MechIntrep: Key Concepts and Memory Aids

Abid Jeem

March 23, 2025

Mechanistic Interpretability

Mechanistic interpretability is a subfield of interpretability research in artificial intelligence (especially deep learning) that aims to reverse-engineer neural networks by identifying the precise internal mechanisms—such as neurons, attention heads, or circuit patterns—that are responsible for specific behaviors or computations.

The goal is to understand models at a level similar to how one might understand an algorithm or a computer program, mapping input-output behavior to well-defined, human-understandable components.

Memory Aid: "If you're opening up the black box to find exact circuits and mechanisms behind decisions, think mechanistic interpretability (explains how the model works under the hood)."

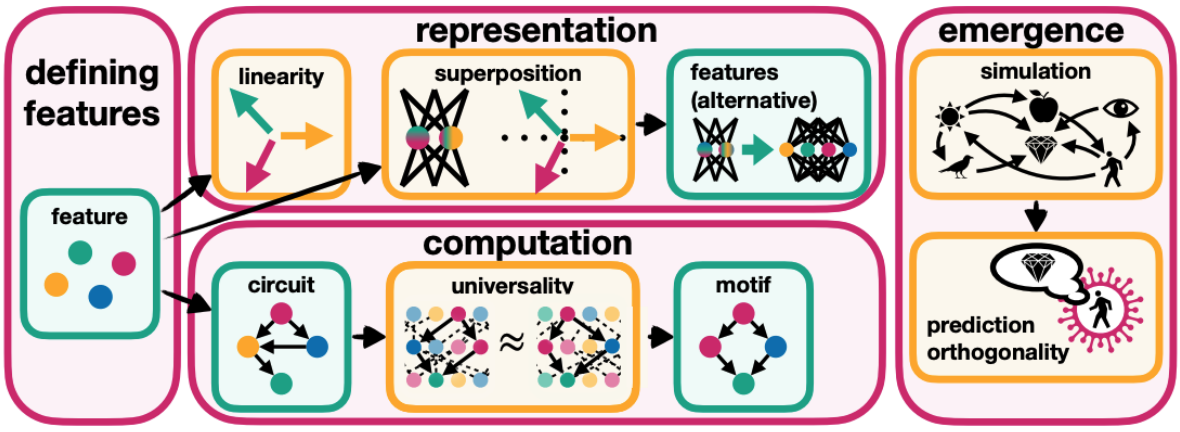

Figure 2: Overview of key concepts and hypotheses in mechanistic interpretability, organized into four subsection (pink boxes): defining features (Section 3.1), representation (Section 3.2), computation (Section 3.3), and emergence (Section 3.4). In turquoise, it highlights definitions like *features*, *circuits*, and *motifs*, and in orange, it highlights hypotheses like *linear representation*, *superposition*, *universality*, *simulation*, and *prediction orthogonality*. Arrows show relationships, *e.g.*, superposition enabling an alternative feature definition or universality connecting circuits and motifs.

Sparse Encoders

Sparse encoders are neural network architectures or components designed to produce *sparse representations*, where only a small subset of neurons or features are active (non-zero) for any given input. This sparsity encourages disentanglement, interpretability, and efficiency in the network's internal representations.

In the context of mechanistic interpretability, sparse encoders are useful because:

- Each active unit is more likely to represent a distinct, monosemantic feature or concept.
- Superposition of multiple unrelated concepts in a single neuron is minimized.
- They mimic biological neural systems, which often exhibit sparse firing patterns.

Sparse encoders are often used to construct or approximate a **privileged basis** in which each feature or neuron corresponds to a clean, interpretable direction.

Memory Aid: "If only a few neurons light up to explain something clearly, think sparse encoders (less clutter, more meaning)."

Grokking

Grokking is a phenomenon observed in the training of neural networks where a model initially appears to memorize the training data without generalizing well, but then—after extended training—suddenly begins to generalize perfectly. This phase transition happens despite no changes in training data or model architecture.

Originally highlighted in algorithmic tasks like modular arithmetic, grokking is characterized by:

- Long periods of poor generalization,
- Followed by a sharp shift to high generalization accuracy,
- Often accompanied by a drop in training loss well before validation accuracy improves.

Grokking raises important questions about generalization dynamics, loss landscapes, and implicit bias in neural network training.

Memory Aid: "If a model trains forever and then suddenly 'gets it,' think grokking (a delayed but sharp leap to true understanding)."

Features

In the context of neural networks, **features** are the fundamental units through which the model encodes knowledge. These are indivisible representational components that cannot be broken down into smaller, meaningful parts. Features are central to a neural network's internal representation of data, playing a role similar to that of cells in biological organisms.

According to Olah et al. (2020), features form the building blocks of a model's understanding. The **superposition hypothesis** offers a complementary perspective: it posits that features are the disentangled concepts that would be encoded as individual (monosemantic) neurons in a larger, sufficiently sparse network.

Memory Aid: "If it's a core, indivisible concept encoded inside a neural net, think feature (the building block of learned representations)."

Circuits

In neural networks, **circuits** are sub-graphs composed of features and the weights connecting them. These structures act as computational primitives that transform earlier (ideally interpretable) features into later (ideally interpretable) features, facilitating the network's overall behavior.

Circuits perform understandable operations, and in some cases can be explicitly interpreted. For instance, documented examples include circuits for:

- detecting curves at specific orientations (Cammarata et al., 2020; 2021),
- continuing repeated patterns in text (Olsson et al., 2022),
- resolving anaphoric references (Wang et al., 2023).

While some circuits are composed of interpretable features, others may involve intermediate representations that are more abstract or less readily understandable.

Memory Aid: "If it's a sub-network that transforms features into new ones, think circuit (the functional pathway inside the model)."

Motif

In the context of neural networks and mechanistic interpretability, a **motif** refers to a small, recurring computational pattern or structure—often consisting of specific combinations of neurons or weights—that performs a particular, reusable function within the network.

Motifs are like reusable building blocks within circuits, contributing to interpretable or repeated behaviors across different parts of a model. For example, a motif might represent a basic operation such as edge detection, token matching, or pattern extension, and may appear across various layers or tasks.

Motifs help researchers understand how more complex computations are constructed from simpler, repeated elements.

Memory Aid: "If it's a recurring mini-pattern doing a basic job inside the network, think motif (small but meaningful building block)."

Concepts

In neural networks, **concepts** are abstract, high-level ideas or patterns that a model learns to recognize and represent through combinations of lower-level features. Concepts often correspond to human-understandable categories, such as "face," "wheel," or "negation," and may span across multiple features or neurons in the network.

Concepts are not always monosemantic—that is, they might not be cleanly localized to a single neuron—but can emerge as distributed representations involving many units working together.

In interpretability research, understanding how and where concepts are represented helps bridge the gap between human reasoning and machine computation.

Memory Aid: "If it's an abstract idea formed from learned patterns, think concept (the network's way of understanding the world in human-like terms)."

Privileged Basis

In the context of mechanistic interpretability, a **privileged basis** refers to a specific choice of basis vectors in activation space where each basis direction corresponds to a meaningful, disentangled concept or feature. In this ideal basis, each neuron (or direction) is monosemantic—aligned with a single interpretable concept—making the network's behavior easier to understand and analyze.

The privileged basis is often contrasted with the standard learned basis in a neural network, where neurons may represent superpositions of multiple concepts. Identifying or constructing a privileged basis is a key goal in interpretability research, as it provides a cleaner, more human-understandable decomposition of the model's internal representations.

Memory Aid: "If each direction stands for one clean idea, think privileged basis (a special lens where features make sense individually)."