

Analysing the decision making of machine learning models

Kārlis Zars

Dr.sc.comp., Prof. Guntis Bārzdīņš

Research topic

- Core problem (unchanged): modern Machine Learning (ML) systems make high-impact decisions, but internal logic is often opaque ("black box").
- What's new (this year): shift from post-hoc Explainable Artificial Intelligence (XAI) towards Mechanistic Interpretability (MI) –opening the model and tracing features + circuits + causal pathways.
- Thesis statement: move from "explain the output" → to "identify the internal mechanism that produced the output".

Kārlis Zars

Academic & Teaching

- 3rd year PhD Student
- Riga Business School
 - Lecturer: Probability, Visualisation, Bachelor thesis
 - AI Studio consulting
- Lecturer at Baltijas Datoru Akadēmija - AI corporate trainer
- Educator/Content creator on Coursera platform
 - GenAI, Coding, Business etc. 33+ courses
 - 55000+ students

Industry Experience

- CTO at SmartSol
- Lead developer / CTO at DoctoWell (medical field software)
- AI Master Labs: consulting

Why it matters

Why interpretability is now a core engineering + research problem

High-stakes use cases

credit scoring, fraud detection, medical triage, legal risk assessment (decisions must be explainable, contestable, auditable).

Failure mode

a model can be right for the wrong reasons (shortcuts / spurious correlations), hard to detect with surface-level explanations.

Decision-making gap

stakeholders ask "why?", but the model outputs "what".

- **Research framing:** interpretability supports trust, debugging, safety, and control (finding and fixing the mechanism, not just describing it).
- **Example:** a loan rejection should reveal whether the model used a proxy (e.g., ZIP code) rather than legitimate risk factors.

Last year's baseline: Post-hoc Explainable Artificial Intelligence (XAI)

Year 2 focus: explaining model behaviour from the outside

Definition: post-hoc Explainable Artificial Intelligence (XAI) explains a trained model after the fact by relating inputs to outputs, without accessing the true internal mechanism.

Methods covered:

- SHapley Additive exPlanations (SHAP): local feature attribution
- Local Interpretable Model-agnostic Explanations (LIME): local surrogate explanations
- Partial Dependence Plot (PDP) / Individual Conditional Expectation (ICE): global + local response curves
- Counterfactual explanations: "minimal change to flip the decision"

Strengths

model-agnostic, fast to apply, useful for reporting and first-pass debugging.

Limitations

can be incomplete or misleading for deep models; often explains correlations around the output—not the computation inside.

Comparison:

- Post-hoc Explainable Artificial Intelligence (XAI): "Which inputs influenced the output?"
- Mechanistic Interpretability (MI): "Which internal features/circuits computed the output?"

What changed since last year (field pivot)

2025–2026: from "feature importance" → to "features, circuits, and causal control"

Mechanistic Interpretability (MI) became the frontier: reverse-engineer internal representations and computations; focus on causal understanding.

Enabling ideas:

- Sparse Autoencoder (SAE): sparse "feature dictionaries" from activations (interpretable internal features)
- Circuit tracing / interventions: edit activations and observe controlled changes in behaviour
- Interpretability-by-design: train models with sparse circuits so internal pathways are smaller and more readable

Why it's exciting for a PhD: open problems remain—scaling, automation, evaluation of faithfulness, usability of outputs.

Concrete anchors:

- Anthropic: "Tracing the thoughts of a large language model"
- OpenAI: "Understanding neural networks through sparse circuits" + released tooling
- TransformerLens: practical tooling for transformer Mechanistic Interpretability (MI) workflows

References:

- <https://www.anthropic.com/research/tracing-thoughts-language-model>
- <https://openai.com/index/understanding-neural-networks-through-sparse-circuits/>
- <https://github.com/TransformerLensOrg/TransformerLens>

Mechanistic Interpretability (MI)

Definition: Mechanistic Interpretability (MI) aims to understand internal computations that produce a model's behaviour by identifying internal features and circuits and validating them with causal interventions.

Post-hoc Explainable Artificial Intelligence (XAI)

input → output relationships (useful, often correlational).

Mechanistic Interpretability (MI)

input → internal mechanism → output (causality + internal structure).

What you do in Mechanistic Interpretability (MI):

- find features inside activations
- map circuits (subgraphs) that implement behaviours
- intervene (edit activations / ablate edges) to test causal impact

Example line: instead of "feature X mattered," ask "which internal pathway turned feature X into decision Y?"

Why interpretability got harder: superposition (polysemantic neurones)

Problem

models often represent many concepts in the same neurones/dimensions, making direct neurone inspection misleading.

Polysemanticity

one neurone activates for multiple unrelated concepts.

Superposition

the model "packs" more features than it has dimensions, accepting interference.

Why it matters: if features are mixed, "this neurone means pneumonia" is usually false—neurones are not clean concepts.

So the field moved to: feature dictionaries with Sparse Autoencoder (SAE) + causal testing.

Sparse Autoencoder (SAE): the current workhorse algorithm

Definition: a Sparse Autoencoder (SAE) learns a sparse "dictionary" of features that reconstruct a model layer's activations—often producing more interpretable, monosemantic features than neurones.

Mechanism: train Sparse Autoencoder (SAE) on activations of a target layer → many latent features, only a few active per input.

Why Sparse Autoencoder (SAE) is important:

- feature A fires on pattern X
- suppress feature A → output changes (causality)

Concrete example: feature fires on "quotation boundaries" / "table headers" / "device marker" → test if it's real signal or shortcut.

Reference: <https://arxiv.org/abs/2309.08600>

Circuits + causal interventions: from "features exist" to "features compute"

Core idea: a circuit is a small set of model components (edges/heads/Multi-Layer Perceptron (MLP) paths) that causally implement a behaviour.



Activation patching / causal tracing

edit internal activations and measure behaviour change.



Anthropic

feature → circuit linking (interpretability "microscope").



Automated Circuit Discovery (ACDC)

attempts to automatically recover sparse subgraphs responsible for behaviours (e.g., rediscovering circuits for "greater-than" in Generative Pre-trained Transformer 2 (GPT-2)).

Comparison:

- Sparse Autoencoder (SAE): "what concepts exist inside the layer?"
- Circuit work: "which concepts + paths cause the output?"

References:

- <https://www.anthropic.com/research/tracing-thoughts-language-model>
- <https://arxiv.org/abs/2304.14997>

Interpretability-by-design: sparse-circuit models (OpenAI direction)

Definition: instead of interpreting a dense model after training, train (or prune) models to have sparse, readable internal wiring—making circuits smaller and more inspectable.

OpenAI: weight-sparse transformers for circuit-level interpretability ("Understanding neural networks through sparse circuits").

Tooling: OpenAI circuit_sparsity repository for inspection and dashboards.

Story pivot:

2024

explain after training

2025–2026

design models to be explainable internally

References:

- <https://openai.com/index/understanding-neural-networks-through-sparse-circuits/>
- https://github.com/openai/circuit_sparsity

The major open problem: scale + automation

Why it's trendy: Mechanistic Interpretability (MI) works well on case studies, but we still lack scalable, push-button methods for large models.

Scaling gap: Mechanistic Interpretability (MI) often requires heavy manual effort; scaling needs better algorithms + tooling.

Automation pipeline target: behaviour selection → activation analysis → candidate circuit → causal validation.

Representative algorithm: Automated Circuit Discovery (ACDC) for recovering sparse circuits in transformer models.

"Can we trust the explanation?" (evaluation + identifiability debate)

Key debate: Mechanistic Interpretability (MI) may be non-identifiable—multiple decompositions can explain the same behaviour.

Identifiability: does a behaviour have a unique mechanistic explanation under Mechanistic Interpretability (MI) criteria? Often not.

Sparse Autoencoder (SAE) features may not be "canonical": feature sets may vary across training runs/configurations.

Why this is a PhD opportunity: evaluation protocols are needed:

- stability across seeds / Sparse Autoencoder (SAE) widths
- faithfulness checks via interventions (does editing change behaviour?)
- agreement across methods (Sparse Autoencoder (SAE) vs circuit discovery vs probes)

References:

- <https://arxiv.org/abs/2502.20914>
- <https://arxiv.org/abs/2502.04878>

Beyond text: vision & video Mechanistic Interpretability (MI) (Prisma)

What changed: vision/video Mechanistic Interpretability (MI) is becoming practical due to dedicated tooling.

Prisma: open-source toolkit for Mechanistic Interpretability (MI) in vision and video transformers, including activation caching + circuit analysis + Sparse Autoencoder (SAE) training.

Notable: supports many vision/video transformers, includes pre-trained Sparse Autoencoder (SAE) assets, and highlights differences between vision Sparse Autoencoder (SAE) and language Sparse Autoencoder (SAE).

Why it fits my original topic: "decision-making of Machine Learning (ML) models" includes vision systems (quality control, medical imaging, surveillance, robotics).

References:

- <https://arxiv.org/abs/2504.19475>
- <https://github.com/Prisma-Multimodal/ViT-Prisma>

Updated research question

Old framing (Year 2)

"How can we explain a prediction?"

New framing (Year 3)

"What internal mechanism produced the prediction, and how can we validate it causally?"

📄 **Research question:** "How can we reliably extract internal features and circuits that causally implement a model's decisions, and evaluate their faithfulness and stability?"

Why now: roadmaps prioritise scalable, goal-driven Mechanistic Interpretability (MI) methods and stronger evaluation foundations.

Key ingredients:

- Sparse Autoencoder (SAE) feature dictionaries
- circuit discovery + interventions (Automated Circuit Discovery (ACDC) + activation patching)
- evaluation lens (identifiability + non-canonical feature risks)

Reference: <https://arxiv.org/abs/250116496>

Year 3 plan

Main track: Automated, reliable Mechanistic Interpretability (MI) pipelines

Goal: reproducible pipeline from behaviour → mechanism → causal proof → reliability score.

Work packages:

1. Baseline behaviours: choose 2–3 behaviours (syntax tracking, retrieval choice, shortcut classification).
2. Mechanism extraction: Sparse Autoencoder (SAE) features + candidate circuits (Automated Circuit Discovery (ACDC) -style).
3. Causal validation: interventions (ablation/patching) prove mechanism → behaviour.
4. Reliability evaluation: stability vs seeds / Sparse Autoencoder (SAE) sizes; report non-identifiability risks.

Extension option: replicate pipeline in vision/video using Prisma.

Industry anchors:

- Anthropic: circuit tracing case studies
- OpenAI: sparse circuits + released tooling

References:

- <https://www.anthropic.com/research/tracing-thoughts-language-model>
- <https://openai.com/index/understanding-neural-networks-through-sparse-circuits/>

External drivers: safety + governance + auditability

Regulatory pressure

EU Artificial Intelligence Act (EU AI Act) requires sufficient transparency for high-risk systems so deployers can interpret outputs (Article 13).

Timeline

European Commission timeline indicates 2 Aug 2026 as the point when most rules apply, including high-risk obligations.

Policy uncertainty

late-2025 discussions about delaying parts of high-risk obligations, but baseline compliance planning still centres on 2026.

Interpretability implication: "nice explanations" aren't enough—stakeholders want traceable mechanisms + reproducible evidence.

References:

- <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-13>
- <https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act>

Standards ecosystem: from research → compliance artefacts

ISO/IEC 42001 (Artificial Intelligence Management System (AIMS)): governance standard emphasising systematic AI risk management, including transparency.

CEN-CENELEC Joint Technical Committee 21 (JTC 21): European standardisation committee supporting implementation via harmonised standards.

Why this matters for Mechanistic Interpretability (MI): raw outputs (features, circuits) must become audit-ready artefacts:

- feature cards (triggers, failure modes, domain meaning)
- circuit reports (causal components, intervention tests)
- reproducibility logs (stability across seeds/settings)

Thesis-sized gap: make Mechanistic Interpretability (MI) outputs usable for auditors + engineers.

References:

- <https://www.iso.org/standard/81230.html>
- <https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/>

Proposed thesis contributions

Three contributions aligned with 2026 research + deployment needs

1. Goal-driven Mechanistic Interpretability (MI) pipeline

behaviour → candidate features/circuits
→ causal validation → report artefact
aligned with scaling needs and roadmaps

2. Reliability & evaluation framework for Mechanistic Interpretability (MI)

stability across seeds, sensitivity to Sparse Autoencoder (SAE) width, agreement across methods
explicitly addresses identifiability / non-canonical feature concerns

3. Open-source report generator

feature cards + circuit diagrams + intervention results
targets language (TransformerLens) and optionally vision/video (Prisma)

References:

- <https://github.com/TransformerLensOrg/TransformerLens>
- <https://github.com/Prisma-Multimodal/ViT-Prisma>
- https://github.com/openai/circuit_sparsity

Conclusion

- **Key update:** interpretability is moving from post-hoc Explainable Artificial Intelligence (XAI) to Mechanistic Interpretability (MI) with internal features, circuits, and causal validation.
- **Core message:** to trust and control Machine Learning (ML) decisions, we need evidence of which internal mechanism caused the output, not only surface-level explanations.
- **Thesis direction:** build scalable, reliable Mechanistic Interpretability (MI) pipelines that produce audit-ready artefacts (feature cards, circuit reports, reproducibility logs).

Next steps

1 Select target behaviours

Select 2–3 target behaviours (e.g., syntax tracking, retrieval choice, shortcut classification) and define evaluation metrics.

3 Implement MI pipeline

Implement Mechanistic Interpretability (MI) pipeline:

- train/apply Sparse Autoencoder (SAE) feature dictionaries,
- discover candidate circuits (Automated Circuit Discovery (ACDC) where applicable),
- run causal interventions (ablation/activation patching).

2 Build baseline comparisons

Build baseline post-hoc Explainable Artificial Intelligence (XAI) comparisons (SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME)) for the same behaviours.

4 Produce concrete outputs

- 1 end-to-end "mechanism report" per behaviour (feature cards + circuit diagram + intervention evidence),
- reliability evaluation results (stability across seeds / Sparse Autoencoder (SAE) settings),
- open-source repository with reproducible scripts and documentation.

Practical value

Outcome: a reusable Mechanistic Interpretability (MI) "audit pipeline" that turns a model into inspectable evidence: what internal features and circuits caused a decision, verified with causal interventions.

Business benefits:

- Faster debugging (detect shortcuts and failure modes earlier) → lower deployment risk and fewer costly incidents.
- Stronger trust and adoption (engineers, managers, regulators can review concrete mechanism reports, not just accuracy charts).
- Better governance readiness (supports transparency documentation workflows for high-risk use cases).

Ideas what possibly to build

A lightweight open-source "Interpretability Report Generator" (toolkit + templates) that outputs:

Feature cards

Sparse Autoencoder (SAE) feature name, top triggers, suspected meaning, risk flags.

Circuit reports

key components, causal tests, minimal subgraph summary.

Reproducibility logs

stability across seeds, Sparse Autoencoder (SAE) settings, model versions.

Optional add-on: a simple web dashboard to browse reports per model/behaviour (useful for teams and audits).

Positioning: a bridge between research-grade Mechanistic Interpretability (MI) and practical model governance workflows.