

K-AI

Chapter 1

+2

Everything I ever learned about AI
Kai Bergin October 2025

When things got dark you were a lighthouse

Thankyou

Writing a series of books on the subject of AI over the span of several years wasn't something I ever consciously planned on doing. It was a spontaneous idea born in a moment of quite introspection. To be specific, I was lying on my couch and it was Wednesday night, to be even more specific, it was a particularly lazy moment of doom scrolling social media slop and pondering which of my many hobbies I should engage in, the answer seemed to be: perhaps, none at all. It was there, in that grey zone between lethargy and latent activity that my brain latched onto the idea of writing a book, from a couch bound slumber thoughts and memories coalesced into two narrative branches: One branch was about myself: The person Kai, who over 5 turbulent years had experienced professional triumphs and emotional failures in equal proportions, and the second was K-AI, my professional alter-ego, who had been locked into a rapturous and perpetual learning loop, where AI played the leading role. From these two concepts, my own cognitive pattern recognition was able to glean a singular insight that tied both branches together, defined the foundation of this technology and finally linked it to my personal life and my journey of learning:

Artificial Intelligence = change.

Predictably my couch driven lethargy continued for longer than this "eureka moment" of inspiration would have suggested. There were valid reasons for this delay in translation: By the time I started writing this book, in the spring of 2025, I had spent years managing an AI development team. These talented individuals built custom solutions for everyone from European financial institutions to Hollywood film studios, and while they tirelessly worked, my morning, noon and night's were filled with a ravenous consumption of academic papers, tech press and random social media buzz. Yes, I worked hard, but they were right there with me sharing that energy and forward motion. The second delay was because my learning process became vocalised: In 2024, and thanks to New Scientist, I found myself standing on the stage with the giants of the Dutch scientific world at the Dutch Science Gala and again on World Earth Day, combine this with countless workshops for tech teams, consultancy agencies and enthusiastic people just wanting to understand more about the technology as it seemingly changed the world. I didn't feel it at the time, but in hindsight being at the centre of the AI revolution was like white water rafting in a three-piece business suit.

My burn out at the end of October 2024 was a testament to how crazy everything was getting, its no exaggeration to say that a decade of innovation had been crammed into two short years. To make any sense of what I was learning and what was happening around me, I wrote everything I learnt down, and converted it into articles and chapters of this book. The challenge I set myself was simple: If I couldn't write about it fluently then I didn't understand it and, as an extension, I probably shouldn't be advising people about it. I'd spent decades studying music and I knew all too well the difference between talking about playing an instrument and actually being fluent enough to perform on a stage. This process created an anchor of equilibrium as I descended into the maelstrom of a professional depression. Kai stopped functioning for a while, K-AI continued writing and learning.

One final note before we kick things off: I am writing this in the most authentic tone of voice possible, this is just me, the title and tone of this book is about a personal journey of one person finding his way and learning in an extraordinary moment of human history. How I write is how I speak and also how I think. This is all intentional, potentially annoying, and part of the deal. Just bear with me and hopefully I can pass on to you, the reader, some of the inspiration I've felt as AI changes our world.

I'm K-AI and this is my story.

Chapter 1: What is intelligence?

My journey starts with the story of a genius, a war hero, the tragedy of inhumane sexual discrimination and an impossibly complicated question he asked the world:

“Can a machine think” ?

The person asking was the father of computer science and artificial intelligence: Alan Turing. Music has Bach to thank for its harmonic foundations, literature has Shakespeare, astronomy Galileo, in the field of computer science we have Turing.

I'm not going to pretend to know the answer to this question, but I do know about Turing, his tragic life and the beautiful paradox his question presented. To add a little context and background: The idea that a machine could think started in the early 1940's as a logical extrapolation of rampant scientific optimism. Born from echoes of theoretical mathematics, like The McCulloch-Pitts neuron, the predecessor of the neural network and embryonic architectures of early computer systems where “memory” interacted with “computing” units, all theorised technological mirrors of human-like cognition. These early academics wondered what the future would bring when these computational units weren't the size of a basketball court and they could fluently simulate and perhaps surpass human thought. In 1950 those sparks of synthetic intelligence were solidified into concepts that were scientifically plausible and more importantly testable. The person responsible for this giant leap from imagination to computer science was Alan Turing. In a decade this man founded this field of computer science, broke the Nazi Enigma code and defined computer architectures that are still in use today. As if that wasn't enough, in the last years of his life he published the seminal paper “Computing machinery and intelligence” in a philosophy journal.

As if the question he raised in this paper wasn't enough of a paradox, there was also an inherent issue of publishing anything in a philosophy magazine, which unfortunately leads to recursive sub-questions, where every verb and adjective

committed to paper threaten the reader and writer with a spiralling descent into navel gazing vagueries. Turing as a mathematician realised this immediately, and he also knew that, with the open phrasing of the question, that he could never prove the answer was correct. He needed something a more scientific starting point:

So his rephrased:

“Can a machine imitate a humans intelligent responses ?”

Now this was a question that he could apply scientific analysis to. Genius works in strange ways though; Instead of publishing a paper on this, he created a game that anyone with a computer and two people could play (and they still do). The “Imitation game”, a foundational test of AI capabilities that would later be rebranded into the “Turing Test” and has grown into a standard benchmark for measuring the extent of AI capabilities compared to basic human cognition. The rules of the Turing Test are fairly simple: The game has three participants, one is a machine, the other human, and the last one is a Judge. In a blind game of questions and answers the judge has to guess which of the participants answers came from a machine and which ones were from a human. Once the machine is able to consistently fool the judge into thinking its human the test has been passed and a human-like level of cognition has been achieved. The questions were originally chess and maths related (perhaps a reflection of the games creator) but have evolved over time into a much broader syllabus. This game is still used as a way of measuring the intelligence of machine systems, however since modern language models have become so proficient in answering these challenges, the modern version focusses on the human elements of interaction, so the speed of the answer, pausing between answers, making mistakes, being a little “too perfect” and using colloquialisms or em dashes. As an interesting sub text of the game: Many modern historians see parallels between Turing’s homosexuality and the original imitation game, where respondents also had genders and challenged the judge to guess the gender and nature of the respondent. Machine, female or male? Turing’s landmark article was published in 1950, tragically two years before his inhumane chemical castration for acts of “gross indecency” (he accidentally mentioned his same-sex partners name to a police officer while reporting a burglary). Turing, one of the greatest minds of the 20th century, a war hero, and a gay man was discovered by his housekeeper at the age of 41, lying in his bed with a half eaten apple next to him. The now

ubiquitous, tragic and defiant symbol of one of the largest technology companies on the planet: Apple.

Final note: History has been kind to Turing, his legacy is very different than the horrific last years of his life, where he was injected with a cocktail of experimental hormones and chemicals on a weekly basis to “cure” his homosexuality. A genius who I could argue was slowly murdered by the country and system that he saved. The Turing Award is now considered the highest honor that computer scientists can achieve and the British government formally apologised for in 2009. British Prime Minister Gordon Brown acknowledged Turing’s role in cracking the enigma code during world war 2, stating that "Without his outstanding contribution, the history of world war two could well have been very different".

Additional note: Everyone loves a good hero-story, in reality there are more dimensions to this story Like the Polish cryptographers, soldiers, spies and mathematicians who cracked the German armies Enigma code, they were on the front line of a relentless and brutal invasion and deserve as much praise for heroism, as anyone involved in this project

Turing wasn’t just trying to create a mechanical gollum though, his question and study was also about our human understanding of thinking and intelligence. Which I personally find to be a circular question, because it is both the starting point and end state of working in the field of artificial intelligence. The challenge is a catalyst and a dead end rolled into one. For some strange reason this paradox inspired and challenged me to learn more about the definition, in these early stages of my journey I started to form questions. Big ones.

- What is it that we are actually trying to model ?
- Can something we haven’t clearly defined even be modelled?
- Are we witnessing the biggest scam in human history ?
- Or, is intelligence going to become a new commodity that will revolutionise our society?

To even partially answer these queries I continued to study the history of AI, following Turing's seminal question as it posthumously crossed the Atlantic to North America.

Specifically: The Dartmouth Conference in 1956, where the titans of computer science: John McCarthy (Dartmouth), Marvin Minsky (Harvard), Claude Shannon (Bell Labs) and Nathaniel Rochester (IBM) kicked off the study of AI. (Both McCarthy and Minsky directly acknowledged Turing's work at the conference and later). In later chapters of this book I'll explain why these people were so important, for now, let's just say that these were the brains behind most of the digital technology we use today. After a week of generalizing in the direction of what machine intelligence could be, these people scientists determined that: If you could describe the brain's inner workings in enough detail, you could (theoretically) create a synthetic version of intelligence.

To quote them more accurately:

“Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”

It was John McCarthy who coined the title “Artificial Intelligence” and his definition goes much further than these two words. McCarthy, a young mathematician from Dartmouth University, wanted a title for this new branch of computer science that was as neutral as possible, one that could avoid at least some of the inevitable semantic discussions that stop progression in its tracks, he was also a rationalist and a staunch believer in the concept that with enough time and effort any challenge or process could be turned into a mathematical model. Later in his life McCarthy pivoted towards engineering artificial intelligence a logical step from rationalising into the art of construction.

From Turing to Dartmouth, the 50's ended in a blur of techno-optimism as humans ventured beyond Earth's atmosphere for the first time, Hollywood special effects gave birth to science fiction, televisions entered living rooms and North America rose out of the ashes of WW2 as a super power of computer science and digital innovation.

Fortunately for my initial goal of understanding intelligence, a lot has changed since the 1950's and the study and definition of intelligence has evolved over the decades into something completely different than it was in the time of Turing, Shannon and McCarthy. Let's follow that study to the present day. For the next half of this chapter I'll briefly summarize the definitions and study of intelligence over the last hundred years and then write a little about its features, wrapping things up with the can of worms of cognitive bias, by way of "The Dunning Kruger Effect". This example of cognitive bias is especially relevant because it tackles our tendency to overestimate our own competence, which in turn raises fundamental questions about the rationalist basis of this study and adds another big question to my growing list: Who are we to assess intelligence, machine or otherwise?

My study starts with the great grandfather of measuring intelligence: The IQ Test.

In 1904 Psychologist Charles Spearman created a series of tests that measured people's abilities to complete common tasks. What he found was that people scoring highly in a specific area tended to also have higher scores across all areas of the test. Spearman aggregated all the results and came up with a single score for general intelligence: (creatively named) "G". An easy metaphorical way of understanding what he means is to use the idea fitness. In broad strokes: If you are a professional footballer, there is a good chance that you'll perform better at other sports than someone who isn't. Even if it's just an extension of your blunt stamina and endurance, hand eye coordination or competitive mindset, the professional sports person is going to perform at a high level across all sports.

This concept was augmented a year later in 1905 France when Alfred Binet and Théodore Simon were commissioned by the French Education department to design a scale to flag pupils who might need extra help. This scale plotted "mental age," not an all-purpose label but one that was applied to specific aspects of learning. And to complete the scoring process into the one we know today, Lewis Terman at Stanford university translates, expands, and renames the Binet-Simon test for American English speakers, coining the now-familiar acronym IQ. Which is simply $(\text{mental age} \div \text{chronological age}) \times 100$

In the 40's Raymond Cattell added the ideas of "fluid intelligence" and "crystalized intelligence", the brains we are born with (fluid) and cultural intelligence (crystalized) and slowly but surely the study of intelligence started to be subdivided into smaller and more specific components or features. Which makes perfect sense when matched back to the fledgling computer scientists of the 50s and the core theory of AI that an individual feature of intelligence could be synthesized.

In the 1960s the pure rationalism of a score lost ground to studies of attention mechanisms, mental representations and other internal processes that science and neurology were beginning to be able to glean through new technologies and studies. This all links to new concepts of learning and teaching that I'll go into in a later chapter, but each one of these concepts ties into the advancement of the field of artificial intelligence and they progress in lock step towards the present day.

The present day study of intelligence is a venn diagram of behavioural psychology, neurology, cognitive science and data science. Intelligence in its academic form advanced from a single score to multiple scores to full blown CATT scans and high resolution imaging of neurons as they connect and form new thoughts.

Following the circular path of my questions, I found myself returning to philosophy, and specifically Paul Thagard, who in his book "Bots and Beasts, what makes machines, animals and people smart" attempts to find correlations between intelligence in animals, people and machines. Thagard quickly concludes that defining the human variant is a form of intellectual entrapment, and we should be wary of people who ask us to do this. Thagard takes the longer, scenic route in his exploration. To start with he follows a simple concept from psychologist Greg Murphy and expands upon it. Murphy suggests that if you want to glean intelligence you just take:

- Examples of intelligent people
- Find typical features in those people
- Explanations provided by the concept

If you go down this rabbit hole you should have at least one or two features of intelligence. At this early stage in my study I was consuming vast amounts of

psychology, both the pop and the academic variants. So I decided to apply this concept to one of my heroes the psychologist and relationship therapist Esther Perel.

What makes her intelligent? Well, she seems to have a deep understanding of the subject matter, a lot of experience, she can translate that into being perceptive in her therapy sessions. She is also communicating and acting on those perceptions in a way that is understandable and relatable for people who aren't professors in psychology. .

If you continue down this path of analysing intelligent people you can quickly create a list of features. Here's a few ones that I came up with while following this logic:

- Perceptual intelligence
- Emotional intelligence
- Communicating skills
- Acting/doing

There are some interesting features in this list, some of them spring out at you and others require a little bit of explanation. For example:

Perceptual intelligence could be a feature of a talented police officer, that person has their radar tuned to what is going on around them and can quickly make distinctions on where they are needed.

Acting Intelligence is for the people I deeply envy. Individuals who know when to cut the chatter, the analysis and the study of a problem. They have a understanding of when they need to step forward and actually do something. Being able to judge when the time is (and is not) right to act.

Feeling or emotional intelligence resonates very strongly in our modern brains. We are constantly interacting and connected with other human beings and our ability to solve problems in this domain is in some ways more valuable and important than solving a mathematics puzzle. We could break it down even further into social intelligence, empathy or self-regulation, all of which are important features of intelligence that deserve their own mini-IQ scores.

In modern times we can also add the concept of “collective intelligence” to the list, the hive or swarm intelligence we have observed in bees or swallows is one I find particularly interesting. Working in teams and solving problems as a group is so rewarding and hopefully everyone reading this has been fortunate enough to experience this. I won’t go into this in too much detail, but I believe that working with AI models could be a form of collective intelligence.

Defining intelligence is tough and it’s a moving target; to make it even tougher it also has an inherent weakness in assessment: cognitive bias.

This is a study in itself and one that is almost as rich and deep as the broader topic of intelligence. I won’t even attempt to describe all the forms of bias, what I will do is take one of my favorite forms “The Dunning-Kruger Effect” and use this as an example of just how great the challenge is when it comes to answering any question regarding intelligence. Cognitive bias suggests that even if we come up with an answer, we as humans can’t really evaluate if its correct. In a nutshell the Dunning Kruger effect says that we are lousy at estimating our own incompetence. The theory started with lemon juice and a failed bank robbery.

Justin Kruger and David Dunning became famous with a hilarious story of a young man and an attempted bank robbery. The bank robber McArthur Wheeler walked into a bank without any form of mask or disguise, except for a faint glowing tint on his face. Employees of the bank were able to describe his features and person in detail to police and his subsequent arrest was assured. The logical, and tried and true, way of robbing a bank would of course be to wear some form of disguise, at the very least a balaclava or a stocking, but on the security camera’s, and from the bank workers testimonies he wasn’t wearing a mask or stocking. So, what happened?

Wheeler had smeared his own face in lemon juice, this was his only disguise and he was convinced it would work. The short version is that Wheeler had been told by someone (confident and convincing) that the lemon juice would create a form of reflection that would blur his face from security cameras. He didn’t need those annoying stockings that just got in the way of your sight anyway. Wheeler’s intensely stupid theory was one that, even after his arrest, he still believed (in his defence, Wheeler had robbed two banks using his juice mask and no one had arrested him yet).

Dunning and Kruger's conclusion was that incompetent individuals often lack the metacognitive insight or tools to recognize their own incompetence. They started to run evaluations to see if they were onto something. They were. In multiple tests participants who had scored the lowest genuinely believed that their competence was above average. The cognitive bias went two ways though, participants who scored in the top percentiles underestimated their abilities assuming that the "easy" task would be easy for others as well. You can expand on this theory in two ways:

We are ignorant of our own ignorance. Or to put it another way we don't know what we don't know, and we fill this hole in our meta cognition with confidence.

We can't monitor our own competence. We need someone else, external indicators, teachers, managers or mentors to calibrate our self-assessments or at least burst the bubble of confidence/ignorance.

Over the last 5 years the Dunning Kruger effect has been applied to other biases as well, and it kind of holds up.

West and Eaton (2019, "Personality and individual differences") suggest that this concept extends to social attitudes like racial and gender prejudice. In their study of almost 400 participants they concluded that people who have a large prejudice towards a gender or race completely overestimate how egalitarian they are, believing themselves to be much less prejudiced than the average person. Unfortunately, the dual burden of ignorance can also be applied to social biases.

Things get interesting when neuroscientists start looking at this theory, especially when it is framed with the idea of biased prior beliefs about their skill. In Jansen, Rafferty, & Griffith's study (2021) – "A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers." In this study they start with the assumption that most people think their cognitive capabilities are somewhat above average to start with. Another variant assumed that low-skill individuals are worse at monitoring their own accuracy on each question (a genuine metacognitive deficit). Findings suggest over- and under-estimators use differing cognitive processes when assessing their performance: Under-estimators rely on memory recollection when asked

how they did on the test, they can literally remember specific cognitive struggles they had during taking it (“I thought the test was really tough!”). Over-estimators draw upon excess familiarity and don’t seem to remember the specifics of experiences they had during taking the exam, (“I’ve done this a million times before, I’m fine”) when over-estimating their performance. From this study we can assume that memory and our strength of memory seems to play a key role in how we evaluate or estimate our performance.

This concept is expanded upon and made into something tangible in Muller, Sirianni, & Addante (2021, *European Journal of Neuroscience*) – “Neural correlates of the Dunning–Kruger effect.” This innovative study explored the brain processes associated with Dunning–Kruger judgments. The question was whether differences in how people encode and recall information might be behind why some underestimate or overestimate their performance. Researchers had participants perform a memory task (an item-recognition test) and then periodically checked what percentile the participants believed they were in relative to others. The participant’s brain activity was recorded using electroencephalography (EEG). The sample was later divided into those who overestimated their performance (low scorers with inflated self-assessment) versus those who underestimated (high scorers with modest to low self-assessment), to compare their brain responses. Good news for our theory of bias, they replicated the classic Dunning–Kruger pattern: the bottom 25% performers overestimated their relative percentile rank, while the top performers tended to underestimate themselves. The results were the same as the previous test I mentioned, but now we can add that over-estimators had faster response times when confidently (and incorrectly) rating themselves highly, suggesting a quick and shallow self-assessment, whereas under-estimators took longer when analysing and downplaying their performance, perhaps due to more analytic reflection, they were used to this process and took their sweet time wallowing in self analysis and criticisms.

My take away from this is that memory clearly plays a key role in all features of intelligence and it obviously plays a crucial role in our assesment of our own intelligence and cognitive bias. I'll return to the concept of bias in a later chapters and explain its machine translation and how it has been turned into a crucial and necesarry constructive process for AI. For now I'm going to round this first chapter off with what I see as a failure of pure rationalism, simply because it doesn't exist in a pure distilled form, perhaps one day we will have a machine judge and two humans, turning us into the lab rats of the future Turing test. Given the branching of intelligence over multiple disciplines I believe that there is much more going on in the foundations of artificial intelligence than simply breaking functions of intelligence into component parts and modelling them into synthetic representations. Like theoretical physics I believe that the more we are able to measure intelligence the more complicated and mutifaceted it becomes. But not all hope is lost, there is another ingredient that we can add to reinforce our foundational understanding of AI though: learning. In the same way as intelligence studies and bias have progressed over the last 100 years, so to have the studies of learning, and it's a crucial next step to understanding the AI.

Chapter 2 : Human vs Machine Learning

If there's one thing I love doing, it's learning. When everything in the fatty grey lump of tissue, that I lovingly call my brain, is working as it should, and the optimal level of caffeine has been added, I can happily process information into new behaviours, skills or knowledge. Whether anything I've learnt actually sticks depends on all kinds of internal and external factors, but it's a moment of cognitive beauty, and one that the entire tech sector is desperately trying to synthesise.

So, in this random moment of musing, I'll:

- 1) Capture the modern definition and study of human learning.
- 2) Introduce the foundations of the non-biological variant.
- 3) Share some of the latest studies of AI and learning.

So let's start learning. Perhaps your lumps of brain matter will enjoy this meta-process of learning about learning as much as I enjoyed writing it.

Let's start by saying that learning isn't a singular action. It's a ridiculously complicated process of:

- Interaction and adaption with our environment;
- Acquisition of new skills, behaviors and knowledge;
- Resulting in change to a system with the new or "learned".

In my humble opinion, it's not an overstatement to say: Learning is the most beautiful lens to view the world through, because it fundamentally alters perception, and through perceptions, action and interaction.

The only problem with defining the biological process of learning is that it isn't just a singular cognitive function of an entity. It's a process that occurs on multiple levels, in parallel:

- Our bodies component parts all have an inherent ability to learn; Whether it's our immune system, our organs, muscles or bones, they are all like individual classrooms of learning and adaption to the challenges we face.
- Our DNA has a software layer built around it that changes the system, sometimes over generations or decades depending on what it has learnt.
- Viruses are even cooler; these ancient pathogens learn and adapt on the fly, they can also scale this learning process up to system level behaviours, like: sneezing, to help them hop over to a new host. (a simplification of many, many amazing learning-like behaviours)

These are just a few “rapid fire” examples, but multiply these levels of learning over multiple entities in an environment filled with biological life and this (already complicated) process scales in all kinds of freakish directions.

In a recent study (2022) psychologists Jan De Houwer and Sean Hughes wrote a fascinating paper that took the entity out of the equation and suggested

“We define learning as changes in the way a system responds to its environment — that is, as learned behavior.” ([Source](#))

I liked this way of framing it because it scales nicely and allows us to zoom in or out without pretending the biological complexity isn't there.

AI Side note: When you hear qualified critics of current AI systems intelligence like Dr. Yann LeCun talking about how Large Language Models are not intelligent, this is the lens he is viewing it through. The pre trained model (you know, the P and T in the “GPT”) is just a giant information retrieval system. LeCun argues that it's impossible for the LLM to interact with a real environment, which in turn discounts learning and therefore real intelligence.

Here's a great interview where he expands on this:

In much the same way as a language model presents an illusion of intelligence, my definition of learning is an illusion of understanding. To really understand

learning we need to look into the mechanics of the process. Fortunately for me (and you as a reader), smarter people than myself are studying this process and its been categorised into:

Behaviorism,

This is world of Pavlov and Skinner, it's the granddaddy of learning studies and its all about conditioning. Just wash, rinse and repeat and eventually it sticks. As someone who has spent years learning how to play music instruments, I can confirm it works. Conditioning also introduces the concept of rewards and punishment. Skinner showed that learning followed by rewards led to repeated positive outcomes. At a high level this has been directly translated into modern AI and the carrot and stick approach is an inherent part of training these systems before they are deployed. Reward good outcomes and ruthlessly punish wrong answers until the AI can model and recognize patterns towards good outcomes.

Cognitivism:

We continue down the line to the next mechanical explanation. Cognitivism suggests that learning is active information processing. It's all about encoding, storing and representing. It's an active and conscious mental process that requires concentration, memory and attention. From this approach the idea of attention and cognitive load arise. Just think of all those cool techniques you learnt at school, like: mind maps, memory hacks and study tricks, they are all part of this school of thought. Here we again see a direct translation to the core concepts of modern AI, the terms used in cognitivism have almost and identical meaning in the study and mechanics of AI, and there's a good reason for that, remember this is a field of study that is actively attempting to simulate the human process of learning.

Constructivism:

This is a way of seeing learning as construction. Learning is something that is an incremental process where new experiences are added to existing mental frameworks that develop as we grow older. The system learning isnt a blank slate in this concept. A simple example would be a young child seeing an exotic bird at the zoo, even though the child has never been to the Amazon he or she recognises that it's a bird, because it already has a mental representation of what

a bird is (you know..wings, beak, feathers etc etc). In the constructivist concept of learning interaction is king and memory (long and short term) plays a key role.

And finally Social learning:

Where learning happens through observation and imitation. During our lives we are constantly looking at how parents, teachers, friends and family perform tasks and we are learning the whole time. This process of modelling behaviour is closely related to AI systems although we perform this in a much richer and more complicated form.

I'm not going to pretend that I understand the neurological processes at play here, but at a high level:

Attention is crucial. We process much more input than is necessary or relevant, there's just a lot going on at any given moment and we don't need it all. So we have a neurological "zoom in and filter" process of paying attention.

Our emotions also play a key role in our learning process, pleasurable emotions act as a reward mechanism, which stimulates learning (just like Skinner and Co described). The flip side of this is that stress or a lack of sleep has a negative effect on our brains reward mechanisms. Another important process that psychologists and neurologists agree on is motivation, a motivated person learns many times more effectively, which in turn is why teachers, employers and parents spend so much time and energy trying to motivate us to learn.

There have been some recent breakthroughs in understanding learning, like:

Glial Cells Storing Memories (2024): For decades, neurons took center stage in neurological learning research. A breakthrough Nature study in 2024 by scientists at Baylor College of Medicine revealed that astrocytes, star-shaped glial cells previously thought to play only a support role, actively participate in memory storage and retrieval. This was a striking paradigm shift because it showed that the brain cells traditionally seen as mere support "glue" are in fact integral to the learning process.

Multiple Learning Rules in the Brain (2025): A long-standing mystery in neuroscience is how the brain's billions of synapses know how to adjust during learning, often termed the "credit assignment problem." In April 2025, a study published in *Science* (by a University of California San Diego team) made a significant breakthrough: it found that individual neurons can follow multiple learning rules simultaneously. This helps explain how complex learning is possible without causing instability in neural networks. Neuroscientists hailed this as a "breakthrough" because it provides a more nuanced picture of plasticity and could inspire more sophisticated artificial intelligence algorithms that mimic the brain's multi-rule learning approach

AI Systems and Learning:

Now finally we arrive at the first part of this book where we step away from humans, history and philosophy and start to look into how an AI system learns through data and feedback. What do I mean by this? Since this is an article about learning I'll skip a deep dive on the technology and just take you through the foundational concepts of how an AI system learns.

Before we start though its absolutely crucial to understand that an AI system is "trained", its not computing conditional logic in circuits of a CPU, its reading data and looking for patterns that create a link between input and output. An AI system needs to see the answers as much as it needs the questions. From that point it can pattern recognise its way to understanding how we get from A to B or from X to Y and then (hopefully) repeat this process in a unique situation without the B or the Y and only based on a new question.

A few definitions before we begin.

A model: an algorithm or system that learns patterns from data to make predictions or decisions.

Data: Raw information — numbers, text, images, or sounds — that machines can process. (I'll expand on this crucial concept in the next chapter, since it's a concept that stretches far beyond machines)

Dataset: A dataset is a structured collection of related data, used to train or evaluate an AI model.

Data labeling: Data labeling is the process of “tagging” data with meaningful information, so categorising or attributing values to data, to help a model learn.

The classics and foundational terms and concepts of AI teach us how these models learn:

Supervised learning: An AI model is trained on labeled data, and the AI system and engineers evaluate its accuracy based on these labels. Let’s imagine we have built an AI system to sort through a massive image library and fish out only photos of dogs. In a supervised learning system all the photos in that library have been labelled so we can train the AI system really directly and accurately, a photo with “dog” as a label is positive (1) and one without this label is negative (0) and anything in between is something that might look like a dog but isn’t one (the classic and hilarious example of “is this a dog or a croissant” photo would mean that a croissant might score 0.8 in this test, suggesting that these two are somehow closely related). This learning process has also been the focus of many AI ethics discussions, because it involves underpaid human data labelers who are tasked with the thankless and mind-numbing job of tagging massive data sets. (Thankfully the internet is now full of prelabelled and organized data sets these days, so “Thankyou AI boom”)

Unsupervised learning: In this form the data has no labels, and the AI model has to find patterns, structures and groupings in the input on its own. Here the AI system is looking for underlying information that isn’t immediately evident. This kind of learning is often used with massive datasets, where there is no single right answer. Going through millions of bank transactions for example and finding patterns based on dates and locations. This kind of learning system is fantastic for exploratory analysis.

Reinforcement Learning: Here’s where Skinner and co. come back into play and an AI agent learns through feedback, the AI’s goal is to maximize its reward through trial and error. You see this form of learning happening with AI learning how to play games, in robotics and in self driving cars. In a nutshell: There is a separate reward system (it could also be a person) that is providing feedback to the learning system in the form of a reward score, after many

attempts the learning system starts to figure out what causes the reward system's scoring to increase or decrease and follows those pathways.

I've written before on the underlying concepts of how a neural network learns, so I'll skip it here for the sake of the reader (and writer). I'll just briefly describe one of the core concepts or the "training loop" that forms the spinal column of machine learning.

The nuts and bolts of learning for an AI system is an iterative process of trial and error guided by a loss function (also called a cost function) that measures how wrong the network's predictions are. Again I'll just go high level and use the Dog photo example again:

Step	What happens	Tiny dog example
1 Feed Forward	The picture's pixel values move through all layers to produce a guess.	The network sees the photo → outputs class probabilities: dog 0.30, cat 0.60, car 0.10.
2 Loss Calculation	A loss function measures the gap between the guess and the truth.	True label = dog (probability 1.0). Cross-entropy loss = $-\log(0.30) \approx 1.20$ → a sizable error because $0.30 \ll 1.0$.
3 Back-propagation	The error signal is sent backward to find how each weight affected the loss.	Working from the output layer to the first layer, the network computes gradients showing which connections lowered dog probability and by how much.
4 Gradient Descent (weight update)	Each weight is nudged in the direction that makes the loss smaller.	Weights that suppressed "dog" are decreased; weights that supported "dog" are increased. After the update the same image might yield dog 0.45, cat 0.45, car 0.10—an immediate improvement.

This cycle of learning is repeated many times until the network figures out how to get the probability of the dog as high as possible for dog photos and as low as possible for everything else. I'll gently introduce the mathematics involved in these steps in later chapters, it's really not difficult and with a bit of visualisation and some simple metaphors anyone can understand what is going on.

So this is the foundation, and on the surface it may not seem particularly human like, but if you take the key concepts I shared from the human learning process you'll see them echoed in each of these steps. AI is inspecting and adapting itself and finally settles on a new changed and learned state where it is able to process and predict unstructured, unsorted or new information.

How are we progressing in our study of how AI learns in 2025? Here are two fascinating studies and I highly recommend them both, if you aren't academically orientated you can find many YouTube videos or LinkedIn articles to explain these simply. Meta's study suggests a new pathway of learning and Anthropic's opens the hood on the current wave of AI models and allows us to look within the machine room of the LLM learning process.

Meta (2024) – Learning at the Level of Concepts: Researchers at Meta have proposed new ways to think about training AI at a higher level of abstraction. In late 2024, Meta's AI research team introduced Large Concept Models (LCMs), an experimental architecture that tries to model language at the level of whole concepts rather than just individual words or tokens arxiv.org. This is a big shift from the usual approach and suggests an alternative path for AI learning: instead of building meaning step by step from tiny pieces (characters or words), the model works with bigger chunks of meaning at once.

Anthropic (2024) – This company headlines in 2024 by achieving the first detailed interpretability analysis of a state-of-the-art large language model. In a report titled "Mapping the Mind of a Large Language Model," published May 2024, Anthropic researchers identified how millions of concepts are represented inside Claude, one of their large language models anthropic.com. This was a breakthrough because it provided a glimpse into what a trained LLM knows and how that knowledge is organized internally. The team used a technique to find emergent features in the neural network's activations. A "feature" here means a pattern of neuron activity that consistently represents some concept or trait. They discovered an astonishing range of human interpretable features in the

model's layers everything from concrete entities like specific cities (e.g. San Francisco), famous individuals (e.g. Rosalind Franklin), or chemical elements, to more abstract themes like "a bug in code" or "the concept of an inner conflict". This study is an amazing read and shows that these language models (in contrast to what critics are saying) build internal representations, which seem like synthetic versions of human learning.

Now we are finally inside the machine and to really understand what's going on here we need to understand and respect the atomic building blocks of these models: Data.

