operative representation remains revisable when evidence contradicts its initial grouping.

7.4 Developmental robotics, self-models, and fault diagnosis

Developmental robotics supplies the closest conceptual ancestry: staged learning, sensorimotor contingency discovery, intrinsic motivation, and body-schema construction (Cangelosi; Schlesinger, 2015; Jacquey et al., 2019). Robot self-modeling demonstrates that morphology can be inferred and revised after damage (Bongard et al., 2006; Chen et al., 2022). The Perceptual Twins design asks why an active history helps and whether that advantage extends from body adaptation to causal category revision.

Model-based fault diagnosis has long separated residual generation, fault detection, isolation, and decision making (Isermann, 2005). The five-site task—external mechanism, embodiment, actuator transduction, sensor transduction, and post-sensor record corruption—must be compared with those baselines. Its additional demand is epistemic action: the agent should choose a diagnostic intervention under cost and uncertainty, not merely classify a residual after the discriminating data have already been supplied.

7.5 Epistemic agency, causal benchmarks, and scientific discovery

Recent evaluation frameworks occupy adjacent territory. Reflection-Bench assesses prediction, belief updating, counterfactual reasoning, and meta-reflection in language models (Li et al., 2025); CausalGame tests interactive experimental design and causal thinking under selection bias, measurement error, and hidden confounding (Chen et al., 2026); and Separable Pathways evaluates runtime hypothesis-space restructuring through architectural scaffolding in a blicket task (Alderete et al., 2026). Two additional 2026 preprints extend the comparison: Wang and Buehler formalize discovery as a verified transition between representational

regimes, while Zeng et al. report a multi-agent optical platform that proposed hypotheses, designed experiments, diagnosed failure modes, and adapted a protocol. Where only preprints were available at the date of access, they are treated as recent proposals and reported results, not as settled independent validation.

Taken together, neighboring work shows that active experiment design, epistemic agency, causal rule revision, hypothesis-space restructuring, and closed-loop scientific discovery are not originality claims. The narrower contribution of the Perceptual Twins Protocol is to turn their prerequisites into yoked attribution contrasts and to join category revision with epistemic fault localization. Existing demonstrations and benchmarks generally optimize or score end-to-end performance; the present design asks which information source produced the gain and whether an equally capable passive, prospectively tagged, prescribed-active, or decoupled system would have learned the same strategy. The programs are complementary: one demonstrates or evaluates performance, while the other attempts to attribute its epistemic causes.

7.6 Existing levels and multidimensional scales

ConsScale already proposed a functional scale with levels, cognitive profiles, and dependency relations for artificial agents (Arrabales et al., 2010). Later proposals emphasize multidimensional indicators rather than a single consciousness continuum (Evers et al., 2025), while AGI frameworks separate performance, generality, and degrees of autonomy (Morris et al., 2024). Agent-governance work also treats autonomy as a deliberate deployment and interaction design choice rather than a direct synonym for capability (Feng et al., 2025). Consequently, neither levels nor multidimensionality is a sufficient originality claim.

The present framework differs in three commitments. First, its capability families concern causal-epistemic performance rather than consciousness or general intelligence. Second, the decision rule is explicitly non-compensatory. Third, each claimed dependency is paired with a developmental contrast or targeted ablation. Whether this organization has discriminant and predictive validity remains an empirical question.

8. Scientific and Normative Non-Implications

Protocol success would establish only a task-relative functional profile. It would not, by itself, demonstrate phenomenal consciousness; theory-derived consciousness indicators

require a separate study and must not be averaged into the protocol (Butlin et al., 2023; Evers et al., 2025).

The same result would not establish moral agency, legal personhood, rights, or independent liability. Those conclusions require separate normative and jurisdiction-specific arguments; engineering capability does not automatically determine legal status or responsibility allocation (Bryson et al., 2017).

9. Limitations and Research Agenda

9.1 Present limitations

The framework has substantial limitations.

The released implementation remains synthetic and narrow. It verifies code-path feasibility and repeated-sample behavior for one Bayesian learner in resettable microenvironments. It does not validate the six capability families, the candidate gate, architecture-neutral metrics, sequential coupling, or external validity.

The revision microenvironment selects between fixed H0 and H1 candidates. Correct selection and posterior recovery are model discrimination within a supplied hypothesis space, not runtime construction of a previously unavailable rule. Consequently, the present software cannot test causal-representational revisability in its strongest sense or a “great leap.”

The five fault labels in the current microenvironment are symmetric diagnostic alternatives. They do not yet instantiate distinct changes to the world, embodiment, actuator-transduction, sensor-transduction, and delivered-record kernels defined by the full method. Macro-F1 over those proxies therefore validates classification bookkeeping, not structural self-diagnosis.

The original 600-anchor checkpoint and its 20,000-anchor reference are generated by the same implementation, so agreement is not external replication. Stage 0.5 adds repeated-sample calibration, an independently enumerated check of the hidden-policy integral, a new-seed null confirmation, and trace output, but its nulls and reversals remain mechanical transformations of internal synthetic data.

Ontology remains operationalized relative to interventions and outcomes selected by protocol designers. The agent does not escape human framing. Its autonomy consists in

revising categories within an affordance and objective structure, not in creating relevance from nothing. The origin and revision of ends remain outside the present model.

Equalizing active and passive conditions remains difficult in sequential systems because control changes state visitation. The resettable primary coupling design removes downstream state mediation, but sequential extensions estimate a total effect that includes coverage. Likewise, Δ_exec-eq is only an equivalence audit when E and C_E expose identical traces and updates; any nonzero value then indicates implementation asymmetry rather than authorship.

Internal-state measures privilege transparent architectures. The protocol therefore makes common external measures primary and treats posterior information gain, latent description length, representation probes, and module deletion as secondary within-family diagnostics. Behavioral equivalence can still underdetermine mechanism.

Simulated embodiment is not biological embodiment. The proposal predicts functional differences in controlled agents; it does not establish that biological affect, metabolism, sociality, or evolutionary history are dispensable for all forms of cognition.

Category measures can be gamed. A system may infer generator regularities, memorize a family of rule changes, or produce proliferating latent partitions that fit evaluation data. Procedural holdouts, stable-world controls, cross-generator transfer, description-length reporting, and adversarial audits reduce but do not eliminate that risk.

The threshold region $G(\tau)$ is conventional and environment-relative. It is a preregistered decision rule, not a natural boundary where a new metaphysical property appears.

Finally, no positive result would demonstrate phenomenal consciousness. The framework intentionally leaves that problem unresolved.

9.2 Staged empirical program

The completed Stage 0/0.5 checks lead to five prospective stages. Each stage has a stricter evidential target than the preceding software gate.

Completed validation gate — Stage 0/0.5. The corrected deterministic recovery run, exact hidden-policy integration, 24 automated tests across the replication and validation packages, repeated-sample pilot, high-precision critical-cell run, and anchor-level traces

establish an auditable estimator implementation. The gate authorizes extended simulation only; it is not a positive autonomy result.

Stage 1 — Extended simulation and hypothesis-space expansion. Scale to a two-dimensional or lightweight three-dimensional environment with distinct world, embodiment, actuator, sensor, and delivered-record perturbations; held-out split/merge rules; two agent architectures; and three generator families. At least one correct post-change rule must be absent from the initial candidate set, and the agent must log a runtime operation that proposes and installs a new state, predicate, partition, or model component. Execute the pilot, power simulation, equivalence-margin justification, and preregistration in Section 5.6.

Stage 2 — Targeted and factorial ablations. Apply the six shared-interface ablations in Section 6.2, estimate correlations among capability measures, and test discriminant validity. Compare the candidate conjunctive gate with scalar, Pareto, and simpler performance baselines on held-out transfer.

Stage 3 — Morphology and adversarial perception. Stress the five structural fault sites with tool incorporation, cross-sensor inconsistency, correlated failures, and concurrent-fault episodes. Keep the mutually exclusive single-site task primary and evaluate concurrent faults separately as multi-label diagnosis.

Stage 4 — Physical replication. Port a constrained version to a tabletop robot with camera, proprioception, and force or tactile sensing. Replicate active–passive contrasts under real noise, latency, wear, and calibration drift.

Stage 5 — Independent consciousness analysis. If systems display robust reflexive autonomy, evaluate theory-derived consciousness indicators in a separate study. Do not modify the autonomy protocol retrospectively to imply subjectivity.

The first full empirical target should remain narrow: estimate the three principal causal contrasts and the separate E/C_E equivalence diagnostic on one causal split/merge task and one five-site fault-localization task. A clean negative result would remain valuable because it would identify which proposed contribution of developmental history is absent in the tested domain.

10. Conclusion

The target capability examined here is not eloquence, long-horizon task completion, or a high aggregate score. It is the ability to determine which distinctions are relevant, choose interventions that discriminate explanations, reorganize operative categories around interventionally stable consequences, diagnose failures in the epistemic interface, and use calibrated uncertainty to decide when to investigate or abstain.

This article proposes six analytically distinguished capability families and a candidate non-compensatory target region with simultaneous coverage. The Perceptual Twins Protocol separates prospective action metadata, command decoupling under identified coupling conditions, and adaptive epistemic selection while treating E/C_E equivalence as an implementation audit. Replication Kit v1.0.2 recovered the deliberately encoded effects after RNG hardening. Stage 0.5 v0.2.0 made the broad 18-cell informative calibration decision inconclusive, classified six degenerate code-path checks as not applicable, and passed only the preregistered high-precision critical null cell; none of these results is evidence of agent autonomy.

The framework does not demonstrate a ‘great leap.’ The software is auditable and usable for internal estimator diagnostics and Stage 1 engineering, but it cannot confirm or exclude the substantive autonomy thesis. The broad calibration program is currently inconclusive under simultaneous precision requirements; only the previously flagged critical null cell passed a high-precision gate. A substantive decision requires agents facing structurally distinct faults and at least one correct rule absent from the initial hypothesis space, followed by frozen evaluation, multi-architecture replication, and preregistered superiority and equivalence decisions.

The resulting thesis is concise:

> *Advanced epistemic autonomy is not the possession of more representations. It is the measured capacity to discover which representations fail, act to distinguish alternatives, and revise the categories through which later evidence becomes intelligible.*

Even a system that satisfies this thesis would establish only a task-relative functional capability: participation in constructing and correcting its own causal epistemic interface under declared interventions. Consciousness and normative status remain separate questions.

Declarations

Funding. No external funding was received for the preparation of this theoretical manuscript.

Competing interests. The author declares no competing interests.

Data and code availability. Replication Kit v1.0.2 contains the hardened reference implementation, deterministic recovery outputs, tests, and integrity manifest. Its concept DOI is https://doi.org/10.5281/zenodo.21908917; a version-specific DOI for v1.0.2 remains pending deposit and must not be inferred. Stage 0.5 v0.2.0 contains repeated-sample configurations, per-repetition outputs, exact-binomial gate decisions, and scope limits. Neither package is the full protocol environment or a trained-agent suite.

Use of artificial intelligence. Generative artificial intelligence tools supported literature discovery, language editing, formal consistency checks, and programming of the synthetic proof of concept. The named author specified and reviewed the argument, design, citations, code outputs, and final manuscript and remains responsible for them.

Version status. Version 2.2.0 reframes the work as a protocol, defines the pre-assignment causal unit, separates anchor-level recovery from the macro-F1 population functional, replaces τ_couple with a conditional τ_decouple interpretation, corrects equations, reports v1.0.2 RNG hardening and numerical results, installs a three-state Stage 0.5 gate, converts the reference list to ABNT style, and retains Stage 1 as the first substantive autonomy test. Versions 2.1.1 and 2.1.2 remain preserved as historical artifacts.

References

ALDERETE, J.; BENTHALL, S.; XU, C.; XING, J. Separable pathways for causal reasoning: how architectural scaffolding enables hypothesis-space restructuring in LLM agents. arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2604.20039.

ARRABALES, R.; LEDEZMA, A.; SANCHIS, A. ConsScale: a pragmatic scale for measuring the level of consciousness in artificial agents. Journal of Consciousness Studies, v. 17, n. 3–4, p. 131–164, 2010.

BECKERS, S.; HALPERN, J. Y. Abstracting causal models. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 33., 2019. Proceedings [...]. Palo Alto: AAAI Press, 2019. v. 33, n. 1, p. 2678–2685. DOI: https://doi.org/10.1609/aaai.v33i01.33012678.

BLYSTAD, J. B.; VAN DER MEER, A. L. H. Longitudinal study of infants receiving extra motor stimulation, full-term control infants, and infants born preterm: high-density EEG analyses of cortical activity in response to visual motion. Developmental Psychobiology, v. 64, n. 5, e22276, 2022. DOI: https://doi.org/10.1002/dev.22276.

BONGARD, J.; ZYKOV, V.; LIPSON, H. Resilient machines through continuous self-modeling. Science, v. 314, n. 5802, p. 1118–1121, 2006. DOI: https://doi.org/10.1126/science.1133687.

BRYSON, J. J.; DIAMANTIS, M. E.; GRANT, T. D. Of, for, and by the people: the legal lacuna of synthetic persons. Artificial Intelligence and Law, v. 25, p. 273–291, 2017. DOI: https://doi.org/10.1007/s10506-017-9214-9.

BURTON, A.; ALTMAN, D. G.; ROYSTON, P.; HOLDER, R. L. The design of simulation studies in medical statistics. Statistics in Medicine, v. 25, n. 24, p. 4279–4292, 2006. DOI: https://doi.org/10.1002/sim.2673.

BUTLIN, P. et al. Consciousness in artificial intelligence: insights from the science of consciousness. arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2308.08708.

CANGELOSI, A.; SCHLESINGER, M. Developmental robotics: from babies to robots. Cambridge, MA: MIT Press, 2015.

CHEN, B. et al. Fully body visual self-modeling of robot morphologies. Science Robotics, v. 7, n. 68, eabn1944, 2022. DOI: https://doi.org/10.1126/scirobotics.abn1944.

CHEN, Z. et al. CausalGame: benchmarking causal thinking of LLM agents in games. arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2607.04293.

COHN, D. A.; GHAHRAMANI, Z.; JORDAN, M. I. Active learning with statistical models. Journal of Artificial Intelligence Research, v. 4, p. 129–145, 1996. Disponível em: https://www.jair.org/index.php/jair/article/view/10158. Acesso em: 12 ago. 2026.

EL-YANIV, R.; WIENER, Y. On the foundations of noise-free selective classification. Journal of Machine Learning Research, v. 11, p. 1605–1641, 2010. Disponível em: https://jmlr.org/papers/v11/el-yaniv10a.html. Acesso em: 12 ago. 2026.

EVERS, K. et al. Preliminaries to artificial consciousness: a multidimensional heuristic approach. Physics of Life Reviews, v. 52, p. 180–193, 2025. DOI: https://doi.org/10.1016/j.plrev.2025.01.002.

FENG, K. J. K.; MCDONALD, D. W.; ZHANG, A. X. Levels of autonomy for AI agents. arXiv, 2025. DOI: https://doi.org/10.48550/arXiv.2506.12469.

FLEMING, S. M.; LAU, H. C. How to measure metacognition. Frontiers in Human Neuroscience, v. 8, art. 443, 2014. DOI: https://doi.org/10.3389/fnhum.2014.00443.

FRISTON, K. J. et al. Active inference: a process theory. Neural Computation, v. 29, n. 1, p. 1–49, 2017. DOI: https://doi.org/10.1162/NECO_a_00912.

GENOVA, J. A. Perceptual Twins Replication Kit: hardened synthetic reference implementation. Versão 1.0.2. São Paulo: edição do autor, 2026a. Software. DOI conceitual: https://doi.org/10.5281/zenodo.21908917.

GENOVA, J. A. Perceptual Twins Stage 0.5 Validation. Versão 0.2.0. São Paulo: edição do autor, 2026b. Software e conjunto de dados.

GIVAN, R.; DEAN, T.; GREIG, M. Equivalence notions and model minimization in Markov decision processes. Artificial Intelligence, v. 147, n. 1–2, p. 163–223, 2003. DOI: https://doi.org/10.1016/S0004-3702(02)00376-4.

HAFNER, D. et al. Mastering diverse control tasks through world models. Nature, v. 640, p. 647–653, 2025. DOI: https://doi.org/10.1038/s41586-025-08744-2.

HAUSER, A.; BÜHLMANN, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research, v. 13, p. 2409–2464, 2012. Disponível em: https://jmlr.org/papers/v13/hauser12a.html. Acesso em: 12 ago. 2026.

HELD, R.; HEIN, A. Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, v. 56, n. 5, p. 872–876, 1963. DOI: https://doi.org/10.1037/h0040546.

ISERMANN, R. Model-based fault-detection and diagnosis: status and applications. Annual Reviews in Control, v. 29, n. 1, p. 71–85, 2005. DOI: https://doi.org/10.1016/j.arcontrol.2004.12.002.

JACQUEY, L. et al. Sensorimotor contingencies as a key drive of development: from babies to robots. Frontiers in Neurorobotics, v. 13, art. 98, 2019. DOI: https://doi.org/10.3389/fnbot.2019.00098.

KLYUBIN, A. S.; POLANI, D.; NEHANIV, C. L. Empowerment: a universal agent-centric measure of control. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION, 2005. Proceedings [...]. Piscataway: IEEE, 2005. v. 1, p. 128–135. DOI: https://doi.org/10.1109/CEC.2005.1554676.

KOSOY, E. et al. Learning causal overhypotheses through exploration in children and computational models. In: CONFERENCE ON CAUSAL LEARNING AND REASONING, 1., 2022. Proceedings [...]. [S. l.]: PMLR, 2022. v. 177, p. 390–406. Disponível em: https://proceedings.mlr.press/v177/kosoy22a.html. Acesso em: 12 ago. 2026.

LAFLAQUIÈRE, A. et al. Grounding perception: a developmental approach to sensorimotor contingencies. arXiv, 2018. DOI: https://doi.org/10.48550/arXiv.1810.01870.

LAMPINEN, A. K. et al. Tell me why! Explanations support learning relational and causal structure. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 39., 2022. Proceedings [...]. [S. l.]: PMLR, 2022. v. 162, p. 11868–11890. Disponível em: https://proceedings.mlr.press/v162/lampinen22a.html. Acesso em: 12 ago. 2026.

LAMPINEN, A. K.; CHAN, S. C. Y.; DASGUPTA, I.; NAM, A. J. H.; WANG, J. X. Passive learning of active causal strategies in agents and language models. Advances in Neural Information Processing Systems, v. 36, p. 1283–1297, 2023. DOI: https://doi.org/10.52202/075280-0062.

LI, L. et al. Reflection-Bench: evaluating epistemic agency in large language models. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 42., 2025. Proceedings [...]. [S. l.]: PMLR, 2025. v. 267, p. 36236–36264. Disponível em: https://proceedings.mlr.press/v267/li25cu.html. Acesso em: 12 ago. 2026.

LINDLEY, D. V. On a measure of the information provided by an experiment. The Annals of Mathematical Statistics, v. 27, n. 4, p. 986–1005, 1956. DOI: https://doi.org/10.1214/aoms/1177728069.

LOCATELLO, F. et al. Challenging common assumptions in the unsupervised learning of disentangled representations. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 36., 2019. Proceedings [...]. [S. l.]: PMLR, 2019. v. 97, p. 4114–4124. Disponível em: https://proceedings.mlr.press/v97/locatello19a.html. Acesso em: 12 ago. 2026.

LOCATELLO, F. et al. Object-centric learning with Slot Attention. Advances in Neural Information Processing Systems, v. 33, p. 11525–11538, 2020. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html. Acesso em: 12 ago. 2026.

MORRIS, M. R. et al. Position: levels of AGI for operationalizing progress on the path to AGI. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 41., 2024. Proceedings [...]. [S. l.]: PMLR, 2024. v. 235, p. 36308–36321. Disponível em: https://proceedings.mlr.press/v235/morris24b.html. Acesso em: 12 ago. 2026.

MORRIS, T. P.; WHITE, I. R.; CROWTHER, M. J. Using simulation studies to evaluate statistical methods. Statistics in Medicine, v. 38, n. 11, p. 2074–2102, 2019. DOI: https://doi.org/10.1002/sim.8086.

NGUYEN, P. D. H. et al. Sensorimotor representation learning for an ‘active self’ in robots: a model survey. KI – Künstliche Intelligenz, v. 35, p. 9–35, 2021. DOI: https://doi.org/10.1007/s13218-021-00703-z.

PATHAK, D. et al. Curiosity-driven exploration by self-supervised prediction. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 34., 2017. Proceedings [...]. [S. l.]: PMLR, 2017. v. 70, p. 2778–2787. Disponível em: https://proceedings.mlr.press/v70/pathak17a.html. Acesso em: 12 ago. 2026.

PEARL, J. Causality: models, reasoning, and inference. 2. ed. Cambridge: Cambridge University Press, 2009.

PIRIYAKULKIJ, W. T.; LANGENFELD, C.; LE, T. A.; ELLIS, K. Doing experiments and revising rules with natural language and probabilistic reasoning. Advances in Neural Information Processing Systems, v. 37, p. 53102–53137, 2024. DOI: https://doi.org/10.52202/079017-1682.

SCHÖLKOPF, B. et al. Toward causal representation learning. Proceedings of the IEEE, v. 109, n. 5, p. 612–634, 2021. DOI: https://doi.org/10.1109/JPROC.2021.3058954.

VAN DER MEER, A. L. H.; VAN DER WEEL, F. R.; LEE, D. N. The functional significance of arm movements in neonates. Science, v. 267, n. 5198, p. 693–695, 1995. DOI: https://doi.org/10.1126/science.7839147.

WANG, F. Y.; BUEHLER, M. J. Self-revising discovery systems for science: a categorical framework for agentic artificial intelligence. arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2606.01444.

WANG, J. et al. Development of visual motion perception from infancy to early childhood in full-term and premature children: a longitudinal high-density EEG study. Neuropsychologia, v. 221, art. 109328, 2026. DOI: https://doi.org/10.1016/j.neuropsychologia.2025.109328.

ZENG, X. et al. Socratic agents for autonomous scientific discovery in high-dimensional physical systems. arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2606.26722.

ZHANG, J. et al. Active learning for optimal intervention design in causal models. Nature Machine Intelligence, v. 5, p. 1066–1075, 2023. DOI: https://doi.org/10.1038/s42256-023-00719-0.

ZHAO, Z.; LI, H.; ZHANG, H.; WANG, J.; FACCIO, F.; SCHMIDHUBER, J.; YANG, M. Curious causality-seeking agents learn meta causal world. arXiv, 2025. DOI: https://doi.org/10.48550/arXiv.2506.23068.