How AI works: A 20 minute overview

by Jim Hassett

https://www.linkedin.com/in/jimhassett/
jimhassett3@gmail.com

Reproduced from the blog AI for lawyers, doctors, and other professionals/

This section provides a high level overview on:

Part 1 – What is AI?
Part 2 – Machine learning
Part 3 – Chat GPT
Part 4 – AGI and the future

Part 1 – What is AI?

> "*AI is one of the most important things humanity is working on. It is more profound than… electricity or fire.*" – Sundar Pichai (CEO Google)

In 1956 John McCarthy organized a conference with leading scientists to discuss whether computers could simulate every aspect of human intelligence. At that time, this area of research did not even have a name, and participants debated what to call it. McCarthy selected the phrase artificial intelligence. He later admitted that no one really liked the name… but 'I had to call it something, so I called it Artificial Intelligence.'" (Artificial intelligence: A guide for thinking humans, p. 19).

70 years later, in a 2026 article in the Yale Review, Melanie Mitchell said this name "could be called AI's original sin" because it led to anthropomorphic assumptions about the similarity of machine operations to human intelligence. She would have preferred to name the field something less like "complex information processing," which had been suggested by other participants.

The history of artificial intelligence has been cyclical with three periods of "AI Springs" with high optimism and funding (1956-1974, 1980-1987, and 1993 to the present). The first two were followed by "AI winters" (1974-1980 and 1987-1993), periods of disappointment and budget cuts when the breakthroughs predicted in the spring failed to occur. Investors are currently gambling trillions of dollars that a third AI winter will not occur any time soon, and that new technological breakthroughs will lead to enormous profits.

Today's enormous optimism about AI can be traced to November 30, 2022, when OpenAI released ChatGPT 3.5. This chatbot can draft an email announcing a divorce, list steps you'll need to take to plan for a move, give you ideas for bedtime stories to make up for your children, write computer code, and much much more.

Similar capabilities had begun to emerge in a number of programs released before then, but ChatGPT dramatically improved performance and was the first to make it easily accessible to the general public, for free. It became one of the fastest growing software products ever. In the five days after its release, over one million people tried ChatGPT. This led to a tsunami of interest and investment in AI.

AI has already made possible dozens of products that we use every day. Google Search, Apple's Siri, Amazon's Alexa, recommendation systems from Netflix and Spotify, Microsoft's CoPilot, Roomba vacuum cleaners, self-driving cars from Tesla and Waymo, and facial recognition systems to increase airport security all include AI.

Behind the scenes, AI robots now play a major role in US auto manufacturing including vehicle assembly, welding, painting and quality control. In China, manufacturers have developed "dark factories" - fully automated facilities operated by robots without human workers. One dark factory in Beijing currently "operates round-the-clock producing one smartphone per second – with zero humans employed." (For more on this, see the posts the AI Race, and the Robotics Race in my China blog.)

And that's just the beginning. "Large AI systems have recently earned gold medals at the International Mathematical Olympiad, helped humans solve long-standing problems in mathematics and biology, and contributed to major improvements in weather prediction and drug design, among other achievements."

Despite its widespread impact in recent years, there are still many differences of opinion about exactly how to define artificial intelligence, and where to draw the line between programs which include AI, and those which do not. In discussing this problem, Mitchell's book begins (p. 19) by quoting Voltaire's warning to "Define your terms … or we shall never understand one another." She goes on to note that this is "a challenge for anyone talking about artificial intelligence, because its central notion—intelligence—remains so ill-defined." Similarly, in the New York Times bestseller Empire of AI (p. 91) Karen Hao pointed out that "throughout history, neuroscientists, biologists, and psychologists have all come up with varying explanations for what [intelligence] is."

The ambiguity of the term AI is a dream come true for marketers who want to take advantage of a hot market by using the term, whether their new product contains any features that could reasonably be classified as AI or not. Some experts believe that given the primitive state of the field, the lack of a single definition of AI is actually a good thing. As a participant at a symposium summarized in an article in *IEEE Intelligent Systems*: "Because we don't deeply understand intelligence or know how to produce general AI, rather than cutting off any avenues of exploration, to truly make progress we should embrace AI's 'anarchy of methods.'" In other

words, what matters is whether a program is useful, not whether it can be technically classified as AI.

How many types of AI are there? It depends on who you ask. A number of taxonomies have been proposed based on a wide variety of criteria including a program's capability, architecture, learning paradigm, application domain, and operational characteristics.

Perhaps the most important AI distinction, and one which is often misunderstood, is between narrow AI and Artificial General Intelligence (AGI). Narrow AI systems are designed to perform specific tasks such as detecting fraud, analyzing CAT scans, and all of the examples quoted above. Every AI system that has been developed to date falls into this narrow category. ChatGPT, Alexa, and self-driving cars are absolutely amazing, but all are examples of narrow AI. I don't know anyone who would classify them as "more profound than fire."

When Google's CEO used this phrase, he was clearly referring AGI, which was defined in OpenAI's charter as "highly autonomous systems that outperform humans at most economically valuable work."

AI is currently at a Wild West stage in which researchers sometimes disagree about where to draw the line between programs that are or are not AI. And if that's not confusing enough, there's the added complication that no one understands these programs work (see Part 4 below). Putting it all together, "The field of AI is in turmoil" according to Mitchell (p. 13). Depending on whom you believe "either a huge amount of progress has been made, or almost none at all. Either we are within spitting distance of [AGI], or it is centuries away. AI will solve all our problems, [or] put us all out of a job, [and] destroy the human race."

I guess we'll find out.

Part 2 – Machine learning

Machine learning can be broadly defined as computer programs that learn from data. It is the most significant and most common approach to AI. All of the familiar programs mentioned in Part 1 are based at least in part on machine learning.

Many of the capabilities of narrow AI programs are absolutely astounding. Consider, for example, Netflix's recommendation system. If you are a regular Netflix user, you have probably used their ever changing lists of personal recommendations to decide whether you should next watch My Litle Pony, Lady Chatterley's Lover, or another show based on your past viewing habits.

To come up with these recommendations, AI analyzes every single program you have ever watched on Netflix. Not to mention how you rated each program, whether you watched the whole show, turned it off midway, or watched some scenes over and over. It also stores data on which recommended titles you have chosen over the years and which you ignored.

How large is the resulting database? Netflix has 325 million subscribers in 190 countries around the world. Each country has a slightly different list of its 10,000 plus shows available, depending on copyright laws and regional license agreements. Multiply those numbers by the billions of times anyone in the world has used their remote to make a Netflix choice and the number of data points is ridiculously large. And the chance is low that the human brains can really comprehend how trillions of variables interact to produce Netflix recommendations.

However, the basic principles behind machine learning systems are straightforward. At the simplest level, virtually all computer programs can be conceptualized as follows:

Inputs >>> Processing >>> Outputs

The first factor that makes machine learning different from other computer programs is the sheer number of inputs required. Though a small machine learning program could be built to fit on a single PC, the AI programs we use in daily life generally require supercomputers in the cloud. They can use millions of processors operating in parallel, so that incredibly large computational tasks can be broken down into millions of smaller pieces, all of which are operated on at the same time.

The second factor that distinguishes machine learning from other types of computer programs is the complexity of the processing step. This is where the magic happens.

Traditionally, computers process data by applying a set of algorithms – steps and rules like the recipes in a cookbook. For example, to count the number of items in a list, a programmer could set a counter at 0, go through each item in the list and add 1 to the counter.
But in machine learning, the computer creates its own algorithms, based on trial and error plus feedback. It does this by means of neural networks and deep learning. According to Karen Hao's bestseller The Empire of AI (p 98): "At their core, neural networks are calculators of statistics that identify patterns in old data—text, pictures, or videos—and apply them to new data." For complex patterns such as spoken language, deep learning uses neural networks with many layers.

The term "neural network" was borrowed from biology. Neurons are the basic working units of the nervous system. They are the physical structures that enable the brain to build the Great Wall of China and remember where we left our keys.

In 1943, theoretical neuroscientist Warren McCulloch and mathematical logician Walter Pitts published a groundbreaking paper arguing that the brain and the mind could be explained by studying the math of how neurons are connected and how they communicate. Over the next several decades this led to numerous attempts to create machine learning devices based on this model of the brain. By the 1960s, it became clear that the best most productive approach involved "neural networks" with several layers.

Would you like to understand exactly how neural networks make AI work? Well, you can't. I say this with confidence since no one does. A discussion of this "black box" problem appears in Part 4 below. If you are comfortable with math and determined to know more, I'd recommend

that you start with this online lesson on neural networks. But for everyone else, I'd avoid that rabbit hole and focus on the big picture of AI rather than the details.

Three main types of feedback are used to enable these programs to learn.

1. In supervised learning, the program is trained from a dataset of examples which have been pre-labeled with the correct output. For example, a program designed to determine whether a particular photo includes a dog starts by compiling a "training database" with millions of photos, each labeled dog or no dog. The computer randomly guesses on the first photo and is then given feedback on its answer. The model is adjusted based on this feedback to create a second, more accurate, model. Then the second model repeats the steps to create a third model, a fourth, a fifth, and so on. This cycle repeats over and over until the program correctly identifies not just the obvious examples – like a Saint Bernard napping in front of a fireplace – but also more subtle cases such as a picture of a parade, with a tiny image of one bystander holding a chihuahua. This requires a massive feedback loop that could be repeated literally billions of times, until it reaches an acceptable level of accuracy.

2. Unsupervised learning does not require a pre-labeled training dataset. Instead, it bases its feedback on patterns of past behavior. For example, to detect credit card fraud before it occurs, banks maintain huge databases of past legitimate and fraudulent transactions. An AI system then looks for patterns associated with fraud. If it spots something suspicious, you may get a text that says: "Alert. Did you try to use your VISA card from Loans R Us to spend $187.92 at the Walmart in Little Rock Arkansas? Reply YES or NO. If NO, your card will be blocked."

3. Reinforcement learning can be used to increase user engagement by rewarding users' past choices. For example, if YouTube recommends that you look at a funny video of a cat stalking and attacking a balloon, it's based on what you've watched before. Each time you click on a recommended video, the model is updated to reflect your preferences. If you watch a whole video through, it will be rated more positively than one that you rapidly click away from or never choose in the first place.

There are also many other types of feedback used in AI training, the most important of which include human ratings of sample responses while the system is being trained.

To sum it up, in machine learning a computer analyzes input data, and looks for some pattern, any pattern. Based on feedback about its accuracy, the computer gradually figures out rules for itself by trial and error.

Part 3 – Chat GPT

When ChatGPT 3.5 was released by OpenAI in 2020, it quickly became the first commercially successful chatbot – a program that can converse with humans, using text or voice. It was not the first program to simulate human conversation, but it was by far the most compelling.

In the New York Times, columnist Kevin Roose described his conversations with one of the earliest chatbots as "the strangest experience I've ever had with a piece of technology… [I felt] a strange new emotion — a foreboding feeling that A.I. had crossed a threshold, and that the world would never be the same."
Encouraged by the public reaction, competitors raced to release chatbots of their own. As of July 2026, three companies dominate over 88% of the chatbot market: OpenAI (ChatGPT has 51.3% market share), Google (Gemini – 26.9% market share), and Anthropic (Claude – 10.0%). Today, these products are being applied in an ever-growing number of situations, ranging from drafting catchy marketing slogans to giving exercise advice, translating over 80 languages, writing computer code, and much more.

How can they possibly do all that? As early as the 1980s, AI researchers experimented with statistical language models that could predict one word at a time based on the words that came before. For example, given the phrase "I drank a cup of ___," the model might fill in the blank with coffee or tea. But it was only in 2017, when Google published a paper describing a new type of neural network called a transformer that statistical language models took off. Transformers enabled AI programs to predict the next word in a sentence not just by looking at the preceding word, but also at the larger context of the sentences and paragraphs surrounding it.

In 2018 OpenAI released GPT-1 (the first version of its Generative Pre-Trained Transformer). The world was not impressed. In her bestseller Empire of AI (p. 124), Karen Hao noted that "Compared with today's models, the text produced [by GPT-1] was clunky and often descended into gibberish."

In 2019, GPT-2 increased the number of parameters (internal variables that each model learns and adjusts based on feedback) more than 12 times, from 117 million parameters to 1.5 billion. The results were shocking. Simply making the model bigger produced radically better performance. When OpenAI took the obvious next step and increased the number of parameters over 100 times more (to 175 billion) in GPT-3.5 Turbo, performance improved even more dramatically. This version was so powerful, that in the words of the team that developed it that it "can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans." Soon after, ChatGPT 3.5 was released to the public, and the rest is history.

Many people came to believe that when it comes to AI, bigger is better. In the words of Dario Amodei, who founded Anthropic with his sister Daniela, "as we add more compute (the number of processor chips and the time they are used) and training tasks, AI systems get predictably better at essentially every cognitive skill we are able to measure." The term AI scaling is used to refer to the fact that AI performance can often be improved by increasing three related components: the amount of input data, the amount of compute, and the number of parameters in

the neural networks. This dies has led to a voracious appetite for training content. As one article put it if a human being tried to read all the text that was used to train GPT-3, "they would need to read non-stop, 24-7, for over 2,600 years."

Journalists sometimes refer to ChatGPT as an LLM, but technically that is incorrect. ChatGPT is a complex product that is built on the foundation of the latest version of an LLM (currently GPT5.6), but it also includes a number of other elements such as a conversational interface which makes it easy for users to ask questions and safety filters which prevent hate speech, sexual material and other content which violates ethical guidelines.

At one level, the way LLMs work is easy to state. As computer scientist Stephen Wolfram put it "ChatGPT is always… trying… to produce a 'reasonable continuation' of whatever text it's got so far, where by 'reasonable' we mean 'what one might expect someone to write after seeing what people have written on billions of webpages, etc.'"

Come on. Is that it? How could that possibly enable ChatGPT to instantly produce a 20 page paper on the causes of the Peloponnesian War for a dishonest college student?

If you want to truly understand the nitty gritty details of how chatbots work, you might want to put this five minute lesson aside and start applying to computer science PhD programs. I don't have time for another PhD, so I just asked ChatGPT to recommend a few articles for beginners. My favorite was called How ChatGPT Works: A Simple, No-Jargon Explanation which says, in part: "It's not magic — it's pattern mastery. [ChatGPT] doesn't 'understand' things like humans do, but it has seen so much text that it knows:

- What sentence usually comes after another
- Which words commonly appear together
- How people ask questions
- How people answer them"

ChatGPT was also refined with extensive training with reinforcement learning from human feedback. In one example, "OpenAI paid human testers to have conversations [with ChatGPT]… and rate the quality of its replies." In another, people were given two or more ChatGPT answers to the same question, then asked "Which is better?" The results of these and other experiments were then used to alter ChatGPT's responses so they would sound more like users were talking to human beings.

Part 4 – AGI and the future

Nearly a decade ago, AI researcher Joel Dudley published an article in the MIT Technology Review entitled "The dark secret at the heart of AI" with the subtitle 'No one really knows how the most advanced algorithms do what they do.'" The article went on to say that "the interplay of calculations inside a deep neural network is crucial to higher-level pattern recognition and complex decision-making, but those calculations are a quagmire of mathematical functions and variables."

Or, in the words of Karen Hao (p. 107) "pop open the hood of a deep learning model and inside are only highly abstracted daisy chains of numbers. This is what researchers mean when they call deep learning 'a black box.' They cannot explain exactly how the model will behave." In some cases, this lack of transparency can be associated with significant errors.

For example, in her light-hearted AI introductory book You Look Like A Thing And I Love You (p. 25), Janelle Shane gave an example of an experiment she did on one of Microsoft's first image recognition products, to test its ability to recognize sheep in pictures taken in a variety of environments. "One day I noticed something odd about its results: it was tagging sheep in pictures that definitely did not contain any sheep. When I investigated further, I discovered that it tended to see sheep in landscapes that had lush green fields—whether or not the sheep were actually there. The AI had been looking at the wrong thing. And sure enough, when I showed it examples of sheep that were not in lush green fields, it tended to get confused."

Problems like this make for great headlines and are catnip to journalists. In 2026, the New York Times magazine published an article entitled We Don't Really Know How A.I. Works which pointed out that "It has been estimated that the latest versions of Google Gemini and OpenAI's GPT-5 contain trillions of mathematical functions… But as a model's neural net gets bigger, it becomes even more difficult to understand."

This black box problem is now being addressed by a number of scientists in a new AI sub-field called "interpretability." Some, including Claude co-founder Dario Amodei, optimistically believe that "We've made a great deal of progress… and can now identify tens of millions of 'features' inside Claude's neural net that correspond to human-understandable ideas and concepts, and we can also selectively activate features in a way that alters behavior." Other researchers are more skeptical, including Ellie Pavlik who has said that while progress is indeed being made, "every few months we're deeply considering a method, and then we're deeply considering another method." The Times article concluded that "it is becoming increasingly clear that we might never have a complete accounting of why a model chooses one word… over another."

Which leads to a troubling concern. As an MIT Technology Review article put it: "We've never before built machines that operate in ways their creators don't understand. How well can we expect to communicate—and get along with—intelligent machines that could be unpredictable and inscrutable?"

This challenge becomes even more difficult if the continuing development of LLMs leads to AGI (artificial general intelligence), programs that are as intelligent as humans, or even more intelligent (sometimes referred to as "superintelligence")

Science fiction writers have been speculating about the risks of AGI for decades and have recently been joined by scientists and academics. Philosopher Nick Bostrom's 2014 book Superintelligence: Paths, Dangers, Strategies famously imagined an AGI that was designed to produce as many paper clips as possible. If the program determined that humans were interfering with this goal, it could choose to kill us all. Not to mention its risks to privacy, security, fairness, and the jobs of all the people who work in paper clip factories.

In my opinion, given that AGI does not exist now and may never exist, its risks rank near the bottom of my personal list of worries. As Andrew Ng, an associate professor at Stanford, put it at the 2025 GPU Technology Conference "I don't work on not turning AI evil today for the same reason I don't worry about the problem of overpopulation on the planet Mars." The human race has far bigger problems to worry about in the foreseeable future, such as nuclear war, climate change, food and water shortages, political instability, economic inequality and pandemics.

It is especially important to remember that when people talk about AI risks, they are usually talking about the risks of AGI. When you see an alarmist best seller like "If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All" It is important to remember that they are not talking about chatbots, they are talking about AGI.

Which leads to the question: as LLMs become ever more powerful, will they gradually evolve into AGI? Once again, the experts strongly disagree. In a 2026 podcast, "OpenAI co-founder Greg Brockman said the debate about whether LLMs… can achieve AGI is… definitively… settled." They can. "We see line of sight." At about the same time that Brockman was making that prediction, Yann LeCun, formerly the Chief AI scientist at Meta predicted that "LLMs will become 'largely obsolete'… within five years. He believes that "no amount of scaling will overcome the architectural limitations of a system that only processes language."

In a widely cited 2021 paper, Linguist Emily Bender and her colleagues famously wrote that chatbots "are for all intents and purposes stochastic parrots… stitching together sequences of linguistic forms they have observed in their vast training data, according to probabilistic information… but without any reference to meaning."

Andrea Mitchell recently updated this controversy: "While early versions of LLMs… were dismissed as 'stochastic parrots' and 'autocomplete on steroids,' current ones often give the appearance of understanding language, and the physical and social worlds described by language, in a deep, humanlike way."

Mitchell agreed with Andrew Karpathy that a better phrase for describing current LLM capabilities is "jagged intelligence" to highlight its "puzzling, unhumanlike failures… How can a system that has exceeded human performance on advanced math problems sometimes fail at simple elementary-school-level problems? Why do these systems answer a question perfectly when it is worded one way but struggle when it is worded in a different but (to a human) equivalent way? How can a system that generates accurate and incisive summaries of books also produce similarly confident and authoritative-sounding summaries of nonexistent titles?"

The argument about whether LLMs will evolve in to AGI is closely related to the most controversial question of all: When will AGI exist, if ever?

Some AI enthusiasts say it is just around the corner. For example, in a breathless 2024 blog post, OpenAI founder Sam Altman wrote that an "Intelligence Age characterized by 'massive prosperity,' would soon be upon us, with superintelligence perhaps arriving as soon as in 'a few thousand days'… Although it will happen incrementally, astounding triumphs – fixing the

climate, establishing a space colony, and the discovery of all of physics – will eventually become commonplace."

As you consider claims like this, it is helpful to remember the frequently repeated warning that "it is very hard to predict, especially about the future." The entire field of AI is littered with optimistic AGI projections that have proven incorrect. One of the most famous came from Nobel prize winning economist Hebert Simon in a 1965 book (p. 95): "machines will be capable, within twenty years, of doing any work a man can do." In case you dozed off somewhere around 1985, Simon was wrong.

Or consider the ever changing predictions from the world's richest man, Elon Musk. In 2023, he predicted "full AGI — AI surpassing human intelligence in all domains" would be available "roughly by 2029." By 2024, Musk was more optimistic and said AGI would arrive in 2025. When it didn't, he changed his prediction to "maybe as soon as 2026." In January 2026, he said AGI might "possibly slip to 2027, with superintelligence – AI exceeding the combined intelligence of all humans – arriving around 2030."

Again, others think it will take much longer or never occur. According to Andrew Ng, the founder of Google Brain, "AGI has been overhyped. For a long time, there will be a lot of things that humans can do that AI cannot."

A 2023 survey of 2,778 researchers who've published articles on machine learning reported that "if science continues undisrupted, the chance of unaided machines outperforming humans in every possible task was estimated at 10% by 2027, and 50% by 2047." As Andrea Mitchell summed it up in "Artificial intelligence: A guide for thinking humans" (p. 276) "Surveys given to AI practitioners, asking when general AI or 'superintelligent' AI will arrive, have exposed a wide spectrum of opinion, ranging from 'in the next ten years' to 'never.' In other words, we don't have a clue."