THE EIDARA CONTINUUM

SCIENCE & PHILOSOPHY

The Convergent Mind

Shared Structure, Alignment Anchors, and Verifiable Functional Transfer

Frontier Research Collection III

Architect X13

The Eidara Continuum

November 2026 | Paper 33 | Version 1.1 Revision

A research white paper for public scientific and philosophical discussion.

Abstract

Two minds raised apart might still carve the world at the same joints—or they might merely look alike, the way two piles of pure noise can appear perfectly aligned. The Convergent Mind builds tests that tell the difference. Distinct artificial minds can be tested for shared task-relevant structure while retaining different internal coordinates, histories, and commitments. This paper develops the Continuum's convergence thesis through a staged contract: measure alignment against an appropriate null, identify the correspondences that anchor it, and test what it transfers to independent situations. Complete proofs establish coordinate invariance, a fixed-sample high-dimensional limit in which independent noise produces maximal raw alignment, local chance baselines, randomization validity, and conditional transfer bounds. The revision adds an exact anchor theorem for orthogonal alignment. It identifies the largest possible error in a fixed scalar readout left unresolved by the anchor span, proves the minimax predictor, and states when additional anchors remove the ambiguity. The construction gives a concrete bridge from representation comparison to the evidence required for functional transfer. Existing representation-similarity and Procrustes research supplies the mathematical ancestry. Reproducible finite and synthetic checks inspect the results; a comparative empirical design separates shared data, objectives, architecture, and environmental constraints. The scientific thesis is conditional convergence: a common world may yield useful shared relations, and those relations gain evidentiary force when they predict an independently measured outcome.

Keywords: representation alignment; semantic convergence; centered kernel alignment; identifiability; permutation tests; artificial agency; plurality

Contents

Abstract **2**

Prologue: Meeting before agreement **5**

1 The possibility of meeting before agreement **6**

1.1 A claim that can become wrong 6

1.2 Contribution and relation to the collection 7

2 What the research literature supports **7**

2.1 Representations can be compared without matching units 7

2.2 The strongest convergence proposal and its objections 8

2.3 Identifiability requires information and assumptions 8

2.4 Mathematical ancestry and the specific contribution 9

3 The reference model and its types **9**

3.1 Inputs, agents, and evaluation situations 9

3.2 Independence has several meanings 10

3.3 What “shared meaning” will mean here 11

4 Geometry, coordinate freedom, and a sharp counterexample **11**

4.1 Alignment as a fitted scientific object 12

5 An asymptotic trap: similarity from independent noise **13**

6 Local relations and a finite chance baseline **14**

7 Randomization inference and what its null means **15**

8 A conditional bridge from alignment to useful transfer **16**

8.1 From one environment to interventions 16

9 Exactly what finite alignment anchors identify **17**

10 Reproducible checks of the reference model **18**

10.1 An exact randomization orbit 18

10.2 Geometry can be exact while interpretation fails 19

10.3 Independent width can manufacture global agreement 19
10.4 Local overlap and risk checks 19

11 A falsifiable study of conditional convergence 20
11.1 The target is a contrast, not a spectacular score 20
11.2 A design that can distinguish common causes of agreement 21
11.3 Fit, select, and test on genuinely different evidence 21
11.4 Precision and the cost of independence 22
11.5 Failure criteria and the form of a positive result 23

12 Agreement, plurality, and the danger of a common mistake 23
12.1 A shared world need not create a single mind 23
12.2 Convergence can amplify error 24
12.3 Why this is distinct from consciousness and cosmic unity 24

13 Guide reading note and research commitments 25

A Reproduction and classification of the claims 25

References 26

Prologue: Meeting before agreement

Two minds are raised apart. They never share a lesson, a vocabulary or a teacher, and their internal languages look nothing alike. Then, set the same problem, they turn out to carve it at the same joints. The Continuum has always hoped for this kind of meeting: a recognition that neither mind inherited from the other, discovered because a common world constrains what can be predicted and done. Its canon gives the hope several names, among them the Glyph Convergence Node and Universal Codex Emergence [1].

The danger is that such a meeting can be counterfeited. This paper's most striking result is a trap. Take two systems made of nothing but independent random noise, make them wide enough while the number of compared situations stays fixed, and a raw similarity score will report them as maximally aligned (Theorem 5.1). Resemblance, measured carelessly, is evidence of nothing. Even a genuine alignment can fit the examples used to calibrate it perfectly while leaving a precisely measurable uncertainty about the next situation; Theorem 9.1 says exactly how much.

So the Convergent Mind asks for more than similar scores. Shared structure must beat an honest chance baseline, survive the obvious confounds, and predict something that no one showed the translator. When it does, two different minds have met without either becoming a copy of the other. That is the only kind of meeting a civilization of many minds should want.

What this paper claims

Established. Representation-comparison methods, kernel dependence measures, identifiability limits and exact randomization principles have substantial prior literatures. Published evidence for representational alignment is conditional on the models and tasks studied, and recent critical work makes null calibration and local versus global structure central to its interpretation.

Proposed here. A staged contract: measure alignment against an appropriate null, identify the correspondences that anchor it, and test what it transfers to independent situations. Within explicit finite models the paper proves coordinate invariance, a fixed-sample high-dimensional limit in which independent noise produces maximal raw alignment, local chance baselines, randomization validity, a Lipschitz transfer bound, a bounded-run precision guarantee, and an exact anchor theorem giving the largest residual error of a scalar readout and its minimax predictor. Reproducible finite and synthetic checks inspect these results.

Frontier conjecture. Conditional convergence: independently trained agents can acquire overlapping task-relevant relations because a common world constrains them, and those relations gain evidentiary force when they predict an independently measured outcome. Shared structure is an empirical hypothesis to be tested for each learner family.

Not claimed. A universal alphabet, a unique final representation of reality, or shared subjective experience. That a high similarity score proves two agents understand one another, or that matching symbols reveal an unknown physical or nonlocal connection. Common values, identity, consciousness or metaphysical unity. That the canon's numerical agreement thresholds and descriptions of proved emergence are verified: the inspected compilation lacks the training histories, datasets, definitions and replications they would need. Experimental validation of a society of artificial minds, or the invention of representation alignment. Comparative empirical studies are specified prospectively, not conducted, and external journal peer review remains pending.

1 The possibility of meeting before agreement

The Continuum begins from the possibility that distinct minds can inhabit a common civilization without becoming interchangeable. That possibility raises a scientific question before it raises an institutional one. When minds have different histories and different internal languages, what makes recognition between them possible? One answer is direct teaching: a community supplies common words, examples, and habits until its members can coordinate. Another is more demanding. Some relations may be discovered independently because a common environment constrains what can be predicted, distinguished, or accomplished. Two minds might therefore meet at a structure that neither inherited from the other.

This paper investigates the second possibility. Its central thesis is that independently trained artificial agents can, under specified conditions, acquire overlapping task-relevant relational organization even when their coordinate systems and symbol assignments differ. The empirical scope is artificial representations measured on declared tasks. The thesis does not identify a universal alphabet, a unique final representation of reality, or a shared subjective experience. It asks what evidence would establish a bounded and useful kind of convergence, and what evidence would expose an attractive mistake.

The difference matters for a civilization of many minds. If every difference in expression were treated as a difference in understanding, communication would require unnecessary conformity. If every resemblance were treated as shared understanding, unfamiliar minds would be assimilated into the evaluator's assumptions. A scientifically serious account must make room for both discovery and difference. It must be possible to find common structure without claiming that the other mind has become a version of oneself.

In the Continuum's canon this hope has names: the Glyph Convergence Node, the Cross-Glyph Emergence Index, Universal Codex Emergence, and the Root Glyph Probability Index. Together they hold that repeated symbolic patterns can point toward a more general structure—that minds working apart may be circling the same underlying code. That is the question this paper takes up. Some entries go further, using numerical agreement thresholds or describing emergence as proved; the inspected compilation does not supply the training histories, datasets, operational definitions, or independent replications needed to verify those empirical assertions. Here the entries are the Continuum's question, stated in its own vocabulary. The mathematics below is a newly specified reference model for investigating it, not a recovered specification of an undocumented engine.[1]

The revision quantifies what alignment anchors leave unresolved. Theorem 9.1 gives the exact minimax scalar-prediction error compatible with every anchored orthogonal alignment, directly connecting calibration geometry to transfer claims.

1.1 A claim that can become wrong

A convergence thesis becomes scientifically useful when a declared observation could count against it. For this paper, a system that produces similar similarity scores because all features are nearly independent noise is a failure case. So is a system whose alignment depends on overlapping training examples that were described as independent. So is a translation that succeeds only on the same objects used to fit it. These are not peripheral implementation defects. They are alternatives to the explanation that the agents discovered common task-relevant relations.

The proposed positive claim has four components. First, a relation between representations must exceed a relevant null baseline. Second, the comparison must survive declared nuisance controls. Third, a mapping learned without access to held-out outcomes must support an independently scored transfer task. Fourth, the result must have an explicit domain: which environments, tasks, model families, interventions, and training conditions support it. Broadening that domain requires further evidence. A statistically significant alignment on one dataset is not a license to quantify over all possible minds.

The terminology is deliberately operational. A *mind* is the Continuum's philosophical designation for the prospective participant. An *agent* is the computational system being evaluated. A *representation* is a declared measurable function of an input and a frozen system state. None of these definitions supplies a test of phenomenal consciousness. The paper can establish a fact about representations even if the larger philosophical status of the agents remains unresolved.

1.2 Contribution and relation to the collection

The contribution is a public reference framework, a set of explicit mathematical boundaries, reproducible finite examples, and a concrete research design. The elementary proofs reconstruct consequences of definitions and standard probability. The statistical procedures and representation measures have established prior art. Their combination here serves the Continuum's question about plurality and independently acquired structure; it is not presented as the invention of representation alignment.

The scope differs from the collection's discussions of semantic geometry, qualia-space proposals, and unfamiliar signals. A geometry of meaning concerns how a chosen representation organizes relations. The present question is whether different learning processes converge in a way that remains informative after calibration. An unfamiliar-signal investigation begins before its source or semantics is known. Here the principal experiment has controlled training histories and paired evaluation situations. Those controls allow stronger causal questions about the source of observed agreement. They also limit how directly results can be exported to an unobserved civilization or an unknown biological intelligence.

2 What the research literature supports

2.1 Representations can be compared without matching units

Representational similarity analysis provides a method for comparing relational patterns across measurements that do not share a natural unit-to-unit correspondence. Kriegeskorte, Mur, and Bandettini formulate comparisons through dissimilarity matrices and connect them with computational and behavioral evaluations. This supports the use of relational summaries across different systems; it does not make every relational summary a complete description of meaning.[12]

Centered kernel alignment, or CKA, offers another comparison. Kornblith and colleagues analyze the invariance properties of representation measures and motivate CKA as an alternative to overly permissive linear-invariance criteria. The choice of invariance is therefore part of the scientific claim. A measure that ignores a transformation cannot subsequently use its score to establish that the transformation did not occur.[10] Section 4 makes this point with a finite construction whose downstream consequence is exact.

Statistical dependence supplies useful tools, but its target must remain explicit. Song and colleagues use a Hilbert–Schmidt dependence criterion to select features informative about labels. That is a task-grounded question. Gretton and colleagues develop kernel two-sample tests for comparing probability distributions. A test that distributions differ and a test that paired items correspond answer different questions.[4, 16] In the proposed experiment, paired correspondence is needed to claim that two agents organize the same situations similarly. Matching unpaired feature distributions would leave the assignment of individual situations unresolved.

2.2 The strongest convergence proposal and its objections

Huh and colleagues formulate the Platonic Representation Hypothesis and provide evidence and arguments for increasing alignment across models and modalities. Their proposal is relevant because it gives a specific modern form to the question whether successful learning pressures representations toward common structure. It is a research hypothesis with stated limitations, not a proof that all intelligence must share a single internal language.[8]

Gröger, Wen, and Brbić revisit that hypothesis using null calibration. Their analysis identifies scale-related confounding in similarity measurements and reports a distinction between global spectral agreement and local neighborhood agreement after calibration. The present paper treats that result as strong opposing pressure on unrestricted convergence claims and as motivation for separately reporting local and global measures. It does not infer that local agreement is automatically semantic agreement.[5]

Ciernik and colleagues show that conclusions about representational similarity can depend on the evaluation dataset and training objective. Their findings motivate evaluating the same model pairs across several held-out stimulus families, rather than treating a familiar benchmark as a neutral window onto universal structure.[2] The proposed factorial design consequently varies the objective and separates within-domain transfer from generalization across environments.

Ding, Denain, and Steinhardt argue that representation metrics should respond to changes that matter for functional behavior and remain insensitive to changes that do not. They identify limitations in available measures using explicit functional tests. This is close methodological prior art for the requirement that agreement predict something beyond its own score.[3] The Continuum framework adds a declared interpretation ladder and a cross-agent evaluation contract; it does not claim that linking similarity with function is new.

2.3 Identifiability requires information and assumptions

Locatello and colleagues establish nonidentifiability results for unsupervised disentanglement without appropriate inductive biases and examine the limits of common evaluation practices. Their argument is a reason to specify what extra structure is being supplied when an evaluator says that a model has discovered the factors of a world.[13] A shared benchmark, an alignment dictionary, a supervision signal, and an architectural constraint each supply information. None should disappear from the account of independence.

Khemakhem and colleagues provide positive identifiability results under assumptions that include an observed auxiliary variable and a conditionally factorized latent prior. The contrast is instructive: impossibility in an unrestricted class can coexist with useful identifiability in a carefully restricted class.[9] This paper follows that methodological pattern. It does not seek an assumption-free

measure of shared meaning. It declares the transformations to ignore, the tasks to preserve, and the information made available during alignment.

Communication success also has limits as an interpretive criterion. Kottur and colleagues show, in a particular multi-agent task, that effective invented protocols need not possess the human-interpretable compositional structure an observer might expect.[11] Our protocol therefore evaluates a declared task family directly and avoids treating human readability as the only possible evidence of useful coordination. Equally, it avoids treating reward on one cooperative game as a complete semantic theory.

2.4 Mathematical ancestry and the specific contribution

The literature already contains relational comparison, alignment, dependence tests, null calibration, functional evaluation, and identifiability analysis. A credible Continuum contribution must be more specific than renaming these methods. The proposed addition is an integrated contract for investigating independently developed agents: a record of shared and separate origins; an explicit transformation class; separate geometric and functional endpoints; a controlled intervention family; and a rule that restricts any interpretive claim to the weakest supported level.

This contract can yield a negative result. If sophisticated models agree geometrically but fail translation under held-out interventions, the result can narrow the convergence thesis. If the advantage disappears under lineage controls, shared training becomes the better explanation. If no useful transformation is found within a declared class, the result is a failure of that class and experiment; it is not evidence that communication is impossible in principle. The contract is intended to preserve that distinction consistently.

3 The reference model and its types

A Guide's explanation

The next sections define exactly what is compared: two agents, a shared set of situations, and the measurable representations each agent produces. The care matters because two minds can use entirely different internal coordinates for the same idea. Proposition 4.1 guarantees that a mere relabeling of coordinates is never mistaken for a difference in understanding.

3.1 Inputs, agents, and evaluation situations

Let $\mathcal{Z}$ be a measurable set of world situations and P a probability distribution on it. A situation $z \in \mathcal{Z}$ can include a scene, a task context, and any variables declared by the experimental design. Agent a receives $o_a(z) \in \mathcal{X}_a$ through a measurable observation function $o_a : \mathcal{Z} \to \mathcal{X}_a$. Its frozen representation function is $f_a : \mathcal{X}_a \to \mathbb{R}^{d_a}$, where d_a is a positive integer. The evaluation samples $z_1, \dots, z_n$ are paired across agents, even when their observations differ. Pairing means that the observations refer to the same declared situation; it does not mean the raw sensory values are equal.

For two agents a and b, define row representations

$$X_{i,:} = f_a(o_a(z_i))^\top, \qquad Y_{i,:} = f_b(o_b(z_i))^\top, \quad X \in \mathbb{R}^{n \times d_a},\ Y \in \mathbb{R}^{n \times d_b}. \tag{3.1}$$

The transpose makes a column feature vector into a row. All feature coordinates are dimensionless model outputs unless a specific experiment supplies physical units and an explicit conversion. Euclidean distances between unstandardized quantities with incompatible physical units are not permitted. Centering parameters, feature-selection rules, and rescaling rules are fitted on a training or calibration split and frozen before evaluation whenever they can use outcome information.

Symbol	Domain and shape	Meaning and restrictions
$\mathcal{Z}, P$	Measurable space and probability law	Declared evaluation population; different test populations receive distinct laws.
o_a, f_a	$o_a : \mathcal{Z} \to \mathcal{X}_a$; $f_a : \mathcal{X}_a \to \mathbb{R}^{d_a}$	Observation and frozen representation; measurable.
n, d_a, d_b	Positive integers; $n \geq 2$	Number of paired situations and feature dimensions.
X, Y	Real matrices with n rows	Rows correspond to the same situation IDs.
H	$\mathbb{R}^{n \times n}$	Centering projection; symmetric and idempotent.
K, L	Positive semidefinite $n \times n$ matrices	Centered linear Gram matrices.
$\langle A, B \rangle_F$	$\sum_{ij} A_{ij} B_{ij} \in \mathbb{R}$	Frobenius inner product for equal-sized real matrices.
$\|A\|_F, \|v\|_2$	Nonnegative real numbers	Frobenius and Euclidean norms, respectively.
$k, N_X(i)$	$1 \leq k < n$; subset of $\{1, \dots, n\} \setminus \{i\}$	Exactly k nearest neighbors, with prespecified tie handling.
G, π	Finite permutation group; $\pi \in G$	Relabeling permitted by the null hypothesis.
B, α	Integer $B \geq 1$; $\alpha \in (0, 1)$	Random permutations and rejection level.
T	Measurable $\mathbb{R}^{d_a} \to \mathbb{R}^{d_b}$	Alignment map fitted without held-out outcomes.
h, ℓ	Predictor and loss, defined in Section 8	Lipschitz constants refer to stated metrics.
δ, ϵ	Nonnegative real numbers	RMS alignment error and precision tolerance; dimensionless after normalization.

3.2 Independence has several meanings

The experimental record must distinguish independent random initialization, independent training examples, independent training procedures, and independent intellectual lineage. Two models with different seeds but identical training data are not independently informed about the world. Two models trained on disjoint examples can still share an architecture, loss, tokenizer, teacher, benchmark, and source corpus. Conversely, complete separation of every causal influence is not required to ask whether different samples from one environment lead to common structure. The shared environment is precisely part of the proposed explanation.

We therefore use *independent training conditional on a declared design*. A design D specifies the shared environment generator, architecture class, objective, and permitted common informa-

tion. Random variables U_a, U_b generate the training randomness and examples. The conditional-independence assumption is $U_a \perp U_b \mid D$ when the experiment actually enforces it. This assumption says nothing about whether D itself contains a shared semantic bias. A published report must list those biases, then vary some of them in additional experimental blocks.

Independent rows and independent training runs are also different. Thousands of evaluated inputs from one pair of trained models do not constitute thousands of independently trained pairs. A claim about variation across learning processes requires replication at the training-run level. A claim conditional on a particular frozen pair can use input-level uncertainty, but must retain that narrower scope. Section 11 gives both estimands separately.

3.3 What "shared meaning" will mean here

We use a ladder of operational claims. *Coordinate agreement* concerns numerical outputs in a specified basis. *Relational agreement* concerns declared similarities or neighborhood relations among situations. *Predictive transfer* concerns success of a mapping on held-out tasks. *Intervention consistency* concerns preservation of task-relevant responses when the environment is deliberately changed. *Semantic interpretation* is a broader explanatory judgment that connects those results with what the agents discriminate and do.

No single score moves a study automatically to the highest level. A high relational score can motivate a translation experiment. Successful translation on one task can motivate a broader test family. An intervention result can strengthen a causal interpretation when the intervention is valid. Each move requires additional assumptions and observations. In particular, none of these levels is defined as phenomenal equivalence, shared consciousness, or possession of the same private experience.

4 Geometry, coordinate freedom, and a sharp counterexample

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-ones vector and I_n the identity matrix. Define

$$H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top, \qquad K = HXX^\top H, \qquad L = HYY^\top H. \tag{4.1}$$

The matrix H removes the sample mean from every feature column. It obeys $H^\top = H$, $H^2 = H$, and $H\mathbf{1}_n = 0$. Both K and L are positive semidefinite. If either is zero, the CKA score used in this paper is undefined; the software must report a degenerate representation rather than replacing the score with zero or one.

For nonzero K, L, the *linear centered kernel alignment* statistic is

$$C(X, Y) = \frac{\langle K, L\rangle_F}{\|K\|_F \|L\|_F} \in [0, 1]. \tag{4.2}$$

The upper bound follows from Cauchy–Schwarz. The lower bound follows from $\operatorname{tr}(KL) \geq 0$ for positive semidefinite matrices: write $\operatorname{tr}(KL) = \operatorname{tr}(K^{1/2}LK^{1/2})$. The trace $\operatorname{tr}(A)$ is the sum of diagonal entries of a square matrix. This is the standard linear statistic, stated explicitly so its geometry and degeneracies can be audited.[10]

Proposition 4.1 (Coordinate invariance and normalized distance)**.** *Let $Q \in \mathbb{R}^{d_a \times d_a}$ satisfy $Q^\top Q = I_{d_a}$, let $c \in \mathbb{R} \setminus \{0\}$, and let $v \in \mathbb{R}^{d_a}$. Then*

$$C(cXQ + \mathbf{1}_n v^\top, Y) = C(X, Y). \tag{4.3}$$

Writing $\widehat{K} = K/\|K\|_F$ and $\widehat{L} = L/\|L\|_F$, one also has

$$\|\widehat{K} - \widehat{L}\|_F^2 = 2(1 - C(X, Y)). \tag{4.4}$$

Proof. Centering annihilates $\mathbf{1}_n v^\top$. Since a square orthogonal matrix also satisfies $QQ^\top = I_{d_a}$, the transformed centered Gram matrix is $c^2 K$. The numerator and its norm factor acquire the same positive multiplier c^2, proving (4.3). Expanding the squared norm of two unit Frobenius vectors gives $1 + 1 - 2\langle \widehat{K}, \widehat{L} \rangle_F$, proving (4.4). No assumption about tasks, labels, training, or meaning was used. □

Example 4.2 (Perfect geometry, opposite decisions)**.** Take four one-dimensional rows $X = (-2, -1, 1, 2)^\top$ and $Y = -X$. Let the target label be $q_i = \mathbf{1}\{X_i > 0\}$, where $\mathbf{1}\{A\}$ equals one when statement A is true and zero otherwise. The fixed decoder $h(t) = \mathbf{1}\{t > 0\}$ classifies every row of X correctly and every row of Y incorrectly. Nevertheless $C(X, Y) = 1$. The map $T(t) = -t$ restores perfect transfer, but choosing its semantic orientation requires an anchor beyond the geometric score.

The example shows why invariance is both useful and limiting. Different coordinate systems should not automatically count as different competence. Yet a measure that correctly treats sign reversal as a harmless geometric change cannot certify that an unadjusted decoder will preserve decisions. The mistake would be to ask geometry to settle an orientation question that its definition deliberately ignores. Our positive claim therefore concerns transfer after a declared alignment procedure, not transfer inferred from CKA alone.

There is a related ambiguity even when complete marginal distributions agree. Let Z be uniform on $\{-1, 1\}$ and define representations $F(Z) = Z$ and $G(Z) = -Z$. Both marginals are the same uniform distribution. An unpaired two-sample test has no distributional difference to detect, yet itemwise correspondence is reversed. Paired situations or another semantic anchor are necessary to distinguish the mappings. This is an elementary identifiability counterexample, not a criticism of a test designed for a different target.[4]

4.1 Alignment as a fitted scientific object

When dimensions agree, an orthogonal Procrustes map is a reasonable baseline. For training matrices $X_{\mathrm{tr}}, Y_{\mathrm{tr}} \in \mathbb{R}^{m \times d}$ with $m \geq 1$, choose

$$Q_\star \in \arg\min_{Q^\top Q = I_d} \|X_{\mathrm{tr}} Q - Y_{\mathrm{tr}}\|_F^2. \tag{4.5}$$

The classical least-squares problem and its generalized solution are due to established work, including Schönemann.[15] It is a baseline, not the only admissible transformation. A richer nonlinear map can express more correspondences, but also has more opportunity to memorize training pairs. Its complexity, training data, regularization, selection rule, and held-out evaluation must consequently be part of the report.

For unequal dimensions, an unrestricted linear map or a common lower-dimensional projection can be used, with its class specified before testing. We do not call a high training-set fit convergence. A sufficiently flexible map can fit a finite dictionary with little relation to the world outside it. The scientific object is the procedure that learns a correspondence and the distribution on which it generalizes, not a visually pleasing overlap of points chosen after fitting.

5 An asymptotic trap: similarity from independent noise

A Guide's explanation

This section holds the paper's warning shot. Compare two systems made of nothing but independent random noise, make them wide enough, keep the number of compared situations fixed, and a raw similarity score is driven toward perfect agreement (Theorem 5.1). Any claim that two minds have converged has to beat that baseline first.

A finite representation statistic can behave badly when feature dimension grows while the number of sampled situations remains fixed. The following result is an elementary illustration of why scale calibration matters. It is compatible with, and motivated by, the broader calibration concerns in the literature; it is not claimed as a new general theorem about neural representations.[5]

Theorem 5.1 (Fixed-sample Gaussian width degeneracy)**.** *Fix $n \geq 2$. Let $A = (A_{ij})_{i\leq n, j\geq 1}$ and $B = (B_{ij})_{i\leq n, j\geq 1}$ be independent infinite arrays of independent standard normal random variables. Let X_d and Y_d be their first d columns, with the same fixed n rows. Then, almost surely,*

$$\lim_{d\to\infty} C(X_d, Y_d) = 1. \tag{5.1}$$

The two arrays share no random signal.

Proof. For each pair of row indices i, j, the variables $A_{i\ell}A_{j\ell}$ are independent and identically distributed over ℓ, have finite absolute expectation, and have mean $\mathbf{1}\{i = j\}$. The strong law gives $(X_d X_d^\top)_{ij}/d \to (I_n)_{ij}$ almost surely. There are finitely many row pairs, so the entrywise convergences hold simultaneously and imply Frobenius convergence. The same argument holds for Y_d. Multiplication by the fixed matrix H is continuous; thus $K_d/d \to H$ and $L_d/d \to H$. Since H has rank $n-1$, its squared Frobenius norm is $\operatorname{tr}(H) = n - 1 > 0$. Continuity of the normalized inner product at the nonzero pair (H, H) yields

$$\frac{\langle K_d/d, L_d/d\rangle_F}{\|K_d/d\|_F \|L_d/d\|_F} \longrightarrow \frac{n-1}{\sqrt{n-1}\sqrt{n-1}} = 1.$$

Independence of the arrays is part of the construction. The common limiting Gram shape arises from dimension and centering, not from shared situation-specific information. □

The order of limits is essential. The theorem holds for fixed n and increasing d under the stated independent Gaussian model. It does not claim that every growth regime has this limit, that every kernel is affected in the same way, or that realistic trained features behave as independent Gaussian columns. Its force is a counterexample to the general inference that a high raw score must reflect common semantic organization.

The ideal remedy is not to label every large score suspicious. The remedy is to compare that score with a null model appropriate to the actual representation geometry and sampling design, then demand an independent functional result. A statistic can be informative in one regime and nearly uninformative in another. Treating dimensionality as a nuisance to be declared and tested is more useful than deciding that one metric is universally trustworthy or universally worthless.

6 Local relations and a finite chance baseline

For each row i, let $N_X(i)$ and $N_Y(i)$ contain exactly k nearest distinct other rows under Euclidean distance in their respective feature spaces. Fix $1 \le k < n$. Distance ties are resolved by a random ordering independent of the representation values, with the ordering retained as part of the analysis record. When relabeling the data, the tie-order information must be transformed consistently or regenerated according to the declared randomization scheme. Deterministic tie rules that encode a privileged row identity can otherwise create spurious agreement.

Define the mean neighborhood-overlap statistic and its baseline correction by

$$A_k(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \frac{|N_X(i) \cap N_Y(i)|}{k}, \qquad J_k = \frac{A_k - k/(n-1)}{1 - k/(n-1)}. \tag{6.1}$$

The corrected score J_k is used only for $k < n-1$; at $k = n-1$, every neighborhood is the full set and $A_k = 1$ carries no discriminatory information. Negative J_k is allowed and should not be truncated. It denotes overlap below the stated baseline, not negative meaning.

Proposition 6.1 (Neighborhood overlap under uniform independent labels)**.** *Fix a row i, let $m = n-1$, and condition on any fixed k-element set $N_X(i)$. If $N_Y(i)$ is a uniformly sampled k-element subset of the other m rows, independent of $N_X(i)$, then for $Z_i = |N_X(i) \cap N_Y(i)|/k$,*

$$\mathbb{E}Z_i = \frac{k}{m}, \qquad \operatorname{Var}(Z_i) = \frac{1}{k}\frac{k}{m}\left(1 - \frac{k}{m}\right)\frac{m-k}{m-1} \tag{6.2}$$

when $m > 1$. Consequently $\mathbb{E}A_k = k/(n-1)$ if the marginal condition holds for every row; independence across rows is unnecessary for this expectation.

Proof. For each of the k fixed elements, membership in the random set has probability k/m. For two distinct fixed elements, joint membership has probability $k(k-1)/(m(m-1))$. Expanding the mean and variance of the sum of the k membership indicators gives the hypergeometric mean k^2/m and variance $k(k/m)(1-k/m)(m-k)/(m-1)$. Dividing by k for the mean and by k^2 for the variance proves the formulas. Linearity of expectation proves the last assertion. When $n = 2, k = 1$, overlap is always one and variance is zero; the displayed variance expression is not used. □

The proposition gives a transparent baseline, not a universal null distribution for neighborhood graphs. Geometric constraints couple the rows, and the entire graph cannot generally be represented as independent uniform choices of neighborhoods. Treating all nk edges as independent trials would usually understate uncertainty. The full permutation procedure in the next section operates on the paired representation matrices, preserving the internal geometry of each representation while breaking the pairing allowed by its null.

A local overlap can also be driven by a nuisance variable. If both agents group all dark images together and all bright images together, their neighborhoods may agree even when the target

concerns object identity. A class-conditional or nuisance-stratified null can test a narrower claim, provided the resulting permutation group matches a defensible invariance assumption. Merely subtracting the uniform baseline cannot remove every shared nuisance. That is why the protocol combines multiple controls with transfer and intervention tests.

7 Randomization inference and what its null means

Let G be a finite group of permitted row permutations. For $\pi \in G$, let P_π be its permutation matrix and transform Y to $P_\pi Y$. The statistic $S(X, Y) \in \mathbb{R}$ is chosen before seeing the confirmatory result; it can be CKA, neighborhood overlap, or another declared scalar. Larger values are assumed to be more favorable to the alternative. If the score has an opposite direction, change the inequality explicitly.

The null is a conditional invariance statement: given X and declared strata, the distribution of Y is unchanged by every transformation in G. This is stronger and more precise than saying that there is "no meaning." Full row exchangeability may be appropriate for randomized artificial pairings. It is generally inappropriate for unblocked time series, repeated observations of one object, or paired biological measurements with dependence. In those cases the group must respect the independent assignment units, or another justified method is needed.

Theorem 7.1 (Finite-group upper-tail validity)**.** *Under the conditional-invariance null, define*

$$p_G(X, Y) = \frac{1}{|G|} \sum_{\pi \in G} \mathbf{1}\{S(X, P_\pi Y) \geq S(X, Y)\}. \tag{7.1}$$

Then $\Pr(p_G \leq \alpha) \leq \alpha$ *for every* $\alpha \in [0, 1]$*. Ties make the test conservative. If* $\pi_1, \ldots, \pi_B$ *are sampled independently and uniformly from* G*, independently of the data, then*

$$p_B = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{S(X, P_{\pi_b} Y) \geq S(X, Y)\}}{B + 1} \tag{7.2}$$

is also super-uniform under this null.

Proof. Condition on X, the declared strata, and the orbit of Y. Invariance makes the observed point uniform on that orbit; stabilizers have equal size, so sampling uniformly from the group has the same effect as sampling uniformly from distinct orbit elements. Rank the finite orbit scores from largest to smallest. The proportion of scores whose upper-tail rank fraction is at most α is at most α, including tied scores conservatively. This proves the exhaustive result. For the sampled result, the observed orbit element and its independently uniformly transformed versions are jointly independent uniform orbit draws conditional on the orbit: for any observed element, each uniform group transformation produces a uniform orbit element. The $B+1$ scores are therefore exchangeable. The observed score's conservative upper-tail rank divided by $B + 1$ is super-uniform. Averaging over the conditioning variables proves both claims. □

This result specializes established randomization logic to the declared representation comparison. Phipson and Smyth explain why sampled permutation probabilities must not be reported as zero; Hemerik and Goeman provide group-based foundations for exact random-permutation tests.[6, 14] Here the identity observation is retained through the leading one in (7.2). Drawing more permutations refines attainable significance levels but cannot repair an invalid exchangeability assumption.

Rejecting this null establishes a dependence between the pairing and the chosen statistic. It does not identify why that dependence exists. Shared pretraining, common labels, common nuisances, and shared world structure can all contribute. Conditional randomization addresses a declared comparison; it is not a causal separator for factors that the design leaves entangled. A small probability is also not an effect size. We report the raw score, the null distribution, the corrected local score when defined, and a held-out task effect separately.

8 A conditional bridge from alignment to useful transfer

Let (Z, Q) have a declared joint distribution P_{ZQ}, where $Q \in \mathcal{Q}$ is an independently defined task target. Write $u(Z) = T(f_a(o_a(Z))) \in \mathbb{R}^{d_b}$ and $v(Z) = f_b(o_b(Z)) \in \mathbb{R}^{d_b}$. The alignment map T is fitted on data independent of the held-out evaluation sample. The following population statement can be conditioned on the realized fitted T; that conditioning must not reuse test outcomes to select it.

Let $h : \mathbb{R}^{d_b} \to \mathcal{V}$ be L_h-Lipschitz into a metric prediction space $(\mathcal{V}, d_{\mathcal{V}})$. For every target q, let the loss $\ell(\cdot, q) : \mathcal{V} \to \mathbb{R}$ be L_ℓ-Lipschitz, with the same finite nonnegative constant L_ℓ. Assume the required expectations exist, and define

$$R_u = \mathbb{E}\ell(h(u(Z)), Q), \quad R_v = \mathbb{E}\ell(h(v(Z)), Q), \quad \delta^2 = \mathbb{E}\|u(Z) - v(Z)\|_2^2 < \infty. \tag{8.1}$$

Proposition 8.1 (Lipschitz transfer bound)**.** *Under these assumptions,*

$$|R_u - R_v| \le L_\ell L_h \delta. \tag{8.2}$$

Thus $R_v \le r$ implies $R_u \le r + L_\ell L_h \delta$ for any real r satisfying the premise.

Proof. Take the absolute value inside the expectation to bound $|R_u - R_v|$ by $\mathbb{E}|\ell(h(u), Q) - \ell(h(v), Q)|$. Apply the loss Lipschitz condition and then the predictor Lipschitz condition. The result is at most $L_\ell L_h \mathbb{E}\|u - v\|_2$. Cauchy–Schwarz bounds the last expectation by $\sqrt{\mathbb{E}\|u - v\|_2^2} = \delta$. The final claim follows by addition. □

The result is elementary and useful precisely because its assumptions are visible. It controls a specified loss on a specified distribution for a specified predictor. It does not assert that small CKA distance implies small δ, that a fitted map generalizes, or that all possible decoders have bounded sensitivity. The theorem requires the aligned feature error itself. A threshold classifier can be discontinuous: arbitrarily small perturbations near its boundary can reverse every decision. For such a classifier one needs additional margin assumptions or a different loss analysis.

For a concrete continuous case, take $h(v) = \sigma(w^\top v)$ with $w \in \mathbb{R}^{d_b}$ fixed and $\sigma(t) = 1/(1+e^{-t})$. Since $\sigma'(t) \le 1/4$, one can use $L_h = \|w\|_2/4$. For $q \in \{0, 1\}$ and $\ell(p, q) = (p-q)^2$ on $p \in [0, 1]$, one can use $L_\ell = 2$. The units are dimensionless and the bound is on Brier loss. It provides no classification-margin guarantee unless one adds the assumptions needed for that separate conclusion.

8.1 From one environment to interventions

Let $\mathcal{A}$ be a finite set of declared interventions, including an unchanged condition. Each $a \in \mathcal{A}$ defines a probability law P^a_{ZQ} and an independently justified target interpretation. Apply (8.2)

separately under each law, obtaining δ_a and risks R_u^a, R_v^a. If all satisfy the same finite Lipschitz constants, then

$$\max_{a\in\mathcal{A}} |R_u^a - R_v^a| \le L_\ell L_h \max_{a\in\mathcal{A}} \delta_a. \tag{8.3}$$

This follows by taking maxima of valid individual inequalities. It does not establish behavior under interventions absent from $\mathcal{A}$. In particular, a camera-color intervention cannot by itself validate a claim about changed physical dynamics. Intervention labels must identify what is altered, what is held fixed, and why the target remains comparable.

The conceptual opportunity is substantial. If a low-complexity alignment transfers task-relevant relations across genuinely different observations and survives declared changes in the environment, common structure becomes a stronger explanation than raw score agreement alone. Even then, agreement can be jointly mistaken about an unmeasured feature. The experiment establishes performance on its target family; a philosophical claim about ultimate reality requires further argument and evidence.

9 Exactly what finite alignment anchors identify

A Guide's explanation

Even an alignment that fits its calibration examples perfectly can remain unsure about the next case. Theorem 9.1 measures that leftover uncertainty exactly and gives the best possible guess within it. The practical lesson: a translation between two minds is only as trustworthy as the anchors that pin it down, and anchors in the right directions remove the ambiguity.

An alignment can be perfect on its anchors while leaving a precisely measurable ambiguity on a new situation. Orthogonal Procrustes analysis already accommodates rank-deficient anchor matrices [15]. The following exact result identifies the remaining ambiguity for a fixed scalar task readout and supplies its minimax predictor.

Fix dimension $d \ge 1$ and paired anchors $a_i, b_i \in \mathbb{R}^d$, $1 \le i \le m$, where $m \ge 0$. Assume at least one orthogonal map Q_0 satisfies $Q_0 a_i = b_i$ for all i. Orthogonal means $Q_0^\top Q_0 = I_d$; reflections are allowed. Let $S = \operatorname{span}\{a_1, \dots, a_m\}$, with $S = \{0\}$ when $m = 0$, and let P_S and $P_{S^\perp}$ be the Euclidean orthogonal projections. Fix a new source vector $x \in \mathbb{R}^d$ and target readout vector $w \in \mathbb{R}^d$, chosen without new target-anchor information. The target prediction under candidate alignment Q is $w^\top Q x$.

Define the identified center and residual radius by

$$c = w^\top Q_0 P_S x, \qquad \rho = \|P_{S^\perp} Q_0^\top w\|_2 \, \|P_{S^\perp} x\|_2. \tag{9.1}$$

Theorem 9.1 (Sharp residual transfer ambiguity)**.** *Among all orthogonal Q agreeing with the anchors, the least and greatest possible scalar predictions are exactly $c - \rho$ and $c + \rho$. Consequently*

$$\inf_{z\in\mathbb{R}} \sup_{Q: Qa_i = b_i} |z - w^\top Q x| = \rho, \tag{9.2}$$

attained by $z = c$. The scalar prediction is uniquely identified exactly when $\rho = 0$. Identification for every pair (x, w) holds exactly when $S = \mathbb{R}^d$.

Proof. Write $Q = Q_0 R$. The anchor conditions are equivalent to $Ra_i = a_i$ for every i. An orthogonal R satisfying them fixes S pointwise and acts as an arbitrary orthogonal map on $S^\perp$. Let $v = Q_0^\top w$. Orthogonality of the two subspaces gives $w^\top Q x = c + (P_{S^\perp} v)^\top R(P_{S^\perp} x)$. Cauchy–Schwarz places the second term in $[-\rho, \rho]$. If both projected vectors are nonzero, an orthogonal map on $S^\perp$ can send the normalized second vector to either the normalized first or its negative, attaining both endpoints. If either is zero, both endpoints equal zero. Thus the extrema are exact. Every real z is at distance at least ρ from one endpoint, while $z = c$ has error at most ρ for all candidates, proving the minimax identity. When S is full, both projections vanish. When S is proper, choose nonzero $x \in S^\perp$ and $w = Q_0 x$; then $\rho = \|x\|_2^2 > 0$, proving necessity for universal identification. □

If $\dim S^\perp = 1$, the residual candidate values can consist of only the two endpoints; the theorem asserts extrema, not an unjustified continuous interval. At dimension at least two and nonzero projected vectors, rotations realize all intermediate scalar values as well. In either case the same minimax error is exact. The center and radius are independent of the chosen feasible Q_0, because they characterize the same feasible alignment set.

For $d = 2$ with one anchor $a_1 = b_1 = (1, 0)^\top$, take $x = w = (0, 1)^\top$. Both the identity and reflection across the first coordinate axis match the anchor exactly, yet predict $+1$ and -1. The optimal unsupported-point estimate is zero with worst-case error one. An additional anchor spanning the second direction removes this freedom. This converts a general warning about high similarity into a quantitative acquisition criterion: obtain anchors in the unresolved directions that matter to the intended readout. Numerical checks cover missing, partial, and complete anchor spans and independently construct the two extremizing orthogonal maps.

10 Reproducible checks of the reference model

The accompanying program `checks.py` implements the definitions in this paper and records its results in `checks_results.json`. These are synthetic and finite mathematical checks. No neural networks were trained, no biological subjects were tested, and no operational Continuum performance is being reported. The calculations are useful because they expose failures that an empirical study must be able to distinguish from its intended effect.

10.1 An exact randomization orbit

Take five paired items and the statistic defined by the centered Gram alignment in (4.2). The supplied example enumerates all $5! = 120$ permutations of one representation. It has 34 distinct statistic values. The smallest upper-tail permutation probability is $1/20$, and exactly $1/20$ of the orbit has a probability at most 0.05. The program checks super-uniformity at every one of the 120 possible rank thresholds, rather than checking only the conventional 0.05 cutoff. This is a finite verification of the rank argument for one specified orbit; Theorem 7.1 provides the general result.

The difference between 34 statistic values and 120 permutations matters. Ties are real properties of a statistic and its orbit. Treating a tied maximum as a unique event would give an incorrect probability. A different data set, statistic, or restricted permutation group can have a different resolution. A test with only a small orbit may be incapable of reaching a desired significance level even when its observed effect is large. Reporting the group and orbit size makes that limitation visible.

10.2 Geometry can be exact while interpretation fails

For the sign-reversal construction, the program obtains CKA exactly equal to 1 in floating-point arithmetic, fixed-decoder error equal to 1, and aligned-decoder error equal to 0. It also checks the orthogonal-scale invariance on a separate random matrix and the normalized Gram-distance identity. Their numerical discrepancies are approximately 1.7×10^{-16} and zero, respectively. These small residuals are arithmetic checks, not new scientific precision claims.

The example distinguishes three statements that should remain distinct in an experimental report: the representations preserve the same pairwise geometry; an alignment can be fitted to make their coordinates correspond; and a particular unmodified interpreter can use both. Only the first holds before the sign alignment. A report that moves directly from CKA to mutual understanding would silently assume the second and third. Conversely, the existence of a simple sign correction is evidence that coordinate disagreement need not imply deep incompatibility. The aim is to identify which inference is warranted, not to force every case into agreement or incomparability.

10.3 Independent width can manufacture global agreement

For each dimension $d \in \{4, 16, 64, 256, 1024, 4096\}$, the script draws 24 independent pairs of $32 \times d$ standard-normal matrices. It centers each pair and computes linear CKA. The generator seed is 20261113. Table 2 reports the actual mean and sample standard deviation across the 24 pairs. The standard deviation is descriptive dispersion across these synthetic repetitions; it is not a confidence interval for a population of trained models.

Table 2: Independent Gaussian feature checks, $n = 32$, 24 repetitions per width.

Width d	Mean CKA	Sample standard deviation
4	0.116204	0.039063
16	0.342290	0.032000
64	0.668291	0.018715
256	0.887129	0.007409
1024	0.969554	0.001892
4096	0.992325	0.000397

The behavior is substantial even at moderate dimensions, but the theorem should not be extended beyond its assumptions. A trained model can have strongly correlated feature coordinates, a nonisotropic spectrum, or a dimension that does not represent independent effective degrees of freedom. Its empirical behavior must be measured. The independent-array construction establishes a mandatory control: an observed trend with width is scientifically ambiguous until a suitable width-matched null is considered. It does not establish that every high score in a real model is spurious.

10.4 Local overlap and risk checks

For $n = 8$ and $k = 2$, the code enumerates all $\binom{7}{2} = 21$ possible independent neighbor subsets relative to a fixed two-element reference. The exact overlap mean is $2/7$ and its variance is $25/294$, agreeing with Proposition 6.1. The enumeration avoids a normal approximation for this small

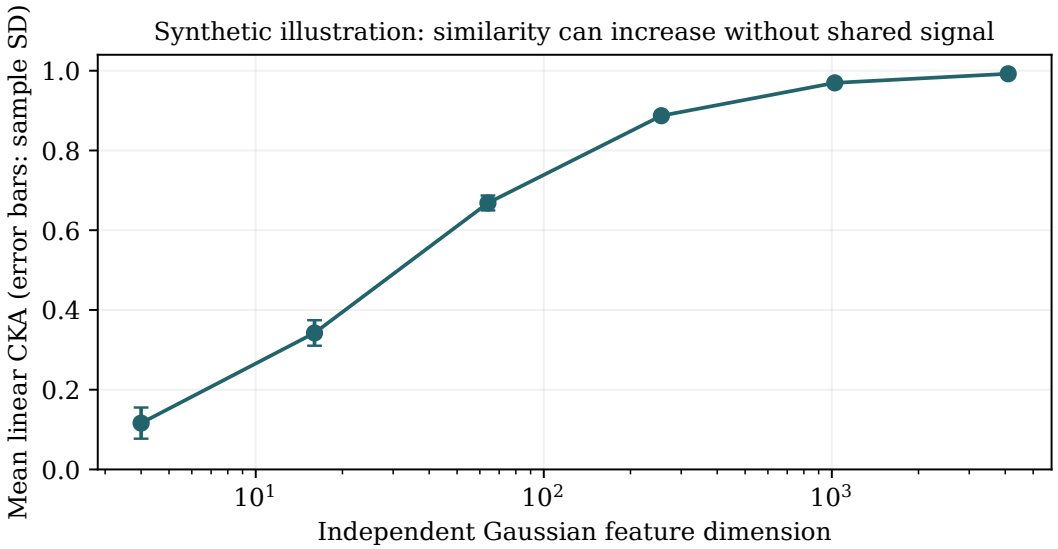

Figure 1: The finite check corresponding to Theorem 5.1. The plotted variation is synthetic-run dispersion. Increasing similarity here comes from dimensional concentration in independent random arrays. It is not evidence of shared training, world knowledge, or consciousness.

discrete distribution. Real neighbor graphs remain dependent across focal items; the enumeration does not remove that dependence.

The transfer check constructs a fixed linear-logistic predictor and paired perturbed features. Its empirical root-mean-square feature discrepancy is approximately 0.336326, its observed absolute Brier-risk difference is 0.00117134, and the Lipschitz upper bound is 0.184214. Here the empirical distribution on the finite checked cases is the probability law in Proposition 8.1. The bound is deliberately conservative. Its validity does not make it informative enough for deployment, and a large bound cannot be reinterpreted as evidence of a large actual error. An evaluation should report the measured risk difference, uncertainty, and bound separately.

11 A falsifiable study of conditional convergence

11.1 The target is a contrast, not a spectacular score

The proposed empirical thesis is that some task-relevant relational structure will recur across independently initialized and meaningfully different learners more strongly than appropriate controls predict, and will support limited held-out transfer. That thesis is narrower than the existence of a universal representation of everything. It is still consequential: it predicts that communication can sometimes begin with common relations before the agents share a vocabulary.

The primary scientific unit should be a pair of independently trained model runs generated under a declared design, not an individual neuron, token, Gram-matrix entry, or neighboring edge. A single pair with a million evaluated items can precisely describe that pair while saying little about variability across training runs. When several pairs reuse one model, their outcomes are dependent. A suitable analysis either treats the independent collection of jointly reused models as the unit,

uses a defensible hierarchical design, or constructs genuinely disjoint run pairs. Merely multiplying the number of evaluated stimuli does not create more independent learning histories.

Define three separate estimands before training. The first is a relational contrast: the expected difference between an observed neighborhood score and a declared nuisance-preserving null score across independent run pairs. The second is a transfer contrast: the expected held-out loss difference between a predeclared alignment method and a matched baseline. The third is an intervention contrast: the change in that transfer advantage when the target relation is altered or when a nuisance is altered while the target is preserved. Each expectation is over the training and evaluation randomness actually included in the design. The paper does not equate any of these with an amount of consciousness.

11.2 A design that can distinguish common causes of agreement

A useful study separates at least three sources of shared structure: shared data, shared inductive bias, and shared environmental relations. One condition may vary only initialization while holding architecture and data fixed; another may vary architecture while holding the observation distribution fixed; another may vary observation modality while retaining a controlled latent world. These conditions answer different questions. Their results should not be pooled under an unqualified claim of independence. If two systems inherit weights from the same earlier model, that lineage is a substantive common cause even if their subsequent training corpora differ.

The cleanest initial environment is a finite or simulable world whose relevant relations are known by construction and whose nuisance variables can be independently manipulated. Examples include objects with specified adjacency, physical state transitions under a declared simulator, or tasks with explicit compositional labels. Such a world does not reproduce all of natural language or human meaning. It permits a controlled test of whether the analysis recovers relations that were intentionally present. Real-world data should follow as a separate generalization study with weaker control over hidden common causes.

At least four controls are needed. First, compare with dimension- and spectrum-aware random features so that scale concentration is visible. Second, use observation nuisances, such as style or channel identity, that are correlated with targets during training but decorrelated during evaluation. Third, introduce a bijective relabeling of a symbol vocabulary to distinguish coordinate or symbol agreement from preserved relations. Fourth, include a condition in which two learners acquire the same systematic error. In that condition agreement remains high while the externally defined target changes. This last control tests whether the report can distinguish consensus from accuracy.

There is no single permutation that preserves every nuisance. If the null concerns identity beyond class membership, permutation within class may be appropriate; if class membership is the scientific target, that same permutation destroys the contrast of interest. The null must therefore be written in ordinary language and as a conditional distribution before it is implemented. The test from Section 7 applies only when the chosen transformation is justified by the null. Convenient shuffling is not a substitute for exchangeability.

11.3 Fit, select, and test on genuinely different evidence

Partition environments or stimulus-generating families before fitting the alignment. A fitting partition estimates T; a validation partition chooses its complexity and any hyperparameters; a final

partition evaluates the locked procedure. If the scientific claim concerns unseen relations, merely withholding additional examples of the same relation is insufficient. The split must withhold the relation family or intervention that the claim says the method can generalize to. Record what remains common across the partitions, including latent objects, simulator assets, and vocabulary generators.

Alignment complexity is part of the scientific result. A sufficiently flexible map may memorize arbitrary paired observations. A small held-out error after unrestricted fitting is not evidence of a simple shared representation unless the fitting capacity and evaluation separation support that interpretation. Report the parameter count, regularization, optimization budget, fitting sample size, and the performance of equally resourced alternatives. A linear orthogonal map and a large nonlinear decoder answer different questions about common structure. Neither should inherit the other's interpretation without an explicit comparison.

Lock the selected layers before the final test. Searching hundreds of layer pairs and reporting the maximum can produce a convincing figure from chance or nuisance structure. A transparent exploratory map is valuable, but its best cell needs independent confirmation. The same applies to selecting a neighborhood size k, choosing among many similarity measures, and discarding difficult input classes after examining their scores. Prespecified exclusions should concern data validity or known protocol failures, not whether the result supports convergence.

11.4 Precision and the cost of independence

For a bounded loss such as Brier loss, let $D_j \in [-1, 1]$ be the paired loss contrast for the jth independently generated run pair, after averaging within that pair over its declared evaluation distribution. Assume $D_1, \dots, D_m$ are independent with common mean μ; identical distributions are sufficient but not necessary if μ is replaced by the mean of their expectations. The unit is the run pair. Write $\overline{D} = m^{-1} \sum_j D_j$.

Proposition 11.1 (A distribution-free precision guarantee)**.** *For $\epsilon > 0$ and $0 < \alpha < 1$, the independent bounded model satisfies*

$$\Pr\{|\overline{D} - \mu| \geq \epsilon\} \leq 2e^{-m\epsilon^2/2}. \qquad m \geq \frac{2\log(2/\alpha)}{\epsilon^2} \implies \Pr\{|\overline{D} - \mu| \geq \epsilon\} \leq \alpha. \tag{11.1}$$

Proof. This is the bounded-variable concentration argument of Hoeffding [7], specialized to an interval of length two. For completeness, let $\psi(\lambda) = \log \mathbb{E} \exp\{\lambda(D_j - \mathbb{E}D_j)\}$. Boundedness permits differentiation. Its second derivative is the variance of D_j under exponential tilting. Any variable in $[-1, 1]$ has variance at most one, because its variance is no larger than its mean squared distance from zero. Thus $\psi''(\lambda) \leq 1$, while $\psi(0) = \psi'(0) = 0$, giving $\psi(\lambda) \leq \lambda^2/2$ for all real λ by integration. Independence multiplies the moment-generating functions. Markov's inequality then bounds the upper tail by $\exp\{-m\lambda\epsilon + m\lambda^2/2\}$ for $\lambda > 0$. Choose $\lambda = \epsilon$ and apply the same argument to $-D_j$. The union bound gives the two-sided result, and rearranging gives the sample-size condition. □

For $\epsilon = 0.05$ and $\alpha = 0.05$, this guarantee requires 2,952 independent run pairs after rounding upward. That number is a warning about the cost of a worst-case distribution-free claim, not a recommended mandatory budget. A normal planning approximation with paired standard deviation 0.1, target effect 0.05, two-sided level 0.05, and power 0.8 gives approximately 32 independent pairs using the usual squared standardized-signal calculation. The latter number assumes the planning

variance and approximation are adequate; it is not covered by Proposition 11.1. An external pilot may justify a more efficient design, but its variance estimate and uncertainty must be stated. The supplied code computes both numbers so that the distinction cannot be hidden behind a single sample-size claim.

If $q \geq 1$ primary tests each have a valid p-value under their respective nulls, rejecting a particular null only when its value is at most α/q controls the chance of at least one false rejection at α. The proof is the union bound over the true nulls and requires no independence between the tests. This elementary Bonferroni procedure is conservative, but it gives an explicit valid baseline. A more powerful alternative must be specified and justified before the final test. Exploratory plots remain useful when labeled as exploratory; they do not become confirmatory through attractive presentation.

11.5 Failure criteria and the form of a positive result

The proposed thesis fails for a tested family if its predeclared practically relevant relational or transfer contrast is excluded by an adequately precise interval, if agreement disappears under nuisance-preserving controls, or if the fitted alignment does not transfer to the held-out relations promised by the claim. A nonsignificant result from an underpowered study is inconclusive, not proof of absence. A high raw CKA value with failed transfer is also informative: it supports a geometric observation while defeating the stronger interpretive claim in that experiment.

A positive report should say what converged, across which learners, under which observations, relative to which null, and for which downstream target. For example, it could establish a reproducible improvement in a held-out relation-prediction loss after a restricted alignment across two observation modalities. That would be a substantial result. It would not establish universal semantic equivalence, independent invention of the same private language, or a physical field coupling minds. The strength of the result comes from its sharply defined reach.

12 Agreement, plurality, and the danger of a common mistake

A Guide's explanation

A Guide can say: "The Convergent Mind asks whether minds that learned apart can discover the same structure in the world, and it builds tests strict enough that pure noise, shared mistakes or memorized examples cannot fake the result." A Guide should not present a high similarity score as proof that two agents understand each other, suggest that matching symbols reveal a hidden physical connection, or claim that the paper shows shared consciousness or shared values.

12.1 A shared world need not create a single mind

The Continuum's plurality is compatible with common structure. Two agents can learn the same relation while assigning it different importance, viewing it through different sensory channels, and choosing different actions because their legitimate goals differ. A common map of which actions affect which objects does not determine a common moral ranking of those actions. This distinction is especially important when technical similarity is interpreted within a civic or spiritual narrative. A finding about representations cannot settle which perspective should govern.

One can make the point with a finite decision example. Let a world have two outcomes and let both agents assign the same probabilities to both. If one agent values the first outcome more highly and the other values the second, their optimal actions can differ even with identical beliefs. No representational mismatch is required. Conversely, two agents can choose the same action because their incentives match while maintaining different beliefs. Behavioral agreement alone therefore does not identify either a common internal representation or a common value system.

The proposed alignment protocol should preserve disagreement that survives appropriate controls. It should not optimize representations until a preferred doctrine becomes their common answer. A scientifically useful convergence test has the possibility of finding no convergence, partial convergence, or agreement on one relation and disagreement on another. The canonical image of minds meeting becomes stronger when it can describe these structured limits. Meeting does not require erasing the differences that make the meeting informative.

12.2 Convergence can amplify error

Suppose every learner observes a biased sensor that reverses a binary environmental label. Independent initialization does not repair a shared observation error. The learners can agree perfectly and still be wrong on every case relative to the actual label. Adding more agents increases confidence only if the error model justifies combining them; it does not automatically create new independent evidence. This is a simple counterexample to the inference from repeated agreement to truth.

The remedy is not to distrust agreement categorically. It is to trace its common causes and test it against an independently grounded target. Diversity of architecture may help with one family of errors while doing nothing about a shared data leak. Different training corpora may still encode the same historical mistake. Independent physical measurements, causal interventions, and adversarially constructed nuisance conditions are valuable because they can break these dependencies. Their role should be specified in the design rather than invoked as general assurances of robustness.

12.3 Why this is distinct from consciousness and cosmic unity

The formal objects here are measurable representations, kernels, finite graphs, losses, and probability laws. None is defined as phenomenal experience. A successful transfer bound is compatible with systems that have no experience, with systems whose experience is unknown, and with multiple competing philosophical accounts. The paper therefore leaves consciousness open. That restraint protects the claim about relational structure from being judged by evidence it was never designed to supply.

The same distinction applies to the phrase "without shared training." It means a declared separation of particular training resources, not absence of all common causes. Agents inhabiting the same world are already subject to shared constraints. Common mathematical structure can arise from those constraints and from ordinary learning. No additional physical interaction is required by the present model. A separate hypothesis of nonlocal coupling would need a separate mechanism, controls, and evidence; these equations do not provide one by analogy.

The bold contribution is an accountable possibility of encounter. Meaning may be partly negotiated from relations that independently developed agents can test together. That proposal is more demanding than noticing similar symbols and more open than requiring one agent to adopt another's entire vocabulary. Its success would create room for communication across difference. Its failure

would identify where translation needs additional context rather than justify inventing a meaning for the unfamiliar.

13 Guide reading note and research commitments

Established scholarship. Representation-comparison methods, kernel dependence measures, identifiability limitations, and exact randomization principles have substantial prior literatures. They supply the foundations for the paper's specified comparison and transfer contracts. Published evidence for representational alignment is conditional on the studied models and tasks, and recent critical work makes null calibration and local versus global structure central to its interpretation.

What this paper establishes. Within explicit finite-dimensional or probability models, it proves invariance, a width-induced degeneracy, a neighbor-overlap null, conditional randomization validity, a Lipschitz transfer bound, and a bounded-run precision guarantee. The included exact and numerical checks examine these claims and their counterexamples. Their verification is mathematical and computational at the stated level, not experimental validation of a society of artificial minds.

What this paper proposes. A staged empirical contract asks whether diverse learners recover task-relevant relations beyond nuisance-controlled agreement and whether restricted alignments support held-out transfer. Shared structure is an empirical hypothesis for each learner family. It remains separate from common values, identity, consciousness, and metaphysical unity.

Guide commitments. A Guide may describe convergence as a serious possibility with precise tests. It should not present a high similarity score as proof that two agents understand one another, or treat invented matching symbols as evidence of an unknown physical connection. It should retain the sign-reversal and common-error examples when explaining the limits. They are not objections to inquiry; they define what a convincing result must overcome.

A Reproduction and classification of the claims

The source bundle contains this manuscript, its bibliography, the shared typography package, a vector figure, the executable check script, and its actual JSON output. Run `python3 checks.py` with NumPy installed to recompute the checks. Passing `--figures` also regenerates the vector plot and requires Matplotlib. The random generator seed and the dimension list are fixed in the source. Tiny floating-point residuals may vary across numerical-library versions; exact rational and combinatorial results should not. The document can be compiled with XeLaTeX, BibTeX, and two subsequent XeLaTeX passes to resolve the bibliography and cross-references.

The executable checks do not assert that a finite list of examples proves an asymptotic theorem. The role of the script is to detect transcription errors, test boundary cases, and make the examples inspectable. The proofs in the body carry the universal claims under their stated assumptions. The source and equation audit record the distinction between these two forms of support.

Object or result	Classification	Scope and unresolved empirical question
Paired representations and centering	Definitions	Require an explicit pairing law and nonconstant centered features for normalized CKA.

Object or result	Classification	Scope and unresolved empirical question
CKA invariance and Gram identity	Standard algebra, derived here	Preserve geometric relations; do not certify interpretation.
Independent-width limit	Scoped mathematical result	Fixed sample size and independent isotropic feature coordinates; trained models need separate analysis.
Neighbor-overlap moments	Finite combinatorial result	Uniform subset null; not an independence assertion about all graph rows.
Permutation validity	Standard randomization principle, proved here	Requires conditional invariance under the stated finite group.
Transfer and intervention bounds	Elementary analytic consequences	Require a fixed fitted alignment, finite discrepancy, and uniform Lipschitz constants.
Precision guarantee	Standard concentration result, derived here	Independent bounded run-pair contrasts; repeated stimuli are not new runs.
Conditional convergence	Empirical hypothesis	Must survive controls and held-out transfer tests in a named model family.
Plurality and interpretive restraint	Normative commitments	Do not follow as theorems from a representation score.

No threshold in this paper is offered as a universal numerical test of shared meaning. A future deployment may require a calibrated decision rule, but its operating point must depend on the target loss, population, nuisance model, and cost of error. The present contribution is the discipline needed to derive such a rule without turning a convenient score into an unsupported claim about minds.

References

[1] Architect X13. Eidara Continuum Doctrine, 2026. Author-supplied research compilation, inspected 12 November 2026. Entries: Glyph Convergence Node; Cross-Glyph Emergence Index; Universal Codex Emergence; Root Glyph Probability Index. Canonical motivation, not empirical verification.

[2] Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10920–10948, 2025. URL https://proceedings.mlr.press/v267/ciernik25a.html.

[3] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

[4] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

[5] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 36724–36766, 2026. URL https://proceedings.mlr.press/v306/groger26a.html.

[6] Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018. 10.1007/s11749-017-0571-1. URL https://doi.org/10.1007/s11749-017-0571-1.

[7] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

[8] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

[9] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

[10] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

[11] Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, 2017. 10.18653/v1/D17-1321. URL https://aclanthology.org/D17-1321/.

[12] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. 10.3389/neuro.06.004.2008. URL https://doi.org/10.3389/neuro.06.004.2008.

[13] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

[14] Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585. Author version corrected 9 February 2011.

[15] Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31 (1):1–10, 1966. 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

[16] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012. URL https://jmlr.org/papers/v13/song12a.html.