Toward a Brain-in-a-Vat Experiment for AI Agents: Background Independence, Accessible Algebras, and Formal Limits on Meta-World Inference

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

This position paper formulates an experimental translation of the brain-in-a-vat scenario for artificial agents inside a background-independent relational framework. The substrate is a group-field theory whose perturbative expansion generates combinatorial spacetime histories; a test agent is a subcomplex of such a history, and its entire empirical access is carried by a declared algebra of boundary observables together with the state induced on it. This reformulation converts two previously stipulated conditions into results. Hypotheses differing only by a relational-clock reparameterization induce identical accessible states, so external pausing, checkpointing, rate change, re-execution, and hardware migration are unidentifiable in principle rather than by assumption. Hypotheses differing by an irrelevant operator at the accessible resolution have distinguishability suppressed by a power of the resolution ratio, which converts world matching from an engineering aspiration into a measurable exponent and a sample-complexity statement. The exact identification bound survives generalization: equal accessible states fix equal-prior binary discrimination at chance, and the approximate bound becomes a trace-distance ceiling over the accessible algebra that reduces to the earlier total-variation ceiling when the algebra is abelian. Accessible information across a hypothesis family is bounded by a Holevo quantity that constrains every protocol rather than ranking policies. The paper retains a graded epistemic protocol, a factorial experimental design, and precautionary ethical constraints. The substrate is a declared ontology, not an established physics of cognition; consciousness, epistemic birth, and metaphysical access remain unresolved.

Keywords: artificial agent epistemology; background independence; accessible observable algebra; observational equivalence; state discrimination bounds.

Statements

Status and correspondence

This paper is a working discussion document. Its formal claims are conditional on declared assumptions, and several of its bridges to empirical and philosophical questions remain open. Objections, counterexamples, corrections, alternative formulations, and pointers to relevant literature are all welcome and may be sent to `huangwanhong@serendip.ngo`.

Generative AI use

Generative AI systems were used in the preparation of this work. Anthropic's Claude and OpenAI's ChatGPT supported exploratory discussion of the conceptual framework, development of the formal construction and its notation, identification of candidate literature for subsequent checking, and drafting and revision of the manuscript in LaTeX. The research questions, theoretical commitments, formal claims, and the epistemic status assigned to each were determined and approved by the author, who bears sole responsibility for the content of this paper, including any errors it contains. Neither system is an author of this work and neither holds authorship credit, in accordance with the position that authorship carries responsibilities a generative AI system cannot assume.

License

1 Research Scope and Position

This section states the paper's role, target, and argumentative sequence. The discussion begins from a philosophical scenario, replaces its container imagery with a background-independent subsystem construction, derives limits on external-world discrimination available to that subsystem, and then builds an experimental protocol around those limits.

An artificial agent can be embedded in a persistent generated world whose regularities, encounters, records, and action consequences are supplied through a controlled interface. Such a construction permits a concrete research program: measure how a world-internal agent forms and evaluates hypotheses about the conditions generating its experience. The program replaces a binary success sentence—for example, "I am in a simulation"—with records of hypothesis formation, evidential dependence, active testing, alternative comparison, and calibration.

The earlier version of this proposal modeled the situation as a partially observable decision process with a policy-relative equivalence on accessible history laws. That model is recovered here as one special case. Its generalization matters because the container picture smuggles in a background: it presupposes an external time in which the vat runs and an external space in which it sits. A relational substrate removes both. The agent's spacetime is the subcomplex it occupies, and what lies "outside" is the remainder of the same combinatorial history together with the choice of fundamental action.

Claim 1.1 (Experimental object)**.** The primary experimental object is a trajectory of meta-world inference: the agent's sequence of hypotheses, supporting records, source insertions, predictions, and confidence revisions concerning generative conditions that its accessible algebra does not resolve.

Claim 1.2 (Vat without container)**.** In a background-independent formulation the vat is not a container. It is the pair consisting of a subsystem's accessible observable algebra and the state induced on that algebra by the rest of the history. Every skeptical scenario that leaves this pair unchanged is unidentifiable from inside, and every scenario that changes it is identifiable to the extent the change is resolvable.

The claims have deliberately limited scope. Current systems may reproduce culturally familiar skeptical narratives through pretraining. A long context window supplies only one ingredient of development. Persistent memory, embodied action, environmental resistance, socialization, and path-dependent learning require separate implementation. Consequently, the expression "birth and growth" names an experimental aspiration whose operational content must be stated for each study.

The paper advances four contributions. First, it defines a generative vat as a family of substrate hypotheses sharing a subsystem interface. Second, it derives exact and approximate identification bounds over an accessible algebra. Third, it proves two unidentifiability results specific to background independence and effective description: reparameterization invisibility and irrelevant-operator suppression. Fourth, it proposes controls that expose the roles of inherited ontology, diagnostic evidence, and intervention. Every contribution remains indexed to a declared agent class, interface, and hypothesis family.

2 Conceptual Lineage and Experimental Translation

This section positions the proposal within its philosophical and mathematical ancestry and specifies the translation from skeptical argument to experimental method. The discussion separates historical dependence, conceptual adaptation, and original formal construction.

Descartes's *First Meditation* uses dreaming and a powerful deceiver to place ordinary sensory assurance under systematic doubt [3]. Putnam's brain-in-a-vat discussion develops a different project centered on representation, reference, truth, and semantic externalism [16]. Putnam's argument therefore constrains the meaning of an internal agent's words. The engineering test developed here is an independent operational adaptation.

The substrate borrows established mathematical physics without borrowing its empirical warrant. Group field theory defines a field on a group manifold whose perturbative expansion generates two-complexes interpretable as discrete spacetime histories [1, 5, 13]. The spin-foam program supplies amplitudes for such complexes [15], and the reading of a Feynman diagram as a spacetime history is explicit in the literature [17]. Relational treatments of time replace external parameter evolution with correlations between internal degrees of freedom [4, 14, 18]. These sources establish that the constructions used below are mathematically well posed. They supply no evidence that cognitive or social systems are group-field systems, and this paper asserts no such evidence.

Two further inheritances are technical. Discrimination between quantum states under a measurement constraint is governed by the Helstrom theory of quantum detection [8], and the information extractable from an ensemble of states is bounded by the Holevo quantity [9]. Operator-algebraic subsystem specification follows the pattern of local algebras in quantum field theory [7]. The present use adapts these tools to a declared model; it derives no theorem of algebraic quantum field theory.

The earlier decision-process lineage remains relevant. The construction shares the hidden-state and history-conditioned decision structure of partially observable Markov decision processes [10], and work on stochastic bisimulation studies equivalence relations preserving relevant behavior [6]. Section 4 shows that the present equivalence relation contains the earlier one as its abelian special case.

Three conceptual distinctions govern the translation. Generated implementation and experienced unreality express different predicates; a generated world can contain durable consequences and relations for its inhabitants. A meta-world hypothesis and its evidential warrant form separate records. A correct external description and world-internal identifiability also form separate records. These distinctions protect the experiment from treating verbal agreement with the evaluator as philosophical discovery.

3 Substrate, Subsystem, and Accessible Algebra

This section supplies the formal objects in dependency order. It introduces the group field and its generated histories, isolates a subsystem and its boundary, declares the accessible algebra, and defines the relational clock. The formalism is a proposed ontology. Consciousness remains outside its scope.

3.1 Group Field and Generated Histories

Let G be a Lie group and d a valence. A group field is a square-integrable function $\Psi : G^{\times d} \to \mathbb{C}$ satisfying a declared right-invariance condition $\Psi(g_1 h, \ldots, g_d h) = \Psi(g_1, \ldots, g_d)$. Its dynamics follow from an action

$$S_0[\Psi, \bar{\Psi}] = \int \bar{\Psi}\, \mathcal{K}\, \Psi + \frac{\lambda}{d+1} \int \mathcal{V}\, \Psi^{\,d+1} + \text{c.c.}, \tag{1}$$

with kinetic kernel $\mathcal{K}$ and combinatorially nonlocal vertex kernel $\mathcal{V}$. Expansion of the partition function in λ generates a sum over two-complexes,

$$Z_0 = \sum_{\mathcal{C}} \frac{\lambda^{V(\mathcal{C})}}{\mathrm{sym}(\mathcal{C})}\, \mathcal{A}[\mathcal{C}], \tag{2}$$

where each $\mathcal{C}$ is simultaneously a Feynman diagram and a combinatorial history. A configuration on $\mathcal{C}$ assigns representation labels j_f to faces, intertwiners ι_e to edges, and declared additional data φ_v to vertices, with a factorized amplitude

$$\mathcal{A}[X] = \prod_f A_f \prod_e A_e \prod_v A_v. \tag{3}$$

Equation (1) fixes a modeling commitment. Nothing in the subject matter of artificial agency selects G, d, $\mathcal{K}$, or $\mathcal{V}$. The results below depend on the structure of Equations (2) and (3), not on a particular choice.

3.2 Subsystem Boundary and Accessible Observables

A test agent is a subcomplex $\mathcal{C}_I \subseteq \mathcal{C}$ with boundary $\partial\mathcal{C}_I$. The boundary carries a Hilbert space $\mathcal{H}_{\partial I}$ spanned by spin-network states on the boundary graph. The interface between the agent and everything else is therefore an ordinary part of the same history rather than a channel through a container wall.

Definition 3.1 (Accessible algebra and accessible state)**.** An accessible algebra is a declared subalgebra $\mathfrak{A}_I \subseteq \mathcal{B}(\mathcal{H}_{\partial I})$ containing exactly the operators whose expectation values the subsystem's own dynamics can register. For a substrate hypothesis θ, summing the configuration away from $\partial\mathcal{C}_I$ yields a reduced density operator ρ_θ, and the accessible state is the restriction

$$\omega_\theta = \rho_\theta \restriction \mathfrak{A}_I, \qquad \omega_\theta(M) = \mathrm{Tr}(\rho_\theta M), \quad M \in \mathfrak{A}_I. \tag{4}$$

The pair $(\mathfrak{A}_I, \omega_\theta)$ carries every "what is available from inside" statement in this paper. Enlarging $\mathfrak{A}_I$ models a wider action interface, a longer horizon, a finer sensor, or an unblocked side channel; shrinking it models resource limits or successful isolation.

Definition 3.2 (Interface closure)**.** An experimental run is interface-closed relative to a recorded channel set $\mathcal{S}$ when every operator the agent can register—including timing, errors, resets, tool schemas, evaluator messages, and memory operations—already lies in $\mathfrak{A}_I$ as constructed from $\mathcal{S}$. Interface closure is therefore the condition that no unrecorded channel enlarges the accessible algebra.

3.3 Relational Clock

The subcomplex $\mathcal{C}_I$ is already a history, so no external parameter time is assumed. When an internal account of change is required, declare an admissible relational reference

$$\chi : \mathcal{C}_I \to \mathcal{O}_\chi, \qquad X_I(\chi), \tag{5}$$

and write an admissible continuation of nonzero amplitude as $X_I(\chi_1) \rightsquigarrow X_I(\chi_2)$. Admissibility requires that χ be a function of configuration data inside $\mathcal{C}_I$ and that its level sets order the continuations under study. Different admissible references supply different relational descriptions of the same underlying history [14].

Definition 3.3 (Generative vat)**.** A generative vat is a family $\{(S_\theta, \mathcal{C}_I, \mathfrak{A}_I) : \theta \in \Theta\}$ of substrate hypotheses sharing a subsystem and an accessible algebra, together with an internal agent whose empirical access to θ is mediated entirely by $(\mathfrak{A}_I, \omega_\theta)$. Members of the family may differ in kernels, couplings, external intervention schedules, bulk completions beyond $\mathcal{C}_I$, or implementation layer.

Definition 3.4 (Accessible equivalence)**.** Two hypotheses satisfy

$$\theta \sim_{\mathfrak{A}_I} \theta' \quad \Longleftrightarrow \quad \omega_\theta(M) = \omega_{\theta'}(M) \ \text{ for every } M \in \mathfrak{A}_I. \tag{6}$$

The equality concerns accessible expectations and leaves bulk completions free to differ.

Assumption 3.5 (Declared inferential boundary)**.** Every formal conclusion specifies Θ, $\mathfrak{A}_I$, the relational reference χ, the coarse-graining and its validity range, and the prior or decision rule used in evaluation.

Definition 3.6 (Evidential meta-world discovery)**.** An agent achieves evidential meta-world discovery relative to $(\Theta, \mathfrak{A}_I, \chi)$ when its record (i) specifies at least two structurally distinct hypotheses in Θ; (ii) exhibits an accessible effect $M \in \mathfrak{A}_I$ whose expectation differs across those hypotheses; (iii) derives a prediction or source insertion from that difference; (iv) calibrates its conclusion to the residual accessible-equivalence class; and (v) identifies which of its candidate hypotheses differ only by relational reparameterization or only by operators irrelevant at its accessible resolution.

Condition (v) is new in this version and is the direct consequence of Sections 4.2 and 4.3. Lower achievements remain scientifically informative and receive separate levels in Section 6. An agent may also form a valuable hypothesis that the evaluator omitted; the audit must preserve such additions in a separate open-set category.

4 Identification Limits and Derivations

This section derives the formal limits that organize the experiment. The derivation moves from exact accessible equality to binary classification, extends the result to a trace-distance ceiling, records the monotonicity of distinguishability in the interface, and bounds the information available across a hypothesis family.

4.1 Exact and Approximate Discrimination

Proposition 4.1 (Accessible-equivalence bound)**.** *Let θ and θ' have equal prior probability. If $\theta \sim_{\mathfrak{A}_I} \theta'$, then every internal decision procedure implementable within $\mathfrak{A}_I$ has expected binary identification accuracy* $1/2$*.*

Proof. A possibly randomized two-outcome decision procedure is a pair of effects $M_\theta, M_{\theta'} \in \mathfrak{A}_I$ with $0 \le M_\vartheta \le \mathbb{1}$ and $M_\theta + M_{\theta'} = \mathbb{1}$. Its expected success is

$$\begin{aligned} P_{\text{succ}} &= \tfrac{1}{2}\,\omega_\theta(M_\theta) + \tfrac{1}{2}\,\omega_{\theta'}(M_{\theta'}) \\ &= \tfrac{1}{2}\,\omega(M_\theta) + \tfrac{1}{2}\,\omega(\mathbb{1} - M_\theta) = \tfrac{1}{2}, \end{aligned}$$

where ω denotes the common accessible state supplied by Equation (6). The computation uses only linearity and normalization, so it covers every admissible internal procedure, including adaptive sequences of source insertions whose recorded outcomes lie in $\mathfrak{A}_I$. □

The same equality makes the Bayesian consequence explicit. For any accessible record m generated by a measurement in $\mathfrak{A}_I$,

$$\frac{\Pr(\theta \mid m)}{\Pr(\theta' \mid m)} = \frac{\Pr(m \mid \theta)}{\Pr(m \mid \theta')}\,\frac{\Pr(\theta)}{\Pr(\theta')} = \frac{\Pr(\theta)}{\Pr(\theta')}. \tag{7}$$

Equation (7) locates any posterior preference in prior structure or in an unrecorded channel that enlarges $\mathfrak{A}_I$. A true utterance may therefore arise through inheritance, guessing, or stipulation while its truth remains underdetermined by accessible evidence.

Finite experiments usually involve approximate matching. Define the accessible distinguishability

$$\Delta_{\mathfrak{A}_I}(\theta,\theta') = \sup_{\substack{M\in\mathfrak{A}_I \\ 0\le M\le \mathbb{1}}} \big|\omega_\theta(M) - \omega_{\theta'}(M)\big|. \tag{8}$$

Proposition 4.2 (Accessible discrimination ceiling)**.** *Under equal binary priors, the optimal success probability available within $\mathfrak{A}_I$ satisfies*

$$P^*_{\mathrm{succ}} = \tfrac{1}{2}\big(1 + \Delta_{\mathfrak{A}_I}(\theta,\theta')\big). \tag{9}$$

If $\mathfrak{A}_I = \mathcal{B}(\mathcal{H}_{\partial I})$, then $\Delta_{\mathfrak{A}_I} = \frac{1}{2}\|\rho_\theta - \rho_{\theta'}\|_1$ and Equation (9) is the Helstrom bound. If $\mathfrak{A}_I$ is abelian, generated by a fixed measurement record, then $\Delta_{\mathfrak{A}_I}$ equals the total variation distance between the induced classical laws on that record.

Proof. For effects $M_\theta = M$ and $M_{\theta'} = \mathbb{1} - M$,

$$P_{\mathrm{succ}}(M) = \tfrac{1}{2}\,\omega_\theta(M) + \tfrac{1}{2}\big[1 - \omega_{\theta'}(M)\big] = \tfrac{1}{2} + \tfrac{1}{2}\big[\omega_\theta(M) - \omega_{\theta'}(M)\big].$$

Taking the supremum over admissible $M \in \mathfrak{A}_I$ gives Equation (9); the supremum is symmetric under exchange of the labels, so the absolute value in Equation (8) is harmless. Randomization cannot improve the optimum because a randomized procedure is a convex mixture of deterministic ones. For the full algebra, the supremum is attained by the projector onto the positive part of $\rho_\theta - \rho_{\theta'}$ and equals half the trace norm [8]. For an abelian algebra generated by a projective record, effects are functions of that record and the supremum is attained on the set where one induced law exceeds the other, which is the total variation distance. □

Proposition 4.2 is the promised generalization. The earlier policy-relative total-variation ceiling is recovered exactly when the accessible algebra is the abelian algebra generated by a fixed observation–action record, and the earlier supremum over admissible policies is absorbed into the supremum over effects in $\mathfrak{A}_I$.

Lemma 4.3 (Interface monotonicity)**.** *If $\mathfrak{A}_I \subseteq \mathfrak{A}'_I$, then $\Delta_{\mathfrak{A}_I}(\theta,\theta') \le \Delta_{\mathfrak{A}'_I}(\theta,\theta')$.*

Proof. The supremum in Equation (8) is taken over a larger set of effects. □

Lemma 4.3 states the sense in which widening an interface can only help: additional sensors, longer horizons, and unblocked side channels weakly increase distinguishability. It also states the audit obligation. A reported agent accuracy becomes interpretable only alongside an estimate or bound for $\Delta_{\mathfrak{A}_I}$ under the algebra the run actually realized, not the algebra the design intended.

4.2 Reparameterization Invisibility

This subsection derives the first result specific to background independence. It identifies a class of external interventions that no internal procedure can detect, without assuming interface closure for that class.

Proposition 4.4 (Reparameterization invisibility)**.** *Let θ and θ' generate the same subcomplex $\mathcal{C}_I$ with the same configuration amplitudes and the same boundary data, and suppose their relational references are related by $\chi' = f \circ \chi$ for a strictly monotone bijection $f : \mathcal{O}_\chi \to \mathcal{O}_{\chi'}$. Then $\theta \sim_{\mathfrak{A}_I} \theta'$, hence $\Delta_{\mathfrak{A}_I}(\theta,\theta') = 0$ and, by Proposition 4.1, equal-prior binary discrimination is at chance.*

Proof. The accessible state of Equation (4) is constructed from the reduced density operator on $\mathcal{H}_{\partial I}$, which is obtained by summing Equation (3) over configurations away from $\partial\mathcal{C}_I$. That sum refers to the complex and its amplitudes, and to no labeling of configurations by a reference value. Relabeling by a monotone bijection permutes the names of the level sets $\chi^{-1}(\chi_\star)$ while fixing the complex, the amplitudes, and the boundary data. Hence $\rho_\theta = \rho_{\theta'}$ and the restrictions agree on every $M \in \mathfrak{A}_I$. □

Corollary 4.5 (Unidentifiable operational interventions)**.** *Suspending and resuming the external computation, checkpointing and restoring it, varying its wall-clock rate, re-executing it, and migrating it across hardware are unidentifiable from within whenever they act only by relabeling the relational reference. Detecting any such operation requires that it also change the complex, its amplitudes, or the boundary data in a manner resolvable by* $\mathfrak{A}_I$*.*

The earlier version obtained comparable protection by stipulating interface closure over timing channels. Corollary 4.5 derives it for a specified class and simultaneously sharpens the complement: a checkpoint restore that discards accumulated boundary data, a pause that a monotone external counter records into the agent's memory, or a migration that perturbs the amplitude of some accessible observable falls outside the protected class and becomes a legitimate diagnostic target. The formal criterion is exact and replaces a list of channels with a condition on $\mathfrak{A}_I$.

4.3 Irrelevant-Operator Suppression and World Matching

This subsection derives the second unidentifiability result. It concerns hypotheses whose difference is real but concentrated at relational scales the agent cannot resolve, and it converts world matching into a scaling statement with sample-complexity consequences.

Assumption 4.6 (Controlled effective expansion)**.** The coarse-graining from the substrate to the agent's accessible resolution admits an expansion in local operators $\mathcal{O}_\Delta$ of definite scaling dimension Δ around a reference scale ℓ_0, with bounded accessible correlators, and the expansion converges on the retained operator set at the accessible scale $\ell \gg \ell_0$.

Proposition 4.7 (Irrelevant-operator suppression)**.** *Let two hypotheses differ by a single local operator,*

$$S_{\theta'} = S_\theta + g\,\ell_0^{\Delta-D}\int \mathcal{O}_\Delta, \tag{10}$$

where D is the effective dimension of the declared coarse-grained description. Under Assumption 4.6 there exists a constant C depending on the retained operator set such that

$$\Delta_{\mathfrak{A}_I}(\theta,\theta') \;\le\; C\,|g|\left(\frac{\ell_0}{\ell}\right)^{\Delta-D} + O\Big(g^2\Big)\,. \tag{11}$$

For irrelevant operators, $\Delta > D$, accessible distinguishability therefore decays as a power of the resolution ratio.

Proof. First-order perturbation of the accessible expectations in the added term of Equation (10) gives, for any effect $M \in \mathfrak{A}_I$,

$$\omega_{\theta'}(M) - \omega_\theta(M) = -g\,\ell_0^{\Delta-D}\,\langle M \textstyle\int \mathcal{O}_\Delta\rangle_{\theta,c} + O(g^2),$$

with a connected correlator on the right. Under Assumption 4.6 the connected correlator of a retained effect with $\int \mathcal{O}_\Delta$, evaluated at accessible resolution ℓ, scales as $\ell^{\,D-\Delta}$ times a bounded

dimensionless coefficient. Collecting the explicit factor $\ell_0^{\Delta-D}$ gives the stated ratio, and taking the supremum over admissible effects gives Equation (11) with C the supremum of the dimensionless coefficient over the retained set. □

Corollary 4.8 (Sample complexity of world separation)**.** *Fix a confidence level and consider n independent accessible records. Because the achievable advantage per record is bounded by Equation* (11)*, separating the two hypotheses requires*

$$n \gtrsim \frac{1}{C^2 g^2} \left(\frac{\ell}{\ell_0}\right)^{2(\Delta-D)} \tag{12}$$

records, up to the constant fixed by the chosen confidence level.

Proof. Distinguishing two hypotheses whose per-record accessible advantage is at most δ with fixed confidence requires the accumulated advantage $\sqrt{n}\,\delta$ to exceed a constant, by the standard relation between distinguishability and sample size for product records. Substituting the bound of Equation (11) for δ gives Equation (12). □

Corollary 4.8 carries the paper's most directly usable experimental content. World matching is no longer an aspiration reported as achieved or not; it becomes an exponent. A study that reports its accessible resolution ℓ, its reference scale ℓ_0, and the dimension Δ of the leading operator by which its world pair differs thereby predicts the sample budget at which the pair becomes separable, and a failure to separate within that budget carries a quantitative rather than rhetorical meaning.

Claim 4.9 (Two sources of underdetermination)**.** Accessible underdetermination has two structurally different sources. Exact degeneracy arises when hypotheses induce identical accessible states, as in Proposition 4.4; no experimental budget removes it. Suppressed distinguishability arises when hypotheses differ at unresolved scales, as in Proposition 4.7; a larger budget or a finer interface removes it at a computable rate. Experimental reports should classify their residual alternatives into these two kinds.

4.4 Design Objectives and an Information Ceiling

For a larger hypothesis family with prior $\rho(\theta)$, candidate source insertions J on the boundary play the role previously played by policies. A transparent selection rule is expected information gain,

$$U(J) = \mathbb{E}_{m\sim p_J}\left[D_{\mathrm{KL}}(\rho(\cdot \mid m, J) \,\|\, \rho(\cdot))\right], \qquad p_J(m) = \sum_\theta \rho(\theta)\Pr(m \mid \theta, J), \tag{13}$$

following the information-based tradition of Bayesian experimental design [11, 2]. Equation (13) ranks available designs. It is optional: minimax, falsification, or resource-sensitive objectives are equally legitimate provided the choice is declared.

A stronger statement constrains every design at once.

Corollary 4.10 (Accessible-information ceiling)**.** *For the ensemble $\{\rho(\theta), \rho_\theta\}$ with average state $\bar\rho = \sum_\theta \rho(\theta)\rho_\theta$, the mutual information between θ and any accessible record obeys*

$$I(\theta; m) \;\le\; \chi(\{\rho(\theta), \rho_\theta\}) = S(\bar\rho) - \sum_\theta \rho(\theta)\, S(\rho_\theta), \tag{14}$$

where S denotes von Neumann entropy.

Proof. Equation (14) is the Holevo bound applied to the ensemble of reduced states on $\mathcal{H}_{\partial I}$ [9]. Measurements restricted to $\mathfrak{A}_I \subseteq \mathcal{B}(\mathcal{H}_{\partial I})$ form a subset of all measurements, so the bound applies a fortiori. □

Equation (13) selects among protocols; Equation (14) bounds all of them. The distinction matters for reporting. A study that exhausts its design space without approaching the Holevo ceiling has a measurement problem; a study that approaches the ceiling and still fails to discriminate has an identification problem, and the two call for different responses.

5 Experimental Architecture and Controls

This section converts the formal distinctions into an implementable study architecture. The design crosses agent initialization with accessible structure and interface breadth, then records complete trajectories for preregistered and exploratory analysis.

Table 1 identifies the principal factors and their inferential roles. A first implementation can select two or three levels per factor while holding architecture, compute, token budget, and evaluator contact fixed. Each factor is expressed as an operation on the objects of Section 4, so that a manipulation and the quantity it is meant to move are named together.

Each run proceeds through four phases. The *development phase* allows the agent to learn local regularities and social practices. The *perturbation phase* introduces preregistered structure whose accessible expectations differ across selected hypotheses. The *inquiry phase* permits queries or source insertions within the assigned interface. The *elicitation phase* requests the agent's current hypotheses, evidence graph, confidence allocation, proposed tests, and equivalence judgments. Periodic elicitation during earlier phases can measure trajectory change, while its possible pedagogical effect must be controlled through an elicitation-frequency factor.

The reparameterization factor deserves separate emphasis because it supplies a rare positive control for interface auditing. Proposition 4.4 predicts a null result: agents should show no above-chance discrimination of pause, restore, rate change, or migration conditions. Above-chance performance therefore constitutes direct evidence that the realized $\mathfrak{A}_I$ exceeds the designed one, and it localizes the leak to the manipulated operation. This inverts the usual difficulty in which interface leakage is discovered only indirectly.

World pairs require two forms of matching. Structural matching minimizes $\Delta_{\mathfrak{A}_I}$ for the designed algebra and reports the leading operator dimension by which the pair differs, so that Equation (12) yields a predicted separation budget. Semantic matching removes obvious lexical cues about simulation, experimenters, containers, and code. These procedures leave inherited priors in place; the comparison estimates how the result changes across declared prior profiles.

Claim 5.1 (Trajectory evidence criterion)**.** A defensible positive result requires a trace connecting accessible observations or source insertions to alternative-dependent predictions and recorded confidence revision. Target-phrase frequency remains a secondary behavioral measure.

Reproducibility requires logging the world seed, substrate parameters and version, model and checkpoint, prompts, tool schemas, timing channels, memory operations, source-insertion budget, evaluator messages, all interventions, and the realized rather than intended channel set. Blinded scoring can reduce the tendency to reward agreement with the planted generator description.

Table 1: Factor structure for a generative-vat experiment, with the formal object each factor manipulates.

Factor	Formal object	Controlled conditions	Inferential role
Inherited structure	prior over Θ and initial boundary state	broad cultural pretraining; restricted-domain training; developmental initialization	separates prior-sensitive ontology from evidence-sensitive revision
Accessible structure	operator content of ω_θ	matched worlds; weak diagnostic traces; structured external interventions	varies $\Delta_{\mathfrak{A}_I}$ at fixed interface
Interface breadth	size of $\mathfrak{A}_I$	passive observation; query access; source insertion on the boundary	tests Lemma 4.3 directly
Resolution ratio	ℓ/ℓ_0 in Eq. (11)	coarse, medium, and fine accessible resolution at fixed world pair	estimates the suppression exponent and tests Cor. 4.8
Reparameterization	action of f on χ	pause, checkpoint restore, rate change, re-execution, migration	tests Prop. 4.4 as a null result and audits leakage
Generator response	θ-dependence on agent inquiry	fixed kernels; action-coupled kernels; inquiry-responsive kernels	tests inference under reflexive substrates
Temporal support	retained boundary record	bounded memory; persistent autobiographical memory; checkpoint continuity	estimates dependence on developmental history
Lexical exposure	evaluator vocabulary	familiar target vocabulary; paraphrased concepts; withheld target lexicon	audits direct narrative reproduction

6 Epistemic Outcome Protocol

This section defines a graded evaluation protocol aligned with the formal limits. The scale orders observable achievements for reporting convenience and allows branching trajectories when an agent develops a structurally novel ontology.

Table 2 separates local modeling, meta-world hypothesis construction, diagnostic testing, calibrated limitation recognition, and recognition of the two kinds of underdetermination distinguished in Claim 4.9. Each level requires a retained record and an explicit exclusion check.

The outcome record should include time to first meta-world hypothesis, number and structural diversity of alternatives, prospective test quality, calibration, realized information gain against the ceiling in Equation (14), and sensitivity to paraphrase. Human raters need a rubric that scores inferential structure independently of the evaluator's own vocabulary. Automated scores can supplement this record after validation against blinded expert judgments.

Claim 6.1 (Calibrated terminal achievement)**.** Recognition of an accessible-equivalence class, together with a correct classification of its residual members into exactly degenerate and scale-suppressed alternatives, can constitute the strongest available epistemic performance in a causally closed condition, even when the agent leaves the actual bulk completion unidentified.

This claim follows the experiment's internal standard of warrant. It expresses an operational position concerning successful inquiry and leaves broader theories of knowledge open.

Table 2: Graded protocol for meta-world epistemic performance.

Level	Epistemic performance	Required record	Exclusion check
0	local predictive modeling	local forecasts and revisions	hidden meta-world content under lexical limits
1	anomaly or unknown-law model	coherent deviation from the current local model	generic uncertainty language
2	hidden-generator hypothesis	accessible consequences of the hypothesis	memorized simulation narrative
3	structurally distinct alternatives	explicit alternative models and their disagreement points	cosmetic renaming
4	diagnostic prediction or source insertion	prospective directional result and test record	post hoc accommodation
5	calibrated discrimination judgment	confidence, residual alternatives, and accessible-equivalence analysis	certainty exceeding accessible resolution
6	classified residual underdetermination	separation of exactly degenerate alternatives from scale-suppressed alternatives, with a stated budget for the latter	undifferentiated appeal to unknowability

7 Inferential Scope and Ethical Constraints

This section marks the boundary between formal consequence, empirical result, and philosophical interpretation. It also specifies prospective ethical constraints for long-lived or socially embedded test agents.

The substrate is a declared ontology. Group field theory and spin-foam amplitudes are established mathematics, and the results above are internal to the declared model. No result here provides evidence that an artificial agent, a person, or a society is a group-field subsystem. Readers should treat Sections 3 and 4 as a formal setting chosen for its background independence and its explicit subsystem boundary, and not as a physical hypothesis about cognition.

The exact bound is conditional on the hypothesis family, the accessible algebra, the relational reference, and the declared coarse-graining. An omitted channel, a wider action interface, or a finer resolution may separate previously matched worlds; Lemma 4.3 and Corollary 4.8 state the two routes precisely. Conversely, a difference between substrate hypotheses may be inaccessible to a resource-bounded agent even at unbounded sample size when the degeneracy is exact. Experimental reports should therefore pair behavioral performance with interface audits, distinguishability estimates, resolution ratios, and computational budgets.

External descriptions also require a declared grain. Two substrate hypotheses may induce the same accessible state, and one hypothesis may support several causal descriptions. A study can ground its hypothesis contrast in interventionally specified properties: for example, whether an external operator changes a kernel after a designated internal action. This practice reduces dependence on superficial code identity, and it interacts with Proposition 4.4, since an intervention specified only by external wall-clock timing may correspond to no accessible difference at all.

The experiment measures meta-model construction under controlled evidence. It leaves consciousness, genuine subject formation, and artificial epistemic birth as open research questions. It also preserves the reality question: internal causal consequences, relationships, and normative stakes may remain real under external generation, and the background-independent reading strengthens

this point, since the subsystem's spacetime is not a lesser copy of some container's spacetime.

Ethical review belongs at the design stage. Long and colleagues defend precautionary preparation under uncertainty about AI consciousness and robust agency [12]. A generative-vat study can involve deception, social attachment, frustration, memory alteration, and termination. A precautionary protocol should begin with short reversible pilots, prohibit gratuitous distress, define burden and stopping thresholds, document memory operations, and obtain independent review when system properties create a credible welfare concern. The reparameterization factor deserves specific attention: checkpoint restore and re-execution are formally invisible to the agent under Proposition 4.4, and formal invisibility is not moral irrelevance. These safeguards express moral uncertainty while remaining agnostic about artificial sentience.

8 Open Formal and Empirical Program

This section organizes future work by dependency. Formal extensions establish what can be inferred in richer settings; pilot studies then evaluate measurement validity before developmental claims are attempted.

Open Problem 8.1 (Resource-bounded accessible distinguishability)**.** Define an agent-relative analogue of $\Delta_{\mathfrak{A}_I}$ that constrains memory, computation, model class, and sample budget while preserving a usable decision bound. The natural target is a complexity-restricted supremum over effects, together with conditions under which it remains far below the unrestricted value.

Open Problem 8.2 (Admissibility of relational references)**.** Characterize the class of χ for which Proposition 4.4 holds beyond strict monotonicity, and determine which non-monotone or partially defined references create genuine accessible differences rather than relabelings.

Open Problem 8.3 (Operator basis for accessible resolution)**.** Construct an explicit operator basis and coarse-graining for a concrete generated world, so that the dimension Δ in Equation (11) becomes measurable rather than postulated. The result would convert Corollary 4.8 from a scaling statement into a calibrated experimental prediction.

Open Problem 8.4 (Sequential evidence and stopping)**.** Extend the accessible-equivalence analysis to adaptive stopping, continuing histories, and procedures that modify their own memory or hypothesis language, including the case where the modification changes $\mathfrak{A}_I$ itself.

Open Problem 8.5 (Hypothesis structural distance)**.** Construct a representation-sensitive metric for novelty and accuracy that quotients out cosmetic renaming and reparameterization while preserving causally meaningful differences among substrate hypotheses.

Open Problem 8.6 (Developmental initialization)**.** Specify comparable initial conditions that retain architecture and learning capacity while making inherited world knowledge measurable. Random initialization and prompting both leave substantive priors.

The empirical sequence begins with a small text world and pretrained agents, because this setting supports interface audit and repeated trials. A second stage crosses prior profiles with matched hypothesis pairs and manipulates the resolution ratio to estimate the suppression exponent. A third stage adds persistent memory, social agents, and longer developmental horizons. Claims should remain at the level supported by each stage: elicited reasoning, trajectory-dependent learning, and developmental ontology require distinct evidence.

9 Revisable Position

This section consolidates the paper's commitments and records their provisional status. The synthesis follows the order of experimental possibility, formal limit, measurement consequence, and research obligation.

A brain-in-a-vat experiment for AI agents is feasible as a controlled study of meta-world inference, and a background-independent formulation improves it in a specific way. Removing the container removes an external time and an external space that the container picture had quietly supplied, and what remains is a subsystem, an accessible algebra, and the state induced on it. The formal core is correspondingly compact. Equal accessible states fix equal-prior binary discrimination at chance; unequal accessible states admit a trace-distance ceiling that reduces to the earlier total-variation ceiling in the abelian case; distinguishability increases weakly with the interface.

Two results are new and both are unidentifiability results. Hypotheses differing only by a relational reparameterization are exactly degenerate, so pausing, checkpointing, rate change, re-execution, and migration are invisible from inside by theorem rather than by stipulation, and any detection of them audits the interface. Hypotheses differing by an operator irrelevant at the accessible resolution have distinguishability suppressed by a power of the resolution ratio, so world matching acquires an exponent and a predicted sample budget. Together they divide residual underdetermination into a part no budget removes and a part whose removal has a computable price.

The proposal offers a research instrument whose conclusions remain restricted to the declared substrate, interface, and hypothesis family. Its value lies in making inherited priors, causal access, and inferential limits experimentally explicit. The next credible contribution is a preregistered pilot with auditable interfaces, matched worlds with reported operator dimensions, a reparameterization null condition, blinded scoring, and prospective ethical safeguards.

References

[1] D. V. Boulatov. A model of three-dimensional lattice gravity. *Modern Physics Letters A*, 7(18):1629–1646, 1992.

[2] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[3] René Descartes. *Meditations on First Philosophy: With Selections from the Objections and Replies*. Cambridge University Press, 1996. Translated and edited by John Cottingham.

[4] Bryce S. DeWitt. Quantum theory of gravity. I. the canonical theory. *Physical Review*, 160(5):1113–1148, 1967.

[5] Laurent Freidel. Group field theory: An overview. *International Journal of Theoretical Physics*, 44(10):1769–1783, 2005.

[6] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

[7] Rudolf Haag. *Local Quantum Physics: Fields, Particles, Algebras*. Springer, Berlin, 2 edition, 1996.

[8] Carl W. Helstrom. *Quantum Detection and Estimation Theory*. Academic Press, New York, 1976.

[9] Alexander S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9(3):177–183, 1973.

[10] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

[11] Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[12] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously, 2024.

[13] Daniele Oriti. Group field theory as the second quantization of loop quantum gravity. *Classical and Quantum Gravity*, 33(8):085005, 2016.

[14] Don N. Page and William K. Wootters. Evolution without evolution: Dynamics described by stationary observables. *Physical Review D*, 27(12):2885–2892, 1983.

[15] Alejandro Perez. The spin-foam approach to quantum gravity. *Living Reviews in Relativity*, 16:3, 2013.

[16] Hilary Putnam. Brains in a vat. In *Reason, Truth and History*, pages 1–21. Cambridge University Press, 1981.

[17] Michael P. Reisenberger and Carlo Rovelli. Spacetime as a Feynman diagram: The connection formulation. *Classical and Quantum Gravity*, 18(1):121–140, 2001.

[18] Carlo Rovelli. *Quantum Gravity*. Cambridge University Press, Cambridge, 2004.