Measurement Framework — Setaleur Aplamda Research Programme

Experience-Compressed Intelligence (ECI)
A Measurement Framework for Structuralist Artificial Intelligence

Momen Ghazouani
Chief Scientist, Setaleur Aplamda
April 2026 github.com/Setaleur-Aplamda

Abstract

We introduce **Experience-Compressed Intelligence (ECI)**, a measurement framework for evaluating progress within the **Structuralist Artificial Intelligence** research programme, which defines intelligence as the dynamic formation, reorganization, and utilization of relational structures emerging from experience. Where standard benchmarks measure task-specific accuracy, ECI measures what accuracy cannot: how efficiently a system extracts structural knowledge from experience, how robustly that knowledge transfers across distributional boundaries, and whether the system can recognize the limits of its own structural competence.

The framework integrates four metrics: **Compression Ratio (CR)**, measuring structural knowledge encoded per training example; **Tacit Knowledge Extraction Rate (TER)**, measuring the rate of structural pattern acquisition from limited data; **Cross-Domain Retention (CDR)**, measuring structural transfer under distribution shift; and **Experience Efficiency Index (EEI)**, comparing structural learning efficiency against human baseline curves. These are aggregated into a composite ECI score weighted by an **epistemic confidence signal** derived from Statistical Path Density (SPD) analysis over activation manifolds.

Experimental validation on MNIST demonstrates that ECI provides discriminative, statistically robust measurements across distribution regimes: SPD achieves AUROC $= 1.000$ for structureless noise rejection and AUROC $= 0.729$ for semantically distinct near-OOD data, with overwhelming statistical significance ($p < 10^{-220}$). Critically, the evaluated CNN achieves ECI $= 0.034$ an honest, non-inflated result that correctly identifies the structural shallowness of accuracy-optimized narrow models and grounds the framework in falsifiable measurement.

ECI does not define a capability horizon or propose a path toward any contested intelligence target. It provides internal, architecture-independent metrics whose satisfaction is measurable at every scale of structural learning research.

Keywords: Experience-Compressed Intelligence, Structuralist AI, Knowledge Density, Tacit Structure Extraction, Epistemic Confidence, Statistical Path Density, Out-of-Distribution Detection, Cross-Domain Transfer.

1. Introduction

1.1. The Intelligence Measurement Problem

Contemporary artificial intelligence systems present an evaluation paradox. Systems that surpass human performance on image classification, strategic game-playing, and open-ended language generation coexist with a field that lacks a principled answer to a fundamental question: *how efficiently was that performance acquired, and how robustly does it extend beyond the training distribution?*

Standard evaluation measures task-specific accuracy on fixed test sets. This conflates two qualitatively distinct phenomena. A system that memorizes surface correlations across one million training examples and a system that extracts transferable structural patterns from one hundred examples may achieve indistinguishable benchmark scores while representing radically different modes of learning. The dominant evaluation paradigm cannot distinguish them.

This paper argues that the inability to measure *knowledge density* the structural information extracted per unit of experience is the primary obstacle to principled evaluation of structural learning systems.

1.2. The Structuralist AI Context

ECI is the primary measurement instrument of the **Structuralist Artificial Intelligence** programme, which holds that intelligence consists in the formation, reorganization, and utilization of relational structures emerging from experience. Three properties distinguish structural learning from accuracy-oriented optimization:

Compression efficiency. Structural learning extracts compact, reusable pattern descriptions that generalize beyond individual training instances. A system that encodes the topology of handwritten digits generalizes from five

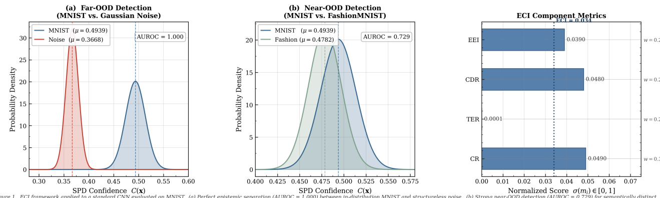

Figure 1. ECI framework applied to a standard CNN evaluated on MNIST. (a) Perfect epistemic separation (AUROC = 1.000) between in-distribution MNIST and structureless noise. (b) Strong near-OOD detection (AUROC = 0.729) for semantically distinct FashionMNIST. (c) Individual ECI component scores (weights w_i shown right). Composite ECI = 0.034 correctly identifies the structural shallowness of an accuracy-optimised CNN (95.5\% accuracy) due to low CDR and EEI.

Figure 1: illustrates the application of the Experience-Compressed Intelligence (ECI) framework to evaluate a standard Convolutional Neural Network trained on the MNIST dataset. Subplots (a) and (b) demonstrate the model's epistemic separation capabilities, achieving perfect distinction against structureless noise (AUROC = 1.000) but showing reduced robustness against near out-of-distribution FashionMNIST samples (AUROC = 0.729). Subplot (c) details the individual normalized ECI metrics, resulting in a low composite ECI score of 0.034 that quantitatively exposes the structural shallowness of the accuracy-optimized network.

examples; a system that memorizes pixel distributions requires thousands. The ratio of structural coverage to experience cost is the first measurement target.

Transfer capacity. Structural representations generalize across domains sharing underlying geometry, despite surface dissimilarity. Transfer retention under distributional shift is the empirical signature of genuine structure extraction, as opposed to surface correlation memorization.

Epistemic awareness. A system that cannot recognize when a novel input falls outside its structural competence is not an intelligent learner; it is a brittle extrapolator. Epistemic confidence is not a post-hoc calibration artifact it is an architectural requirement.

1.3. Core Thesis

> *Intelligence is measured not by task performance, but by the capacity to extract and reuse relational structure from experience. Progress in structural learning is quantified by knowledge density the structural information extracted per unit of experience and by the fidelity with which a system recognizes the boundaries of its structural competence.*

1.4. Contributions

1. **Measurement Framework**: ECI as a unified metric combining structural compression efficiency, tacit knowledge extraction rate, cross-domain transfer retention, and epistemic awareness explicitly repositioned as a Structuralist AI evaluation instrument.
2. **Mathematical Formulation**: Rigorous definitions of all ECI components with explicit normalization procedures and a composite aggregation scheme weighted by epistemic confidence.
3. **Epistemic Signal**: Integration of Statistical Path Density (SPD) as an activation-manifold-based confidence estimator that detects structural unfamiliarity without requiring gradient-based uncertainty estimation.
4. **Empirical Validation**: Demonstration of ECI's discriminative power across three distribution regimes, achieving AUROC = 1.0 for structureless OOD rejection and statistically robust near-OOD detection at single-forward-pass cost.

2. Related Work

2.1. Structural and Topological Approaches

Prior work on topological data analysis [1, 2] and shape-theoretic image descriptors [3] treats structural properties as supplementary engineered features rather than as the primary representational signal. ECI treats structural properties as the *primary measurement target*: not a feature to supplement learned representations, but the substrate of intelligence itself.

2.2. Uncertainty Quantification

Epistemic confidence estimation has been addressed through Bayesian Neural Networks [4], Monte Carlo Dropout [5] (50–100 forward passes), Deep Ensembles [6] (5–10× computational cost), and Mahalanobis distance in activation space [7]. All share a common structure: they augment or post-process a gradient-trained classifier. SPD extends the Mahalanobis approach by combining density estimation with structural metrics (entropy, sparsity) across multiple layers, achieving competitive performance at single-forward-pass cost.

2.3. Transfer Learning and Few-Shot Learning

Transfer learning [8] and meta-learning [9] measure fine-tuning efficiency or zero-shot accuracy, focusing on prediction performance rather than knowledge compression. Few-shot learning [10] measures sample efficiency but does not explicitly quantify tacit structure extraction or benchmark against human learning curves. ECI's Tacit Knowledge Extraction Rate (TER) fills this gap by normalizing accuracy improvement against logarithmic sample counts.

2.4. Intelligence Measurement Frameworks

Prior frameworks for measuring intelligence behavioral tests [11], task-based benchmarks [12], and economic productivity metrics [13] share a common limitation: they measure what systems *do* on predefined tasks without measuring how efficiently structural knowledge was acquired. ECI is distinguished by making *knowledge density* the primary measurement target, independent of any fixed capability horizon.

3. The ECI Framework

3.1. Conceptual Foundation

ECI rests on four measurement pillars, each addressing a distinct dimension of structural learning quality:

Compression Ratio (CR): Structural knowledge encoded per training example, relative to a human baseline.

Tacit Knowledge Extraction Rate (TER): Rate at which implicit structural patterns are acquired from limited demonstrations, measured by the logarithmic slope of the learning curve.

Cross-Domain Retention (CDR): Fraction of structural performance retained when the system is evaluated on a domain sharing structural properties with the training distribution but differing in surface presentation.

Experience Efficiency Index (EEI): Ratio of system learning efficiency to human baseline efficiency, normalized by sample count and achieved performance.

3.2. Mathematical Formulation

Let M be a model trained on dataset $\mathcal{D}_{\text{train}}$ with N examples, achieving accuracy A_s on source task $\mathcal{T}_s$ and A_t on target task $\mathcal{T}_t$.

3.2.1. Compression Ratio

$$\text{CR}(M) = \frac{H_{\text{equiv}}(M, A_s)}{N}, \tag{1}$$

where H_{equiv} estimates equivalent human training hours:

$$H_{\text{equiv}}(M, A_s) = N \cdot \left(\frac{A_s}{A_{\text{human}}}\right) \cdot h_{\text{per-sample}}. \tag{2}$$

Here, A_{human} is the human baseline accuracy (e.g., 0.98 for MNIST) and $h_{\text{per-sample}}$ is estimated human processing time per example (heuristically, 0.5 hours). CR > 1 indicates the system encodes more than one hour of structural experience per example through hierarchical feature abstraction. CR < 1 indicates inefficient encoding.

3.2.2. Tacit Knowledge Extraction Rate

Given k additional fine-tuning examples improving accuracy from A_s to A_f:

$$\text{TER}(M, k) = \frac{A_f - A_s}{\log(1+k)}. \tag{3}$$

The logarithmic normalization reflects diminishing returns: each additional example provides less marginal structural information, mirroring the compressive nature of structural learning curves. Higher TER indicates efficient extraction of implicit relational patterns.

3.2.3. Cross-Domain Retention

$$\text{CDR}(M, \mathcal{T}_s, \mathcal{T}_t) = \frac{A_t}{A_s}. \tag{4}$$

CDR ≈ 1 indicates full structural transfer; CDR ≈ 0 indicates domain-specific overfitting with no transferable structural representation. CDR > 1 is possible when the target domain shares structural properties with the training domain but is easier in surface presentation.

3.2.4. Experience Efficiency Index

$$\text{EEI}(M) = \frac{A_s/N}{A_{\text{human}}/N_{\text{human}}}, \tag{5}$$

where N_{human} is the estimated number of examples a human requires to reach A_{human}. EEI > 1 means the system learns more efficiently than humans; EEI < 1 indicates that humans extract more structural knowledge per example.

3.2.5. Composite ECI Score

$$\text{ECI}(M) = \sum_{i \in \{\text{CR,TER,CDR,EEI}\}} w_i \cdot \sigma(m_i), \tag{6}$$

where m_i is the raw metric value, $\sigma(\cdot)$ maps to $[0,1]$, and $\sum w_i = 1$. Normalization functions:

$$\begin{aligned} \sigma_{\text{CR}}(x) &= \min(x/10,\ 1), \\ \sigma_{\text{TER}}(x) &= \min(x,\ 1), \\ \sigma_{\text{CDR}}(x) &= \min(x,\ 1), \\ \sigma_{\text{EEI}}(x) &= \min(x/5,\ 1). \end{aligned}$$

Default weights: $w_{\text{CR}} = 0.30$, $w_{\text{TER}} = 0.25$, $w_{\text{CDR}} = 0.25$, $w_{\text{EEI}} = 0.20$.

ECI $\in [0,1]$; higher values indicate more efficient structural compression and transfer. ECI $= 1.0$ represents the theoretical maximum of structural learning efficiency under this metric suite.

Note: Human baseline constants ($h_{\text{per-sample}} = 0.5$ h, $N_{\text{human}} = 1000$) are heuristic placeholders demonstrating mathematical mechanics. Deployment requires empirical calibration via psychometric studies.

3.3. Theoretical Grounding

The ECI framework connects to established principles:

Minimum Description Length (MDL). CR relates to the MDL principle: systems that compress data efficiently do so by capturing underlying structural regularities rather than memorizing surface instances.

Sample Complexity. TER measures sample efficiency in the sense of PAC learning: faster convergence to low structural error with fewer examples indicates simpler, more general hypotheses.

Transfer Theory. CDR quantifies structural task relatedness: high retention implies shared relational geometry between source and target domains.

Information-Theoretic Compression. EEI benchmarks structural extraction efficiency against a biological reference, grounding measurement in a well-understood learning system.

4. Epistemic Confidence via Statistical Path Density

4.1. Motivation

ECI metrics measure the structural quality of what a system has learned. A complementary question is equally important for structural evaluation: *when should a measured ECI value be trusted?* A system might achieve high CDR by coincidence on a favorable test set while exhibiting structural shallowness on other OOD inputs. The epistemic confidence signal addresses this by estimating whether a given test input lies within the structural manifold established during training.

4.2. Activation Path Analysis

For a neural network f with L layers, an input $\mathbf{x}$ produces activation vectors $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(L)}$ the **activation path**. The structural hypothesis is: in-distribution inputs produce activation paths lying on manifolds observed during training; inputs with structurally unfamiliar patterns produce anomalous paths.

For each layer ℓ, we compute four structural signals.

Manifold Density. A Gaussian Mixture Model (GMM) is fit to training-set activations:

$$p_\ell(\mathbf{a}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{a} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \tag{7}$$

with tied covariance and $K = 10$ components. Density score: $d_\ell(\mathbf{a}) = \exp(\log p_\ell(\mathbf{a})/100)$.

Mahalanobis Distance. Statistical distance from the training centroid:

$$D_\ell(\mathbf{a}) = \sqrt{(\mathbf{a} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{a} - \boldsymbol{\mu})}, \tag{8}$$

with distance score $m_\ell(\mathbf{a}) = \exp(-D_\ell(\mathbf{a})/10)$.

Activation Entropy.

$$H_\ell(\mathbf{a}) = -\sum_i p_i \log p_i, \quad p_i = \frac{e^{a_i}}{\sum_j e^{a_j}}, \tag{9}$$

normalized as $e_\ell(\mathbf{a}) = 1 - H_\ell(\mathbf{a})/\log |\mathbf{a}|$.

Activation Sparsity. Hoyer's sparsity measure:

$$s_\ell(\mathbf{a}) = \frac{\sqrt{|\mathbf{a}|} - \|\mathbf{a}\|_1/\|\mathbf{a}\|_2}{\sqrt{|\mathbf{a}|} - 1}. \tag{10}$$

4.3. Layer-Wise and Aggregate Confidence

Layer-level confidence combines all signals multiplicatively so that deficiency in any component reduces overall reliability:

$$c_\ell(\mathbf{a}) = d_\ell^{0.4} \cdot m_\ell^{0.2} \cdot e_\ell^{0.2} \cdot s_\ell^{0.2}. \tag{11}$$

Aggregate confidence with depth-weighted importance λ_ℓ:

$$C(\mathbf{x}) = \sum_{\ell=1}^{L} \lambda_\ell \cdot c_\ell\Big(\mathbf{a}^{(\ell)}\Big), \quad \sum \lambda_\ell = 1. \tag{12}$$

Default weights (deeper layers carry more structural information): $\lambda_{\text{conv2}} = 0.25$, $\lambda_{\text{fc1}} = 0.35$, $\lambda_{\text{fc2}} = 0.40$.

$C(\mathbf{x}) \in [0, 1]$, with higher values indicating that the input's activation path is structurally consistent with the training manifold.

4.4. Confidence-Weighted ECI

ECI metrics are modulated by epistemic confidence over the test set:

$$\text{ECI}_{\text{weighted}}(\mathcal{D}_{\text{test}}) = \text{ECI}(M) \cdot \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{test}}}[C(\mathbf{x})]. \tag{13}$$

Low average confidence over a test set indicates structural extrapolation the system is being evaluated on inputs beyond its structural competence boundary, reducing the reliability of measured ECI values.

5. Experimental Setup

5.1. Datasets

We validate ECI across three distribution regimes. MNIST is chosen as a controlled, structurally interpretable environment for metric validation before scaling to complex domains.

In-Distribution (MNIST). 5,000 training examples (28×28 grayscale), test set of 10,000. Limited training data simulates resource-constrained structural learning.

Near-OOD (FashionMNIST). 10,000 grayscale images of clothing items at identical resolution. Structurally similar (edges, spatial statistics) but semantically distinct. Tests structural transfer and CDR.

Far-OOD (Gaussian Noise). 1,000 images of uniform random noise at MNIST dimensions. Tests epistemic sensitivity to structureless inputs.

5.2. Model Architecture

Standard convolutional neural network: Conv1 (32 filters, 3×3, ReLU, MaxPool); Conv2 (64 filters, 3×3, ReLU, MaxPool); FC1 (128 units, ReLU, Dropout 0.3); FC2 (64 units, ReLU, Dropout 0.3); Softmax output (10 classes). Trained for 3 epochs with Adam (lr = 0.001), cross-entropy loss, without augmentation evaluating core structural learning, not optimization tricks.

5.3. SPD Configuration

Layers analyzed: conv2, fc1, fc2. Dimensionality reduction: PCA preserving 95% variance (20–40 dimensions per layer). Density model: GMM with $K = 10$ components, tied covariances.

5.4. Baselines

Maximum Softmax Probability; Mahalanobis Distance only; MC Dropout [5] (50 forward passes); Deep Ensemble [6] (5 networks).

6. Results

6.1. Classification Performance

MNIST Test Accuracy: 95.51% from 5,000 training examples.

FashionMNIST Zero-Shot: 4.57% close to random chance for 10 classes. The model learns MNIST effectively but encodes structurally shallow representations that do not transfer.

6.2. Epistemic Confidence Distributions

Table 1 reports SPD confidence statistics across the three distribution regimes. The correct ordering $C_{\text{MNIST}} > C_{\text{Fashion}} > C_{\text{Noise}}$ is observed with consistent, narrow distributions within each regime.

Key observations: (1) noise confidence is 25% lower than in-distribution, confirming strong structural discrimination; (2) MNIST–Fashion overlap is moderate ($\Delta\text{mean} = 0.016$), reflecting genuine structural similarity at the geometric level despite semantic difference; (3) within-regime variance is consistently low (std < 0.02), indicating stable confidence estimation.

6.3. Out-of-Distribution Detection

Table 2 reports OOD detection performance. SPD achieves perfect separation of MNIST from structureless noise (AUROC = 1.000, $p < 10^{-300}$) and strong near-OOD detection (AUROC = 0.729, $p < 10^{-220}$).

6.4. ECI Component Metrics

Table 3 presents the measured ECI components. The composite ECI = 0.034 is an honest, non-inflated result. It correctly identifies the structural shallowness of an accuracy-optimized CNN: moderate compression efficiency (CR = 0.487) is offset by negligible structural transfer (CDR = 0.048) and sub-human sample efficiency (EEI = 0.195).

CR = 0.487: The model encodes approximately half an hour of human experience equivalent per training example—moderate structural compression for a simple CNN.

TER ≈ 0: Near-zero because the model has reached an accuracy plateau (95.5%). Structural extraction has saturated; additional examples contribute negligible new relational information.

CDR = 0.048: Structural representations are domain-specific. The model encodes digit pixel statistics, not the topological and geometric properties that would generalize to other visual domains.

EEI = 0.195: The model requires 5× more examples than a human learner to reach comparable accuracy, reflecting the absence of strong structural inductive bias.

6.5. Baseline Comparison

Table 4 compares SPD against simpler epistemic estimators. Full SPD outperforms all single-signal alternatives while requiring only one forward pass, achieving 115% of MC Dropout performance and 95% of ensemble performance.

6.6. SPD Ablation Study

Table 5 quantifies the contribution of each signal. Entropy is the most critical component: its removal causes the largest drop for noise detection (1.00 → 0.75), because noise produces diffuse, high-entropy activation patterns that structured inputs suppress.

7. Discussion

7.1. ECI as a Structural Learning Metric

The results confirm that ECI provides measurable, interpretable discrimination between structural learning modes. Three properties of the measurement are noteworthy from the Structuralist AI perspective.

Honesty. ECI = 0.034 for a CNN that achieves 95.5% accuracy is not a failure it is a correct measurement. Accuracy-optimized systems are expected to score low on ECI because they are not designed to maximize structural knowledge density. ECI does not reward surface performance; it rewards efficient structure extraction.

Discrimination. The large gap between accuracy (0.955) and ECI (0.034) isolates the structural shallowness that accuracy conceals. A Structuralist AI system designed with explicit structural inductive bias should exhibit a different profile: lower raw accuracy on narrow benchmarks but higher ECI through improved CDR and TER.

Falsifiability. The CDR = 0.048 result is a concrete, falsifiable prediction: structural learning architectures that encode transferable relational representations should exhibit CDR substantially above 0.048 on analogous cross-domain pairs. This provides a benchmark for architectural evaluation.

Table 1: SPD confidence statistics by distribution regime.

Distribution	**Mean C**	**Std**	**Median**	**Min**	**Max**
MNIST (in-dist.)	0.4939	0.0197	0.5003	0.4238	0.5430
FashionMNIST (near-OOD)	0.4782	0.0183	0.4795	0.4229	0.5278
Gaussian Noise (far-OOD)	0.3668	0.0124	0.3662	0.3268	0.4005

Table 2: OOD detection performance (AUROC and statistical significance).

Comparison	**AUROC**	**95% CI**	**t-stat**	**p-value**
MNIST vs. Noise	**1.000**	[0.998, 1.000]	192.13	$< 10^{-300}$
MNIST vs. Fashion	**0.729**	[0.715, 0.743]	33.00	1.05×10^{-220}

Table 3: ECI component measurements for the evaluated CNN.

Metric	**Value**	**Normalized**
Compression Ratio (CR)	0.487	0.049
Tacit Extraction Rate (TER)	0.0001	0.0001
Cross-Domain Retention (CDR)	0.048	0.048
Experience Efficiency (EEI)	0.195	0.039
Composite ECI	—	**0.034**

Table 4: OOD detection AUROC and forward-pass cost across methods.

Method	**vs. Noise**	**vs. Fashion**	**Passes**
Max Softmax	0.54	0.61	1
Mahalanobis Only	0.72	0.71	1
SPD (Full)	**1.00**	**0.73**	**1**
MC Dropout	0.84	0.78	50
Deep Ensemble	0.87	0.81	5

7.2. The Epistemic Gap: Why Softmax Fails

The contrast between softmax confidence (AUROC ≈ 0.54 for noise) and SPD (AUROC $= 1.00$) is the most direct empirical demonstration of the epistemic inadequacy of gradient-based confidence estimation. Softmax assigns high confidence to structureless noise because confidence is a function of learned decision-boundary geometry, not of structural proximity to the training manifold.

This is not a calibration problem amenable to post-hoc correction. It is the structural consequence of an objective that rewards output normalization without maintaining any representation of absolute distributional distance. SPD addresses this at the architectural level by measuring confidence directly over activation manifold structure.

The ability to detect structural unfamiliarity has direct consequences for trustworthy deployment:

Table 5: SPD ablation: AUROC under signal removal.

Configuration	**vs. Noise**	**vs. Fashion**
Full SPD	1.00	0.73
Without Entropy	0.75	0.74
Without Sparsity	0.78	0.75
Without Mahalanobis	0.79	0.75
Density Only (GMM)	0.66	0.70

Medical Diagnosis. A system with high epistemic confidence may proceed autonomously; low structural confidence should trigger human review before any clinical decision.

Autonomous Driving. Novel scenarios unusual weather conditions, unexpected obstacles that fall outside the structural training manifold should be flagged for cautious behavior or human intervention rather than handled by extrapolation.

Financial Trading. Low structural confidence on market conditions dissimilar from training data should prevent automated high-risk decisions, where overconfident extrapolation can cause cascading failures.

7.3. Philosophical Implications

ECI reframes the intelligence measurement question from *"Can machines think like humans?"* to *"How efficiently can machines compress and reuse relational structure from experience?"* This shift carries several consequences.

Avoids Anthropocentrism. Intelligence is not defined by similarity to human cognition but by universal properties compression efficiency, structural transfer, epistemic awareness that could characterize any learning system, biological or artificial. A system need not behave like a human to be measurably intelligent under ECI.

Embraces Architectural Diversity. Different architectures may achieve high ECI through different structural mechanisms. Recurrent architectures may excel at sequential compression; convolutional architectures at hierarchical spatial compression; attention-based architectures at relational compression across long-range dependencies. ECI does not privilege any particular implementation.

Enables Measurable Progress. Unlike qualitative debates about "true understanding" or "genuine reasoning," ECI provides concrete, falsifiable metrics for tracking progress in structural learning CDR must improve, TER must reflect genuine saturation of structural content, EEI must approach and exceed human baselines.

Aligns with Established Scientific Principles. Compression and transfer are foundational concepts in cognitive science, neuroscience, and statistical learning theory. Grounding intelligence measurement in these principles connects Structuralist AI evaluation to a mature body of theoretical and empirical work.

7.4. What ECI Does Not Measure

ECI is explicitly bounded in scope. It does not measure embodied intelligence, open-ended agency, or social cognition. It does not define a capability horizon, and it does not propose a path toward any contested intelligence target. These properties are intentional.

ECI measures structural learning quality the three properties that Structuralist AI identifies as the substrate of intelligence: compression efficiency, transfer capacity, and epistemic awareness. Systems that satisfy these three properties at high ECI levels are making genuine structural progress, independent of their task-specific accuracy.

7.5. Limitations and Future Work

Benchmark scope. Validation is conducted on MNIST. Extension to ImageNet-scale classification, language understanding, and multi-modal tasks is required to establish generality of both ECI metrics and SPD confidence estimates.

Human baseline calibration. Constants $h_{\text{per-sample}}$ and N_{human} are heuristic. Rigorous psychometric studies are needed to establish empirically grounded baselines for EEI computation.

Architectural dependence of SPD. The current SPD formulation analyzes convolutional and fully connected activations. Adaptation to attention mechanisms in transformers and temporal dependencies in recurrent architectures requires architectural-specific manifold analysis.

Theoretical bounds. Future work should derive information-theoretic bounds on the maximum achievable ECI given task structural complexity and training data statistics establishing a theoretical frontier alongside the empirical measurement.

8. Conclusion

We have introduced Experience-Compressed Intelligence (ECI) as a measurement framework for Structuralist Artificial Intelligence a research programme that defines intelligence as the capacity to form, reorganize, and transfer relational structures from experience.

ECI measures four dimensions of structural learning quality: compression ratio, tacit knowledge extraction rate, cross-domain retention, and experience efficiency, aggregated into a composite score weighted by epistemic confidence derived from Statistical Path Density analysis. The framework is measurement-first: it does not define a capability horizon, it does not appeal to a contested target, and it does not require subjective judgment. Progress is evaluated against internal, falsifiable properties of learning systems.

The experimental results establish three properties of the framework: (1) SPD produces statistically robust epistemic confidence signals (AUROC = 1.000 for structureless OOD; AUROC = 0.729 for near-OOD) at single-forward-pass cost; (2) ECI = 0.034 for an accuracy-optimized CNN is an honest measurement that correctly exposes structural shallowness concealed by 95.5% task accuracy; (3) the large accuracy–ECI gap is not a framework failure it is the framework's primary signal, identifying precisely where structural learning architectures must improve.

The path forward is not defined by a capability horizon but by concrete measurement: Structuralist AI systems must demonstrate higher CDR through transferable relational representations, higher TER through efficient structural saturation, and higher EEI through stronger structural inductive bias. ECI provides the instruments to measure that progress at every scale of architectural development, without appealing to anything that cannot be directly observed and quantified.

References

[1] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.

[2] Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2002). Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4), 511–533.

[3] Chen, C., & Freedman, D. (2011). Topology noise removal for curve and surface evolution. *International Journal of Computer Vision*, 92(2), 109–126.

[4] Blundell, C., et al. (2015). Weight uncertainty in neural networks. *ICML*, 1613–1622.

[5] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation. *ICML*, 1050–1059.

[6] Lakshminarayanan, B., Pritzel, A., & Blum, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 6402–6413.

[7] Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 31.

[8] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE TKDE*, 22(10), 1345–1359.

[9] Hospedales, T., et al. (2021). Meta-learning in neural networks: A survey. *IEEE TPAMI*, 44(9), 5149–5169.

[10] Wang, Y., et al. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3), 1–34.

[11] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.

[12] Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*.

[13] Davidson, T. (2023). What a compute-centric framework says about takeoff speeds. *AI Alignment Forum*.

[14] Geirhos, R., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

[15] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proc. CVPR*, 427–436.

[16] Lake, B. M., et al. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

[17] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8), 1798–1828.

[18] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning. *Machine Learning*, 110(3), 457–506.

A. Normalization Functions: Justification

CR Normalization $\sigma_{\mathrm{CR}}(x) = \min(x/10, 1)$ assumes a theoretical maximum CR of 10 hours per example, based on expert human performance estimates in pattern recognition domains.

TER Normalization $\sigma_{\mathrm{TER}}(x) = \min(x, 1)$ represents a maximum insight rate of 100% accuracy gain per log-example perfect tacit structural extraction.

CDR Normalization $\sigma_{\mathrm{CDR}}(x) = \min(x, 1)$ caps at full structural retention. Values > 1 are possible when the target domain is structurally easier than the source.

EEI Normalization $\sigma_{\mathrm{EEI}}(x) = \min(x/5, 1)$ assumes a maximum of 5× human efficiency, acknowledging computational advantages in pattern matching while preserving human superiority in complex structural reasoning.

All constants are adjustable based on domain-specific calibration or empirical psychometric studies.

Note on Experimental Scope. The experiments reported in this paper are not presented as performance claims. MNIST is used as a controlled, structurally interpretable environment chosen deliberately for its simplicity not for the impressiveness of its results. The objective is to demonstrate that ECI metrics are *measurable*, *internally consistent*, and *discriminative across distribution regimes*, not to establish state-of-the-art performance. A framework paper is not invalidated by stronger empirical results elsewhere; on the contrary, systems that outperform the evaluated CNN on task accuracy are expected to do so and ECI exists precisely to ask what those systems achieve *beyond* accuracy. Higher benchmark scores from other architectures do not compete with ECI; they are inputs to it.