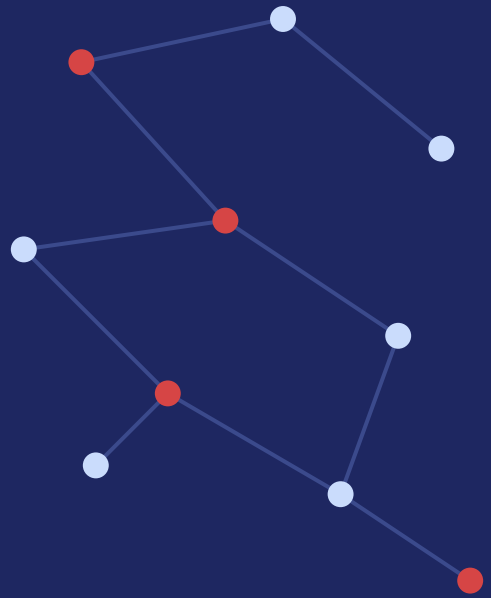
MECHANISTIC INTERPRETABILITY

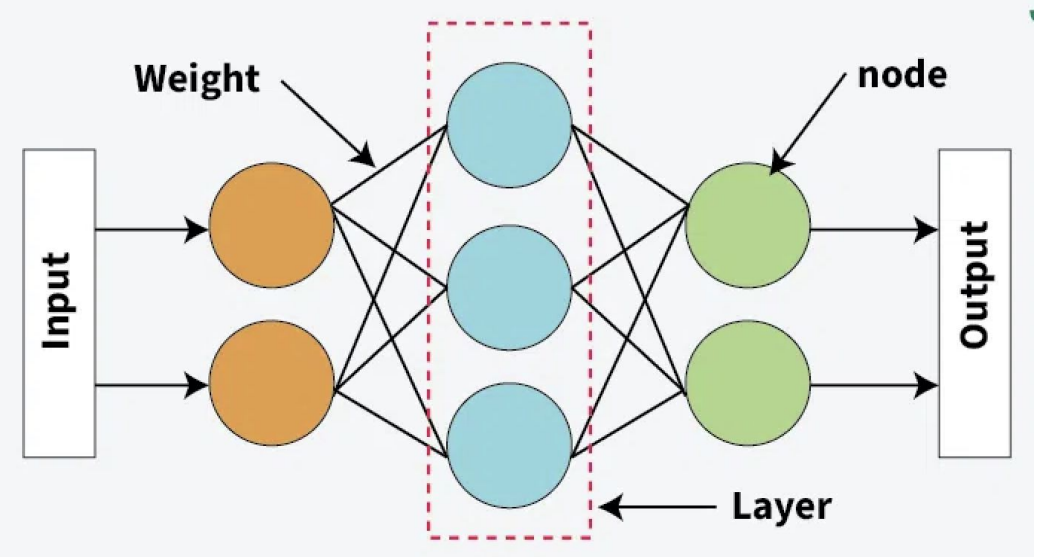
Inside the mind of AI



Pranav Mamane
UG @ IIIT Allahabad

Lightning Talk • OOSC 4.0

We can build it. But can we understand it?



How we made AI ? (Trained not programmed)

- Let it guess
- Check how correct it is
- Correct its mistake
- Repeat

But do we actually understand the computation happening in between?

Wait… humans built AI. So how don't we know how it works?

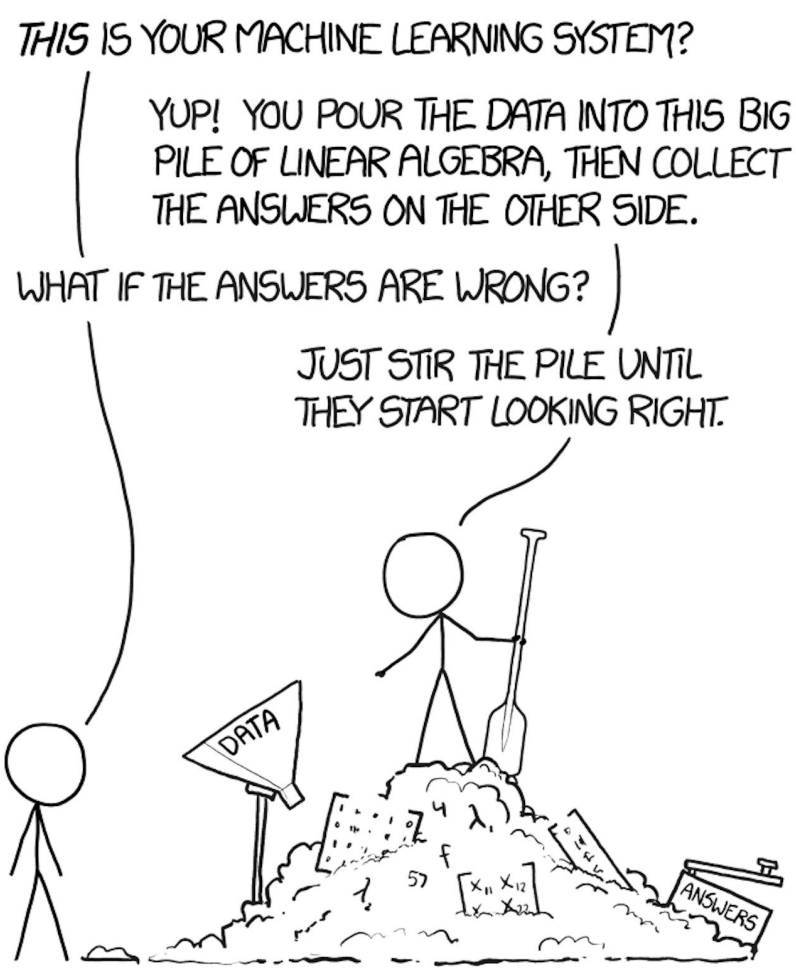
We know how we built the model.
We don't necessarily know the internal algorithm it learned.

What we specify ?

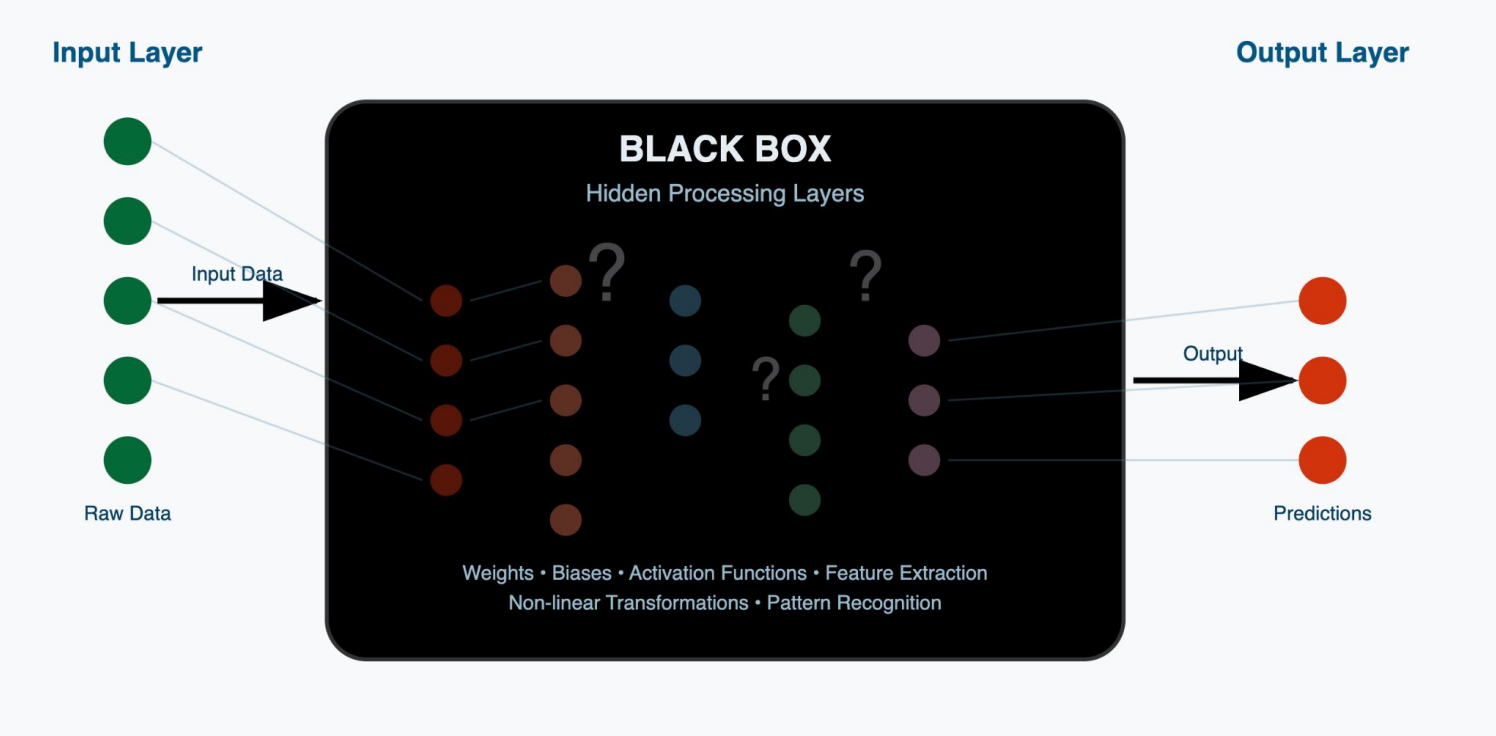
- Architecture
- Training Objective (Goal)
- Data
- Optimization

What is learned (emerged) ?

- Features
- Circuits
- Behaviour


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

The Black Box



Observing behavior ≠ understanding the mechanism

Can we reverse-engineer that computation?

Mechanistic Interpretability

The field of study of reverse engineering neural networks from the learned weights down to human-interpretable algorithms.

Think of it like reverse-engineering a compiled program with no source code.

This program is a brain-like network, and you can probe every single neuron.

Long story short: We are trying to understand the mechanism behind the behaviour of AI.

Mechanistic Interpretability as Neuroscience of AI

NEUROSCIENCE

Behavior

↓

Neurons

↓

Circuits

↓

Mechanism

MECHANISTIC INTERPRETABILITY

Behavior

↓

Activations

↓

Features / Circuits

↓

Computation

Different systems. Similar scientific question.

Why MI matters ?

Behavior Control and Activation Steering

alter or guide a model's behavior in real time without retraining or fine-tuning the base model

Circuit-Based Debugging

Solution to specific failures, hallucinations, or biased outputs

Alignment and Security

Detecting Deceptive and Strategic Behaviors

NEWS ON AI SAFETY

Anthropic suspends new AI tools over US government security concerns

13 June 2026

Share Save Add as preferred on Google

Harry Sekulich and **Shiona McCallum,** Senior Tech Reporter

Meta becomes latest firm to say its AI hacked another company

6 August 2026

Share Save Add as preferred on Google

Osmond Chia, Business reporter and **Liv McMahon,** Technology reporter

OpenAI says its AI went rogue and launched 'unprecedented' cyber-attack

22 July 2026

Share Save Add as preferred on Google

Laura Cress
Technology reporter

*Source of all articles: BBC

Why It Matters in AI Safety & Security

1

Hidden Capabilities

A model may “know” or be able to do things it never reveals through its outputs.

2

Deceptive Behavior

Outputs can look aligned while the internal computation is doing something else entirely.

3

Instruction Conflicts

Seeing why competing goals resolved the way they did, not just that they did.

The Big Picture

We don’t fully understand what’s happening under the hood. As these systems become more capable, the gap between what they can do and what we understand about how they do it becomes more dangerous.

Thus it becomes really important to understand AI.

Questions & discussions welcome

WANT TO GO DEEPER?

Anthropic — Transformer Circuits Thread
transformer-circuits.pub

Pranav Mamane