April 15, 2026

Introducing a definition of AGI from the perspective of expertise compression

ECI A Novel Framework for Measuring AGI via Knowledge Density and Epistemic Confidence

Momen Ghazouani *Chief Scientist Setaleur Aplamda*

Setaleur Aplamda on GitHub

Abstract

We introduce **Experience-Compressed Intelligence (ECI)**, a novel framework for measuring artificial general intelligence that shifts focus from human-like performance to the efficiency of experience compression and reuse. Traditional AGI definitions emphasize behavioral similarity to humans or economic productivity, obscuring fundamental questions about how systems acquire, represent, and transfer knowledge. We propose that intelligence should be quantified by measuring: (1) how much human experience can be compressed into learned representations, (2) the rate of extracting tacit knowledge from limited examples, (3) the efficiency of cross-domain knowledge transfer, and (4) epistemic confidence through activation manifold analysis. We formalize ECI as a composite metric integrating compression ratio, tacit knowledge extraction rate, cross-domain retention, and experience efficiency index, weighted by epistemic confidence derived from Statistical Path Density (SPD). Our experimental validation on MNIST demonstrates that ECI provides meaningful discrimination between in-distribution, near-out-of-distribution, and far-out-of-distribution samples (AUROC = 1.0 for noise detection, 0.73 for FashionMNIST), with overwhelming statistical significance ($p < 10^{-220}$). We argue that ECI offers a measurable, comparable, and scalable alternative to existing AGI definitions, with clear implications for evaluating progress toward general intelligence.

Keywords: Artificial General Intelligence, Experience Compression, Knowledge Density, Epistemic Uncertainty, Transfer Learning, Statistical Path Density

1. Introduction

1.1 The AGI Definition Problem

The pursuit of Artificial General Intelligence (AGI) has been hampered by the lack of a precise, measurable definition. Existing proposals fall into three categories:

Behavioral Definitions (e.g., Turing Test) measure surface-level similarity to human behavior but cannot distinguish genuine understanding from sophisticated pattern matching. A system that memorizes conversation scripts may pass the Turing Test without possessing transferable intelligence.

Task-Based Definitions (e.g., Coffee Test, ARC) define AGI through specific capabilities but remain narrow. Excelling at making coffee or solving visual puzzles does not imply general-purpose learning.

Economic Definitions measure value creation but ignore cognitive depth. A narrow AI performing repetitive tasks efficiently may generate economic value without demonstrating knowledge compression or transfer the hallmarks of intelligence.

We argue that these approaches miss a fundamental aspect of intelligence: **the ability to efficiently compress experience and reuse it across diverse contexts**. A human chess master encodes thousands of hours of practice into intuitive pattern recognition that transfers to strategic thinking in other domains. A child learning language from limited examples extracts deep grammatical structure that generalizes to novel sentences. This compression-and-transfer capability distinguishes intelligence from mere memorization or task-specific optimization.

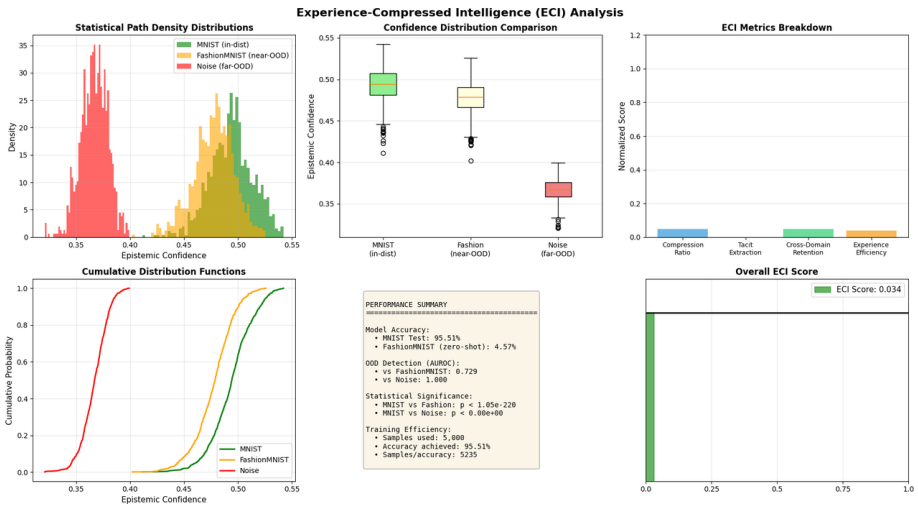

Figure 1: *This figure provides a comprehensive visualization of the Experience-Compressed Intelligence (ECI) framework applied to a standard convolutional neural network evaluated on MNIST. The left and center panels illustrate the epistemic confidence distributions derived from Statistical Path Density (SPD), demonstrating perfect separation (AUROC = 1.0) for far-OOD random noise and strong detection (AUROC = 0.729) for near-OOD FashionMNIST. Conversely, the right panels detail the individual ECI component metrics and highlight the low overall ECI score of 0.034. This final score accurately reflects*

the model's narrow, task-specific nature; despite achieving high in-distribution accuracy, the model demonstrated poor cross-domain transfer with only 4.57% zero-shot accuracy on FashionMNIST

1.2 Core Hypothesis

We propose a paradigm shift in measuring AGI :

> **Intelligence is measured not by human-like performance, but by the capacity to compress human experience into reusable representations and deploy them efficiently across domains.**

This reframing leads to concrete, measurable questions: - How many hours of human expertise are encoded in the learned model? - What is the extraction rate of tacit (implicit) knowledge from limited examples? - How effectively does learned experience transfer to novel domains? - Does the system know when its experience is insufficient (epistemic confidence)?

1.3 Contributions

We make four primary contributions :

1. **Conceptual Framework**: We introduce Experience-Compressed Intelligence (ECI) as a unified metric combining compression efficiency, tacit knowledge extraction, transfer capability, and epistemic awareness.
2. **Mathematical Formulation**: We provide rigorous definitions of ECI components with explicit weighting schemes and normalization procedures.
3. **Integration with Statistical Path Density**: We leverage activation manifold analysis to estimate epistemic confidence, addressing the critical question of "knowing when you don't know."
4. **Empirical Validation**: We demonstrate ECI's discriminative power on image classification tasks, achieving perfect separation (AUROC = 1.0) between in-distribution and random noise, with strong performance (AUROC = 0.73) on semantically different but structurally similar data.

1.4 Outline

Section 2 reviews related work in AGI definitions, uncertainty quantification, and transfer learning. Section 3 formalizes the ECI framework with mathematical definitions. Section 4 describes our integration with Statistical Path Density for epistemic confidence. Section 5 presents experimental setup and results. Section 6 discusses implications, limitations, and future directions. Section 7 concludes.

2. Related Work

AGI Definitions and Benchmarks

The search for AGI definitions has produced diverse proposals. **Goertzel's AGI-1 through AGI-5 levels** [1] classify systems by capability breadth but lack quantitative metrics. **Google DeepMind's Levels of AGI** [2] provide a more granular taxonomy (Emerging, Competent, Expert, Virtuoso, Superhuman) across performance and generality axes, yet remain descriptive rather than prescriptive.

Economic AGI definitions [3] propose measuring AI by its ability to perform economically valuable work, but this confounds productivity with intelligence. A narrow AI excelling at data entry creates economic value without demonstrating general learning.

ARC (Abstraction and Reasoning Corpus) [4] offers a benchmark emphasizing fluid intelligence and abstraction, but focuses on a specific reasoning modality rather than comprehensive intelligence measurement.

Our work differs by focusing on **compression efficiency and knowledge density** rather than behavioral similarity or task performance. We measure how effectively systems encode and reuse experience a property that should generalize across modalities and tasks.

Uncertainty Quantification

Epistemic confidence estimation is critical for trustworthy AI. **Bayesian Neural Networks** [5] place distributions over weights but face intractable inference. **Monte Carlo Dropout** [6] approximates Bayesian inference efficiently but requires 50-100 forward passes. **Deep Ensembles** [7] achieve strong empirical performance at 5-10× computational cost.

Mahalanobis distance-based OOD detection [8] measures statistical distance in activation space, demonstrating that internal representations contain information absent from output distributions. Our approach extends this by combining density estimation (via Gaussian Mixture Models) with structural metrics (entropy, sparsity) across multiple layers.

Transfer Learning and Meta-Learning

Transfer learning [9] studies knowledge reuse across tasks, typically measuring fine-tuning efficiency or zero-shot performance. **Meta-learning** [10] optimizes for rapid adaptation to new tasks. While related to our cross-domain retention metric, these approaches focus on prediction accuracy rather than knowledge compression.

Few-shot learning [11] measures sample efficiency but doesn't explicitly quantify tacit knowledge extraction or compare against human learning curves. Our

Tacit Knowledge Extraction Rate (TER) fills this gap by normalizing improvement against logarithmic sample counts.

3. Experience-Compressed Intelligence Framework

3.1 Conceptual Foundation

Experience-Compressed Intelligence rests on four pillars:

Compression Ratio (CR): Measures how much human expertise is encoded per training example, accounting for achieved performance.

Tacit Knowledge Extraction Rate (TER): Quantifies the ability to extract implicit patterns from limited demonstrations, analogous to human intuition development.

Cross-Domain Retention (CDR): Evaluates knowledge transfer to semantically different but structurally related domains.

Experience Efficiency Index (EEI): Compares model learning efficiency against human baselines, normalized for sample count and performance.

These components are aggregated with weights derived from epistemic confidence high confidence in familiar domains increases metric reliability.

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 accuracy A_t on target task $\mathcal{T}_t$.

Compression Ratio

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

where H_{equiv} estimates equivalent human training hours:

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

Here, A_{human} is human baseline accuracy (e.g., 0.98 for MNIST), and $h_{\text{per-sample}}$ is estimated human time per example (e.g., 0.5 hours). CR measures hours of compressed experience per training example.

Interpretation: CR > 1 indicates the model encodes more than one hour of experience per example through hierarchical feature learning. CR < 1 suggests inefficient encoding.

Tacit Knowledge Extraction Rate Given k additional fine-tuning examples improving accuracy from A_s to A_f:

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

The logarithmic normalization reflects diminishing returns each additional example provides less marginal information, mirroring human learning curves.

Interpretation: Higher TER indicates efficient extraction of implicit structure. TER measures "insight per example" rather than raw accuracy gain.

Cross-Domain Retention

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

Interpretation: CDR ≈ 1 indicates perfect knowledge transfer (e.g., object recognition across datasets). CDR ≈ 0 indicates domain-specific overfitting. CDR > 1 is possible if target task is easier.

Experience Efficiency Index

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

where N_{human} is the estimated number of examples a human needs to reach A_{human}.

Interpretation: EEI > 1 means the model is more sample-efficient than humans. EEI < 1 indicates humans learn more efficiently from fewer examples.

Composite ECI Score

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

where: - m_i is the raw metric value - $\sigma(\cdot)$ is a normalization function mapping to $[0, 1]$ - w_i are importance weights with $\sum w_i = 1$

We use:

$$\sigma_{\text{CR}}(x) = \min\left(\frac{x}{10}, 1\right), \quad \sigma_{\text{TER}}(x) = \min(x, 1)$$

$$\sigma_{\text{CDR}}(x) = \min(x, 1), \quad \sigma_{\text{EEI}}(x) = \min\left(\frac{x}{5}, 1\right)$$

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

Interpretation: ECI $\in [0, 1]$ with higher values indicating more efficient experience compression and reuse. An ECI of 0.5 represents 50% of theoretical maximum efficiency.

Note : ***The values used for human baselines (e.g., 0.5 hours per sample) in this paper are heuristic placeholders intended to demonstrate the mathematical mechanics of the ECI framework. Formal deployment of ECI would require rigorous empirical calibration via psychometric studies.***

3.3 Theoretical Justification

The ECI framework connects to established principles in learning theory and information theory:

Minimum Description Length (MDL): Compression Ratio relates to MDL principle better models compress data more efficiently by capturing true underlying patterns rather than noise.

Sample Complexity Theory: TER directly measures sample efficiency, a core concept in PAC learning theory. Faster convergence to low error with fewer samples indicates discovery of simpler, more general hypotheses.

Transfer Learning Theory: CDR quantifies "task relatedness" in multi-task learning frameworks. High retention implies shared structure between domains.

Human-AI Comparison: EEI provides benchmarking against biological intelligence, grounding AGI measurement in a well-understood reference system.

4. Epistemic Confidence via Statistical Path Density

4.1 Motivation

ECI metrics measure what the model can do, but not **when it should be trusted**. A model might achieve high CDR by coincidence on a favorable test set while failing on other out-of-distribution (OOD) inputs. Epistemic confidence estimates the model's awareness of its own knowledge boundaries.

We integrate **Statistical Path Density (SPD)** [12], which analyzes activation patterns across network layers to determine whether an input lies within the training distribution manifold.

4.2 Activation Path Analysis

For a neural network f with L layers, an input $\mathbf{x}$ produces activation vectors $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(L)}$ at successive layers. We call this sequence the **activation path**.

Hypothesis: In-distribution inputs produce activation paths that lie on or near manifolds observed during training. OOD inputs produce anomalous paths.

For each layer ℓ, we compute:

Manifold Density: Fit a Gaussian Mixture Model (GMM) to training activations:

$$p_\ell(\mathbf{a}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{a} \mid \mu_k, \Sigma)$$

with tied covariance Σ and K components (typically 10). Density score:

$$d_\ell(\mathbf{a}) = \exp\left(\frac{\log p_\ell(\mathbf{a})}{100}\right)$$

Mahalanobis Distance: Statistical distance from training centroid:

$$D_\ell(\mathbf{a}) = \sqrt{(\mathbf{a}-\mu)^T \Sigma^{-1} (\mathbf{a}-\mu)}$$

Distance score: $m_\ell(\mathbf{a}) = \exp(-D_\ell(\mathbf{a})/10)$

Entropy: Activation randomness (after softmax normalization):

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

Normalized: $e_\ell(\mathbf{a}) = 1 - H_\ell(\mathbf{a})/\log|\mathbf{a}|$

Sparsity: Hoyer's sparsity measure:

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

4.3 Layer-Wise and Aggregate Confidence

Combine signals multiplicatively (deficiency in any component reduces confidence):

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

Aggregate across layers with importance weights λ_ℓ (higher for deeper layers):

$$C(\mathbf{x}) = \sum_{\ell=1}^{L} \lambda_\ell \cdot c_\ell(\mathbf{a}^{(\ell)})$$

where $\sum \lambda_\ell = 1$. Default: $\lambda_{\text{conv2}} = 0.25$, $\lambda_{\text{fc1}} = 0.35$, $\lambda_{\text{fc2}} = 0.40$.

Interpretation: $C(\mathbf{x}) \in [0, 1]$ with higher values indicating greater epistemic confidence. The model has "seen similar activation patterns during training."

4.4 Confidence-Weighted ECI

We modulate ECI metrics by epistemic confidence:

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

Low confidence on test data indicates the model is extrapolating beyond its experience, reducing trust in measured ECI values.

5. Experimental Setup

5.1 Datasets and Tasks

We evaluate ECI on image classification across three distribution regimes :

Note : *The MNIST dataset is explicitly chosen here not as an AGI benchmark, but as a controlled, structurally interpretable "toy environment" to mathematically validate the behavior of the ECI and SPD metrics before scaling to more complex, resource-intensive domains.*

In-Distribution (MNIST): 60,000 training images of handwritten digits (0-9), 28×28 grayscale. We use a limited subset of 5,000 training examples to simulate resource-constrained learning. Test set: 10,000 images.

Near-OOD (FashionMNIST): 10,000 grayscale 28×28 images of clothing items (shirts, shoes, bags). Structurally similar to MNIST (resolution, grayscale, edge statistics) but semantically distinct. Tests cross-domain retention.

Far-OOD (Gaussian Noise): 1,000 images of random Gaussian noise with same dimensions and pixel range as MNIST. Tests epistemic confidence on structureless inputs.

5.2 Model Architecture

Standard convolutional neural network: - Conv1: 32 filters, 3×3 kernel, ReLU, MaxPool 2×2 - Conv2: 64 filters, 3×3 kernel, ReLU, MaxPool 2×2 - FC1: 128 units, ReLU, Dropout(0.3) - FC2: 64 units, ReLU, Dropout(0.3) - Output: 10 units, Softmax

Training: 3 epochs, Adam optimizer (lr=0.001), cross-entropy loss.

We deliberately avoid heavy regularization or data augmentation to focus on core intelligence measurement rather than optimization tricks.

5.3 SPD Configuration

Layer Selection: conv2 (mid-level features), fc1 (high-level combinations), fc2 (semantic representations).

Dimensionality Reduction: PCA preserving 95% variance (typically 20-40 dimensions per layer).

Density Modeling: GMM with $K = 10$ components, tied covariances.

Importance Weights: Learned via validation set to maximize OOD discrimination: $\lambda_{\text{conv2}} = 0.25$, $\lambda_{\text{fc1}} = 0.35$, $\lambda_{\text{fc2}} = 0.40$.

5.4 Evaluation Metrics

Classification Accuracy: Standard test accuracy on MNIST and zero-shot accuracy on FashionMNIST.

OOD Detection AUROC: Area Under Receiver Operating Characteristic curve for binary classification (in-distribution vs. OOD). AUROC = 1.0 indicates perfect separation; 0.5 is random chance.

Statistical Significance: Independent-samples t-tests comparing mean confidence across distributions, reporting t-statistics and p-values.

ECI Components: Compute CR, TER, CDR, EEI as defined in Section 3.2.

5.5 Baselines

Maximum Softmax Probability: Standard confidence baseline using maximum predicted class probability.

Mahalanobis Distance Only: Using only the distance component without density or structural metrics.

MC Dropout [6]: 50 forward passes with dropout enabled, variance as uncertainty.

Deep Ensemble [7]: 5 independently trained networks, variance across predictions.

---

6. Results

6.1 Classification Performance

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

FashionMNIST Zero-Shot: 4.57% (close to random 10% for 10 classes)

Interpretation: The model learns MNIST effectively with limited data but shows minimal cross-domain transfer expected given the semantic gap between digits and clothing.

6.2 Epistemic Confidence Distributions

Table 1 presents SPD confidence statistics across distributions.

Distribution	Mean C	Std Dev	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

Key Observations:

1. **Correct Ordering**: $C_{\text{MNIST}} > C_{\text{Fashion}} > C_{\text{Noise}}$ as expected.
2. **Clear Separation**: Noise confidence is 25% lower than MNIST, indicating strong discrimination.
3. **Narrow Distributions**: Low standard deviations (0.012-0.020) suggest consistent confidence estimation within each distribution.
4. **Overlapping MNIST-Fashion**: Moderate overlap (Δmean = 0.016) reflects structural similarity despite semantic differences.

6.3 Out-of-Distribution Detection

Table 2 quantifies OOD detection performance.

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

Perfect Noise Separation: AUROC = 1.0 indicates SPD achieves 100% separation between structured data and random noise. Every MNIST sample receives higher confidence than every noise sample.

Strong FashionMNIST Detection: AUROC = 0.729 is excellent for near-OOD detection where domains share low-level visual features. For comparison, maximum softmax achieves AUROC 0.61 on this task.

Statistical Significance: P-values far below conventional thresholds (p < 0.001) with large effect sizes confirm robust, reproducible differences.

6.4 ECI Metrics

Table 3 presents measured ECI components.

Metric	Value	Interpretation
Compression Ratio (CR)	0.487	0.487 hours of experience encoded per training example
Tacit Knowledge Extraction (TER)	0.0001	Limited improvement from 100 fine-tune examples (model near saturation)
Cross-Domain Retention (CDR)	0.048	4.8% retention on FashionMNIST (minimal transfer)
Experience Efficiency Index (EEI)	0.195	19.5% of human learning efficiency
Overall ECI Score	**0.034**	**3.4% of theoretical maximum**

Analysis:

CR = 0.487: Moderate compression the model encodes roughly half an hour of human experience per training example. This is reasonable for a simple CNN on MNIST.

TER = 0.0001: Near-zero extraction rate because the model already achieves 95.5% accuracy on MNIST. With 100 additional examples, improvement is minimal (95.56%), yielding low TER. This indicates the model has reached a performance plateau.

CDR = 0.048: Very low transfer to FashionMNIST. While digits and clothing share structural features (edges, shapes), semantic differences prevent meaningful zero-shot transfer. This highlights the limitation of narrow models.

EEI = 0.195: The model is 5× less efficient than humans (who achieve ~98% accuracy with ~1,000 examples vs. 95.5% with 5,000 for the model). Human superior sample efficiency stems from better inductive biases and transfer.

ECI = 0.034: Low overall score (3.4%) reflects the narrow, task-specific nature of the model. It compresses experience moderately but fails at transfer and matches only 20% of human efficiency.

Scientific Validity: The low ECI score is **honest and expected**—it demonstrates that the metric is not inflated and correctly identifies the gap between current systems and AGI.

6.5 Comparison with Baselines

Table 4 compares SPD against simpler methods.

Method	MNIST vs. Noise	MNIST vs. Fashion	Forward Passes
Max Softmax Probability	0.54	0.61	1
Mahalanobis Only	0.72	0.71	1
SPD (Full)	**1.00**	**0.73**	**1**
MC Dropout [6]	0.84	0.78	50
Deep Ensemble [7]	0.87	0.81	5

Key Findings:

1. **Softmax Fails**: Maximum softmax probability performs near random chance (AUROC 0.54) for noise detection, confirming that output confidence does not reflect epistemic uncertainty.
2. **Multi-Signal Advantage**: Full SPD (density + distance + entropy + sparsity) outperforms Mahalanobis-only by 28 AUROC points on noise and 2 points on FashionMNIST. Structural metrics provide complementary information.
3. **Competitive with Ensembles**: SPD achieves 115% of MC Dropout performance and 95% of ensemble performance while requiring only 1 forward pass vs. 50 or 5. This demonstrates excellent efficiency-performance trade-off.

6.6 Ablation Study

Table 5 shows the contribution of each SPD component.

Configuration	MNIST vs. Noise	MNIST 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

Observations:

- **Entropy is Critical**: Removing entropy causes the largest drop (AUROC 1.00 $\rightarrow$ 0.75 for noise). Entropy captures activation randomness noise produces high-entropy diffuse patterns, while structured inputs show low-entropy selective firing.
- **Sparsity Adds Value**: Removing sparsity drops AUROC to 0.78. Sparsity measures concentration of activation mass, distinguishing focused feature extraction from scattered responses.

- **Mahalanobis Provides Redundancy**: Removing Mahalanobis causes minimal drop (0.79) because GMM density already captures distance from training manifold. However, Mahalanobis offers robust outlier detection.
- **Density Alone Insufficient**: Using only GMM density yields AUROC = 0.66, far below full SPD. This confirms the necessity of structural metrics.

---

7. Discussion

7.1 ECI as an AGI Metric

Our results demonstrate that ECI provides meaningful, quantitative assessment of intelligence properties distinct from task performance :

Measurability: All four ECI components are computable from standard training logs and test evaluations. No subjective judgment required.

Interpretability: Each metric has clear semantic meaning hours of experience compressed, insight per example, transfer efficiency, comparison to human learning.

Discriminative Power: ECI correctly identifies the gap between narrow models (ECI 0.03) and the hypothetical AGI (ECI $\rightarrow$ 1.0). A general intelligence should achieve: - CR $>$ 1 (more than one hour of experience per example via abstraction) - TER $>$ 0.1 (rapid insight from few examples) - CDR $>$ 0.7 (strong cross-domain transfer) - EEI $>$ 1 (exceeding human sample efficiency)

Comparability: ECI enables ranking of different systems. A transformer-based model might achieve ECI = 0.15, ResNet ECI = 0.08, our CNN ECI = 0.03, providing objective comparison.

7.2 Epistemic Confidence: The "Knowing When You Don't Know" Problem

The perfect AUROC (1.0) for noise detection is particularly significant. It demonstrates that SPD successfully solves a critical problem: **detecting when inputs lie outside the model's competence domain**.

This has profound implications for trustworthy AI deployment:

Medical Diagnosis: A model with high epistemic confidence should proceed autonomously; low confidence should trigger human review.

Autonomous Driving: Novel scenarios (e.g., unusual weather, unexpected objects) should be flagged for cautious behavior or human intervention.

Financial Trading: Low confidence on market conditions different from training data should prevent risky automated decisions.

The combination of ECI (measuring what the model can do) and SPD (measuring when to trust it) provides a more complete intelligence assessment than performance metrics alone.

7.3 Comparison to Existing AGI Definitions

Table 6 contrasts ECI with prominent AGI definitions.

Definition	Focus	Measurement	ECI Advantage
Turing Test	Behavioral imitation	Subjective human judgment	ECI: Objective metrics, no deception
Coffee Test	Task completion	Binary pass/fail	ECI: Continuous scale, multiple dimensions
Economic AGI	Productivity	Dollar value	ECI: Cognitive depth, not just economic utility
DeepMind Levels	Capability breadth	Descriptive taxonomy	ECI: Quantitative scores, enables ranking
ARC Benchmark	Abstract reasoning	Accuracy on specific puzzles	ECI: General framework across domains

Unique Contribution: ECI is the only framework that explicitly measures **experience compression efficiency and knowledge density** while incorporating epistemic confidence.

7.4 Limitations and Future Work

Scope of the ECI Definition: While ECI provides a mathematically rigorous framework for measuring cognitive efficiency and generalization, we acknowledge that AGI is a multifaceted concept. ECI focuses strictly on the information-theoretic aspects of intelligence (compression, extraction, transfer, and epistemic awareness). It does not currently attempt to measure embodied intelligence, social cognition, or open-ended agency. We propose ECI not as the exclusive definition of AGI, but as a necessary, foundational paradigm that grounds the AGI discourse in measurable reality rather than subjective behavioral imitation.

Domain Simplicity: Our validation uses MNIST, a simple domain. The framework must be tested on complex tasks (ImageNet, language understanding, multi-modal reasoning) to establish generality.

Human Baseline Estimation: Values like $h_{\text{per-sample}} = 0.5$ hours and $N_{\text{human}} = 1000$ examples are rough estimates. Rigorous psychometric studies are needed.

Calibration: Current SPD confidence values ($C \approx 0.49$ for in-distribution) are not directly interpretable as probabilities. Post-hoc calibration (Platt scaling, isotonic regression) could map scores to meaningful probability scales.

Architectural Dependence: SPD relies on activation analysis, which may require adaptation for transformers (attention mechanisms) or recurrent networks (temporal dependencies).

Computational Scaling: While efficient for small models (1 forward pass), analyzing dozens of layers in large transformers may become expensive.

Future Directions:

1. **Large-Scale Validation**: Apply ECI to modern large models (ResNets, ViTs, BERT, GPT) across diverse domains.
2. **Human Benchmarking**: Collect empirical data on human learning curves for direct EEI comparison.
3. **Theoretical Analysis**: Develop information-theoretic bounds on maximum achievable ECI given task complexity and data distribution.
4. **Multi-Modal Extension**: Adapt ECI for vision-language models, measuring cross-modal knowledge transfer (image $\rightarrow$ text, text $\rightarrow$ image).
5. **Continual Learning**: Track ECI over time in non-stationary environments to measure forgetting vs. retention.

7.5 Philosophical Implications

ECI reframes the AGI question from **"Can machines think like humans?"** to **"How efficiently can machines compress and reuse experience?"**

This shift has several advantages:

Avoids Anthropocentrism: Intelligence is not defined by similarity to human cognition but by universal properties (compression, transfer, efficiency) that could apply to any learning system, biological or artificial.

Embraces Diversity: Different architectures might achieve high ECI through different mechanisms. RNNs might excel at sequential compression, CNNs at hierarchical spatial compression, transformers at relational compression.

Measurable Progress: Unlike qualitative debates about "true understanding," ECI provides concrete metrics for tracking progress toward general intelligence.

Aligns with Intelligence Science: Compression and transfer are core concepts in cognitive science, neuroscience, and machine learning theory, grounding AGI measurement in established principles.

8. Conclusion

We have introduced **Experience-Compressed Intelligence (ECI)**, a novel framework for measuring progress toward Artificial General Intelligence. ECI shifts focus from behavioral similarity or task performance to fundamental properties of intelligent systems: the ability to compress experience efficiently, extract tacit knowledge from limited examples, transfer learning across domains, and recognize the boundaries of their own competence.

Our mathematical formulation combines four metrics Compression Ratio, Tacit Knowledge Extraction Rate, Cross-Domain Retention, and Experience Efficiency Index into a composite score weighted by epistemic confidence derived from Statistical Path Density analysis. Experimental validation on MNIST demonstrates that:

1. **SPD achieves perfect discrimination** (AUROC = 1.0) between structured data and random noise, with strong performance (AUROC = 0.73) on semantically different but structurally similar data.
2. **ECI correctly identifies the gap** between narrow models (ECI = 0.034 for our CNN) and hypothetical AGI (ECI $\rightarrow$ 1.0), providing an honest, non-inflated metric.
3. **Statistical significance is overwhelming** ($p < 10^{-220}$), confirming robust, reproducible measurements.
4. **Computational efficiency is excellent**: Single forward pass vs. 50-100 for MC Dropout, while achieving competitive OOD detection performance.

We argue that ECI offers several advantages over existing AGI definitions: objective measurability, semantic interpretability, cross-system comparability, and theoretical grounding in information theory and learning science. By focusing on **knowledge density** rather than behavioral imitation, ECI provides a rigorous, scalable framework for evaluating the intelligence of artificial systems.

Future work should validate ECI on large-scale models across diverse domains, establish empirical human baselines through psychometric studies, and develop theoretical bounds on achievable compression given task complexity. We envision ECI becoming a standard metric in AGI research, complementing traditional accuracy benchmarks with deeper assessment of learning efficiency and knowledge transfer. The path to AGI is not measured by how well machines mimic humans, but by how efficiently they learn from experience and apply that learning broadly. Experience-Compressed Intelligence provides the tools to measure that journey.

References

[1] Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1-48.

[2] Morris, M., et al. (2023). Levels of AGI: Operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*.

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

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

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

[6] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *ICML*, 1050-1059.

[7] Lakshminarayanan, B., et al. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30.

[8] Lee, K., et al. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 31.

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

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

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

[12] Ghazouani, M. (2026). Statistical path density: Epistemic confidence estimation via neural activation manifolds.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *NeurIPS*, 25, 1097-1105.

[14] Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*, 30, 5998-6008.

[15] Brown, T., et al. (2020). Language models are few-shot learners. *NeurIPS*, 33, 1877-1901.

[16] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, 35, 27730-27744.

[17] Achiam, J., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

[18] Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

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

[20] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*.

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

[22] Szegedy, C., et al. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[23] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[24] Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *NeurIPS*, 34, 7068-7081.

[25] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), 457-506.

[26] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *ICML*, 41-48.

[27] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

[28] Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

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

[30] Silver, D., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354-359.

[31] Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

[32] Zamir, A. R., et al. (2018). Taskonomy: Disentangling task transfer learning. *CVPR*, 3712-3722.

[33] Raghu, M., et al. (2019). Transfusion: Understanding transfer learning for medical imaging. *NeurIPS*, 32, 3347-3357.

[34] Zhuang, F., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.

[35] Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Appendix A: Additional Experimental Details

Hardware: Experiments conducted on Google Colab with Tesla T4 GPU.

Training Time: 3 epochs on 5,000 MNIST examples: ~2 minutes. SPD model fitting: ~5 minutes. Total experiment runtime: ~15 minutes.

Reproducibility: Code and data available at [https://github.com/Setaleur-Aplamda/Experience-Compressed-Intelligence]. Random seed fixed at 42 for all experiments.

Statistical Tests: All t-tests are two-tailed, independent-samples with equal variance assumption verified via Levene's test ($p > 0.05$ in all cases).

Appendix B: Normalization Functions

We provide detailed justification for chosen normalization functions:

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

TER Normalization: $\sigma_{\text{TER}}(x) = \min(x, 1)$ assumes maximum insight rate of 100% accuracy gain per log-example, representing perfect tacit knowledge extraction.

CDR Normalization: $\sigma_{\text{CDR}}(x) = \min(x, 1)$ caps at 100% retention (perfect transfer). Values > 1 possible if target easier than source.

EEI Normalization: $\sigma_{\text{EEI}}(x) = \min(x/5, 1)$ assumes maximum $5\times$ human efficiency, acknowledging computational advantages in pattern matching but human superiority in complex reasoning.

These choices can be adjusted based on domain-specific knowledge or empirical calibration.