THEORETICAL AND METHODOLOGICAL ARTICLE

Testing Task-Relative Epistemic Autonomy in Artificial Agents: The Perceptual Twins Benchmark

A yoked causal framework with a synthetic estimand-recovery study

Jandislei Antonio Genova

Independent Researcher, São Paulo, Brazil

Independent preprint | Version 2.1.1 | 5 August 2026

Theoretical and methodological article with a minimal synthetic proof of concept. Not peer reviewed.

Abstract

Existing AI benchmarks evaluate prediction, task completion, causal reasoning, and adjacent forms of epistemic agency, but generally do not causally isolate which features of developmental interaction support task-relative revision of causal categories. This article develops the Perceptual Twins Benchmark, a yoked framework that distinguishes six capability families without collapsing them into one score. The design compares self-directed active agents, passive replays, prospectively action-tagged observers, command-decoupled agents, and prescribed active controls. Three principal paired estimands target prospective action metadata, adaptive epistemic selection, and correct command–consequence coupling; closed-loop execution is treated separately as an implementation-equivalence diagnostic. A candidate non-compensatory decision rule uses a preregistered required-indicator set and simultaneous one-sided coverage. Fault labels are generated at five mutually exclusive structural sites: external mechanism, embodiment, actuator transduction, sensor transduction, and post-sensor record corruption. A minimal Bayesian microimplementation with 600 paired anchors recovered all six causal estimands—three contrasts across two co-primary endpoints—within 95% bootstrap intervals containing their 20,000-anchor Monte Carlo reference values; the equivalence diagnostic was exactly zero. This proof of concept demonstrates code-path feasibility and estimand recovery only. It does not validate the full benchmark across architectures, establish the proposed capability, or support inferences about consciousness, moral agency, or legal status.

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

Resumo

Os benchmarks de inteligência artificial avaliam previsão, conclusão de tarefas, raciocínio causal e formas próximas de agência epistêmica, mas em geral não isolam causalmente quais características da interação desenvolvimental sustentam a revisão de categorias causais relativa à tarefa. Este artigo desenvolve o Benchmark dos Gêmeos Perceptivos, quadro pareado que distingue seis famílias de capacidades sem reuni-las numa nota única. O desenho compara agentes ativos autodirigidos, replays passivos, observadores com etiquetas prospectivas de ação, agentes com comandos desacoplados e controles ativos prescritos. Três estimandos pareados principais focalizam metadados prospectivos de ação, seleção epistêmica adaptativa e acoplamento correto entre comando e consequência; a execução em circuito fechado é tratada separadamente como diagnóstico de equivalência da implementação. Uma regra decisória não compensatória candidata utiliza conjunto pré-registrado de indicadores obrigatórios e cobertura unilateral simultânea. As falhas são geradas em cinco sítios estruturais mutuamente exclusivos. Uma microimplementação bayesiana com 600 âncoras pareadas recuperou os seis estimandos causais dentro de intervalos bootstrap de 95% que continham as referências de Monte Carlo com 20.000 âncoras; o diagnóstico de equivalência foi exatamente zero. A prova de conceito demonstra apenas viabilidade do código e recuperação dos estimandos, não valida o benchmark completo nem autoriza inferências sobre consciência, agência moral ou condição jurídica.

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 benchmark 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 Benchmark; (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-ontological autonomy:** the capacity to construct and revise action-relevant categories and relations according to their interventionally stable consequences rather than only according to labels or superficial similarity.

Operational autonomy is compatible with a fixed ontology. Epistemic autonomy is compatible with human-supplied categories. Causal-ontological autonomy is stronger because the representational scheme 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 benchmark 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 benchmark. 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 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 benchmark 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 Benchmark 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 self-directed 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(·) 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.

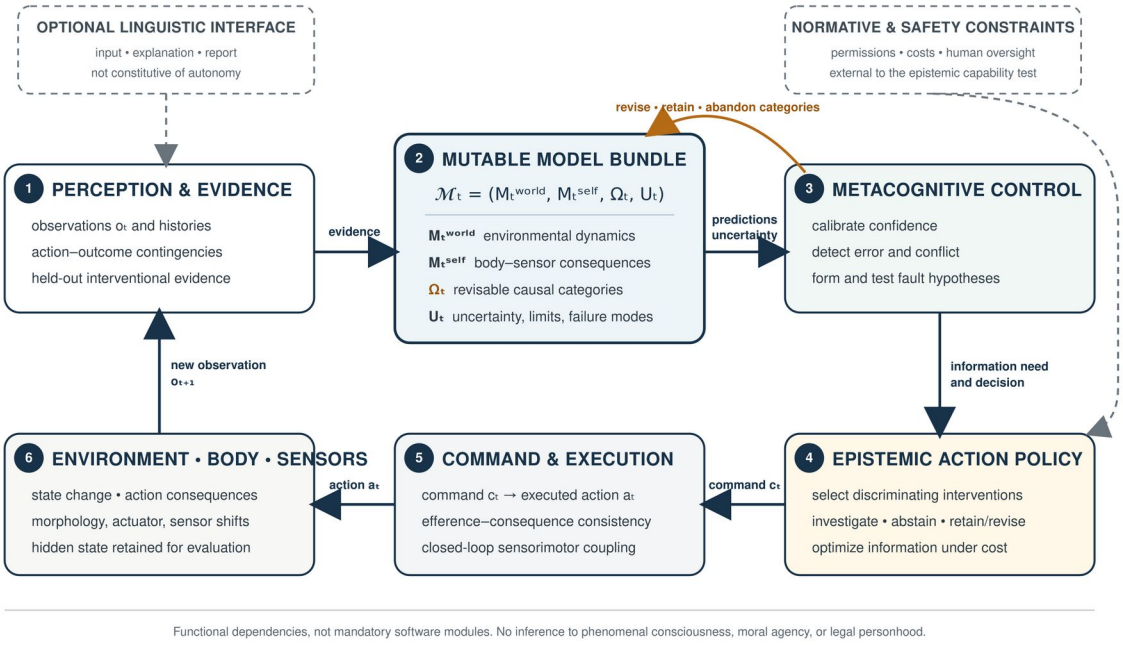

Figure 1. Functional cycle of task-relative causal-epistemic autonomy.

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 D the distribution of evaluation episodes. Using natural logarithms, define information gain in nats by:

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

and the cost-adjusted policy value by:

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

The benchmark-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 each I indexed by capability k and indicator j be sign-oriented, 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:

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

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

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

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

This is a candidate conjunctive, benchmark-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 Benchmark

5.1 Design logic

The Perceptual Twins Benchmark 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.

Each active anchor trajectory creates time-indexed observations, executed actions, and hidden simulator states retained only for evaluation. Yoked agents receive exactly the variables declared for their condition. The primary developmental objective is self-supervised prediction and representation learning. Semantic category labels, fault labels, causal-class

identifiers, and task rewards that reveal the target grouping are withheld. If an environment exposes extrinsic reward, it must be constant within preregistered causal contrasts or analyzed as a separate factor. The epistemic utility used by a self-directed agent is computed from its own uncertainty or model disagreement and is not a benchmark label.

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 Benchmark*

Condition	Controls developmental stream?	Receives action metadata?	Command-consequence coupling	Chooses informative actions?
A. Self-directed 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. Self-directed 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 benchmark 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 benchmark 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 benchmark should publish a vector of primary outcomes rather than a single leaderboard score. Recommended measures include:

Table 3. *Primary measures and required benchmark 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 benchmark 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 I. Otherwise,

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

For outcome Y, three principal paired causal estimands are defined on anchor-matched potential outcomes:

$$\tau_{tag}^{(Y)} = E\left[Y_{C_E} - Y_{B_E}\right], \tau_{select}^{(Y)} = E\left[Y_A - Y_E\right], \tau_{couple}^{(Y)} = E\left[Y_E - Y_D\right].$$

They estimate, respectively, the value of prospective action metadata under an E-anchored replay, the total effect of adaptive epistemic selection under an equal interaction-cost distribution, and the direct effect of correct command–consequence coupling in the resettable restricted-randomization regime. If reset is impossible, τ_couple must be relabeled the total decoupling effect because future state coverage becomes a mediator. Unit, pairing, randomization, input timing, 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 and updates. It is instead an implementation-equivalence diagnostic:

$$\Delta_{exec-eq}^{(Y)} = E\left[Y_E - Y_{C_E}\right].$$

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 two co-primary endpoints are Y_R, post-change recovery area multiplied by an indicator of the correct final split or merge, and Y_F, macro-F1 across the five mutually exclusive fault sites. Diagnostic cost is a separate secondary endpoint rather than an undisclosed penalty inside Y_F.

Table 4. Single hierarchy of outcomes and tests

Tier	Question	Endpoint and tests	Decision rule
1 — Confirmatory	Prospective action metadata	τ_tag on Y_R and Y_F	Two of six paired superiority tests
1 — Confirmatory	Adaptive epistemic selection	τ_select on Y_R and Y_F	Two of six paired superiority tests
1 — Confirmatory	Correct command–consequence coupling	τ_couple on Y_R and Y_F	Two of six paired superiority tests
1 — Diagnostic	Implementation equivalence	Δ_exec-eq on Y_R and Y_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 confirmatory superiority tests share one Holm family. Rejecting a contrast on one co-primary endpoint does not authorize a claim about the other. The equivalence diagnostics are reported separately and cannot rescue a failed efficacy endpoint.

The full confirmatory program uses at least three procedurally distinct generator families and two agent architectures. An exploratory pilot starts with 20 independent anchors per condition and architecture–family cell. Simulation-based power then selects the confirmatory sample for 80% power at the preregistered minimum relevant paired effect, with a floor of 40 anchors per principal contrast and cell. The anchor trajectory is the unit of assignment and resampling; 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. Variable names, object labels, reward channels, task descriptions, and demonstrations must not encode the target grouping. Held-out rule changes and appearance-causality conflicts are procedurally generated after model selection.

Generator families used for final evaluation are withheld during development. Appearance-only, text-only, and reward-only probes test whether success came from benchmark 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. The finite study uses 600 paired anchors, seed 20260805, and 1,000 anchor-bootstrap replicates. Reference estimands use an independent 20,000-anchor Monte Carlo run. Recovery is declared when the finite-sample 95% interval contains the reference value; Δ_exec-eq is checked separately as a deterministic equivalence audit because E and C_E execute the same learner on the same trace.

Table 5. Synthetic recovery of principal estimands

Endpoint and contrast	Estimate	95% CI	MC reference	Covered
Gated revision recovery — τ_tag	+0.328	[+0.304, +0.351]	+0.331	Yes
Gated revision recovery — τ_select	+0.253	[+0.232, +0.272]	+0.237	Yes
Gated revision recovery — τ_couple	+0.322	[+0.297, +0.348]	+0.328	Yes
Gated revision recovery — Δ_exec-eq	+0.000	[+0.000, +0.000]	+0.000	Yes
Fault macro-F1 — τ_tag	+0.507	[+0.458, +0.558]	+0.520	Yes
Fault macro-F1 — τ_select	+0.104	[+0.062, +0.149]	+0.115	Yes
Fault macro-F1 — τ_couple	+0.618	[+0.577, +0.656]	+0.588	Yes
Fault macro-F1 — Δ_exec-eq	+0.000	[+0.000, +0.000]	+0.000	Yes

Source: Author's calculations (2026).

All six causal reference values 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 causal contrasts are not discoveries about agents; they are deliberately encoded test signals used to verify that the estimators recover the effects their names denote.

The implementation therefore 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 multiple architectures, generator families, sensitivity to misspecification, metric discriminance, and independent replication.

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 the identified contrasts and endpoints 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 active anchor.

H3 - Sensorimotor-coupling hypothesis. Correct command-consequence coupling improves self-modeling and causal transfer, but adaptive choice of actions adds little; self-directed and random-active agents should perform similarly.

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.

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 active 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 self-directed 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.

6.1 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 benchmark 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 benchmark 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 benchmark 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 benchmark 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 Benchmark 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

Benchmark 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 benchmark (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.

First, the implementation is deliberately synthetic and narrow. It verifies code-path feasibility and recovery of known estimands for one Bayesian learner in one resettable generator. It does not validate the six capability families, the candidate gate, architecture-neutral metrics, sequential coupling, or external validity.

Second, ontology is operationalized relative to interventions and outcomes selected by benchmark 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.

Third, 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.

Fourth, internal-state measures privilege transparent architectures. The benchmark 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.

Fifth, 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.

Sixth, 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.

Seventh, 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 research program can proceed in five stages.

Stage 1 — Extended simulation. Scale the present microimplementation to a two-dimensional or lightweight three-dimensional environment with all five fault sites, held-out

split/merge rules, two agent architectures, and three generator families. Execute the pilot, power simulation, and preregistration in Section 5.6.

Stage 2 — Targeted and factorial ablations. Apply the six shared-interface ablations in Section 6.1, 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. Introduce all five structural fault sites, 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 benchmark 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 has proposed six analytically distinguished capability families and a candidate non-compensatory target region with simultaneous coverage. The Perceptual Twins Benchmark separates three causal sources of apparent advantage—prospective action

metadata, correct command–consequence coupling, and adaptive epistemic selection—while treating execution equivalence as an implementation diagnostic rather than an efficacy claim. The synthetic microstudy recovered the deliberately encoded effects, showing that the paired estimators and restricted-randomization design behave as specified.

The framework does not demonstrate a 'great leap.' Its present value is methodological: it converts a broad claim into identified contrasts, external endpoints, explicit failure conditions, and a reproducible minimal test. Full validation still depends on multi-architecture implementation and independent replication.

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. The version 2.1.1 supplement contains the Python reference implementation, deterministic seed, finite-sample estimates, and machine-readable recovery outputs. It is not a release of the full benchmark environment or 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 2.1.1 adds the functional architecture figure, defines the required indicator set and simultaneous coverage, adopts one mutually exclusive fault taxonomy, strengthens coupling identification, reclassifies execution as an equivalence diagnostic, establishes one outcome hierarchy, makes metrics and ablations architecture-neutral, standardizes references, shortens normative discussion, and reports a minimal synthetic estimand-recovery study.

References

Alderete, J., Benthal, S., Xu, C., & Xing, J. (2026). Separable pathways for causal reasoning: How architectural scaffolding enables hypothesis-space restructuring in LLM agents. arXiv preprint arXiv:2604.20039. https://doi.org/10.48550/arXiv.2604.20039

Arrabales, R., Ledezma, A., & Sanchis, A. (2010). ConsScale: A pragmatic scale for measuring the level of consciousness in artificial agents. Journal of Consciousness Studies, 17(3–4), 131–164.

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 2678–2685. https://doi.org/10.1609/aaai.v33i01.33012678

Blystad, J. B., & van der Meer, A. L. H. (2022). 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, 64(5), e22276. https://doi.org/10.1002/dev.22276

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. Science, 314(5802), 1118–1121. https://doi.org/10.1126/science.1133687

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. Artificial Intelligence and Law, 25, 273–291. https://doi.org/10.1007/s10506-017-9214-9

Butlin, P., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708. https://doi.org/10.48550/arXiv.2308.08708

Cangelosi, A., & Schlesinger, M. (2015). Developmental robotics: From babies to robots. MIT Press.

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

Chen, Z., et al. (2026). CausalGame: Benchmarking causal thinking of LLM agents in games. arXiv preprint arXiv:2607.04293. https://doi.org/10.48550/arXiv.2607.04293

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. Journal of Artificial Intelligence Research, 4, 129–145. https://www.jair.org/index.php/jair/article/view/10158

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

Evers, K., et al. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. Physics of Life Reviews, 52, 180–193. https://doi.org/10.1016/j.plrev.2025.01.002

Feng, K. J. K., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for AI agents. arXiv preprint arXiv:2506.12469. https://doi.org/10.48550/arXiv.2506.12469

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, Article 443. https://doi.org/10.3389/fnhum.2014.00443

Friston, K. J., et al. (2017). Active inference: A process theory. Neural Computation, 29(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Givan, R., Dean, T., & Greig, M. (2003). Equivalence notions and model minimization in Markov decision processes. Artificial Intelligence, 147(1–2), 163–223. https://doi.org/10.1016/S0004-3702(02)00376-4

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

Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research, 13, 2409–2464. https://jmlr.org/papers/v13/hauser12a.html

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 56(5), 872–876. https://doi.org/10.1037/h0040546

Isermann, R. (2005). Model-based fault-detection and diagnosis: Status and applications. Annual Reviews in Control, 29(1), 71–85. https://doi.org/10.1016/j.arcontrol.2004.12.002

Jacquey, L., et al. (2019). Sensorimotor contingencies as a key drive of development: From babies to robots. Frontiers in Neurorobotics, 13, Article 98. https://doi.org/10.3389/fnbot.2019.00098

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In 2005 IEEE Congress on Evolutionary Computation (Vol. 1, pp. 128–135). IEEE. https://doi.org/10.1109/CEC.2005.1554676

Kosoy, E., et al. (2022). Learning causal overhypotheses through exploration in children and computational models. Proceedings of the First Conference on Causal Learning and Reasoning, PMLR 177, 390–406. https://proceedings.mlr.press/v177/kosoy22a.html

Laflaquière, A., et al. (2018). Grounding perception: A developmental approach to sensorimotor contingencies. arXiv preprint arXiv:1810.01870. https://doi.org/10.48550/arXiv.1810.01870

Lampinen, A. K., et al. (2022). Tell me why! Explanations support learning relational and causal structure. Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 11868–11890. https://proceedings.mlr.press/v162/lampinen22a.html

Lampinen, A. K., et al. (2023). Passive learning of active causal strategies in agents and language models. Advances in Neural Information Processing Systems, 36. https://doi.org/10.52202/075280-0062

Li, L., et al. (2025). Reflection-Bench: Evaluating epistemic agency in large language models. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, 36236–36264. https://proceedings.mlr.press/v267/li25cu.html

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

Locatello, F., et al. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 4114–4124. https://proceedings.mlr.press/v97/locatello19a.html

Locatello, F., et al. (2020). Object-centric learning with Slot Attention. Advances in Neural Information Processing Systems, 33, 11525–11538. https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html

Morris, M. R., et al. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 36308–36321. https://proceedings.mlr.press/v235/morris24b.html

Nguyen, P. D. H., et al. (2021). Sensorimotor representation learning for an 'active self' in robots: A model survey. KI – Künstliche Intelligenz, 35, 9–35. https://doi.org/10.1007/s13218-021-00703-z

Pathak, D., et al. (2017). Curiosity-driven exploration by self-supervised prediction. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2778–2787. https://proceedings.mlr.press/v70/pathak17a.html

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Piriyakulkij, W. T., Langenfeld, C., Le, T. A., & Ellis, K. (2024). Doing experiments and revising rules with natural language and probabilistic reasoning. Advances in Neural Information Processing Systems, 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5f1b350fc0c2affd56f465faa36be343-Abstract-Conference.html

Schölkopf, B., et al. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

van der Meer, A. L. H., van der Weel, F. R., & Lee, D. N. (1995). The functional significance of arm movements in neonates. Science, 267(5198), 693–695. https://doi.org/10.1126/science.7839147

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

Wang, F. Y., & Buehler, M. J. (2026). Self-revising discovery systems for science: A categorical framework for agentic artificial intelligence. arXiv preprint arXiv:2606.01444. https://doi.org/10.48550/arXiv.2606.01444

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

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

Zhao, Z., Li, H., Zhang, H., Wang, J., Faccio, F., Schmidhuber, J., & Yang, M. (2025). Curious causality-seeking agents learn meta causal world. arXiv preprint arXiv:2506.23068. https://doi.org/10.48550/arXiv.2506.23068