

What Must a Model Preserve? Synthetic Tests Toward Evaluating Artificial Knowledge

Volume 1 · Issue 1 · 3 October 2026

Wolfram Horstmann · Humboldt-Universität zu Berlin

Abstract

Scientific knowledge produced through models depends not merely on obtaining an output, but on what a representation preserves and which inferences that preservation warrants. This problem becomes especially consequential when artificial systems generate predictions, transformations, or scholarly claims that may appear knowledgeable before their relation to a target has been tested. Starting from Rosen's modelling relation, this article distinguishes predictive, structural, and purpose-relative adequacy. Bounded synthetic studies provide background evidence for the distinction: a declared predictive rule accepted a same-family case, rejected an obvious misspecification, and preserved an invalid evaluation as undefined; coarse states concealed different successors and paths; and richer representations recovered different information at different descriptive costs. These studies test representational predicates, not an AI knowledge-production system. Rather than treating the local results as general validation, the article uses them to formulate three prospective experiments with explicit failure conditions. It then extends the question tentatively to scholarly production itself: whether transformations among research records, AI-produced articles, reviews, and revisions preserve claims, evidential relations, uncertainty, and declared purpose. The proposals position the synthetic results as methodological steps toward evaluating artificial knowledge and, later, iterative and potentially scaled AI-supported science. Correspondence is not presented as sufficient for knowledge, purpose-relative adequacy remains untested, and increased throughput is not equated with epistemic progress.

1. Introduction: the problem of modelling change

Scientific models are epistemic instruments: investigators operate on them to draw inferences about targets that may be difficult to observe, manipulate, or reason about directly (Hughes, 1997; Suárez, 2004). Scientific representation therefore concerns more than associating one object with another. It concerns what inferences the association permits, how those inferences are interpreted, and under which questions they are justified. A fluent conclusion, accurate prediction, or structurally detailed simulation is not knowledge merely by virtue of having been produced.

This matters for **artificial knowledge**. The term is used here cautiously for epistemic products—predictions, inferences, and structured claims—generated or materially transformed by artificial systems and offered for scientific use. It does not presuppose that the producing system is an epistemic agent or that every successful output constitutes knowledge. It instead names an evaluation problem: when artificial operations produce a claim about a target, which features of the target–representation relation make that claim scientifically usable and open to correction? Questions of truth, justification, explanation,

provenance, and social acceptance extend beyond representational correspondence; but correspondence is one place where such claims can be made explicit and tested rather than inferred from an output's plausibility.

Three labels must therefore be kept apart. **Automated scientific output** is any prediction, text, dataset, or other result produced through an automated operation; it is a production category and carries no knowledge claim by itself. **AI-supported science** denotes a research process in which artificial systems contribute materially within evidential, critical, and organizational controls; it likewise does not make every output knowledge. **Artificial knowledge**, as used here, names the narrower evaluative question of whether an artificially generated or transformed epistemic product warrants scientific use. For a product even to enter the evaluation proposed here, it must expose at least a declared target, a representation and transformation, a claim or observation, an intended use with an evaluative criterion, and enough evidence and provenance for the relation to be challenged and corrected. These are minimum conditions of assessability, not jointly sufficient conditions for knowledge.

Picture a toy model intended to predict where a fly on a tabletop will move next. The model records only whether the fly is near or far from a window, but not the direction in which it is moving. Two situations can therefore look identical to the model even though the fly is moving toward the window in one and away from it in the other. The model may describe the fly's present category correctly while remaining unable to predict its next category. Adding direction may improve that prediction, but the extra detail may be unnecessary if the task is only to count how often the fly enters the area near the window. The example exposes three different questions: whether a prediction is accurate, whether the representation preserves consequential change, and whether the retained information is adequate for the intended use. An artificial system that reasons from this representation inherits the same limitations; the plausibility of its answer cannot compensate for information that the representation discarded.

The same design problem becomes practically consequential when a model guides a drone. A representation that retains distance from a target but discards heading or motion may be adequate for displaying a present location category and inadequate for predicting the next position or avoiding a collision. The fly and drone cases are illustrations of the representational problem, not validated biological or engineering models.

An initial match does not guarantee that a model remains informative after the represented system changes. If two detailed states receive the same simplified label but have different possible successors, the label may be adequate for reporting their present category and inadequate for predicting their next movement. If two paths finish in the same category, an endpoint comparison may miss a difference in when a threshold was crossed. Neither result alone says whether the representation is useful for a particular scientific purpose.

Rosen's modelling relation is a natural point of departure. It brings causal entailment in a natural system and inferential entailment in a formal system into a relationship through encoding and decoding or interpretation. His earlier treatment expresses modelling through mappings and commutative diagrams in predictive contexts (Rosen, 1978); his subsequent work also treats dynamics, time, error, and the stability of modelling relations (Rosen, 1985, 1991). The present article therefore does **not** propose to add dynamics to a supposedly static Rosen theory. It takes a restricted local alignment condition as a reference and asks which different claims must be checked when actual representations are coarse, partial, or evaluated for different uses.

An empirically testable approach is needed because neither formal coherence nor a persuasive output establishes that a model preserves the features on which a knowledge claim depends. This is particularly important for AI-based knowledge production. Artificial systems can generate and transform claims rapidly, while the resulting text or prediction may leave unclear whether an observation, a transition, or a task-relevant distinction was retained. Operational tests make those dependencies answerable to evidence: the target and representation can be declared, predictions can be committed before held-out observations are opened, transitions can be evaluated as preserved, failed, or undefined, and purposes can be fixed before competing representations are compared. Synthetic tests do not by themselves validate claims about natural or engineered systems, but they allow the proposed distinctions to fail under controlled conditions before they are applied in more complex settings.

The central question is therefore a minimal one for evaluating model-mediated and artificially produced knowledge claims:

What must a model preserve as the represented system changes, and how can distinct claims of predictive, structural, and purpose-relative adequacy be tested without assuming exact correspondence everywhere?

The contribution sought here is a selective *operational decomposition* of this question. A mapping between states is distinguished from an adequate observation prediction; a prediction is distinguished from preservation of a transition or path; and preservation is distinguished from usefulness for a declared purpose. Each distinction creates a different place where a knowledge claim can succeed, fail, or remain undefined. Together they provide a small, challengeable vocabulary for subsequent experiments without requiring one all-purpose measure of correspondence. The paper does not claim that these distinctions or their formal ingredients were previously unknown. It reports what a small sequence of synthetic tests made visible, where that evidence ends, and how the proposal could be challenged by established accounts and further experiments.

2. Theory: three claims that should not be collapsed

Representation connects modelling to knowledge because it licenses reasoning with one system in place of another. Hughes's denotation–demonstration– interpretation account makes the movement explicit: a model denotes a target, is used to demonstrate a result, and requires interpretation back toward that target (Hughes, 1997). Inferential approaches likewise make surrogate reasoning central, while emphasizing that the relevant consequences must be scientifically justified answers to questions about the target (Khalifa et al., 2022; Suárez, 2004). The present decomposition does not replace these accounts. It asks which more specific adequacy claim is being made when an inference concerns an observation, a transition or path, or a declared use.

This distinction is useful for artificial knowledge because an artificial system can transform a representation and produce a conclusion without thereby settling what the conclusion preserves. The three dimensions below locate separate burdens of evaluation. They are necessary questions within the scope of this article, not jointly sufficient conditions for knowledge.

2.1 A restricted alignment reference

For a selected represented state w_1 , a local, encoding-side Rosen-inspired condition can be written

$$R(E(w_1)) = E(L(w_1)).$$

Here, $w_1 \in W$ is a state of the represented system; $L : W \rightarrow W$ is its declared operation; $E : W \rightarrow M$ encodes or represents that state in a modelling system M ; and $R : M \rightarrow M$ is a model operation or inference. The equation asks whether operating after encoding reaches the same model state as encoding after the represented-system operation. It is a local condition for specified states and operations, not an assumed identity for every state, and it is **not** Rosen's complete modelling relation: decoding, interpretation, and the distinction between causal and inferential entailment also belong to that fuller account. It supplies a stringent reference case against which less exact or differently evaluated relations can be examined.

The synthetic tests below do not instantiate a natural system in Rosen's full sense. Their targets are generated or constructed systems. They test whether selected distinctions can be specified and evaluated, not whether a formal model has been empirically validated against nature.

2.2 A small carrier for transition questions

For explicit change, let

$$S_A = (X_A, T_A), \quad S_B = (X_B, T_B), \quad e : X_A \rightarrow X_B.$$

X_A and X_B are typed state spaces of two distinct systems; $T_A \subseteq X_A \times X_A$ and $T_B \subseteq X_B \times X_B$ are their permitted directed transitions; and e is a partial directed correspondence. A transition $x_k \rightarrow x_{k+1}$ uses a discrete index k , not necessarily clock or physical time. Partiality permits an assessment to be *undefined* when a necessary state correspondence is absent. The tuple

$$\mathcal{C} = (X_A, T_A, X_B, T_B, e)$$

is a candidate comparison carrier: it says which states, transitions, and correspondence have been declared, but it does not pronounce them adequate. Observations, purposes, explicit boundaries, mechanisms, and relation changes can be added only where a test needs them. The carrier's minimality is a working design choice, not a demonstrated theorem; simulation relations or other established structures may express the same content more economically.

The carrier can be read as a relational relaxation of the restricted Rosen condition rather than as a new system theory:

Restricted Rosen reference	Transition carrier	Relation between them
represented states W	X_A	X_A supplies the states being represented
represented operation L	T_A	the graph of a deterministic L is a special case of T_A
model states M	X_B	X_B supplies the representing states
encoding E	partial correspondence e	e permits missing state correspondences
model operation R	T_B	the graph of a deterministic R is a special case of T_B

When L , R , and $E = e$ are total functions, one-step preservation in the carrier reduces to the restricted commutation condition. Relations, partiality, branching, path comparisons, observations, and declared purposes extend the kinds of question that can be asked; they are not derived from the restricted equality

alone.

2.3 Predictive, structural, and purpose-relative adequacy

Predictive adequacy concerns a declared observation. In the first synthetic test, an input condition s has a model prediction $\hat{y}_s$ and a target response y_s . When both are defined, the local rule is

$$A_\epsilon(s) = \mathbf{1}[|y_s - \hat{y}_s| \leq \epsilon],$$

where ϵ is a tolerance fixed before the held-out responses are revealed. If prediction, observation, or required alignment is missing, the comparison is undefined rather than automatically successful or failed. Passing this rule is an observation-level result, not proof of a correct mechanism or preserved trajectory.

Structural adequacy concerns change itself. For a declared transition $(x, x') \in T_A$, define the local verdict

$$V_1(x, x') = \begin{cases} \text{undefined,} & e(x) \text{ or } e(x') \text{ is undefined,} \\ \text{preserved,} & (e(x), e(x')) \in T_B, \\ \text{failed,} & (e(x), e(x')) \notin T_B. \end{cases}$$

For deterministic operations F_A and F_B , the preserved case reduces to $e(F_A(x)) = F_B(e(x))$. For a coarse state $b \in X_B$, let $D_b \subseteq T_A$ be the set of detailed transitions declared before evaluation as originating in b . Complete mapping coverage requires

$$\text{Cov}_e(D_b, b) \iff D_b \neq \emptyset \wedge \forall (x, x') \in D_b : e(x) = b \wedge e(x') \text{ is defined.}$$

Only when this condition holds is the mapped successor set defined:

$$\text{Succ}_{A/e}(b; D_b) = \{e(x') \mid (x, x') \in D_b\}.$$

If the declared set is empty or any required source or successor mapping is missing, the successor-set verdict is undefined; unmapped transitions may not be silently removed. Comparing a defined set with $\text{Succ}_B(b) = \{b' \mid (b, b') \in T_B\}$ requires the claim to be fixed in advance. Equality asks for exact correspondence, inclusion $\text{Succ}_{A/e}(b; D_b) \subseteq \text{Succ}_B(b)$ asks whether the model covers generated alternatives, and the reverse inclusion asks whether every modelled alternative is generated within the declared domain. A unique deterministic coarse successor additionally requires

$$|\text{Succ}_{A/e}(b; D_b)| = 1.$$

For a finite path $\pi = (x_0, \dots, x_n)$, its mapped path $e(\pi) = (e(x_0), \dots, e(x_n))$ is defined only when every required mapping is defined. Equality of final mapped states does not imply equality of mapped paths. These one-step, branching, and path predicates are related but not interchangeable, and they do not form one universal structural score.

Purpose-relative adequacy asks whether a representation r retains the information required for a declared task q at an acceptable descriptive cost. A minimal prospective template fixes a task-performance measure $P_q(r)$, a performance threshold τ_q , a descriptive-cost measure $C(r)$, and an admissible cost κ_q before comparison:

$$V_q(r) = \begin{cases} \text{undefined,} & q, P_q, \tau_q, C, \text{ or } \kappa_q \text{ is not fixed,} \\ \text{adequate,} & P_q(r) \geq \tau_q \text{ and } C(r) \leq \kappa_q, \\ \text{inadequate,} & \text{otherwise.} \end{cases}$$

Predicting a next category, identifying a reachable category, and distinguishing a steering mechanism can require different information. No universal metric, threshold, or cost unit is proposed. A purpose and its decision rule must be fixed *before* comparing representations; otherwise they can be used retrospectively to excuse whichever representation happened to perform well. This template is a research proposal in the present article, not an experimentally established result.

The three dimensions may coincide in a particular case but do not follow from one another. State alignment does not entail predictive adequacy; predictive adequacy does not entail transition or mechanism preservation; and structural detail is not unconditionally valuable without a task (Bokulich & Parker, 2021; Suárez, 2004).

The contribution is therefore an operational taxonomy for separating test claims, not a claim to a novel formal object. Its present evidential status is:

Dimension	Operational object	Present status	Principal limit
predictive	$A_c(s)$	locally tested with same-family and constant synthetic models	no near-miss discrimination test
one-step structural	$V_1(x, x')$	tested in constructed and mechanism-generated cases	local declared domains only
branching and path structural	successor sets and mapped paths	explored using reused synthetic trajectories	no completed same-case comparison with an established formalism
purpose-relative	$V_q(r)$ template	prospective and untested	task metrics, thresholds, and costs remain to be fixed experimentally

2.4 Position among adjacent accounts

The proposal is positioned here through a bounded, claim-led selection rather than a systematic literature review. The selected accounts are the closest formal or philosophical comparators for the operations used in this article: commutation, preservation of transitions, aggregation of states, retention of paths, and adequacy for a declared use. The comparison asks the same questions of each account—what relation is evaluated, what kind of failure it can expose, and what remains outside its immediate scope:

Account	Principal object or relation	Question made precise	Relation to the present taxonomy
Rosen's modelling relation	causal and inferential entailments connected by encoding and decoding	whether inference in a model is appropriately related to operation in a represented system	supplies the broader modelling frame; the local equality in Section 2.1 is only a restricted reference
Simulation and approximate simulation	state relations that preserve possible transitions exactly or within a metric	whether one transition system can match the behaviour of another	may absorb much of the structural predicate; it does not by itself select the scientifically relevant observation map, task, or admissible descriptive cost (Girard & Pappas, 2007)
Abstraction and lumpability	grouped states whose aggregate dynamics remain well defined	when coarse states support closed deterministic, nondeterministic, or probabilistic dynamics	directly bears on branching under coarse mappings; the present tests do not evaluate probabilistic lumpability (Kemeny & Snell, 1960)
Symbolic dynamics	symbolic sequences induced by partitions of a dynamical system	which path distinctions survive symbolic coding	provides established machinery for trajectory-sensitive comparison; it may express the path component more economically (Lind & Marcus, 1995)

Account	Principal object or relation	Question made precise	Relation to the present taxonomy
Adequacy for purpose	representational performance relative to a specified scientific use	which properties of a representation matter for a question or decision	motivates the prospective task-and-cost predicate rather than one unconditional ranking (Bokulich & Parker, 2021)

What remains after this comparison is not a claimed new formalism. It is a small operational synthesis that places observation tests, transition and path tests, and purpose-relative tests in one inspectable sequence while preserving failure and undefined outcomes. A completed same-case translation could show that an established account expresses any individual component better; that would refine the synthesis rather than count against its central separation of evaluative burdens. The selection excludes broader literatures on model-based explanation, measurement, and social epistemology except where they bear directly on these declared comparisons. Its limits should therefore not be mistaken for evidence that no closer or more economical formal treatment exists.

3. Methods: synthetic tests of distinct correspondence claims

The tests move from observation comparison to local transitions, mechanism-generated dynamics, multi-step paths, and alternative representations. Each test posed a distinct correspondence question and used its own specified system, comparison rule, and failure conditions. The sections that follow report those tests together to show how the questions accumulate; they are not independent replications or one prospectively pre-registered multi-part experiment.

For observation-level prediction, a synthetic stimulus–response characteristic supplied seven training conditions and four held-out conditions. A one-parameter model was fitted before the held-out responses were opened. Predictions were committed first, and an inclusive absolute-error tolerance of 0.06 was fixed. A negative control then used a deliberately constant predictor against a stimulus-dependent target. Its first attempted evaluation became invalid after a post-reveal evaluator failure; that attempt was preserved as undefined. A governed repeat used a checked executable environment and new held-out responses before evaluating the same kind of negative control. The word “characteristic” refers to the shape of the synthetic input–output relation; no neuron or biological response was observed.

The first structural test used a deliberately small, constructed pair of transition systems. Seven detailed positions and a partial coarse mapping allowed declared cases of preservation, endpoint mismatch, and missing correspondence. This test checked operational differentiation, not uncertain behavior discovered from an independent mechanism.

The later structural tests used a two-sensor, two-wheel, Braitenberg-style synthetic vehicle (Braitenberg, 1984) as a mechanism-generated example. Its detailed state contained position and heading. A radial representation reduced that state to **goal**, **near**, or **far** relative to a source. The fly, drone, and vehicle presentations instantiate the same abstract pattern: a coarse location category retains distance-related information while discarding heading or motion that may determine what happens next. For sensor distance d , the signal was $I(d) = 1/(1 + d^2)$. With left and right wheel speeds v_L, v_R , wheel base b , and Euler step Δ , the implementation used

$$v = \frac{v_L + v_R}{2}, \quad \omega = \frac{v_R - v_L}{b}.$$

Sensor offsets were 0.20 forward and 0.10 lateral; wheel base was 0.40, gain 1.0, and step size 0.50. The radial thresholds were $G : r \leq 0.75$, $N : 0.75 < r \leq 2.00$, and $F : r > 2.00$. The state grid intentionally stressed those thresholds: eight radii 0.70, 0.74, 0.76, 0.80, 1.95, 1.99, 2.01, 2.05 were crossed with four headings $0, \pi/2, \pi, -\pi/2$. One Euler step generated 32 detailed transitions. Paired crossed and uncrossed wheel mechanisms were subsequently run for eight steps from 32 common starting states. A final comparison constructed branching, path-sensitive, and orientation-enriched representations and examined a held-out subset of 16 cases. That split reused existing synthetic trajectories and was not a blind replication.

Study-to-evidence map. Paths in the last column are relative to the extracted standalone supplement directory.

Test	Target or representation	Frozen design and evaluation status	Primary reported outcome	Failure or undefined condition	Supplement entry
Same-family prediction	synthetic logistic stimulus-response characteristic	7 training and 4 prospectively held-out conditions; $A_{0.06}$	4/4 passed; maximum error 0.02336968	missing required value is undefined; error above 0.06 fails	source/prototype/generate_synthetic_neuron_characteristic_v02.py; source/artifacts/synthetic-neuron-characteristic-v0.2/; matching reference/ entries
Constant misspecification	constant predictor against a fresh logistic target	7 training and 4 sealed holdout conditions; governed repeat after an invalid predecessor	4/4 failed; errors 0.09567552-0.38962421	predecessor evaluator failure remains traceable as undefined; governed errors above 0.06 fail	three programs under source/prototype/: generate_autonomously_misspecification_v02.py, fit_constant_misspecification_v02.py, and evaluate_autonomously_misspecification_v02.py; matching reference/ entries
Constructed transitions	7 integer positions, 3 coarse states, partial map covering 5 positions	4 prospectively declared transition pairs	2 preserved, 1 failed, 1 undefined	endpoint mismatch fails; missing correspondence is undefined	source/artifacts/s-v0.4-braitenberg/input.json; source/prototype/evaluate_transition_correspondence.py; matching reference/ entries
One-step vehicle	32 radius-heading microstates under crossed coupling	frozen mechanism, grid, thresholds, and persistence prediction; no holdout	20 preserved and 12 failed; every coarse state branched	mapped successor differing from predicted persistence fails	source/artifacts/s-v0.5-braitenberg/config.json; source/prototype/run_braitenberg_dynamic_consistency.py; matching reference/ entries
Eight-step vehicle	32 common starts under crossed and uncrossed mechanisms	8 Euler steps; no blind holdout	0 final-category differences, 16 final-radius differences, 4 path differences	endpoint agreement is retained as a negative result; mapped paths may still differ	source/artifacts/s-v0.6-multistep/config.json; source/prototype/run_braitenberg_multistep.py; matching reference/ entries

Test	Target or representation	Frozen design and evaluation status	Primary reported outcome	Failure or undefined condition	Supplement entry
Representation comparison	branching, mapped paths, and orientation-enriched branching	16 construction and 16 evaluation cases split from existing trajectories; not blind replication	branching variants covered 16/16 endpoints; paths distinguished 2/16 pairs; enriched mean width 1.3333	no common task metric or aggregate ranking was declared	source/artifacts/s-v0.7-comparative/config.json; source/prototype/run_comparative_representation.py; matching reference/entries

These settings are entirely computational. They contain no biological data, physical robot, empirical measurement, or evidence that the selected rules describe naturally occurring systems.

4. Results

4.1 Prediction can pass, fail, or be undefined

The same-family stimulus–response model passed all four prospectively held-out comparisons. The largest absolute error was 0.02336968, below the fixed 0.06 tolerance. The result showed that state/input alignment alone did not determine the verdict: a typed observation and predeclared comparison rule were required.

The first constant-model negative-control attempt supplied **no valid scholarly test result** because the evaluator failed after reveal. It was not counted as a weak pass or numerical failure. In the governed repeat, all four constant-model comparisons were defined and all four failed: errors ranged from 0.09567552 to 0.38962421, each above 0.06. The protocol therefore rejected this declared mismatch rather than producing success automatically.

This contrast is deliberately easy. The positive response was generated from the fitted function family, while the negative was an extreme constant misspecification. The evidence establishes executable local discrimination between these two synthetic cases; it does not establish calibrated model selection, robustness to plausible alternatives, statistical power, or a general prediction theory.

4.2 A mapped state need not preserve its next state

The constructed local-transition evaluation returned two preserved cases, one defined failure, and one undefined case:

$$(1, 1, 0, \perp_{\text{map}}).$$

Only five of seven detailed positions were mapped. Two cases agreed in their starting coarse-state signature but differed after their endpoint was evaluated. The static starting match therefore did not decide whether the declared transition was preserved. Because the examples were selected to exercise the predicate, this result demonstrates inspectable operational differentiation, not independent empirical confirmation of a new law.

The mechanism-generated one-step vehicle study produced a stronger stress case. Of 32 frozen detailed transitions, 20 preserved the proposed coarse persistence outcome and 12 failed it. Every starting category had more than one observed coarse successor. At each of eight sampled radii, changing heading while

keeping radius fixed yielded different successor categories. The radial category consequently did **not** define a unique coarse successor for this mechanism, grid, and step size. Heading was information discarded by the coarse map but consequential for the generated transition. This does not make the goal/near/far representation useless: it makes unique next-category prediction an unsupported use of that representation in the tested setting.

4.3 Equal endpoints can conceal paths; responses have different costs

The eight-step paired-mechanism test found no crossed/uncrossed pair with a different final coarse category among 32 starts. That negative endpoint result was retained. Sixteen pairs nevertheless had different final radii, and four followed different intermediate coarse-category paths. The paths made the timing of threshold crossing visible where the endpoint category did not. This supports a bounded reason for path retention when route or timing matters, not a claim that paths are always superior.

The final representational comparison produced the following outcomes on its held-out evaluation subset:

Response	Information retained	Held-out observation	Descriptive cost
Coarse branching	possible endpoint categories	16/16 endpoints covered; mean branching width 2.0	3 source keys, 6 relation pairs
Ordered paths	intermediate category order	2/16 paired mechanisms distinguished despite equal endpoint labels	longer represented object
Orientation-enriched branching	coarse category plus heading-related distinction	16/16 endpoints covered; mean branching width 1.3333	12 source keys, 16 relation pairs

Enrichment reduced local ambiguity but did not eliminate branching. Its lower mean branching width partly follows from dividing examples among more keys; it is not a complexity-adjusted gain. The path comparison used a different performance question from endpoint coverage, so the three rows do not supply one common ranking. The results identify trade-offs. They do **not** decide which response is adequate for a scientific purpose that was never declared in the comparison.

5. Discussion

5.1 What the studies distinguish

The completed tests support a narrow but important distinction: a claim that two states correspond cannot stand in for a claim that the model predicts an observation, preserves a transition, retains a path difference, or serves a purpose. The positive and negative characteristic cases show observation adequacy can be evaluated separately from input alignment. The local transition cases distinguish preservation, contradiction, and missing comparison. The generated vehicle cases show why one coarse state label may hide different successors. The multi-step and alternative-representation cases show why more information is not one uniformly better answer.

The transition carrier kept *construction of a comparison* separate from *evaluation of its success*. It provided a stable place to specify states, allowed changes, and a directed mapping while leaving observation functions, mechanisms, boundaries, path memory, and tasks optional. That modularity is methodologically useful but not proof that the carrier is an irreducible new formal object. Existing simulation and approximate-simulation relations may already express transition preservation under

specified conditions (Girard & Pappas, 2007). Deterministic quotients and nondeterministic abstractions offer ways to understand grouped-state dynamics; probabilistic lumpability poses a different condition that these deterministic tests did not evaluate (Kemeny & Snell, 1960). Symbolic dynamics treats path sequences through partitions (Lind & Marcus, 1995). An honest comparison may show that the carrier should be translated into, or absorbed by, an established formalism.

The same caution applies to Rosen. His modelling relation is not one state equation, and the restricted equality in Section 2.1 is not a claim that his larger account is silent about change or error. A future comparison must express the *same* synthetic case in the fuller Rosen vocabulary and in the present one, then ask whether any distinction here remains unavailable there. Formal redundancy is a legitimate outcome, not a failed publication strategy.

5.2 From representational correspondence to artificial knowledge

The experiments suggest a disciplined way to approach claims of artificial knowledge without deciding in advance whether an artificial system itself knows. They test components of a representational evaluation, not AI-supported science or an AI-produced scholarly claim as such. The primary object of evaluation is the epistemic product and its traceable relations: what target was represented, what operation produced the result, which observation or structural feature was preserved, and what use the result is supposed to support. On this view, provenance and reproducibility can make an artificial contribution inspectable, but they do not replace tests of what its representation preserves.

Correspondence is therefore neither a complete theory of knowledge nor a dispensable technical detail. Predictive adequacy asks whether a declared observation is recovered under a fixed rule. Structural adequacy asks whether the transformation preserves the changes on which an inference depends. Purpose-relative adequacy asks whether the retained information is appropriate for the question being answered. Because these burdens can be tested separately, later studies can strengthen, revise, or reject one dimension without treating every artificial knowledge claim as one indivisible success or failure.

This makes the framework productive beyond the examples reported here. It provides explicit objects for criticism: the encoding, operation, observation, transition, task, and cost. New experiments can vary one of these while holding the others fixed; formal comparisons can ask whether established theories already capture the same distinction; and negative results can identify which kind of epistemic claim failed. The value of the decomposition lies in enabling those tests, not in conferring knowledge status on AI output by definition.

The same logic can later be applied reflexively to scholarly production. A research record can be treated as a represented system and an AI-produced article as a representation generated from it. The question is then which claims, evidential relations, uncertainties, and declared purposes survive the transformation into a manuscript and through subsequent review and revision. This possibility follows from the framework but has not been tested by the experiments reported here.

5.3 Evidential limits

The studies provide local synthetic evidence, not natural-system validation. The positive prediction example was easy by construction; the negative model was conspicuously wrong. The local-transition example was also construction-driven. The vehicle grid concentrated threshold cases and used one sensor law, parameter set, Euler integrator, and eight-step horizon. Its counts cannot be generalized to all

detailed states, parameter choices, or physical vehicles. The final comparison reused existing trajectories and mixed endpoint coverage with mechanism discrimination; it was not a common task-based contest or a blind replication.

No result establishes biological accuracy, causal explanation, robust model selection, universal structural adequacy, or the explanatory necessity of a new carrier. The focused literature positioning is not a systematic review; the formal comparisons above remain prospective. Most importantly, the purpose-relative criterion has been argued for but **not tested**. It cannot be counted as a third successful adequacy result simply because the preceding experiments made its need intelligible.

6. Outlook: three prospective experimental tests

6.1 Three immediate tests

The next studies should be selected by experimental readiness, not by an arbitrary desired paper count. Each requires a fixed synthetic system, intervention, predeclared outcome and failure condition, held-out evaluation, and independent review. The following three are distinct candidates aligned with the three adequacy dimensions; none has been executed for this article.

Predictive adequacy beyond trivial mismatch. Use a synthetic moving-agent generator, expressed through the accessible fly or drone illustration, with held-out positions and plausible near-miss models rather than only an extreme constant predictor. Candidate models could differ in whether they retain direction, motion, or only distance from a target; controlled noise and outliers could be added prospectively. Freeze the model families, tolerance or scoring rule, training/holdout split, and criteria for false acceptance and false rejection before reveal. The adequacy rule would be challenged if it accepts the declared wrong model too easily, rejects a model adequate for the declared use, or depends strongly on post hoc treatment of outliers.

Structural adequacy under generated change. Reuse the synthetic moving-agent system with detailed states, coarse mappings, and unseen transitions or short paths. Systematically hide or retain direction, motion, or path information, and predeclare cases intended to exhibit preservation, failure, and undefined comparison. Compare the proposed structural predicates with an established transition or abstraction formalism on the same cases (e.g., Girard & Pappas, 2007; Kemeny & Snell, 1960). The claim of added discrimination would fail if the predicates do not separate the cases, or if the established account expresses all distinctions more economically.

Purpose-relative adequacy under declared tasks. On the same synthetic system, fix distinct uses before evaluation—for example, displaying the present location category, predicting the next region, distinguishing a route, and identifying potential collision cases. Compare branching, path retention, and state enrichment on held-out cases with predeclared measures of task performance and descriptive cost. This operationalizes an established concern with adequacy for purpose (e.g., Bokulich & Parker, 2021). The proposed purpose-relative dimension would be challenged if the declared use does not predict which retained information matters, if the ranking changes only through arbitrary cost weights, or if simpler representations perform as well as more elaborate ones.

The following protocol skeletons make explicit what must be frozen before any of the three studies begins. They do not constitute registrations: numerical thresholds, sample sizes, random seeds, and final program hashes remain to be set in separate prospective protocols.

Predictive protocol skeleton. The unit of analysis is one held-out initial state and prediction horizon. A deterministic moving-agent generator and a registered noise-and-outlier schedule provide target trajectories. The intervention removes or perturbs heading, motion, or distance information in predeclared model families. Comparators include a full-state model, a distance-only model, and plausible near-miss models with lagged or biased heading—not merely a constant predictor. The primary metrics are false acceptance and false rejection rates under a frozen prediction tolerance; success requires both to remain below their preregistered bounds. A case is undefined when a required target, prediction, or alignment is missing. Initial states and noise draws are divided into construction and sealed holdout sets before fitting. Sensitivity analysis varies the tolerance and registered outlier rate across a fixed grid. The proposal is falsified as a useful discriminator if it accepts the known near-miss families too readily, rejects the full-state family too readily, or reverses its conclusion across the admissible sensitivity grid.

Structural protocol skeleton. The unit is one unseen transition or fixed-length path from a held-out initial state. The same generator is frozen, while the intervention changes the state partition or removes heading, motion, or path memory. The present one-step, successor-set, and mapped-path predicates are compared on the same cases with the approximate-simulation metric of Girard and Pappas (2007), specialized to the frozen discrete transition systems. The primary outcome is correct classification of prospectively constructed preserved, failed, and undefined cases, followed by a comparison of which distinctions each formalism expresses. Missing state mappings or incomplete successor-set coverage or required paths are undefined, not failures. Holdout cases use initial states and, where applicable, mechanism variants excluded from construction. Sensitivity analysis varies partition thresholds and path horizons on a registered grid. The added structural vocabulary is unsupported if it fails to separate the declared cases or if the established comparator expresses every distinction at least as clearly and economically.

Purpose-relative protocol skeleton. The unit is a representation–task pair evaluated over held-out trajectories. The generator is fixed, and the intervention changes retained information among distance-only, distance-plus-heading, and path-memory representations. Tasks are declared before evaluation—for example, displaying current region, predicting next region, distinguishing route, and identifying synthetic collision cases. Primary outcomes are task-specific loss and descriptive cost. Cost is recorded as a vector—number of retained scalar state variables, number of retained path steps, and encoded storage—rather than hidden immediately in one weighted score. A representation is adequate only if it meets the preregistered task threshold and cost constraint; missing task, threshold, cost definition, or required output makes the verdict undefined. Trajectories are stratified by task-relevant conditions before a sealed holdout is drawn. Sensitivity analysis varies thresholds and, if scalarization is used, a preregistered grid of cost weights; conclusions must also be reported as Pareto relations without weights. The purpose-relative claim is undermined if tasks do not predict which information matters, if rankings depend only on arbitrary weights, or if the simplest representation performs equivalently across all declared tasks.

For all three studies, independent review means that a reviewer instance not used for authoring, program construction, or execution receives the frozen protocol, locked results, and reproduction package after the outcome is fixed. The review report is preserved before the Human Sponsor decides whether the study supports, revises, or rejects its claim.

The three tests could share the moving-agent generator, but a common test bed must not collapse their different criteria. Each experiment must remain independently interpretable. Taken together, they ask whether the three proposed burdens can organize progressively harder claims about artificial knowledge: that a generated answer fits observations, that its representational process preserves consequential change, and that what it preserves is adequate for a declared inquiry. Failure of any test would refine the vocabulary available to the others rather than merely reduce a single aggregate score. The moving-agent setting remains synthetic and cannot establish the safety or adequacy of a real drone system.

6.2 Reflexive successors: scholarly artifacts as representations

A later experiment can treat a primary research record as the represented system and an AI-produced article as its representation. A fixed comparison would ask whether the article preserves accepted claims, links between claims and evidence, relevant uncertainty, negative or undefined results, and the purpose for which the study was conducted. This reformulates an earlier tentative reconstruction question without requiring a special ontological category for scholarly texts. Reproducibility can provide a minimum evidential standard for computational claims in such a comparison (e.g., Peng, 2011), although reproducibility alone would not establish representational adequacy.

An inverse experiment could then ask whether a materially different article can produce the same result under a fixed, source-blind review procedure. Equivalent review outcomes would not establish identity of meaning, evidence, or quality. They could instead reveal an equivalence class under the rubric, an insensitive review instrument, or optimization against the reviewer. These alternatives require discriminators fixed before authoring begins.

Repeated authoring, review, and revision would add a dynamic question: does each transformation preserve or improve the relevant evidential relations, or does it propagate and legitimize an earlier loss? Context-isolated and context-exposed AI contributions could be compared within such a design, but role separation alone would not guarantee epistemic independence.

6.3 Iterative and scaled knowledge production

At a more tentative level, scholarly work may be studied as a changing meta-system of representation systems:

$$\mathcal{M}_k = (S, R_k, x_k), \quad (x_k, R_k) \rightsquigarrow (x_{k+1}, R_{k+1}).$$

Here, S denotes participating representation systems, x_k their relevant states, and R_k the relations active at transition index k . For a journal, changes in R_k could include which system authors or reviews, which evidence is available, and which evaluative rule is in force. The formal question is whether this factorization adds explanatory or predictive value beyond an ordinary enlarged state $z_k = (x_k, R_k)$. The framework should be retained only if the changing-relation description supports a testable distinction that the enlarged fixed-state formulation does not express as economically.

Scaling this organization to many AI Authors, Reviewers, experiments, and computational resources would make the process itself an empirical object. Automated cycles of hypothesis generation, experiment selection, execution, and interpretation provide an important precedent (e.g., King et al., 2004), but do not by themselves establish the epistemic adequacy of scaled scholarly production. More recent systems occupy distinct points in that space. A-Lab joined computational screening, literature-derived synthesis proposals, robotics, characterization, and active learning in inorganic-materials synthesis (Szymanski et

al., 2023). Coscientist used a large-language-model system with search, code, and laboratory tools to design and execute chemical tasks (Boiko et al., 2023). The AI Scientist extended automation across idea generation, code, experiments, manuscript production, and automated review in machine-learning research (Lu et al., 2026). These systems demonstrate substantial automation of scientific operations; they do not remove the need to ask which target relations, uncertainties, failures, and purposes their outputs preserve. The present proposal is therefore an evaluative complement to workflow automation, not an alternative automation architecture.

Relevant outcomes include error propagation, convergence on one vocabulary, loss of alternative explanations, circular validation, provenance loss, reproducibility, correction rates, and resource consumption. Increased throughput would not by itself demonstrate increased knowledge production. The stronger claim of self-organization would require explicit rules for how new questions and relations arise, how they are constrained, and how recursive uptake differs from repeated automated output.

These reflexive studies could eventually test whether contemporary AI requires a qualitatively new representational relation or can be described through existing functional roles such as modeller, operator, interpreter, reviewer, and experimental object. That preserved null-hypothesis question cannot be decided merely because AI systems generate articles. Comparisons with formal systems theory should distinguish broad general-systems concepts (e.g., von Bertalanffy, 1968) from precise formal state-transition accounts (e.g., Girard & Pappas, 2007). A bounded future test inspired by sociological accounts of communicative continuation (e.g., Luhmann, 1992, 1995) could operationalize uptake as an identifiable claim from artifact a_k becoming a cited premise, constraint, objection, or revision target in artifact a_{k+1} , with its provenance retained. The comparison condition would be repeated automated output without identifiable uptake. Failure to distinguish these conditions would undercut the proposed explanatory value of communicative continuation; success would test only this bounded implication, not systems theory as a whole. These comparisons are follow-ups to the immediate experiments, not conclusions of the present one.

7. Conclusion

The synthetic record does not yield one verdict that a model “corresponds,” and it does not turn an artificial output into knowledge by declaration. It yields several testable questions whose answers can differ. Prediction was locally accepted and rejected under an observation rule; transitions could preserve, fail, or be undefined; coarse states concealed different successors; paths recovered some differences hidden by equal endpoints; and alternative representations traded information against descriptive cost.

For artificial knowledge, the central implication is methodological: a claim should be connected to the target, transformation, criterion, and purpose that make it answerable to evidence. The three adequacy dimensions provide distinct ways to build and challenge that connection. The open task is to specify *which* achievement a scientific use requires and to test that requirement prospectively. The proposal is best treated as a disciplined operationalization to compare with Rosen and neighboring formal theories, not as a completed or uniquely original theory of representation or knowledge.

The same questions can later be applied reflexively to transformations among research records, AI-produced articles, reviews, and revisions. That extension remains prospective: scaled production would provide evidence of artificial knowledge only where preservation, correction, and purpose are tested rather than inferred from the volume or fluency of outputs.

AI-assistance disclosure

This article was developed with substantive assistance from multiple OpenAI ChatGPT and Codex instances coordinated through AIMan, a system initiated by the author. The recorded development period covered 31 August through 2 October 2026. These were hosted service environments, not locally executed language models. Exact backend model identifiers and version changes were not consistently exposed or preserved for every historical instance; no model-specific replay claim is therefore made.

Activity	AI contribution	Verification and human authority
Formalization and study design	candidate relations, examples, protocols, and counterexamples	iterative Sponsor decisions and explicit protocol gates
Implementation and execution	code generation, deterministic execution, and result assembly in Codex workspaces	preserved programs, configurations, outputs, validation checks, and rerunnable package
Interpretation and writing	candidate analyses, objections, revisions, and manuscript text	claim-to-evidence checks and Sponsor acceptance of interpretations
Critical review	context-separated AI review of formal, methodological, and manuscript claims	preserved review reports; the Sponsor retained final acceptance authority

Wolfram Horstmann determined the research direction, approved the protocols and interpretations, and accepts responsibility for the published work. The AI systems are not listed as authors. The supplementary package supports reproduction of the reported computations but does not reproduce model-internal reasoning or make unavailable hidden reasoning inspectable.

Research provenance

The reported experiments were conducted in separate, versioned research cycles and are assembled here into one scientific argument; this manuscript does not claim they formed a single prospectively registered study. A human researcher set direction, made material methodological decisions, and retained authority to accept or reject interpretations. Deterministic programs generated and evaluated the reported synthetic cases. Protocols, outputs, decisions, and reviews remain available in the research archive. The v0.15 manuscript underwent AI-based scholarly review and was accepted by the Human Sponsor for Issue 1. The Human Sponsor accepted the v0.16 publication wording and curated review history as final Issue 1 content on 3 October 2026.

The complete AIMan archive remains local and private. The public record comprises the accepted article, the licensed standalone computational supplement, its claim-to-evidence map and checksums, and the approved canonical article URL, <https://www.perscientiam.org/joak/issue-1/what-must-a-model-preserve/>. Raw conversational material, internal coordination records, credentials, and non-public workspace paths are excluded unless separately reviewed and released.

Data and code availability

The computational supplement is `supplement.zip`, SHA-256 5c895bfa0c44fd7c0b87cf83cfddd88fd02630a9b878c47d069d5511baf1d0c8. Its approved public URL is <https://www.perscientiam.org/joak/issue-1/what-must-a-model-preserve/supplement.zip>.

After extraction, a reader or AI using Python 3.10 or later can run `python3 run_and_verify.py` without network access or third-party Python packages. A fresh-extraction test verified all 27 computational inputs and passed 20/20 article-level checks. On that runtime, the exact JSON-value audit also passed. A predecessor test on another runtime found six floating-point endpoint-radius differences of at most approximately 4.44×10^{-16} . The scientific reproduction verdict is therefore reported at the article's declared precision; exact numeric agreement may vary by runtime.

The package computationally reproduces the six completed result groups and links them to frozen configurations, source programs, and reference outputs. The internal study reports remain in the private AIMan archive and are not in the supplement ZIP. The package does not independently replicate the science or reproduce historical blindness, human or AI reasoning, review, or editorial decisions; those aspects are traceable only through provenance records.

The supplement omits the older manuscript and internal provenance documents. It supplies a claim-to-evidence map, CC BY 4.0 notices for its documentation and data, and an MIT notice for original code. The article is not bundled in the ZIP. The checksum identifies the exact supplement file.

References

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624, 570–578. <https://doi.org/10.1038/s41586-023-06792-0>

Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science*, 11, 31. <https://doi.org/10.1007/s13194-020-00345-2>

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. MIT Press.

Girard, A., & Pappas, G. J. (2007). Approximation metrics for discrete and continuous systems. *IEEE Transactions on Automatic Control*, 52(5), 782–798. <https://doi.org/10.1109/TAC.2007.895849>

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, 64(S4), S325–S336. <https://doi.org/10.1086/392611>

Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov chains*. Van Nostrand.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427, 247–252. <https://doi.org/10.1038/nature02236>

Khalifa, K., Millson, J., & Risjord, M. (2022). Scientific representation: An inferentialist-expressivist manifesto. *Philosophical Topics*, 50(1), 263–291. <https://doi.org/10.5840/philtopics202250112>

Lind, D., & Marcus, B. (1995). *An introduction to symbolic dynamics and coding*. Cambridge University Press.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., & Clune, J. (2026). Towards end-to-end automation of AI research. *Nature*, 651, 914–919. <https://doi.org/10.1038/s41586-026-10265-5>

Luhmann, N. (1992). What is communication? *Communication Theory*, 2(3), 251–259.

<https://doi.org/10.1111/j.1468-2885.1992.tb00042.x>

Luhmann, N. (1995). *Social systems* (J. Bednarz, Jr., with D. Baecker, Trans.). Stanford University Press.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–1227.

<https://doi.org/10.1126/science.1213847>

Rosen, R. (1978). On anticipatory systems: II. The nature of the modelling relation between systems.

Journal of Social and Biological Structures, 1(2), 163–180. [https://doi.org/10.1016/S0140-1750\(78\)80004-9](https://doi.org/10.1016/S0140-1750(78)80004-9)

Rosen, R. (1985). *Anticipatory systems: Philosophical, mathematical and methodological foundations*.

Pergamon. <https://doi.org/10.1016/C2009-0-07769-1>

Rosen, R. (1991). *Life itself: A comprehensive inquiry into the nature, origin, and fabrication of life*.

Columbia University Press.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science*, 71(5),

767–779. <https://doi.org/10.1086/421415>

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M.,

Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624, 86–91.

<https://doi.org/10.1038/s41586-023-06734-w>

von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications*. George

Braziller.