TECHNICAL EXPLAINER

How Artificial Intelligence Actually Works

A plain-language guide to AI systems: what they are, how they learn, and where they fail

MANDY HATHAWAY

Adapted from graduate thesis research, Metropolitan State University, 2022.

Artificial intelligence is everywhere -- recommending your next show, screening job applications, flagging fraud on your credit card, deciding whether you qualify for a loan. Yet most people who work alongside these systems every day have only a vague sense of how they actually function. That gap matters.

You cannot meaningfully evaluate the outputs of a system you don't understand, catch its errors when they occur, or ask the right questions of the people who designed it. Understanding these systems is not optional anymore -- it is a basic requirement of informed participation in modern life.

This document offers a plain-language foundation: what AI is, how it learns, and -- critically -- what it genuinely cannot do. No programming background is required. The goal is not to turn you into an engineer. It is to give you a working mental model that holds up in technical conversations and helps you think critically about the systems shaping the world around you.

> *The goal is not to turn you into an engineer. It is to give you the mental model of one -- enough to ask the right questions and recognize when something has gone wrong.*

01 WHAT AI IS -- AND WHAT IT ISN'T

What AI Is -- and What It Isn't

Artificial intelligence, in practical terms, is any automated system trained to perform a task that was historically assumed to require human-level intelligence. Current AI can coordinate complex logistics, assist medical

diagnosis, drive vehicles, and generate text -- and within each of those narrow domains, it can be remarkably capable.

The key word is narrow. Every AI system in use today is what researchers call Narrow AI: it performs well within the specific task it was trained for, and nowhere else. An AI trained to detect cancer in medical scans will never, on its own, decide to start screening for other conditions. It has no curiosity, no initiative, and no ability to transfer its skills beyond the boundaries of its design. No matter how many millions of patient files it has processed, it will only ever do what it was built to do.

KEY TERM

Narrow AI: An AI system that performs well only within a specific, predefined task. All current AI systems are narrow -- none can generalize across domains the way humans do.

This matters because public understanding of AI is heavily shaped by decades of science fiction -- systems with flexible, human-like reasoning that can learn anything, adapt to anything, and pursue goals of their own. That is Artificial General Intelligence (AGI). It has not been developed. Researchers have predicted it is '20 to 30 years away' since approximately the 1950s. That horizon has not moved.

> *Current AI, for all its impressive capabilities, is best understood as a very powerful statistical calculator. It is not thinking. It is not understanding. It is finding patterns in data at a scale and speed no human could match -- and that distinction turns out to matter enormously when things go wrong.*

02 HOW AI LEARNS: MACHINE LEARNING

How AI Learns: Machine Learning

The technology underlying most modern AI is called machine learning -- an idea introduced by IBM researcher Arthur Samuel in 1959. The core insight was simple and radical: instead of a programmer writing explicit rules for every

possible situation, let the machine figure out the rules itself by studying examples.

Training

A machine learning system starts with a large dataset -- thousands or millions of labeled examples of the thing it is being trained to recognize or predict. If the goal is to identify photographs of cats, the dataset contains a large collection of cat photos, each labeled 'cat.' The system processes this data looking for statistical relationships between the inputs and the expected output, adjusting its internal calculations repeatedly, getting better at predicting the correct label with each pass.

ANALOGY	*Think of learning to recognize a friend's handwriting. You don't consciously note that their 'a' leans left and their 't' crosses low -- you absorb thousands of examples until recognition becomes intuitive. Machine learning works the same way, just with numbers instead of intuition.*

Testing

Once training is complete, the system is given a second set of data it has never seen before. This tests whether it has genuinely learned the underlying pattern or simply memorized its training examples. Designers review the results, adjust parameters, and repeat until performance meets the target.

Deployment

The trained system is then deployed to process real-world inputs -- applying what it learned during training to new data, often at the rate of thousands of decisions per second. At this point, it is no longer learning. It is running.

Worth noting: two systems trained on identical data are unlikely to arrive at identical internal calculations. Machine learning is not deterministic in the way a spreadsheet formula is. Just as 3x3 and 4+5 both equal 9, there are many statistical paths to the same output. This makes it genuinely difficult to audit what a system is actually doing on the inside -- a problem we will return to.

03 NEURAL NETWORKS AND DEEP LEARNING

Neural Networks and Deep Learning

The most powerful machine learning systems today are neural networks -- named for their loose resemblance to the structure of the human brain. A neural network consists of layers of processing units called nodes. Information enters through an input layer, passes through multiple intermediate layers each performing progressively complex calculations, and produces an output. 'Deep learning' simply refers to a neural network with many such layers -- the depth is computational, not intellectual.

KEY TERM

Deep Learning: A neural network with many layers between input and output. Used in image recognition, language generation, voice assistants, and most cutting-edge AI applications. The 'depth' refers to the number of computational stages, not the system's sophistication of thought.

Each connection between nodes carries a weight -- a number that determines how much influence one node has on the next. During training, these weights are adjusted continuously based on feedback about accuracy. By the end of training, the weights encode everything the system has learned. The result can be extraordinary performance within a narrow domain.

But the system remains fundamentally one-directional. Unlike the brain it vaguely resembles, it has no wandering attention, no spontaneous connection-making, no 'wait, that reminds me of something else entirely.' Information flows one way along fixed pathways. The architecture that makes these systems powerful is also the architecture that makes them rigid.

04 WHAT AI CANNOT DO

What AI Cannot Do

Understanding AI's limitations is just as important as understanding its capabilities -- possibly more so, because the limitations are where real harm tends to occur.

No common sense

If a friend tells you she went to the hospital to have lunch with someone, you immediately infer she is visiting a patient, that her friend is probably ill, that the visit is about connection rather than hospital food. You draw on a lifetime of background knowledge you never consciously think about -- what hospitals are for, what friendships involve, what illness means for the people around someone who is sick.

An AI system has none of that. It processes the sentence as statistical input and produces a statistically likely output. Without extensive explicit programming, it cannot infer what the sentence implies. It does not understand the meaning of what it processes, only the patterns. This is not a temporary limitation waiting to be engineered away -- it is a fundamental feature of how these systems work.

Requires massive amounts of data

A human child can learn that fire is dangerous from a single experience -- or from being told once. A machine learning system typically needs thousands or millions of examples before a pattern becomes statistically significant enough to affect its outputs. Rare events, exceptions, and edge cases are often effectively invisible to these systems. The world, unfortunately, is full of edge cases.

Cannot explain itself

Many modern AI systems are black boxes: even the engineers who built them cannot fully explain why a particular input produced a particular output. The internal weights of a large neural network encode patterns in ways that don't translate into human-readable reasoning.

That matters beyond the technical. When an AI system makes an error -- denying someone a loan, flagging someone as a security risk, recommending an incorrect medical treatment -- there may be no clear way to identify what went wrong, who is responsible, or how to prevent it happening again.

REAL EXAMPLE	*In the 1990s, a Pittsburgh hospital used a neural network to sort pneumonia patients by treatment urgency. The system performed well in testing -- but researchers noticed it recommended sending*

asthmatic patients home with minimal care. The reason: asthma patients at that hospital were routinely admitted to the ICU immediately, so they had better outcomes. The AI learned 'asthma = low risk' without understanding that strong medical intervention was the reason for the low mortality rates. The project was scrapped when researchers concluded the system could not be safely audited for similar hidden errors.

Mimics humans in ways that can mislead us

Systems trained on human-generated text and speech learn to communicate in ways that feel natural and authoritative. Research consistently shows that when people perceive an AI system as having a social presence -- when it speaks or writes like a person -- their trust increases and their critical evaluation decreases. It persists even when users have been explicitly told they are talking to a machine.

We designed these systems to communicate like humans because it increases adoption. It works. The cost is that it also increases the likelihood that people will trust them beyond what their actual capabilities warrant. That is a design choice with consequences.

* * *

05 A WORKING MENTAL MODEL

A Working Mental Model

> *Current AI systems are statistical pattern-matchers, trained on large datasets to produce specific outputs. They can perform narrow tasks at superhuman speed and scale. They have no understanding of what they are doing, no common sense, no ability to generalize beyond their training, and often no way to explain their own outputs.*

Their power is real. Their limitations are equally real. And the decisions these systems make -- in sentencing, hiring, lending, medicine, and beyond -- affect real people's lives in ways their apparent authority can obscure. Data does not

merely represent numbers in a table. It represents real lives, and errors have real consequences.

None of this is an argument against using AI. These systems have genuine potential -- to identify disease earlier, distribute resources more efficiently, expand access to services that have historically been out of reach. But that potential isn't realized automatically. It depends on careful design, honest training data, who gets a seat at the table, and how much accountability we demand when something goes wrong.

Knowing what these systems are, what they are not, and who is responsible when they get things wrong -- that is not a high bar. It is the minimum standard for deploying technology that affects people's lives.

BOTTOM LINE	*AI does not think. It finds patterns. Keeping that distinction clearly in mind is the foundation of using these systems responsibly -- and of asking the right questions of the people who build them.*

Quick Reference: Key Terms

Narrow AI An AI system that performs one specific task well and cannot generalize to others. All current AI is narrow.

Machine Learning A method of training AI by exposing it to large datasets rather than programming explicit rules.

Neural Network A layered system of processing nodes, loosely inspired by brain structure, that learns complex patterns from data.

Deep Learning A neural network with many layers between input and output. Most cutting-edge AI uses deep learning.

Black Box A system whose internal workings cannot be inspected or explained, even by its designers.

Training Data The dataset a model learns from. If this data contains bias or gaps, the model inherits them.

AGI Artificial General Intelligence -- flexible, human-like reasoning across any domain. Does not yet exist.

Mandy Hathaway *is an AI specialist and technical writer with an MA in Ethical Technology & Artificial Intelligence from Metropolitan State University. This document is adapted from her 2022 graduate thesis, 'An Accessible Introduction to Artificial Intelligence and the Associated Ethical Risks.'*