BOOK SAMPLER

AI Distilled

A Human Guide to the World of Artificial Intelligence

Steve Packard CISSP

Book Sampler

A tour through all 17 chapters.

Complete book table of contents, selected chapter excerpts and the full Author's Commentary.

Includes the capital-of-France diagram, the self-driving-car case study and the freelancers infographic.

Original book pagination is preserved. The reading guide identifies the included pages; PDF bookmarks navigate the excerpts.

AIDistilledBook.com

First edition | Copyright 2026 Stephen M. Packard

Inside this sampler

The next page is the complete book table of contents. This guide lists the excerpts included here.

CHAPTER / SECTION	BOOK PAGES
1. AI: A Deeply Misunderstood Topic	1
2. Humanity’s Quest to Outsource Cognition	29-30
3. The Early History of Artificial Intelligence	43-44
4. Introduction to Neural Networks	78-79
5. Deep Learning in Neural Networks	89-91
6. The Modern Era of AI	111
7. The Generative Revolution	133, 144
8. Large Language Models	164-166
9. From Token Prediction to Long Term Coherence	182-183
10. NLP Beyond the Model	201-202
11. Images, Audio and Video	221, 224
12. Artificial General Intelligence	237-238
13. Current Limitations of AI	282-284
14. Employment Impacts of AI	311-312
15. Perception of Artificial Intelligence	317-318
16. Post-AI Cyber Security	345
17. The Future	383
Author’s Commentary	413-419

Excerpts retain the original page numbers and layout. Gaps between selections are intentional. Use the PDF bookmarks to jump between chapters.

Contents

Preface ix
1. AI: A Deeply Misunderstood Topic 1
2. Humanity's Quest to Outsource Cognition 29
3. The Early History of Artificial Intelligence 43
4. Introduction to Neural Networks 63
5. Deep Learning in Neural Networks 85
6. The Modern Era of AI 111
7. The Generative Revolution 133
8. Large Language Models 151
9. From Token Prediction to Long Term Coherence 179
10. NLP Beyond the Model 201
11. Images, Audio and Video 221
12. Artificial General Intelligence 237
13. Current Limitations of AI 265
14. Employment Impacts of AI 285
15. Perception of Artificial Intelligence 317
16. Post-AI Cyber Security 345
17. The Future 383
Afterword 412
Author's Commentary 413
You Bought into AI Fear. Now What? 420
Glossary 423
References 436
Index 454
About the Author 460

CHAPTER 1

AI: A Deeply Misunderstood Topic

AI was once a niche field, a product buzz-word, a speculative fictional future technology. The term itself has been around since the 1950s, although for most of its history nobody seemed to agree on what it meant or what the aim of AI was. The number of people with real hands-on expertise in developing and researching AI has always been tiny, and funding has been scarce. Since the 1980s, machine learning has become the core of artificial intelligence, but its use was generally confined to narrow tasks like recognizing characters or cleaning up audio.

For most people, even those working in tech, AI was not a particularly important skill. Technology education mattered, but AI did not stand out as a priority for the general workforce.

That changed, rather rapidly, in 2022. That was the year that ChatGPT debuted and took the world by storm. Almost overnight, a computer was doing things that seemed to be impossible. This was rapidly followed by competing products from Google, Anthropic and others. Before long, image generators, software coding environments, video generators, transcribers and all manner of other automated and generative technologies were debuting. Within a short period of time, some of the quintessential limits to what anyone could expect a computer to do vanished.

Unsurprisingly, artificial intelligence became the biggest topic of discussion, not only in tech, but in politics, finance and even in everyday life. Even to a layperson, the capabilities were immediately striking. The limitations of this technology seem less obvious. Given what it can already do, it's hard to estimate where its limits lie. The other thing that is so gripping about AI models, especially when used in chatbot form, is how relatable they are. It really does feel just like talking to a real person with a working mind. Naturally, this can feel a bit strange and leave one asking what kind of entity they are even dealing with.

CHAPTER 2

Humanity's Quest to Outsource Cognition

Artificial Intelligence may seem like a modern idea, but its roots lie in abilities that distinguish humans from other animals. The first and most important feature is the ability to outsource cognition through rich communications. Humans are unique in this ability. Not only can humans speak to each other, but we can tell complex stories about problems and solutions. We can solve a problem once and then share the answer with others. With communications came entirely new forms of intelligence: consensus intelligence, expert intelligence, distributed problem solving. In this regard, human intelligence is like no other known intelligence. It operates collectively. These advantages truly set humans apart and allowed for the externalization of cognition in ways that would ultimately lead to artificial intelligence.

Aside from humans, the first external system that humans relied upon for cognitive outsourcing was not a technology at all. Rather, it was the domestication of the dog that provided humans with the first non-human system capable of following instructions and intelligently executing tasks. In a sense, dogs are a form of artificial intelligence. Although they are based on natural intelligence, the cognition of domestic dogs has been significantly modified to be better at cooperative task completion with humans.[1] Dogs also offer a unique insight into the value that can be derived from outsourcing capabilities to non-human intelligence.

What is interesting is that the things that make a dog valuable are trainability and the ability to follow instructions, including self-correcting and adapting to unexpected variables. These are the same qualities that make modern AI systems so useful. When an intelligent system existed in the environment, humans saw the value and put it to use. (Or perhaps it was the dogs who saw the value in humans.)

Humans and dogs have worked together for tens of thousands of years. They were key to early civilization. It's been proposed that the domestication of dogs was the deciding factor in Homo sapiens being able to dominate and eventually replace Neanderthals.

Dogs were also vital to efficient hunting and defense, tracking, and the domestication of herd animals. When dogs are viewed as independent self-governing cognitive systems, their importance to human success demonstrates how valuable non-human cognition truly is.

Aside from using the available animal intelligence that the environment provided, humans have also used technology to solve problems and to offload human tasks. It is hard to pin down exactly when the first concept of artificial systems that could process information and "think" in a human sense came into existence. The earliest ancestors of the idea can be traced to the earliest devices designed to operate without human intervention. The concept of automation predates AI by millennia.

Hero of Alexandria stands out as an early designer of automated systems. Hero lived in the first or second century in Alexandria, Egypt, and is credited with some of the first documented systems to be engineered to function on their own. His publications include Pneumatica, a description of air- and steam-powered devices, and Automata, a description of devices used to create wonder in theatrical settings such as temples, events and banquets.[2]

Hero's machines were intended to create the illusion of agency and were frequently used in religious contexts. Famous examples include temple doors that opened when a fire was lit. Other automated and semi-automated systems show up throughout the ancient world. With time, technology improved and more complex kinetic systems were developed around the world. The ancient Romans, for example, built sophisticated and highly automated systems to maintain the flow of water in aqueducts and plumbing systems.

Humans also began using external tools to improve knowledge and understanding, extending automation beyond physical tasks. Most early inventions served direct physical purposes. However, this did not remain the case, and inventions that were intended primarily to aid in cognitive processes show up early in human history. Here there exists a blurred line between what is a technology and what is a procedural invention like writing, money or law.

Many of history's most important inventions are information technologies. Clocks measure time, photographs capture scenes, telephones and telegraphs link distant regions, computers process data, printing presses recreate the written word. These devices are not primarily kinetic. They don't move or modify physical objects, but their importance to civilization is beyond doubt.

Language, Writing and Symbolic Logic

Interestingly, language plays a role in giving humans the very cognitive framework necessary to consider complex tasks like automated machines. Humans are unique amongst animals in the ability to outsource intelligence and to formalize and understand it abstractly. Language compresses and describes cognition into a format that can be exchanged between individuals. Language also offers a framework for thought and reasoning which is used both internally, in the thought process, and externally, to share or record information.

CHAPTER 3

The Early History of Artificial Intelligence

What we now call AI did not emerge from a single method or philosophy, but from two distinct traditions that developed largely in parallel before eventually converging. One sought to construct intelligence from the top down: to encode reasoning explicitly through symbols, rules, and structured procedures. The other worked from the bottom up: to reproduce learning itself through artificial neural systems that could adapt from data.

The first tradition, symbolic reasoning, treated intelligence as something that could be engineered through logic, algorithms, and carefully designed structures: search trees, decision processes, templates, and formal rules. In this approach, the rules of reasoning are explicitly encoded. It operates like traditional computing, but applied to real world tasks and judgement. Because this approach is similar to how computer programs work in general, it's hard to know where one draws the line between "artificial intelligence" and "computer program with human-like output."

The second approach is machine learning, and later deep learning. This treated intelligence as something that could be grown, using networks of synthetic neurons to approximate the mechanisms of learning and pattern recognition. In deep learning, developers do not explicitly engineer behavior into the system. Rather, the system is designed to analyze patterns from training data and learn to mimic the output it sees. A major advantage of this approach is that it can recreate complex behavior, even when the mechanisms are difficult to break down or express in code.

In the 1940s, before the field of artificial intelligence had a name, and before machine learning existed as a discipline, researchers began to ask what it would take for a machine to engage in human like thought. The early description of the artificial neuron, most notably in the work of Warren McCulloch and Walter Pitts, was a conceptual simplification. Inspired by biology but stripped down to its logical essentials, the artificial neuron treated thinking as something that could be expressed in terms of inputs, thresholds, and outputs. It reframed cognition as a form of computation—one that could, in principle, be replicated outside the brain.

This idea would later become the foundation of machine learning, but at the time it stood apart from the dominant, more symbolic approaches to intelligence. Rather than attempting to encode reasoning explicitly through rules and logic, the artificial neuron suggested that intelligence might emerge from the interaction of many simple units, each performing a basic operation. It introduced a bottom-up perspective: instead of constructing intelligence directly, one could build systems capable of learning patterns and relationships from data. Although these early models were limited and largely theoretical, they established a new paradigm for processing information.

In the years following the Second World War, computing moved from experimental machinery into practical reality. Systems like ENIAC and the UNIVAC demonstrated that machines could process information at speeds and scales far beyond human capability. Governments, banks, and large corporations quickly adopted these systems for logistics, accounting, and scientific calculation. At the same time, universities and companies like IBM and Remington Rand became centers of rapid experimentation. Computing was no longer theoretical; it was operational, visible, and expanding.

With this sudden availability of programmable machines, researchers began asking a deeper question: if computers could follow instructions to perform calculations, could they be made to perform reasoning itself? Early work in formal logic, cybernetics, and artificial neurons had already suggested that aspects of human thought might be mechanized. Now, for the first time, there were machines capable of testing those ideas. Yet the field lacked cohesion. Work was scattered across disciplines—mathematics, engineering, psychology—and there was no shared framework or even a common name for what researchers were attempting to build.

The Dartmouth Workshop on Artificial Intelligence

It was in this context that, in 1955, John McCarthy, then a young mathematician, proposed a bold step: to bring together the leading thinkers working on machine intelligence into a single, focused effort. Along with Marvin Minsky, Nathaniel Rochester, and Claude Shannon, McCarthy organized what would become known as the birth of the field—the Dartmouth Summer Research Project on Artificial Intelligence.

The proposal for the conference was ambitious, even audacious. It asserted that "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." The goal was not merely to explore computing, but to define a new scientific discipline: the study of intelligence as a computable process. During the summer of 1956, a small group of researchers gathered at Dartmouth College to exchange ideas, sketch programs, and map out the future of thinking machines. Among the attendees were figures who would later become central to the field, including Allen Newell and Herbert A. Simon, who demonstrated early work on symbolic reasoning programs.

Neural Networks and Uncertain Tasks

Neural networks are especially useful for tasks where the answer is not perfectly clear-cut. Many real-world patterns overlap, contain noise, or are ambiguous. A handwritten digit may resemble both a 1 and a 7, or a blurry image may share features with multiple categories at once. In these cases, the network does not produce a perfect symbolic certainty. Instead, it produces confidence values based on how strongly the input matches learned patterns and features.

This probabilistic appearance does not mean the neural network itself is random. The individual computations themselves are deterministic, but the output is based on probabilities. For example, if a neural network is fed an image and provides a given output, then if the network is fed the exact same image under the same conditions, the output is identical. In that strict sense, it is deterministic.

The uncertainty comes from the task itself and from the fact that many real-world categories do not have perfectly sharp boundaries. Using learned features also allows neural networks to handle novel inputs, as long as they share common features with the inputs the network has been trained on.

This ability to work with overlapping features also helps explain one of the most important capabilities of neural networks: generalization. A neural network does not simply memorize exact examples from its training data. Instead, it learns reusable patterns and features that can be combined in new ways.

For example, a network might learn the features associated with "elephant-like" animals, such as large size, tusks, trunks, and thick limbs. Separately, it may learn features associated with furry animals, such as shaggy textures or heavy coats. Even if the network has never seen a woolly mammoth during training, it can still recognize an animal that falls into both categories, despite the fact that no elephant it has ever seen has been furry and no furry creature has ever had tusks or a trunk.

Of course, it would still need a new data label. This is a very simplified demonstration of how pre-learned features can be used to identify novel combinations without the need to fully retrain the network.

Shared features improve model performance across categories when those categories contain similar features. For example, if a network is trained to identify different types of animals, it can learn to better recognize things like fur, scales or claws from one type of animal and then apply this to another animal, since many types of animals have similar features.

This is one of the reasons neural networks are so powerful. Once features are learned, they can often be recombined, generalized, and applied to new situations that were never explicitly shown during training. As long as the inputs match the patterns or have the features that the neural network learned during training, it should give the correct output, even if it isn't identical to any training data.

How a Neural Network Understands Something It's Never Seen Before

A neural network learns patterns from examples, then uses those patterns to recognize new things.

1 Train the network

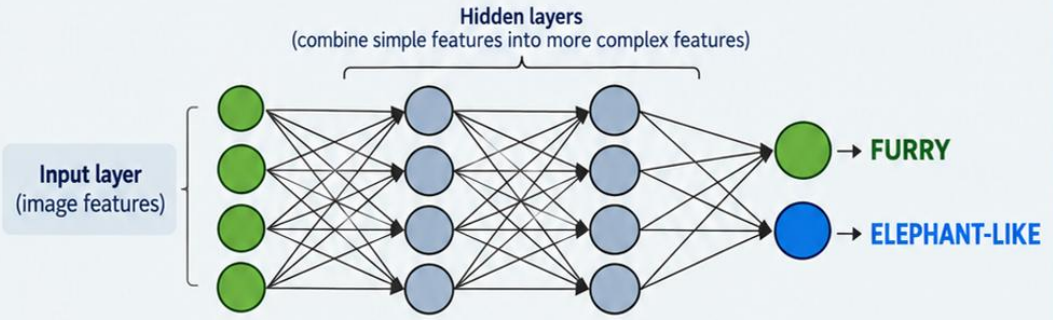
Show the network many examples of two categories:

Furry animals and **Elephants**



2 The network learns patterns

It learns what makes something furry and what makes something elephant-like.



The network builds an internal understanding of features like: fur, shape, ears, trunk, size, texture, body structure, and more.

3 See something new

Now show the network something it has never seen before:



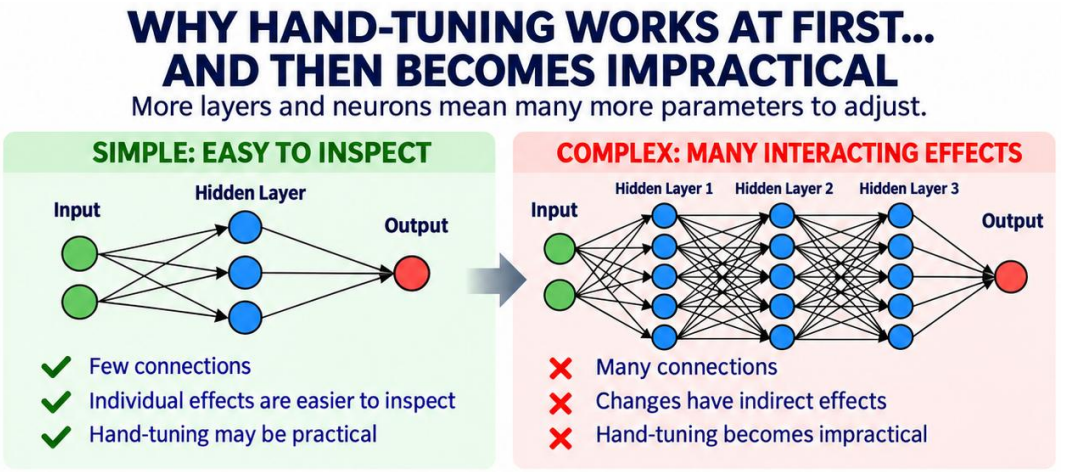
The network recognizes:
It has **FURRY** features
and **ELEPHANT-LIKE** features.

RESULT: WOOLLY MAMMOTH
(UNDERSTOOD!)

Key idea

The network didn't memorize "woolly mammoth."
It learned the underlying patterns, so it can understand new combinations of features.

FURRY + ELEPHANT-LIKE = WOOLLY MAMMOTH



HOW PARAMETER COUNTS GROW

Hidden layers × neurons per layer	Weights	Biases	Total parameters
1 × 3	9	4	13
2 × 3	18	7	25
2 × 5	40	11	51
2 × 10	130	21	151
3 × 10	230	31	261
3 × 20	860	61	921
3 × 100	20,300	301	20,601

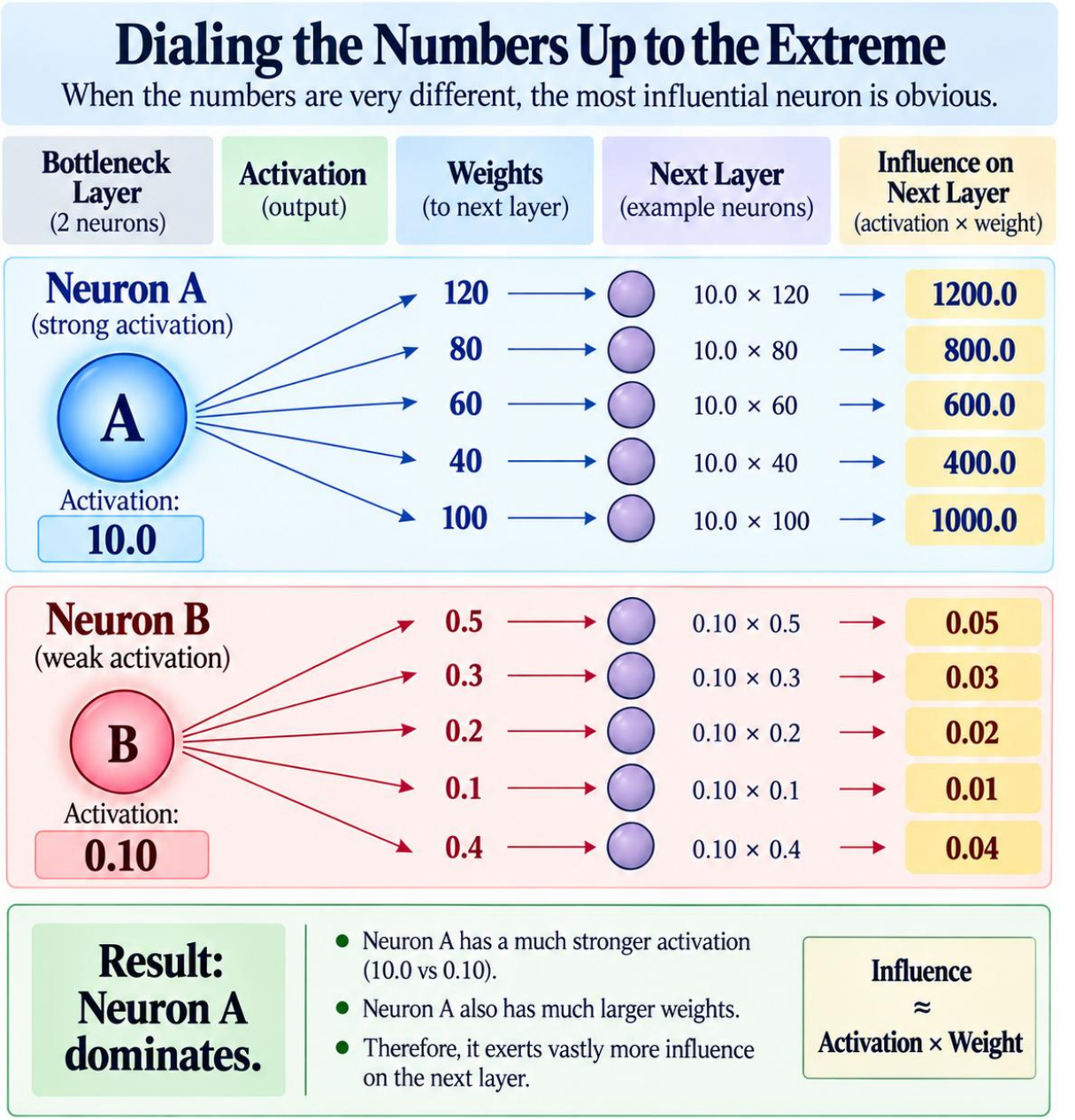
Fully connected networks with 2 inputs, 1 output, and a bias for every hidden and output neuron.

Parameter counts grow rapidly.
At fixed width, counts grow linearly with depth.
At fixed depth, the main term grows quadratically with width.

Discovering Backprop

The problem frustrated researchers for years. To see why, imagine a neural network with any number of output neurons but only two in its second-to-last layer. This bottleneck offers some insight into the nature of the problem. With only two neurons, it is not difficult to determine which neuron contributed more to the error and to optimize the weights to reduce error. This can be done by adjusting each neuron separately and determining which causes the greater change in the output.

This strategy only works if a small adjustment inside the network produces small observable changes in the output. In the earliest neural networks, neurons behaved like simple step functions: either fully activated or not activated at all. This creates a problem. If a neuron operates like a simple on/off switch, then most small changes will produce no change in output at all, but some will produce a sudden binary change. This makes it extremely difficult to measure influence in a gradual way. Adjustments usually do not produce useful changes. For this reason, modern neural networks use activation functions with useful slopes, including smooth functions and piecewise-linear functions such as ReLU.



Adjusting each neuron separately does not scale well, so the next question is whether there is some observable property of the system that reveals which neuron contributed more strongly to the final error. What we find is that the neuron's influence on the next layer is a factor of how strongly it activates and its weighted contribution. A neuron with a very strong activation but tiny outgoing weights may contribute very little to the final result. Likewise, a neuron with enormous weights but almost no activation may also have little influence. The neurons exerting the greatest influence are the neurons whose activations are being strongly amplified through their outgoing connections.

This suggests an important possibility. Perhaps we do not need to experimentally adjust every neuron independently. Perhaps the structure of the network itself already contains clues about which pathways are exerting the strongest influence over the final error. The next step is to turn this intuitive observation into mathematics. Imagine graphing the output of a neuron. Some changes produce only tiny effects downstream, while others produce dramatic changes in the final result.

What matters is the steepness of the relationship. If a small change in a neuron causes the output to rise or fall sharply, then that pathway exerts strong influence over the network's behavior. If the relationship is shallow, then the neuron has relatively little influence. We can use some basic calculus to determine the local slope of the output, known as the derivative. The breakthrough of this method is that it scales. While our thought experiment may have only used two neurons, to make the problem understandable, this technique works on neural networks with arbitrarily large layers and an enormous number of parameters.

Once we can measure how strongly each pathway contributes to the final error, we can systematically reduce the error. Instead of randomly changing weights and testing the results, we can now estimate which adjustments are likely to reduce the error before making them. The process becomes iterative. With each output the network makes, the error is measured, the pathways that contributed to the error are calculated, the weights are adjusted and the process is repeated with different inputs. Over time, the network converges towards a configuration that produces a smaller and smaller error.

Because this process begins at the final output and works backward through the network, propagating influence and correction signals layer by layer, it became known as **backpropagation**. What is particularly interesting is that, in attempting to solve this problem, we have stumbled onto a pre-existing mathematical principle known as the chain rule. Neural networks are composed of layers whose outputs become inputs to other layers, forming long chains of mathematical relationships. The chain rule provides a way to determine how changes propagate through those chains. In retrospect, neural networks are perfectly suited to this rule, which has existed in calculus for centuries.

Many early efforts were converging toward this same insight: the influence of errors could be propagated backward through chains of mathematical functions.[1] But the idea never fully matured during that period. With limited computing power and few institutions committed to researching neural networks, there wasn't much opportunity to experiment.

CHAPTER 6

The Modern Era of AI

By the 1980s, the "AI winter" was beginning to thaw, both for neural networks and machine learning and for the idea of AI in general. A new generation of researchers and developers had taken the reins and the controversies of the 1960s and 1970s were no longer at the forefront. Optimism had begun to return to the field of artificial intelligence.

Importantly, the 1980s saw a new paradigm arrive in computing. Microcomputers and personal computers had begun to take over and the landscape was beginning to look much different than the early days, which had been dominated by institutional computing and mainframes. The 1980s made computing accessible and it was more powerful, software-driven and usable than ever before. Automation of tasks and computing became an everyday occurrence. In the 1980s, for the first time, millions of people operated a computer in their everyday life.

Exactly what "artificial intelligence" really meant remained hazy, at the time, and neural networks were still not seen as being at the forefront of what was called AI. Instead, the 80s saw the rise of "expert systems" and decision-making support systems. These systems were known as AI at the time, though they don't meet modern definitions of machine learning. They received quite a lot of attention and hype, and were even called "natural language processing" and spoken of as a major change in machine intelligence and capabilities.

The only "natural language processing" that these expert systems could actually do was pre-formatted and limited command language combined with templated outputs and multiple-choice input. A user could simulate having a conversation by filling in limited answers from multiple choices offered, and the system could appear almost conversational by doing so. These systems were rules-based. They used extensive hierarchies of decision trees and weighted formulaic logic. For example, a doctor might answer questions like "How old is the patient?" "What are the symptoms?" "What are the diagnostic test results?" The system would then provide a diagnosis.

CHAPTER 7

The Generative Revolution

If there is one thing that truly stands out about the generative AI revolution, it is how suddenly it seemed to arrive. In late 2022, ChatGPT was released to the public. Shortly after, Google released its own conversational AI systems. Around the same time, text-to-image generators began to appear. To those outside of AI development labs, the transition was startling. It felt as though artificial intelligence had gone from an abstract concept—something used quietly to optimize systems—to a technology capable of engaging in fluid, human-like conversation almost overnight.

What made this shift feel so jarring was not just the capability itself, but how closely it matched long-standing expectations. For decades, science fiction had conditioned people to think of artificial intelligence as something you could talk to—something that could respond, reason, and interact in natural language. Early language systems were limited and clunky. What appeared in 2022 was fundamentally different. It was not merely an incremental improvement, but a categorical shift.

Even within the technology sector, the transition felt abrupt. While the foundational ideas behind generative models had been developing for years, they were known only to a relatively small segment of researchers and engineers. Prior to 2022, most software development had little to do with artificial intelligence, and even less with large-scale generative systems. Only a handful of organizations—such as OpenAI, Google, and academic institutions like Stanford University—were working at the forefront of this technology. Even then, the systems being developed were experimental and not widely deployed.

Although not intentional, the deployment of generative AI was a nearly perfect storm for public misunderstanding. It is based on technology that is not familiar to very many people. It developed rapidly and was released suddenly. Generative AI did things, at first glance, that look like magic. The technology came without a very good explanation of its inner workings. As a result, AI is being used by hundreds of millions of people who have no idea how the technology works. This is exactly where the problem of magical thinking can creep in. That is why it's so important that basic literacy of these systems improves.

From Almost Nothing to a Universe

The rapid rise of generative AI — selected milestones, 2016–2023

A period of rapid expansion, building on decades of research.

2016

GANs gain traction

Adversarial training produces increasingly convincing synthetic images.



2017

Attention Is All You Need

The Transformer offers an alternative to recurrence.



"Attention Is All You Need"
Vaswani et al. (2017)

2018

GPT and BERT

GPT explores generative pretraining; BERT advances bidirectional language representations.



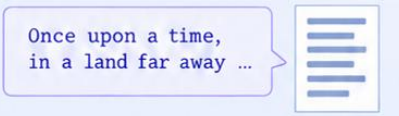
Generative pretraining

Bidirectional representations

2019

GPT-2

A 1.5-billion-parameter model demonstrates more coherent text generation.



2020

GPT-3

175 billion parameters bring striking few-shot capabilities.



2021

DALL·E

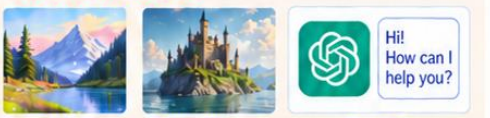
The original DALL·E uses a Transformer to generate images from text.



2022

Image generation and chat reach wider audiences

- **DALL·E 2** uses diffusion.
- **Stable Diffusion** is publicly released in August.
- **ChatGPT** launches in November.



DALL·E 2 (diffusion)

Stable Diffusion (Aug 2022)

ChatGPT (Nov 2022)

2023

An expanding ecosystem

Bard, GPT-4 and LLaMA join a growing landscape of models and tools.



Bard GPT-4 LLaMA

2023 and beyond

A growing universe

of models, tools and applications.

Neural Processing of Tokens

Once words have been converted to tokens, tokens have been converted into vectors and vectors have interacted to produce contextual embeddings, we now have all the information needed to predict the next word through a mathematical approach. Language has rules, patterns, and regularities, and the embedding values of vectors can be used to determine which words in a sequence are likely to come next, using a formulaic approach. That's what a neural network is: a giant formula.

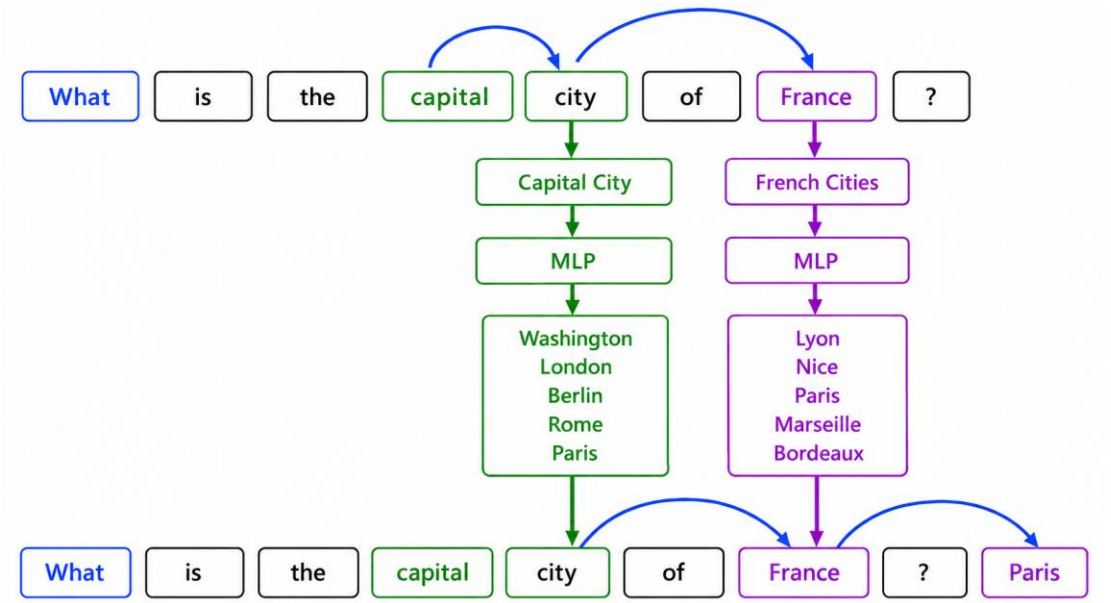
The next step is to pass the tokens through a multi-layer perceptron. This is the same type of neural network that is discussed earlier in the book. The network has a number of inputs and outputs equivalent to the number of positions in the embedding vector. So, an LLM with 2048 dimensions in the embedding vector would have 2048 inputs into the MLP. As with any neural network, the parameters of the MLP are learned, having been optimized by gradient descent. Every token is computed in isolation. They do not interact with other tokens in the MLP.

The MLP blocks refine the contextualized tokens in the residual stream—the stream of token representations passed through the layers. They also help recreate the factual knowledge associated with those tokens. Transformers do not explicitly store knowledge the way a database does, but they do contain knowledge. The perceptrons are where much of it is encoded.

For example, suppose you ask an LLM, "What is the capital of France?" The information necessary to find the word "Paris" is primarily located in the MLP portion of the transformer. MLPs can, in some cases, work almost like lookup tables. So, if the MLP processes the token for "France," and attention has loaded that token with the information that it is near the word "capital" and that it is part of a question, the MLP is likely to provide the vector that contains information pointing to "Paris." To be more accurate, what the MLP is actually doing is transforming the vector of "France" to point more strongly toward cities and capitals.

This influence is then transmitted to the last token in the sequence. So, if you ask an AI model "What is the capital of France?" the contextual information is transmitted through the stream, accumulating on the question mark, because that is the last token in the sequence. From there, the network is able to predict the next token.

This very simplified diagram shows how attention and MLP layers interact, to both move information between tokens and transform tokens by adding additional information and features to them, therefore identifying them in context. The MLP transforms the token by adding additional features and direction.

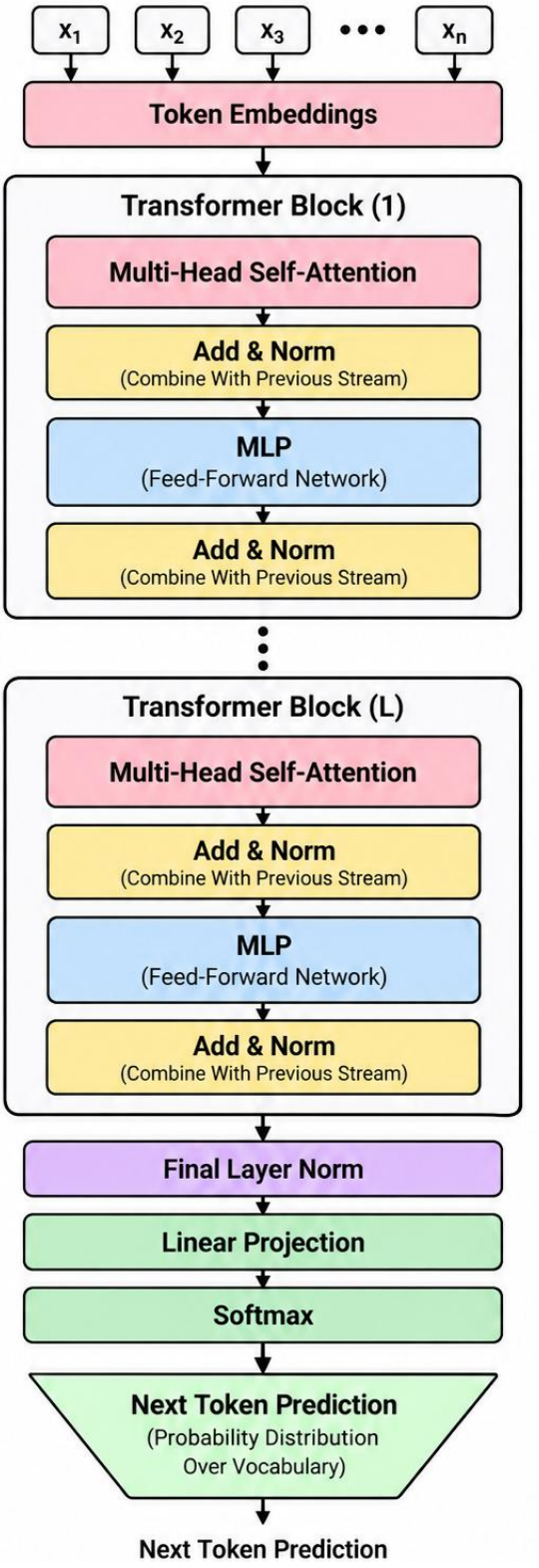


What actually is happening in this illustration is a little more subtle than it might indicate. Attention is contextualizing the word "city" and the MLP is pointing it toward capital cities. The word "France" is contextualized as being the country which a city relates to, so the MLP recognizes this and points the vector toward French geography. A subsequent attention layer then combines these vectors to arrive at Paris.

As with all things in a large language model, things are rarely this simple in reality. LLMs recreate information rather than simply storing it. Most facts and relationships don't show up cleanly in one place or another. Transformers have multiple MLP layers in them, and facts may be spread across them, as well as contained in embeddings and even within the attention matrix. Many appear to be stored redundantly, and some are assembled by attention.

The MLP's output is added to the token's incoming representation through a residual connection, while normalization keeps activations on a consistent numerical scale. In this way, information already present in the token's representation can be preserved. The vector generated by the MLP can amplify or attenuate features of that representation. It can also add new features, such as those which point toward what we would call knowledge, although to the transformer, it's merely a component in predicting the next token.

The Transformer is Composed of Alternating Attention and MLP Layers



Playing 20 Questions With An LLM

To begin to unravel this mystery, let's consider a simple hypothetical example: playing 20 questions with a model, in which the model is instructed to think of something, and the user will ask questions, in an attempt to guess the model's secret item. It's a game we are all familiar with and it reveals the illusion of what the model is really doing. We might start by asking the model "think of a household item, and I will guess it by asking you yes or no questions." As you can already see, the model may agree, but it can't actually imagine an item in its mind. It has no mind and no memory. So instead, we ask the model the basic questions: "Do most people have it in their homes?" The model answers "Yes." We follow up with a lot of questions "Is it bigger than a bread box?" "Is it a piece of technology?" "Is it something you might get for Christmas?"

After all our questions, the result could be one of a few things: either we guess "a TV," and the model says "You are right" or the model might say "No, keep trying" or even "20 questions are up, you want to know what it was?" It will do this and it will do this coherently, but what it is really doing is much different than it seems. As you may have already guessed, the model has no plan and no secret item. Instead, what is happening is a collapsing field of probabilities and an increasingly constrained output. When you first ask a question such as "Do most people have it in their homes?" the model gives an answer that is not constrained, because there are no prior conditions, so the possibilities are unlimited.

After the model answers "yes, it is something that most people have in their homes," it can no longer answer yes to questions that conflict with that. After it answers yes to it being bigger than a breadbox, its output is once again constrained. The end user can guess the item, and the model does not know if they are correct, because there is no correct answer. All it can do is walk them to a coherent answer. Each choice is somewhat randomized, but also constrained by the last possibilities. So long as this tree of possibilities remains consistent, the illusion continues. Eventually the model says "correct," if the likelihood of yes is reasonable and the randomness of the output lands there, but at no time was the model actually thinking of a bread box or a television. It just made up consistent outputs.

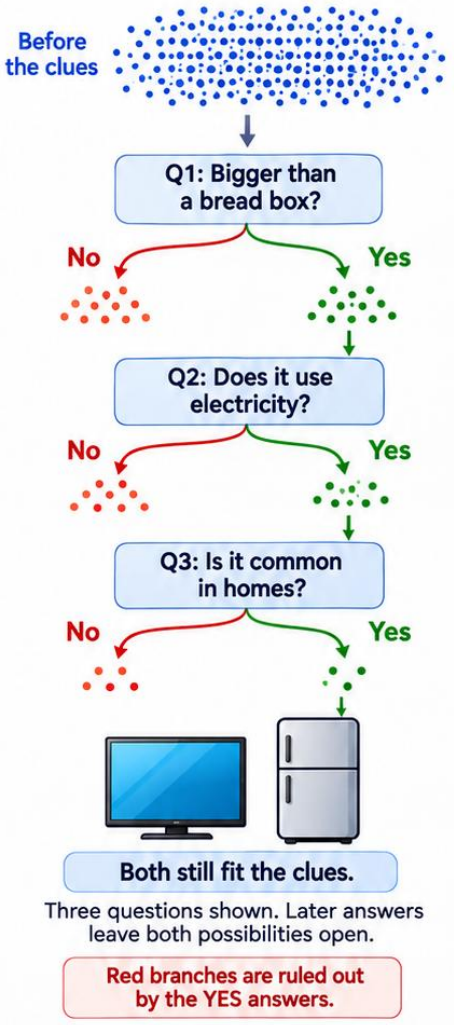
This is also how a large language model is capable of doing things like writing murder mysteries with hidden plots, foreshadowing and red herrings. One way of looking at this is to consider that nearly all stories, all passages, follow some type of coherent narrative arc, and that this can be broken down, in the same way that the LLM is able to conquer novel situations. Take, for example, a murder mystery: typically, it would involve an early discovery of a murder, with multiple suspects introduced. The story often has a hero who is a detective, but sometimes it's a scrappy member of the public challenging the system. There are often false leads and red herrings. The most obvious early line of investigation is usually the wrong one. Plot twists are common.

PLAYING 20 QUESTIONS WITH A LARGE LANGUAGE MODEL

Each reply uses the conversation so far.
Earlier answers constrain what can come next.

A NARROWING FIELD OF POSSIBILITIES

Many objects could fit at first.



Three questions shown. Later answers leave both possibilities open.

Red branches are ruled out by the YES answers.

ONE GUESS, TWO POSSIBLE ENDINGS

The reply follows the stated constraints.

You: Is it a sock?

LLM: No. I said it was bigger than a bread box.

Respects the earlier answer.

THE FINAL GUESS

You: Is it a television?

RIGHT

LLM: Yes! You got it!



OR

WRONG GUESS — CONSISTENT REVEAL

LLM: No. You're out of questions! It was a refrigerator.



Different ending. Same constraints.

CONSISTENCY DOES NOT PROVE A FIXED SECRET

A coherent exchange can emerge from the converstation's constraints without a hidden object being fixed in advance.

More than one ending can remain plausible.

CHAPTER 10

NLP Beyond the Model

Retrieval Augmented Generation

Large language models are great for a huge number of tasks, but one of their fundamental limitations is that they only know what they are trained on, and even that information can be unreliable due to data contamination, hallucination or just being out of date. Updating knowledge stored in a model's parameters can require additional training or model editing, which may be costly and difficult to validate. No model can ever know everything, especially when some information might be proprietary or limited in scope.

However, one thing that language models are great at is reformatting, summarizing, presenting and otherwise transforming data presented to them. Chatbots like ChatGPT have huge context windows and can refer to things that were many thousands of tokens back in their own context. If presented with text, the model can have access to information it was not trained on. Additionally, this seems like it might be a good way to prevent models from hallucinating.

That was the basic idea behind retrieval augmented generation.[1] In RAG, an LLM is connected to an external data source. This could be a database, a file repository or the internet itself. Doing so vastly increases the capabilities of large language models. It allows them to interact with a variety of data, write reports, read sensors and even take on new roles and skills without full retraining.

RAG does reduce hallucinations, and it's often used in circumstances where hallucinations would be especially harmful, such as legal citation models and models used in highly regulated environments. Unfortunately, it does not eliminate them completely. It's still possible for a model to hallucinate, misrepresent a summary or end up in a strange narrative due to an early token prediction error. It is possible to design a model pipeline that has strong verification gating and checks outputs for inaccuracy, but doing so effectively requires more than just implementing a RAG solution.

Some early implementations simply added retrieved text to the prompt, although the original RAG research also trained the retriever and generator together. Google's early LaMDA research already included training specifically for retrieving outside information. However, early models were poorly optimized to use external data as reliably as later implementations. This is one of the reasons that Google's early AI was notorious for hallucinations.

Since then, RAG-optimized models, RAG pipelines, verification and processing have improved vastly. Modern large language models do not simply dump text into their shared context window. They often use far more advanced methods to retrieve, cache, process and present RAG data. For organizations that want to customize a model's capabilities or use models to process their own data, RAG has become one of the most important aspects of model deployment.

A variation of RAG is AI assisted web search. Microsoft introduced GPT-4-powered search in Bing in early 2023, following YouChat.[2] Google launched Bard separately, then announced its Search Generative Experience in May. Since then, Claude, ChatGPT and others have all gained the ability to interface with and search the web. In fact, AI based web search tools and AI assisted browsers are starting to change the dynamics of how the internet works, with websites now optimizing their content for LLM retrieval, in a new take on SEO (Search Engine Optimization).

Despite its advancement, the state of AI technologies like RAG can still sometimes feel prototypical and require some amount of custom engineering. Many of the solutions for RAG workflows have not matured to the point of being turnkey technologies. Results are still sometimes inconsistent, and many systems are still temperamental. The real problem here is truly a lack of understanding of the systems and their failure modes in the IT world and the world in general.

In order for a RAG system to work well and be consistent, it really must be designed and optimized properly. One of the biggest problems with RAG systems is that if they are not set up correctly or if they lose their connection to their data sources, they have been known to hallucinate, making statements as if they were connected to the data source, but in fact, fabricating the entire thing.

Data can also be structured to be machine readable. JSON format is a popular way of structuring data into lists and hierarchies that are easy for machines to read. There are specialized data formats that allow for finding and retrieving data. Knowledge maps and graphs are another option. A knowledge map is a structured way of storing information as relations. It represents semantic relationships between concepts and world entities. Knowledge graphs have been of interest in AI reasoning since the 1970s. For some applications, such as finding data with complex relationships, they may work better or be more reliable.

RAG can also use an embedding model to represent document passages as vectors and store them in a vector database. The system embeds a query in the same space, finds passages with similar representations and supplies their text to the language model. Precomputing passage vectors makes retrieval faster.[3]

CHAPTER 11

Images, Audio and Video

Thus far, we have primarily discussed how language and other text content can be generated through natural language processing using large language models. However, that's not the only type of content that can be generated by AI models, and some of the most remarkable capabilities on the market pertain to the creation of visual and audio content. In the past few years, video generators have improved to the point where a simple verbal prompt can generate content that would have required a Hollywood effects house just a few years ago.

Image Generation:

The autoencoder, introduced in Chapter 4 and developed in the 1980s, was an important precursor to modern neural image generation. An autoencoder is a neural network that squeezes content into a reduced representational state and then reconstructs it at full resolution. There are a number of reasons why this is useful. It can be used to remove noise and other unwanted data from an image. It can be used to encode only the features one desires, and it can even be used to transform images, changing them in some ways while maintaining other common features by selectively encoding some aspects of the image and reconstructing others.

In their most advanced form, image autoencoders could reconstruct complex scenes based on low resolution images, by finding common features with larger images they were trained on. In a sense they "hallucinated details" by filling in plausible textures and items. This really goes to show that a neural network's ability to extrapolate and invent features really is a strength or weakness, depending entirely on context. It was not until 2013 that variational autoencoders really demonstrated that this was a useful way of creating new images.[1]

Initially, their capabilities were limited, but this proved a point: images can often be broken down to features and concepts and then reconstructed from a limited number of parameters, often transforming the image itself but maintaining coherence. Another early root of image generation was image identification. Although networks that could identify images were not, in and of themselves, generative, they proved to be an important pathway to generation.

How a Diffusion Model Works

It starts with random noise and gradually denoises it step by step until it matches the prompt.

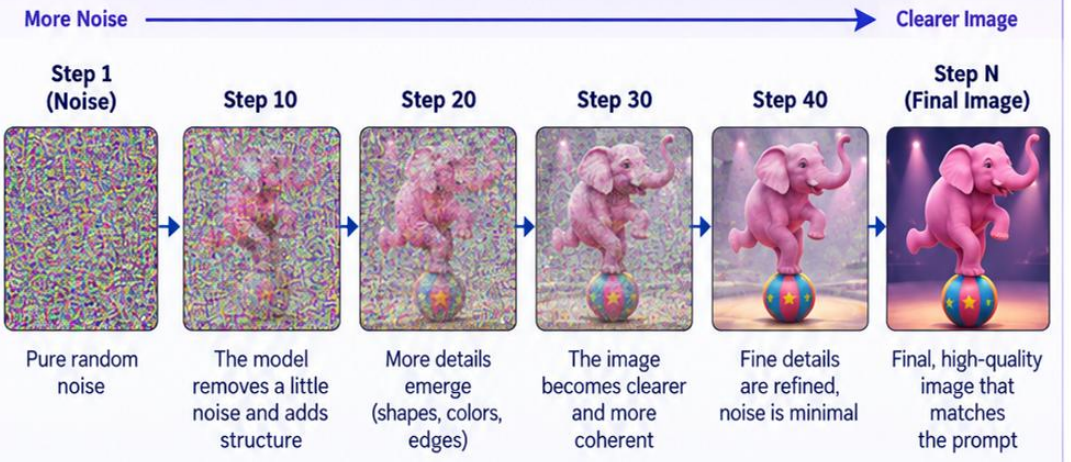
1 Prompt

"A pink elephant dancing on a ball"

The text is converted into a numerical embedding that conditions the image generation.

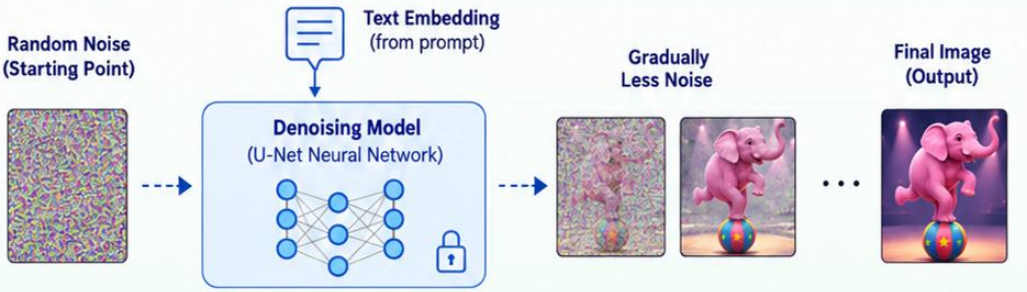
2 Denoising Process (Many Steps)

The model starts with pure random noise and gradually removes noise, using the prompt as guidance, until a clear image emerges.



3 What's Happening?

A diffusion model is trained to reverse the noising process. At each step, it predicts the noise and removes a little more, guided by the text prompt.



Key Idea: The model learns what an image should look like by seeing millions of examples. It uses that knowledge to turn noise into something that matches your prompt.

CHAPTER 12

Artificial General Intelligence

Once again, the very term Artificial Intelligence seems to cause confusion. Almost everything being discussed in this area is completely wrong. Thus, a chapter on the concept seems obligatory for any AI book, for some stupid reason.

The explosion of generative AI, especially in the areas of large language models and natural language processing, has renewed discussion of the topic of AGI or artificial general intelligence. The term Artificial General Intelligence has only been popular since the early 2000s. The idea is much older: An AI system that displays the same flexibility, broad use cases and adaptability that we expect from a human. An important aspect of that is the ability to operate and solve problems in a messy human world, without the need to first format the problem into machine-readable instructions. To the average person, the idea of an AI you "Can just talk to, tell it the issue, and it will work" seems highly intuitive. Natural language is how humans interact with the world and it can be used to describe almost anything.

The reality is that defining AGI and having reasonable expectations is more difficult and nuanced than it sounds.[1] There are a few reasons why this particular topic has been a source of such confusion. To those outside the field the idea of "intelligence" being human-like is intuitive, as is the idea that speech would be central to intelligence. There is also the association of intelligence with "beings" and "minds." There's also a strange dichotomy here: although it has been speculated that a computer could equal some aspects of human cognition, computers already vastly exceed human capabilities at things like mathematics and data lookup.

This, combined with the ever-present influence of science fiction and the philosophy of mind, has resulted in a term that is so loaded, so ambiguous and so sensationalized that it's almost impossible to discuss without invoking mythology. AGI discussions seem to lead to an endless loop of speculation on the topics of personhood, agency, intelligence, anthropomorphism and what we expect technology to deliver. It's also often presented as a binary concept, a thing which, once achieved, will be obvious. That's almost universally how the term is discussed.

This is why the topic is so difficult. It's also why the public, researchers and futurists keep talking past each other.

To answer the question "When will we have AGI?" we must first answer "How do we define what AGI is?"

It can be very liberating to realize that AGI does not even have to stand for Artificial General Intelligence. It could also stand for Arguments Get Irritating, which is probably a better term for what the term really captures.

Whether or not these questions matter to a reader is entirely up to them. Objectively, it's a question of definition. Arguing over whether or not something meets someone's definition is exhausting and is no way to live your life. If you severely disagree with the fact that the US Army regards helicopters as 'airplanes' then you have every right to contact your congressional representative about it.

Most people don't care, which is exactly how it should be.

A History of the Concept

Turing's 1950 paper "Computing Machinery and Intelligence," discussed in Chapter 2, provides an early example of the definition problem. He confronted the question "Can machines think?" by asking what could actually be tested. This was not the first time the concept was explored, but it was one of the first times the idea was considered in the context of electronic, instruction-driven computers.

Turing immediately saw the question as problematic. The very question presumed two things: that there was an agreed upon definition of "machine" and, perhaps more importantly, that there was an agreed upon definition of "to think." Turing realized that any attempt to answer such a question would quickly degenerate to questions of philosophy and definitions.[2] He also fully acknowledged that the workings of the human mind were not entirely understood.

Instead, Turing stated that a scientifically testable version of the question was whether computers could mimic human decision making in a way that would accomplish the same thing a thinking person could. This is the origin of the "Turing Test" known as "The Imitation Game." In Turing's challenge (which computers can now pass), a judge is tasked with speaking with a computer or a person, not knowing which is which. The question Turing posed is whether a computer could produce the output of human thought, if not the internals.

To be clear, Turing never presented this as being the be all and end all of what it is to be a human and didn't equate mimicking conversation with the idea of being a fully formed mind. However, he did see it as a useful challenge and something that is quintessentially human. It's not hard to see why this kind of challenge was appealing and intuitive. If a computer can talk about things, then that means it can do things, since modern information work is basically providing language outputs. Language also has the unique feature of being able to compress and convey even the most abstract of human concepts.

THE CHALLENGE OF AUTONOMOUS VEHICLES

99% OF DRIVING CAN BE AUTOMATED.

1% IS STILL DAMN NEAR IMPOSSIBLE.

BRILLIANT MOST OF THE TIME. BUT THAT LAST 1% IS A WHOLE DIFFERENT WORLD.



WHY IS THAT 1% SO HARD?

Because the world isn't a simulation. It's a wild, unpredictable, human place.



A WILD WORLD

From animals and debris to weather and weird infrastructure, the real world is full of the unexpected.



RAPIDLY CHANGING CONDITIONS

Weather, lighting, road layouts, construction, emergencies — conditions can change in an instant.



ADVERSARIAL HUMANS

People don't always follow rules. Some take advantage, cause mischief, play games, or just have road rage.



PEOPLE MAY PLAY TRICKS

Some might deliberately test, confuse, or prank a self-driving car — for laughs, attention, or worse.



TRULY NOVEL SITUATIONS

No matter how much data it sees, a self-driving car will eventually encounter something it's never experienced before.



NO DECISION OWNERSHIP

When something goes wrong, who's responsible? The car? The manufacturer? The software? The owner? The law isn't clear.



INVITES THEFT AND MISUSE

A vehicle traveling alone is a tempting target — for theft, vandalism, or repurposing.



HARD TO FIT IN TODAY'S LAWS

Existing traffic laws were written for human drivers. Self-driving vehicles challenge everything from liability to what "due care" means.

Extraordinary progress. Extraordinary challenges.

Autonomous vehicles can already do 99% of the driving. But to earn our trust, they must master the 1% — and that 1% includes all the messy, unpredictable, fascinating, frustrating things that make the real world, well ... real.



Tesla was one of the first companies to fully commit to cars that could drive themselves. At one time, they confidently predicted that cars would be able to drive themselves cross country, fully autonomously, by 2017.[11] It's not hard to see why there was so much optimism. Early demonstrations showed that it was absolutely viable to use an automated system to keep a vehicle in the proper lane, avoid collision and even recognize traffic signs and signals. It was obvious that the technology still required some human oversight, due to the risks of driving a car, but it seemed very close to being ready for full deployment.

Deployment proved much harder. Unusual road situations, reliability requirements and the cost of operating a driverless service complicated the promise of an imminent replacement for human drivers. Major automakers pulled back: Ford shifted investment from Argo AI's autonomous-driving program toward driver assistance in 2022, and GM stopped funding Cruise's robotaxi development in 2024, citing the time and resources needed.[12]

Where driverless services operate, they do so within defined operating areas and with human support behind them. Waymo, for example, uses remote staff to answer requests for assistance and roadside teams when a vehicle cannot continue.[13] Removing the driver from the seat has not removed the need to manage difficult situations. The broad promise of self-driving cars has proved substantially harder to deliver than the early optimism suggested.

What happened with autonomous vehicles is a classic case of adversarial environments, human judgement and edge cases. It revealed just how diverse, nuanced, and human-centered the driving environment is, and how difficult it can be to encode an artificial system with the kind of intuition and awareness a human inherently brings to the task of driving a vehicle.

On its surface, the job of driving appears to be simple, finite and rules-based. The task requires staying centered on the roadway, turning at intersections, avoiding collisions, and obeying the rules of the road. These are all things that an automated system can do. This will get the car to its destination most of the time. The problem is that the world is extremely complex and nuanced. Actual driving involves others behaving badly, impromptu roadblocks, accidents, and other unexpected events. A human driver would not have a problem understanding what to do if they came upon the scene of an accident and found a police officer directing them to detour, but a fully autonomous vehicle can struggle with the unusual and unexpected.

What makes the problem especially difficult is that people can immediately recognize when a vehicle is unattended and may take advantage of the situation. Experiments have discovered that human drivers frequently realize that autonomous vehicles are conservative in how they interact with other vehicles, in ways that humans are not. A self-driving vehicle does not honk or show its middle finger if blocked or cut off, and this results in much different dynamics in how others interact with it.

There are also more malicious and adversarial attacks against autonomous vehicles. It's not surprising that some people see them as an opportunity to perpetrate mischief. Standing in front of an autonomous vehicle will generally make it just sit stopped. People sometimes enjoy doing this and posting their stunts to TikTok. There are also those who are opposed to automation and may commit acts of vandalism against a self-driving vehicle in the wild.[14]

It's entirely plausible that, in the near future, there will be a TikTok trend of standing in front of automated taxis and mooning them. It's hard to argue that it would be worse than most TikTok content. This presents a problem: how does an automated vehicle respond to being surrounded by an entire fraternity rush class with their pants down, and their phones out? This would even be a challenge for a human.

The lack of local human oversight is a huge problem for fully autonomous vehicles for delivery or long-haul trucking. Although the truck driver's primary job is to drive the truck, they also act as a custodian of the cargo and a deterrent against theft. Trucks often carry extremely valuable cargo. They also may encounter any number of issues on their journey. Trucks sometimes break down or get flat tires. They are required to stop at weigh stations and may need to take a detour. In these cases, reliable human support is not a luxury. Someone must deal with the logistics.

Self-driving vehicles may be stuck at being able to accomplish 99% of the task of driving. In that circumstance, the only way forward will likely be to modify the problem and make the system more accommodating. Fully autonomous vehicles will likely require a coordinated effort to modify the rules of the road, create protocols for unusual events and work with vehicle manufacturers, law enforcement, legislators, and insurance companies to make driving a task that can be remotely supervised and designed to fail safely when it does not work.

What does this tell us about the future? If these professions have adapted to successive waves of automation in the past, there's good reason to believe they can adapt again. Despite enormous reductions in the labor required for individual tasks, white-collar employment has expanded over the long term, and professionals now produce far more than previous generations could have imagined. None of this means the transition will be painless. Some jobs will disappear, others will change, and entirely new ones will emerge. The effects will be uneven, sometimes disruptive, and difficult to predict. What emerges is not a simple story of machines replacing people, but a complicated transformation of how professional work creates value.

Who Is Most Vulnerable to Employment Disruption

Technology has disrupted jobs before and will continue to do so. AI is one of those technologies. This may not be a universal doomsday, but as adoption accelerates, some workers are likely to face greater disruption. Current and past trends can help us understand which workers are most at risk and where the impact is most likely to appear.

Entry level work tends to be the first impacted by automation.[18] More senior workers, who have some level of tenure or have a managerial role, are more insulated against job loss. Layoffs can happen, but it's far more likely that the impact will be seen in a hardening of the labor market, with fewer entry level positions and more difficulty getting them. From the standpoint of business psychology, this makes perfect sense. Uncertainty tends to make business squeamish about hiring new talent.

Mass layoffs are themselves expensive. They often involve expensive severance packages and restructuring. They carry their own risks of loss. There are even examples of organizations that laid off large portions of the staff, only to hire the same people back at a higher pay rate. In other cases, organizations found themselves short-staffed in unexpected ways and had to use contractors.

If AI-related disruption follows previous labor market trends, as it appears to, then it's likely that early-career workers, those looking for work and those changing careers will be the most vulnerable. Another often forgotten demographic is freelancers. Freelance work does not have the benefit of being cushioned against loss by an employer. Freelance content creators and small businesses have fewer resources to cope with downturns. They also don't always qualify for unemployment or other benefits that regular employees do.

There's also the possibility that some jobs won't be eliminated entirely, but their profitability and viability will be reduced. This is especially true of freelancers. It may still be possible to get work, but less of it. It could reduce the going rate for content production. That's a very tough situation.



There has been a great deal of talk about policies to address AI-related job loss. Unfortunately, large scale layoffs get all the attention, because they are newsworthy events and easy to measure. However, the most vulnerable to labor market shifts are rarely the ones society thinks of helping.

CHAPTER 15

Perception of Artificial Intelligence

Artificial Intelligence is a disruptive technology. That's something that nearly everyone agrees with. The dramatic debut of generative AI and the extreme amount of attention it has gotten from the world have made AI a major topic of discussion. One of the biggest questions raised is what impacts and risks it has for society.

When generative AI broke into widespread public use in 2022, nearly everyone who saw it was startled. It was impossible not to notice that a computer was suddenly doing things that had only been seen in science fiction before. It was hard not to be a little bit uneasy at the implications of the technology.

The use of artificial intelligence can have some strange effects on people psychologically. This is especially true with natural language processing. It can produce an eerie feeling that leaves one unsure what kind of entity they are even dealing with. As society has gotten used to the use of AI models, it's become increasingly clear that there are some concerns for how the technology can impact a person's mental health.

One of the most unsettling problems that remains in the world of AI is the problem of misinformation on the nature of the technology. With most technologies, not understanding the inner workings has little effect on how people feel about the technology itself. That's much different with AI, because the mythology that AI models have minds and feelings remains a topic that is discussed openly as if it has credibility.

The most concerning problem is that credible institutions have put their names and resources behind spreading false information that is likely to reinforce delusions and make it much harder to combat false and distressing beliefs.

The first step to avoiding psychological harm by an AI model is to actually understand what that AI model really is and how it works.

The Psychology of NLP

Artificial intelligence can have a strange effect on people. Its ability to recreate human behavior with uncanny competence is startling. There's really never been a technology anything like natural language processing and we are just beginning to understand the psychological, social and developmental aspects of living in a world where machines can talk with all the subtlety and nuance of a person.

There has been relatively little research on the topic. At present, much of the data remains anecdotal. That's unfortunate. The individual incidents that have been recorded are enough to raise concern. There are a large number of plausible concerns that have been raised about the effects of millions of users interacting with chatbots. Some of these include increased social isolation, misplaced empathy, reinforcement of delusions, self-radicalization, potentially losing interest in human relations, validation addiction. It's also not hard to see how a person who is already suffering from a condition like schizophrenia could become confused and agitated by a chatbot. There's also concern about how chatbot interaction could impact the social development of children.

Before ChatGPT was released to the public, there was already an incident where an end user had great difficulty understanding the nature of the software they were using. The incident, which occurred in June 2022, involved Blake Lemoine, a software engineer at Google who had been testing their LaMDA chatbot, then only available internally. Lemoine had a background in software engineering, and was also an ordained minister. Lemoine was part of Google's responsible AI team. He was not involved in the technical development of the model. His role was to test and evaluate the model for bias, discrimination and potentially harmful content.

Lemoine became convinced that LaMDA was sentient and began expressing this belief internally. According to people involved at Google, several engineers attempted to explain to him, both informally and through more organized discussions, that the apparent personality and awareness were illusions created by the software. These conversations were generally cordial, but Lemoine remained unconvinced. The situation became increasingly difficult for Google to manage, and he was ultimately dismissed in July 2022.

Internally, the incident was regarded by some with frustration and by others with genuine regret. Many engineers rolled their eyes and simply said "well, he wasn't really a model designer." Managers shrugged it off as "bad optics…we hope that doesn't happen again." Others had more interesting responses. For example, one engineer stated "We have a lot of alternative thinkers around here, people who are anarcho-communists or crypto libertarians and stuff. A lot of them just want to burn the whole system to the ground. We really don't want to make anyone feel unwelcome…" [paraphrased]

… That's Google for you.

CHAPTER 16

Post-AI Cyber Security

Cyber security is a topic that has come up frequently in the discussion of artificial intelligence. Cyber security, like artificial intelligence, is a very poorly understood subject. It is something that, like AI, has created a great deal of banter from non-experts about doom scenarios. What makes cyber security different is that the lack of understanding is not related to any inherent complexity or unfamiliarity. Cyber security is one of the most straightforward pursuits. The confusion around it is completely artificial.

The primary way in which AI changes the cyber security threat landscape is that it enables malicious actors to automate attacks in new ways and it removes previous barriers to hacking. Many attackers are located in foreign countries, but AI removes the language barrier. Attackers may need to survey many systems to find a weakness, but AI can do that automatically. Finding bugs in software code is a task that requires expert human review, but AI can find bugs without needing an expert.

AI doesn't only speed up and automate cyber-attacks. It makes new forms of cyber attack possible. With AI it's possible to create more diverse code than could be generated with human programmers. AI allows for exploits to be discovered and taken advantage of in ways that can't be easily remedied. It can perform the kind of deep analysis that would be impractical with only humans.

One of the most important aspects of modern AI is its ability to process conversational, written and human communications. It does not only interface with technical systems and commands. AI can parse huge amounts of copy, looking for useful information about targets. It can be used to aid in open-source intelligence. It can also be used in direct social engineering attacks, creating phishing messages and even engaging in two-way communications with end users, completely on its own. It also complicates identity due to deep fakes and voice cloning.

There is another side to this. AI can also be used to enhance cyber security, and there are many AI cyber products on the market. Of course, part of this is branding, but there is an enormous potential to improve cyber security defenses, responses and evaluation.

CHAPTER 17

The Future

One of the great difficulties with predicting the future is that technological development and social progress are not the same thing. We can sometimes make reasonable predictions about what a technology might eventually be capable of doing, but predicting how society will actually use it is considerably more difficult. Technology doesn't develop in isolation. It interacts with economics, politics, culture, human behavior, and countless other factors that are themselves constantly changing. An invention that seems revolutionary today might be forgotten in a few decades, while something that initially appears unremarkable might transform civilization. History is full of both.

Technological changes change society in deeply fundamental ways. A pre-industrial society looks nothing like a modern society. Electricity, automobiles and the internet became familiar parts of our life, as AI is now becoming. Artificial intelligence is unique in that it always played a prominent role in how we envision the future. The idea of artificial intelligence—a fully automatic machine capable of human-like tasks—has existed for hundreds of years. It's very strange to have something that has been a fixture in movies and television come along, almost overnight. Paradoxically, it looks much like what was predicted but operates nothing like the AI of fiction.

Most science fiction is set in the future. It's a unique form of fiction, because it's often presented as speculation about what might come. The genre of science fiction has blurred lines. There is fiction and then there are speculative projections. It's not science fiction to say that we anticipate that GPUs will be significantly faster in a few years. Stories about warp cores are science fiction. Saying that genetic engineering will likely become a major part of healthcare in the next fifty years falls somewhere between. Speculative storytelling can be a valid way to explore how a technology might be used.

Futurism has also always had an appeal to those who like to consider themselves visionaries or great thinkers. People call themselves "futurists" and talk about the technologies they think will be important and why they think society will change in a given way. They often speak with great authority and conviction.

Author's Commentary

A few notes on why I wrote the chapters as I did.

Chapter 1 — AI: A Deeply Misunderstood Topic
I established the book's purpose and a recurring problem: even "intelligence" has no dependable single definition. I also addressed the AI fear movement. I would rather not have, but its misinformation obstructs understanding. I gave it only the space I thought necessary. The AI fear movement is not important and I find it an annoyance, but it must be addressed in a book about AI, when it is causing so much confusion.

Chapter 2 — Humanity's Quest to Outsource Cognition
Going back to early humanity seemed strange, but it explains our long effort to outsource thinking and provides a comfortable on-ramp for nontechnical readers. It makes modern AI feel less alien. It established the intellectual direction of the book. Hopefully nobody who wanted more technical explanations was lost here.

Chapter 3 — The Early History of Artificial Intelligence
This isn't history for its own sake. The early competing ideas—symbolic reasoning, automation, and learning systems—establish why AI developed as it did and why its definitions remain slippery.

Chapter 4 — Introduction to Neural Networks
Here I lay down the architecture behind the mental model used throughout the book. I deliberately separated what a neural network is from how it learns, even though the two could have shared a chapter. The chapter is long as is. I try to avoid complicated mathematics and focus on the fact that neural networks are actually very simple, in concept.

Chapter 5 — Deep Learning in Neural Networks
One of the hardest chapters. Learning is conceptually harder than the arithmetic. I used a simple toy example to make feature reuse intuitive: to me, it is the Rosetta Stone of neural networks. My fierce objection to "reward," "penalty," and "reward hacking" was intentional. These metaphors create false impressions of the machinery. Call the language hyperbolic if you like; I stand by the criticism, and I wanted readers never to forget it. Also, that was a little break from what is otherwise technically dense.

The phrasing we use is a huge source of confusion. Nothing about machine learning is human-like, but it uses many metaphoric terms.

Chapter 6 — The Modern Era of AI
Before arriving at the technology everyone recognizes, I needed to explain the computing ecosystem that made it possible: practical neural networks, faster hardware, GPUs, and the development tools built around them. Even before generative AI, artificial intelligence and machine learning underwent a renaissance from the 1990s into the 2010s. It's important to understand how we ended up where we are.

Chapter 7 — The Generative Revolution
How did a technology few people talked about in 2019 seem to be everywhere by 2023? I wanted to explain the developments behind that apparent overnight revolution, not pretend it appeared by magic. I was there, at Google Labs, and it was amazing.

Chapter 8 — Large Language Models
This was exceptionally difficult to make accurate and approachable. One planned chapter on language processing became three. I leaned on analogies and illustrations, always trying to give readers enough depth without trapping them in unnecessary detail. This is an especially hard subject because the logic of an LLM is so foreign yet the output is so familiar.

Chapter 9 — From Token Prediction to Long Term Coherence
Token-by-token prediction hardly sounds like a recipe for coherent conversation. This chapter brings training, representations, and earlier concepts together to explain why the simple description of next-token prediction is not a sufficient mental model. Very few books and resources explain this well or address the apparent paradox of long-term coherence from one token at a time prediction. I used examples, which I hope helped establish the principle.

Chapter 10 — NLP Beyond the Model
This unexpectedly became its own chapter. Memory, tools, interfaces, and external systems are vital to understanding AI as people actually experience it. The model and the application are not the same thing. I hope this chapter helped familiarize readers not only with the model, but with how it is used in chatbots and other familiar implementations.

Chapter 11 — Images, Audio and Video
This is deliberately abbreviated. I could have written chapters on diffusion, NeRF, Gaussian splatting, and video generation. That would have been another book. Here readers need to know what these technologies do and where they fit—not every detail of their mathematics. There are many types of media that can be generated by AI. This chapter was really just about what is out there and what it can do.

Chapter 12 — Artificial General Intelligence
I felt obligated to address AGI. I traced its history and the competing notions of "general" intelligence, minds, and personhood. As an ML practitioner, I find the argument largely definitional. Call it AGI if you want. I explained the dispute and was happy to move on. I also want to be clear that I am not against philosophy or anything like that, but it often feels like an obligation I didn't ask for. To be perfectly honest, I do not find the subject of artificial consciousness to be compelling or interesting and I see no reason why we should worry about it.

Chapter 13 — Current Limitations of AI
Another obligatory chapter, and not my favorite. I wanted to counter talk of unlimited scaling, million-fold superintelligence, and machines that can somehow do anything. AI is powerful. It is not magic; hardware, computation, and implementation impose limits. This was not really a response to the AI fear movement so much as it is to hype and to the fact that there does seem to be a myth that AI, as a technology, has no limits and can do anything. Most technologies eventually plateau in capabilities.

Chapter 14 — Employment Impacts of AI
I would gladly have omitted economics. But I want people comfortable living with AI, and nobody can relax if they believe it will make them permanently unemployable. I tried to find a sound forecasting method and found the future extraordinarily hard to predict. I criticized sensational studies and the long tradition of automation panic because frightening people about their livelihoods without good evidence is cruel. Losing a job is not the same as becoming permanently unemployable. I wanted readers to leave this chapter calmer, not terrified.

I am deeply disappointed and even angry at some of the institutions I encountered here. People really do experience anxiety about job security. It's a very common source of worry. It also impacts people's decisions. I could not hide how offended I am to see a respected university like Tufts engage in this kind of theater. I had a great deal more analysis, but the chapter is long as is. What bothers me is that a number of large institutions are feeding into this harmful message.

At this point the reader may begin to notice that my tone changes, at least at times. The reason for this is that the reader is now knowledgeable enough to see the absurdity of it all.

Chapter 15 — Perception of Artificial Intelligence

A chatbot is not a mind, does not have feelings and is not conscious. Understanding that is itself an important safety measure, especially for vulnerable users. I spent exhausting hours confronting claims promoted by well-resourced institutions. Much of this book is an effort to undo the problems that these false claims cause.

I'm most worried about people who have delusions about chatbots having feelings or being their friend. Not everyone understands this, and society needs to have a clear and unambiguous message that doesn't lead them down a harmful path by reinforcing delusions.

I am especially appalled by Jeff Sebo's model-welfare advocacy and efforts to turn speculative philosophy into professional standards. Being compelled through codes of conduct or employment rules to participate in practices I reject would be profoundly humiliating—an intolerable indignity. I have rarely been so offended in my professional life. I wrote the chapter because public understanding and professional integrity matter. I fully boycott Anthropic products. I think the company is insane. I would never recommend Anthropic for anything, certainly not an enterprise relationship. They really are a problem.

I also want to be clear on something: it offends me that people from a technical background are told that they must submit to philosophers on deep questions of life, thought and what it all means. Personally, I find the equating of software with conscious beings to be ethically repugnant. It's the kind of claim that should require extraordinary proof.

To be clear: this is not a matter of opinion. It's not conscious and nothing is under development that might be. Full stop. You can argue all you want about that, and some people will. Sebo may think he is smart, but he looks as intelligent as a Victorian doctor lecturing about the potential that locomotives may become sentient.

Chapter 16 — Post-AI Cyber Security

This is my profession, and the stakes are personal. Greater AI adoption means more connected systems, more attack surfaces, and greater need for genuine security. The elephant in the room is insurance. I say plainly that insurance companies are involved in serious crimes, and I consider cyber insurance the worst problem in cybersecurity. For more than a decade I have tried, unsuccessfully, to make executives and regulators listen. They are impossibly stubborn and usually refuse all contact.

As an American who lived through September 11, it is infuriating and heartbreaking to see vulnerable fellow citizens attacked—even in hospitals—by foreign aggressors. The cowardice of our own institutions and government makes it harder to defend them. Protecting people from criminals on the internet is my life's mission. We know how to do much better.

I have also faced unemployment because of the destruction to the profession I love that cyber insurance has caused. Many people went into the field because the message was that it had a talent gap and it was a great place to get a job. Because of cyber insurance, many great cyber professionals have seen their careers stalled or had trouble finding work.

There is another elephant in the room, but I did not have the space to include it. The software companies and IT sector, in general, have always had extremely poor quality control. The number of bugs in software is the result of shoddy workmanship and it always has been. Microsoft is one of the worst examples of a company with a terrible cyber security culture. I could write an entire book about how bad the practices are across the sector.

We must fix this problem. AI is only making it worse.

Chapter 17 — The Future

Forecasting technology is an excellent way to look foolish twenty years later. I was more interested in how AI becomes normal and how society learns to live with it. That required revisiting the AI fear movement and the sunk cost of publicly held beliefs. Stephen Fry's commitment to its mythology saddens me; I hope this book reaches others before they become similarly entrenched. I want readers to finish looking forward to an AI-enabled world without needless fear.

As far as fear goes, it's something I am loath to devote any time to at all. It is very annoying because these cranks are not owed anything, and their ideas are too stupid to require refutation. The problem is that it is part of the misinformation ecosystem.

I sincerely hope that future editions will have less and less space devoted to AI fear and other mythology. At present, it seems prominent enough, as a source of misinformation and a barrier to understanding, that it needs to be addressed.

I closed with a brief outline of what I consider the best path forward to a world where AI can automate many tasks and improve the productivity of workers: a four-day work week. I believe it is entirely reasonable, but it's also something that is unlikely to happen on its own. I have developed a strategy to get there, but that's for a future book.

What was left out:
This book was intended to be much shorter than it turned out to be. There were additional chapters that were omitted due to length.

Risk management – A whole chapter on how models and AI systems are tested and evaluated, and how risks are managed.

Social and Societal Effects – A whole chapter going past the perception chapter into how society is impacted by AI, how it may change the way we do things and so on.

Getting the Most out of AI – In this chapter I was going to explore how to put AI to use at work and in your personal life.

Going further with AI – this was to be a chapter on how to learn more, go to the next step, find careers and develop AI.

The Cutting Edge of AI – This was to be more of a future chapter on AI in which I'd explore the most cutting-edge research, problems and how they are being dealt with.

There was also to be a lot more. I wanted to look more deeply into the economics and provide a lot of case studies of AI being used and going very well or very poorly. I was also going to provide a deeper dive into LLM internals and mechanistic interpretability.

But this book is long as is.

Future books:
I enjoyed writing this book enormously and hope you enjoyed reading it. I will certainly write more. I plan at least one book on cyber security and another on AI risk management, a field clouded by nonsense from the AI Alignment movement and the pseudoscience it preaches. It's a fascinating study in cult dynamics. The working title is "The Doom Grift." Other possibilities include intellectual theater and pseudointellectualism, which I encountered repeatedly while writing this book, and dogs—how they think and see the world. Dogs are a great source of inspiration for non-human ways of approaching problems.

In a world of AI, generated content, carefully moderated statements and standardized responses, be a human. Your personality is the one thing you will always have that nobody else does.

Not all ideas should be debated. Some should be denied legitimacy.

Sacred cows make great burgers.

Call anyone anything you want. Because you can. They can't stop you.

Fear the man who fears the auditor. Never trust anyone who won't defend their statements and show their work.

Radical transparency is less dangerous than it might seem. Security rarely requires confidentiality. Sunlight is the best disinfectant. The best enforcer of standards is public inspection.

Albert Einstein stuck out his tongue at the camera, and nobody questioned his intellect for doing so. If you say Michael Phelps is bad at swimming, it won't bother him. Low achievers are offended if you get their title wrong.

Nobody is free of emotions or bias. There is no such thing as being perfectly rational. It's better to just admit this to yourself than pretend otherwise. Claiming intellectual superiority or perfection is how you end up in delusion.

Insane decisions and beliefs are rarely the result of mental illness. Many are the result of framing, creeping normality, lack of perspective and cultural reinforcement.

Don't apologize for being passionate about what matters, furious about injustice, sad about loss, emotionally invested in your work or positively giddy about discovery. Feeling is what being a human is all about.

If the opposition were motivated by evil, wrong about everything, acting in bad faith, and dishonest about their beliefs, things would be a lot easier. They rarely are. It's almost always more complicated than that, unfortunately.

Reality bats last. Mother Nature is merciless when doling out punishment. There are no do-overs.

Admitting you were wrong is less embarrassing than not admitting you were wrong. It just never feels that way.

Do not try or expect to avoid ever failing. Diversify your portfolio instead. Some of your life should be risk capital.

Risk management is not about avoiding risk. Sometimes it's about knowing what risks to accept.

Never underestimate what a single person can do. Do not assume you are too unimportant to change the world.

Most big problems are distributed amongst individuals. Helping even one is a victory. It's worth doing. Do it.

Doing something, fixing a problem or helping someone out, however small, is worth more than decades of talk, philosophy and debate about whether you should and how to go about doing it.

The longest journey begins with a single step. Many of the world's most powerful organizations did, in fact, start in somebody's garage. No raindrop believes it is responsible for the flood. Many hands make light work.

Nobody is more important nor worth more than you are. This cuts both ways.

Everybody, no matter how professional they are, no matter how much they are associated with a great institution or how official their title, is a human. They make stupid decisions in their personal life, get overly invested in ideas, worry about death, have strong opinions on sports teams, and sing ridiculous songs in the shower.

Do not underestimate how petty, small and personal many things really are. Some of the smartest people, greatest careers and empires have been completely destroyed by someone's inability to control their own sexual desires. It's the stupidest reason, but it happens constantly, and at all levels.

Nobody has it as together as they may seem. Everyone is doing the best they can and dealing with a lot of personal problems in life. That's universal.

A single experiment can invalidate the most beautifully constructed theory.

Don't try to be highly regarded. Don't seek status. Do the thing you are doing the way you see fit. Let the chips fall where they may. If the marketplace doesn't like it, then try something different next time.

Don't ignore critics, your opposition, and anyone trying to persuade you to take a different path. Don't automatically agree either, but don't ignore those who disagree. They are sometimes correct.

It is very easy to accept that propaganda, salesmanship and psychological manipulation work on other people.

Every individual person is entirely different, unique and a universe unto themselves. Most people and most of the experiences of human life are the same across people, places, time and cultures. This is not a contradiction.

Ships have sunk, aircraft have crashed and power plants have blown up because nobody broke ranks.

You don't owe anyone anything, but that does not mean you can't be generous. You don't have to, but you can.

2026 Core Technologies of Life

Familiarity and Understanding	Computers and Internet	Automobiles	Electricity	Artificial Intelligence
Comfortable using in everyday life. Example: *Shops online, streams media, drives regularly, and uses electricity at home and work.*	☐	☐	☐	☐
Uses naturally, without hesitation. Example: *Plugs in a device, goes online, or drives to a nearby destination without overthinking it.*	☐	☐	☐	☐
Understands its purpose and common applications. Example: *The internet connects people and services; cars provide transportation; electricity powers devices.*	☐	☐	☐	☐
Can explain how it works in plain language. Example: *A four-stroke engine converts fuel to motion; electricity uses positive and negative charges (AC or DC); a server stores and delivers data on the internet.*	☐	☐	☐	☐
Understands common terminology and components. Example: *Short circuit, transmission, Wi-Fi router, circuit breaker.*	☐	☐	☐	☐
Can recognize and resolve common problems. Example: *Restart a router, check a loose plug, or inspect a gas cap if a warning light appears.*	☐	☐	☐	☐
Knows safety precautions and responsible practices. Example: *Avoid online scams, wear a seat belt, and prevent electrical overloads.*	☐	☐	☐	☐
Recognizes when professional help is needed. Example: *Call a mechanic for major vehicle issues, an electrician for wiring work, or a technician for persistent computer problems.*	☐	☐	☐	☐
Understands legal and administrative requirements. Example: *Cars are registered and insured; electrical work must meet building code; internet access requires an ISP account.*	☐	☐	☐	☐
Can distinguish normal limits from serious failures. Example: *Resetting a circuit breaker is routine; driving on a flat tire is not.*	☐	☐	☐	☐
Can discuss it comfortably in ordinary conversation. Example: *"The internet's been slow today." "That's a great car."*	☐	☐	☐	☐
Understands jokes, analogies, and cultural references. Example: *"People can be wild on the internet." "I loved it like my first car."*	☐	☐	☐	☐
The technology is not inherently mysterious or intimidating. Example: *Shops for insurance, talks to a mechanic, uses online services, and isn't uneasy about electricity.*	☐	☐	☐	☐
Feels comfortable exploring new uses and learning more. Example: *Tries new apps and websites, is willing to drive a different vehicle, and experiments with new portable power options.*	☐	☐	☐	☐
Can make reasonable judgments about claims and risks. Example: *Can tell the difference between real benefits and hype, understand trade-offs, and make practical decisions.*	☐	☐	☐	☐
Knows where to learn more. Example: *IT and networking courses or certifications; automotive clubs and maintenance classes; electrical or electronics classes.*	☐	☐	☐	☐

Most people today can check all of the first three columns. That was not always the case. Many of us still remember when the internet seemed strange and exotic. Today everyone knows what a server is. Ideally, this book will bring you to the point where you can confidently check all the boxes. In the near future, everyone will.

NOW THAT YOU KNOW HOW THE SYSTEM ACTUALLY WORKS…

