Foreword

Greg Stuart, CEO MMA Global

Most believe we are in the age of AI, and just in time to make business and marketing (and potentially society) so much better. As CEO of a leading trade organization that includes tech giants such as Google, Meta, and Amazon, and major brands including AT&T, Uber, Unilever, and Verizon, and 800 more, I can attest that AI is of keen focus for all our members.

The reason I am recommending this book to our members is that it explains how AI works so it can be used profitably and responsibly. Most business books gloss over the "how it works," and I think this is a mistake. Without a deeper appreciation for how AI does what it does, we are bound to misapply it. I hope readers take the task of understanding how AI works to heart and spend time with Part 1, which covers the "how to" for businesses in a way they are unlikely to find elsewhere.

This is not the first technological revolution I've experienced—though I suspect it will be the most profound. I vividly recall my initial introduction to the Internet back in 1994. At the time, I was heading an "interactive group" nestled within Young & Rubican, the renowned ad agency based in New York City. We'd invited a consultant to explain this "WWW" thing. thing. We needed to understand how it worked. It was obvious to me right from the beginning that the Internet would be a big deal--and yet I completely underestimated what it has become It has redefined the whole world of commerce, content, community, and collaboration and laid the foundation for the next wave of change: the mobile revolution triggered by the launch of the iPhone in 2007 which put an internet-connected supercomputer in our pocket, with us 24/7. We should learn the lesson of underestimating the transformative power of the Internet and mobile and consider more deeply the implications of AI.

MMA Global, the CMO-led Industry body of which I am CEO, is conducting experiments using AI with prominent brands such as Kroger, ADT, and General Motors. We have seen significant gains in digital page visits—150 percent to 250 percent more visits than without the use of AI—and we are not alone in seeing these types of results. The potential of AI to revolutionize business and society extends beyond advertising. It encompasses social media, customer relationship management. the product experience, and the service delivery. All four Ps of marketing—product, price, place, and promotion—are poised for dramatic improvement with AI. If we use AI responsibly, it just might make customers a little happier or more satisfied in the process.

Building on the capabilities of the consumer internet, and mobile, AI has kicked off the next great wave of innovation. It is going to be exciting for sure.

Greg Stuart
CEO, MMA Global
Industry Association Non-Profit for Major CMOs Rearchitecting the Future of Marketing

Preface

Advances in artificial intelligence are accelerating so quickly, and its presence in our daily lives has become so pervasive, that it is becoming essential for people to develop a solid understanding of how it works. Take our 12-question self-assessment at www.AI-Conundrum.com to see how you score. The fact that you picked up this book suggests you are curious about AI and want to learn more. This book can help you level-up with a foundational understanding of artificial intelligence in a way that is easy to understand yet deep enough to apply AI effectively and responsibly. This volume is not for technical experts—we aren't teaching you to write code. Our approach is to explain the mathematical underpinnings of AI in a way that is accessible to business people, policy makers, and students to address the safety implications in a way that is concrete. The book is filled with real-world case examples.

Mathematicians call artificial intelligence a universal approximator because it has the ability to recognize a pattern from any set of data it is given, making it an extraordinarily powerful tool. But AI is not perfect and in the process of finding patterns, it can find spurious as well as authentic patterns. Biases and imprecision can creep into AI's output, and it can be very difficult to spot them. This can lead to disastrous results for businesses and government entities that rely on the power of artificial intelligence. Hence the AI conundrum: How can business leaders fully exploit the potential of AI in their businesses, when every step towards its full utilization seems to inherently amplify its safety risks? How do we identify the risks so we can effectively mitigate them? As a business, how do we use AI profitably and responsibly?

Understanding AI's strengths and weaknesses requires a basic comprehension of how AI works. Part 1 covers these fundamentals in an accessible manner to provide insight into the workings of AI and develops a framework to analyze the risks inherent in various types of AI applications.

This level of understanding will equip readers with the necessary foundations in order to have a grasp on why results can sometimes go badly awry.

Part 2 provides an abundance of examples that demonstrate the vast array of useful applications for artificial intelligence but also highlights the situations where AI goes wrong. The consequences may be trivial–but they may also cost lives and livelihoods. The risk framework helps to elucidate when AI's failures are relatively inconsequential, such as in marketing and language translations, versus situations where the risk of AI failures can be more serious, such as in certain types of commodities trading, self-driving vehicles, and facial recognition. We conclude with discussion of the risks of artificial intelligence and a discussion of why the industry should embed identity and accountability into AI and avoid anonymous and autonomous use.

Because advances in artificial intelligence are accelerating, we recognize that AI has the power to reshape our world, and any book can become dated quickly. Therefore, we have focused on a foundational explanation of how AI works and a framework for assessing risk that we hope readers will find practical and usable for the foreseeable future.

Introduction

Artificial Intelligence Acceleration

Over the past two decades, research, applications, and interest in artificial intelligence (AI) have exploded. AI's use has significant momentum and is accelerating. Consider these signposts:

- In October 2018, MIT announced a $1 billion commitment to address the global opportunities presented by the rise of AI—to "reorient MIT to bring the power of computing and AI to all fields of study at MIT, allowing the future of computing and AI to be shaped by insights from all other disciplines." Going forward, AI is meant to underpin every course in every subject at MIT.[1]
- In 2020, venture capital investment in AI was $28 billion, according to Protocol, a firm that tracks venture capital. In the first six months of 2021, $30 billion had already been invested, eclipsing the entire prior year's investment in half the time.[2]
- IDC, a technology tracking firm, measured $50.1 billion in global AI spending in 2020, and expected businesses to invest $110 billion in AI software and hardware by 2024.[3] A year later, in 2021, they reported spending had already eclipsed those estimates, growing to an eye-popping $341.8 billion. After their underestimation, IDC revised their forecasts for AI upward, expecting spending will exceed over half a trillion dollars in 2023.[4]
- In 2023, OpenAI, a company with less than $40 million in revenue, set a new start-up valuation record when they received a $10 billion investment from Microsoft, valuing the company at $29 billion.[5]

Artificial intelligence is expanding in part because the barriers to creating and applying AI are decreasing quickly and in part because of the massive profit potential in AI applications. Today, there are "no code" implementations of artificial intelligence that allow nontechnical developers to build their own AI applications. Extensive AI libraries allow software developers to plug AI into a variety of programs without needing to know the underlying math. Amazon, Google, Microsoft, OpenAI, and other major players have made it easy to run AI in their clouds with a credit card and a few clicks. OpenAI's ChatGPT crossed over one million users in less than a week following its public release. Thousands of AI start-ups offer products and services spanning the catalog from A to Z—from advertising optimization to zoo management; from autonomous long-haul trucks to zoonotic disease detection intended to identify the next pandemic before it spreads. Researchers are developing new AI advances every day, and businesses and governments are applying AI at breakneck speed.

However, AI has more weaknesses than most people realize. Given the expected growth in the application of AI and the torrent of AI hype that comes from well-funded companies with an economic interest in celebrating AI's successes, it is important that more people better understand AI's strengths and inherent weaknesses. AI is often anthropomorphized with a human-like ability to take in information and apply human-like reasoning, with human-like motivations to produce answers and insights. But the reality is quite different. In order to function effectively for a given purpose, artificial intelligence requires the acquisition of data and a pipeline to load many orders of magnitude more "experiences" than any human requires to learn similar patterns. Even after "learning" these patterns and producing the answer a human wants the AI to output, the AI may not have really learned the underlying meaning of its input. In many cases, it has developed an uncanny ability to imitate understanding without actually understanding the task at all. AI operates on a programmed reward system, which in and of itself can cause a host of alignment problems. AI can also suffer from bias, which comes from the data the artificial intelligence feeds on. Even if there isn't bias in the data, the mathematics of AI makes it easy for the AI to learn a spurious relationship that doesn't truly match what we are trying to teach the system. Yet, adoption is growing because AI can be useful and profitable in a wide range of domains. By the end of this book, we hope that you will hold two simultaneous impressions of AI:

- Awe: "It is amazing what AI can do."
- Caution: "It is amazing that AI doesn't know what it is doing or why it is doing it."

Ultimately, understanding the fundamental limitations of the current state of the art in artificial intelligence provides the reader with the advantage of a more comprehensive, nuanced, and sober understanding of AI. It allows us to look past the hype and determine the applications where AI can best add value while minimizing risk. In terms of strengths, AI's superpower is its ability to take in any dataset and find patterns that can be used for classification, prediction, or generation of new content. Computer scientists termed this automatic pattern-fitting "learning" and the underlying system of math "artificial intelligence." The downside of AI's learning superpower is that AI transforms data and fits patterns in ways that can produce outputs that aren't precise. AI can be easily fooled when operating in an open environment. AI doesn't offer rationale for the patterns it has learned. And, to learn the patterns in the first place requires massive amounts of data—some would say AI is so data-hungry that it is inefficient. This book will explain these and other limitations of artificial intelligence and suggest how researchers are attempting to expand AI toward an intelligence that is more robust. By the end of Part 1, you will understand how AI operates, and you will have an explanation for AI's weaknesses and an appreciation for its strengths. By the end of Part 2, you will be able to identify areas where AI is a great solution relative to other alternatives as well as areas where AI is risky and where countermeasures are required. You will understand why AI can produce biased output and the countermeasures that can offset the risk (at least in part).

Business decision makers and government policy makers will be well served by investing time to understand the AI risk framework to support responsible use. Some in the field of artificial intelligence today, especially those with an economic interest in AI, tend to gloss over AI's weaknesses. But no one can afford to wear blinders when it comes to artificial intelligence. We all need to study the limits of AI in order to help us avoid risky overestimation of AI now, while potentially building better AI for the future.

Before we explore AI's weaknesses, let's celebrate the strengths of AI. Travel back with the authors to 2017, when we had the privilege of attending MIT's Northern California Artificial Intelligence Conference.[6] One author was an outsider; the other was an insider, invited by the event's headline sponsor. We were at the event together and then, over the next few years, we worked on our separate AI efforts. Five years later, each of our distinct experiences came together to form this book. We will briefly describe each of our experiences with AI so you know where we are coming from and how our experiences influence our perspective in this book.

The Insider's View

As an insider, I had plenty of experience with AI over the past 25 years. In 1996 I applied neural networks to personalize the experience of a prominent website. In tracking the people who visited the website and the stories on which they clicked, the AI learned to predict the content that would lead people to read more, helping the company to sell more ads. I was fortunate that *WIRED* wrote about my work in their May 1998 Cover story entitled, The Promise of One To One (A Love Story).[7]

Over the years, I was keenly interested in applying AI to marketing and market research applications such as sussing out bias in population studies and automatically correcting for it. I founded a company to help marketers evolve from making decisions with their guts to using data and machine learning, and had some very talented and committed people join me in the effort. Together, we applied AI to generate automatic segmentations for personalized messaging and to attribute the influence of advertising to consumer behavior. The team applied AI to a wide range of business needs, including an AI that predicted how the Knicks could best fill Madison Square Garden seats depending on 32 factors such as web traffic, betting odds, social media buzz, day and time of the game, injuries, and more. The model could tell that the surest way to fill up The Garden was for the Knicks to win a lot more—but, alas, we couldn't solve that one for them. We could take the win/loss record as a given, and focus on how best to optimize marketing with our AI. Another artificial intelligence used data to determine how to adjust advertising spending for maximum impact on an auto manufacturers' profits. We had a brilliant graduate student intern develop an AI that could detect a brand logo while watching television to aid in attribution modeling for advertisers.

In my lab, in 2014, the team and I created an AI that could analyze media spending for the top 100 advertisers in the United States and identify opportunities to increase sales and profits by tens of millions of dollars. We anthropomorphized the AI by pairing it with a robot and named it MONICA. We hired a Burning Man costume designer to create a dress to cover the robot's frame. Atop the frame was a Microsoft Surface tablet that displayed a computer-generated female face. MONICA's lips moved in sync with its text-to-voice generated speech. Our goal was to shake up the advertising industry by having the AI roll up to prominent marketers at the Association of National Advertisers (ANA) conference with the question about why they weren't spending more in whatever area the AI identified as the furthest off from optimal. Jack Neff, from the trade magazine *Advertising Age*, wrote a story titled "Hey, ANA: Monica's Coming For You and She Knows Who You Are, What You Spend. Robot Roaming At Conference Will Promote Free 'Marketing Brain'."[8] The case study for my company's work with Warner Brothers on the film *Creed* featured the same virtual MONICA robot explaining how it boosted sales for the film (for reference, the video is available on the books website at www.AI-Conundrum.com). We were grateful for the recognition from The Entrepreneurs Organization for naming our work on marketing attribution as one of the top ten innovations, and to I-Com for their awards for the business impact we produced for our customers.

It was intriguing to see people react to our MONICA robot. At the ANA conference, during the main session, when Kraft's Chief Marketing Officer presented their strategy and opened the floor to questions, the MONICA robot rolled up to the microphone and observed, "Based on my analysis of your advertising spending data, you are not spending enough on digital media. You are underspent by 15.7 percent and that is costing your shareholder's value. You are especially under-investing in social media. Why is that?" It was a curious moment of a human expert having to justify actions to a robot. We hoped to inspire curiosity in AI and what we saw as a coming wave of innovations from artificial intelligence.

Inspiring people to look at data and technology a little differently led me to present a TED Talk on how AI can help humans shift from being knowledge workers to insights workers.[9] We trained AI to become aware of what a person in business is working on so that the AI could offer best-practices recommendations to the human partner. The team and I appreciated the validation that came from the financial backing of one the best early stage AI funds in Silicon Valley, Zetta Ventures. I personally appreciated that Zetta's founder invited me to attend the MIT AI Event, where they were the title sponsor, and to bring my son, Caleb, as he began to experiment with AI.

By the time I sold the controlling interest in my company in 2019, the patent office had issued us five patents for AI data analytics and knowledge management. After exiting the company, I volunteered my data science and technology experience to education, health care, and law enforcement agencies. When COVID struck, I was able to help on a local level with Immunize Nevada, and on a national level with Ad Council and the COVID Collaborative on the vaccination effort. I dove in and created a model that was among the most accurate at predicting hospitalizations and deaths 30 and 90 days out, as well as immunization rates, and the factors that influenced them. We open-sourced the model and shared it daily with our partners, and published it in Research World on a monthly basis. We used the model to help guide Ad Council public health advertising. We published research with our state public health lab, Harvard and Brown University. In these efforts I partnered with ArtsAI, an AI advertising technology firm, to implement AI-based communications—and will be forever grateful for their engagement. We estimate that our collective efforts saved about 3,500 lives and kept over 20,000 out of hospitals—we wish we could have done more.

Considering my experience, it is safe to say I am a strong proponent of AI. I appreciate that AI has power, but I also noticed some unsettling characteristics along the way. Some of AI's challenges were obvious, such as the need to invest lots of human time to label data to develop a controlled vocabulary for media classification, even though our human brains had no problem sorting out the meaning without labels. Or, the challenge of how processor-intensive (and expensive) AI is for simple tasks, like recognizing a company logo in a 30-second TV ad. Or, the fact we needed a human driver to help the MONICA AI robot maneuver, and the human driver to press "enter" when it was time for the AI to speak. The MONICA robot was more of a puppet controlled by a human than it appeared.

Other issues were less obvious, like the way bias could sneak into our datasets in ways that were hard to detect. Or, the way we might be simultaneously predicting and influencing someone's behavior with our segmentation and subsequent personalized advertising—and the way that it might polarize a population over time. For all AI's strengths, I had glimpses of weaknesses. It felt like seeing motion out of the corner of my eye, but by the time I turned to look, I couldn't quite see what had initially caught my attention. I couldn't quite define it yet, but I saw enough glimpses to know something wasn't quite right with AI.

In spring 2021, when I heard my co-author, Caleb, present his thesis paper, "The Fundamental Limitations of AI," he had put his finger on what had been gnawing at me. At the end of his presentation, I turned to my friend, a senior executive from Google who was also interested in Caleb's thesis paper, and we both had the same reaction: "Wow."

Caleb was explaining the shortcomings we were seeing in AI inside our companies, but hadn't yet heard these shortcomings articulated or explained. For all of AI's advantages (and there are many), there are specific areas where AI weaknesses need to be well understood and offset with countermeasures.

The Outsider's View

As an outsider in 2017, the field looked promising. I recently learned to program Lisp, designed in 1958 by John McCarthy of MIT. It is heavy on mathematical notation and became a favored AI language in artificial intelligence's early development. I had tried my hand at creating AI from scratch and started voraciously reading research papers. I was excited to attend the MIT AI conference and absorb the content. A morning session entitled "Ghost in the Machine" discussed AI in vehicles and promised fully autonomous cars in the near future. Venture capitalist Mark Gorenberg gave a compelling explanation of the AI virtuous loop, in which ingesting more data generates more value, which in turn generates more data, thus expanding the value of the AI decisions. Each presentation noted the fast pace of progress and projected an encouraging future.

I read about how artificial intelligence has achieved impressive results in recent decades, which demonstrate AI can exceed human capabilities. AI achieved superhuman levels in chess—laying waste to every human competitor. In the far more complex game "Go," DeepMind's AlphaGo AI beat the 18-time world champion. The AI won four games out of five. An AI called AlphaFold2 managed to achieve success on the notoriously difficult problem of protein folding. It created the first complete database of the known protein universe, covering 200 million proteins from one million species. Human researchers had managed to cover only a fraction of that number, and primarily only for humans, mice and a few other mammalian species.

At the same time, AI has captured the interest of more and more people—Google searches for "AI" have gradually increased since 2004, accelerating even faster since 2016. In the press, AI dazzled reporters. I calculated the number of stories in the New York Times mentioning AI more than doubled in the past five years compared to the prior five years. A natural language processing engine, GPT-2, was even interviewed by *The Economist*[10] and featured in the *New Yorker*.[11] AlphaZero, a chess-playing AI application, was described in this way by Cornell math professor Steven Strogatz:

> Most unnerving was that AlphaZero seemed to express insight. It played like no computer ever has, intuitively and beautifully, with a romantic, attacking style. It played gambits and took risks….Grandmasters had never seen anything like it. AlphaZero had the finesse of a virtuoso and the power of a machine. It was humankind's first glimpse of an awesome new kind of intelligence.[12]

As a fan of Dr. Strogatz's videos that dive deeply into math methods, I wanted to delve into what he learned about the inner workings of AI. My imagination was engaged. I set out to expand my understanding of the underlying math of neural networks so I could build more sophisticated AI from scratch, without the use of AI libraries.

To augment what I learned in biology class, I created genetic algorithms to better understand the evolution of single-cell organisms—I was intrigued by how well the simple goal-seeking AI rules led to artificial life forms with many similarities to our real-world single-cell organisms. I created self-driving bots to navigate randomly generated racing tracks. I was impressed by how the AI utilized an oversight in my code that allowed for negative inputs to "invent" a way to reverse itself when it worked itself into a corner—an outcome I hadn't expected. I built computer vision AI for a robot so the robot could see and grab the right blocks and move them to the right place in the physical world—it worked remarkably well in the predictable confines of a garage. I created a kill bot that uses real-time object detection to align the scope on an enemy target in a first-person shooter game—it killed the other team more efficiently than I ever could. I created AI art—including the design on the table of contents of this book. I created a language AI that scraped Quora for questions, learned from the popularity scores, and developed its own questions—it received over half a million views from humans.

As I worked with math and code, I began to appreciate the strengths and weaknesses of AI. While the AIs that I built could drive, ask questions, produce art, kill, and produce generations of artificial life, I uncovered many weaknesses in the process. As I read more news about AI, I began to worry that there wasn't enough understanding of AI's weaknesses.

Consider the implications from two of my projects. One of my projects emerged when I was invited to the Quora partner program, which allowed users to post questions and get paid based on the ad revenue the views generated. I instantly saw an opportunity: What if I created an AI that could generate the questions automatically?

I wrote a bot that perused Quora, looked for popular questions, and added them to a database. This fed into the training of an AI built to produce new questions that my AI predicted would be popular. In a few weeks, I had amassed over a half million views on "my" questions. It had fooled people—almost. Upon close (and sometimes not-so-close) inspection, it was possible to find the weakness in the AI. The questions often sounded right, but sometimes were absurd. For example: "How should I prepare my two-year old for the SAT?"

At an abstract level, this is a good question—questions about educating kids and getting them into college are highly engaging on Quora. The AI knew the right concepts to ask about, but didn't understand what it meant to be a two-year old; nor did my AI realize the inappropriateness of preparing a two-year-old for the SAT.

The insight I gained from AI's generation of absurd questions is that the meaning in our words is derived from our broader human experience, which AI doesn't yet possess. Like many AI applications, mine was fed only text. My AI lacked the multitude of senses that humans possess that allow us to find meaning in our world. The meaning in our words is deeper than our deepest neural networks can go today.

A second AI project to consider is the AI kill bot. A few of my friends introduced me to the first-person shooter game Counter-Strike: Global Offensive (CS:GO). It is a military-style game that pits two teams, terrorists and counter terrorists, against each other and arms them with weapons like the AK-47 and M16. In one mode, the terrorists attempt to plant a bomb and the counter terrorists attempt to take out the terrorists. Unfortunately, no matter how hard I tried, I was terrible at the game—I was down in the bottom 10 percent of players. Practice barely helped, so I decided that the best way to improve would be to teach an AI to take over the aiming and shooting decisions. I meticulously gathered and labeled thousands of images of the characters in the game and trained the AI to recognize them. I found a way to hijack control of my mouse and slow it down so the AI had human-like speed. It was important that my AI appear human-like in game play because the creators of the game, Valve, have code that attempts to find players using hacks to gain an advantage in the game. The end result: my AI vaulted me to the top 3 percent of players—my AI was a killing machine. However, there were still glimpses of something fundamentally wrong with the AI.

My AI would register a particular oddly shaped small plant it saw with nearly 100 percent certainty that it was an enemy. The AI would fire relentlessly at the offending foliage. A certain lamp it saw on one map was also reliably classified as an enemy. It's not that we can't fix the AI's mistakes in these cases—it's that these mistakes are a tell that the AI doesn't really understand what an enemy, lamp, or plant *is*. AI was fed only images, and lacked the broader understanding of what the objects represent and the context for why it was aiming and firing.

As I considered what I built and AI's weaknesses, I wondered, "What if modern warfare used an AI for targeting and shooting? Might it also occasionally misclassify and destroy civilians?" Without understanding the context, the AI afforded a significant survival and performance advantage—I moved from bottom 10 percent to top 3 percent with the assistance of AI. But the AI doesn't understand the context of war, and if this were applied in the physical world, it could make terrible mistakes. Would society accept the trade-off in performance versus survival as worth the collateral damage? Humans are capable of collateral damage too. How many hunters have mistaken another hunter for a deer? Should we hold AI to a standard of perfection? Or, is a standard of"merely more efficient than humans good enough?

My two projects, with Quora and with the kill bot, are in many ways a microcosm of the weaknesses of AI, which we will cover in Part 1 of this book. The weaknesses raise important questions we need to carefully consider.

Our Perspective

Artificial intelligence is a powerful tool, and it often does a great job at the tasks we ask of it. However, it has also failed to deliver on many of its promises. Its use in image identification in the legal system has led to false arrests. The release of unlimited autonomous cars is continuously delayed due to a range of shortcomings. AI can even fail miserably at playing a video game it has perfected in the past if just a few extra pixels are added to the input—a change that would not affect a human's game play for more than a second. The bottom line is that many AI systems show alarming failures and some could cause serious harm.

To use AI effectively and responsibly, it is important to understand both the strengths and weaknesses. Understanding the limitations of AI today may help us develop better AI for tomorrow and beyond. Accordingly, we have organized the book as follows:

- Chapter 1: Rather than anthropomorphizing artificial intelligence and assuming AI has human-like characteristics, we ask the reader to consider what it is like to be an AI. Artificial intelligence isn't given the context or shared experiences humans often take for granted, which hampers its abilities in many contexts.
- Chapter 2: Here, we explain how AI fits patterns and why AI's structure introduces some fundamental weaknesses that prevent AI from gaining a robust understanding of its inputs.
- Chapter 3: This chapter discusses how AI uses gradient descent and why AI is prone to taking shortcuts to provide the answers we seek. The shortcuts are essentially unfounded correlations that can result in a host of mistakes.
- Chapter 4: AI generates what we would call *associative intelligence*. It can produce impressive output, but it takes massive amounts of data to make the associations, and even after consuming copious amounts of data and compressing it, and finding patterns, the AI may still be found lacking in certain use cases, such as math and engineering.
- Chapter 5: AI has many strengths but also key weaknesses that need to be well understood. This chapter answers why AI struggles with precision, can fail in open environments, and can't easily provide a rationale for its decisions.
- Chapter 6: We offer a framework for analyzing risk that is based on three criteria: the need for precision, amount of control over the inputs the AI uses to operate, and whether a

rationale is required for the system's decisions. This chapter also explains an approach to compare AI to the next best alternative to achieve a business' purposes. And, finally, we conclude with an explanation of why the availability of data, economics, and competitive pressures will likely drive adoption of AI even when its application is risky.

- Chapters 7 through 9: We present specific case examples of AI's strengths and weaknesses in business applications such as language translation, autonomous vehicles and automated trading. To address AI's weaknesses, we suggest countermeasures that can reduce risk.
- Chapter 10: We take a deep dive into AI bias that is inherent in many systems, and suggest approaches to access and reduce bias.
- Chapter 11: We conclude by recommending approaches to training, governance, and accountability as well as a service architecture approach to human workflow that will make applying ever improving AI easier. We suggest approaches to safety, including the use of authenticated identity for AI that has access to the Internet.

The book's website includes the following resources:

- Risk Analysis Checklist: We offer a set of 25 questions that business decision makers should consider as part of AI planning and governance on the book's website. This worksheet can facilitate analyzing risk. For those applying AI in riskier scenarios, the worksheet offers countermeasures that may offset AI's weaknesses, or, that may lead a decision maker to refrain from using AI because the risks are too high.
- Why AI polarizes in social media. The Social Dilemma, a documentary on Netflix, offered a broad explanation of the polarizing effects of social media algorithms, but a deeper understanding of AI and how it interacts with the economic model in media provides a more complete explanation. A discussion of possible solutions is also included.

- A deep dive into AI bias, how to analyze if an AI system is producing biased output, and various strategies to mitigate bias.
- Videos, interactive tools and labs to get hands on with AI.
- Blog of updates to the state of the science in AI and trends we find noteworthy.

In this book we consider, for the most part, the artificial intelligence of today, which is mostly characterized by a single input type, such as text or images. The AI of tomorrow will broaden the range of inputs and introduce new mathematical structures that may lead to a new generation of AI but some of the weaknesses will likely linger. For purposes of readers today, and given the current state of artificial intelligence, we focus on current machine learning techniques such as gradient descent, reinforcement learning, and matrix scaling as well as structures such as neural networks, transformers, and generative adversarial networks. We will touch on the different outcomes AI produces, such as classification, clustering, and regression as well as different techniques such as supervised, unsupervised, and few-shot learning. Our goal is to demystify AI and give you a foundation for understanding AI conceptually.

We look forward to seeing if the challenges described in this book can be overcome in the future. Until then, our hope is this book will be useful to a range of audiences including business decision makers, non-STEM students, and government officials to allow them to:

- Better recognize AI's weaknesses and strengths.
- Ask better questions of their advisors and agents about the AI developed by them or used by their business.
- Aid in evaluating risks and rewards in a business's or government's application of AI throughout their organization.
- Ensure effective and ethical use of AI.

- Avoid risky applications of AI—or at least build in countermeasures to offset the risk.
- Demystify what AI is for everyday citizens and journalists so we can communicate AI's strengths and weaknesses and appreciate why AI can sometimes fail.

Part 1

The Fundamentals of Artificial Intelligence

Part 1 of this book explains the ways in which artificial intelligence learns to associate inputs with desired classifications and outputs. Mathematicians call AI a universal approximator because AI has the ability to fit a pattern to any set of data—even when the data is purely random and there is no true pattern. Universal approximation is both AI's superpower and AI's kryptonite. It's a superpower because AI can be applied broadly, in every domain we can think of. Universal approximation is also AI's kryptonite for several reasons. First, in the process of finding patterns, today's AI can find spurious patterns. Second, compression, which is inherent in AI, and the math of pattern fitting results in AI that lacks precision. In some applications precision is essential—and in these domains, AI poses a meaningful risk. Third, many applications of today's AI rely on data that is sourced from open environments, meaning the AI can be fooled by bad actors intent on tricking the AI. Even when the data is gathered in a closed environment, bias can creep into the AI's output. And finally, AI doesn't provide a rationale for its decisions, making it harder to spot bias, the influence of a bad actor, or the effects of spurious correlations.

Despite these drawbacks, today's AI can be applied in a vast range of situations with pretty good (and sometimes amazing) results, and can generally do so at a lower cost and faster speed than current approaches to addressing the same use cases. Given AI's power, we expect it to continue to expand at a rapid pace. In general, expanding the use of AI will be beneficial, but in some instances, this expansion of AI can be a dangerous formula where the profit motive and hype will lead some users to overlook the shortcomings—sometimes to a disastrous end.

Our objective in Part 1 of this book is for those involved in the technical aspects of AI to internalize an awareness of the weaknesses of AI so they are judicious in applying AI. At the same time, we hope that those in business and government that approve AI applications will read Part 1 to gain a better understanding of the limitations and to learn the critical questions they need to ask in order to ensure adequate governance and countermeasures are in place when using AI.

Chapter 1

Artificial Intelligence Is Not Human Intelligence

At one end of the Las Vegas strip is a famous magic show featuring Penn & Teller. At the other end, in a hotel room near the sprawling Consumer Electronics Show (CES), sits a conversational artificial intelligence attached to a robot modeled after Philip K. Dick, the science fiction writer whose works inspired the dystopian thrillers *Blade Runner, The Minority Report, Total Recall,* and *The Adjustment Bureau.* The artificial intelligence, which feeds on a catalog of Dick's science fiction writings, is interviewed by a parade of reporters as if it were a human. Penn & Teller and Hanson are both performing magic tricks of sorts.

The magician relies on psychological illusions and sleight of hand to exploit gaps in our conscious experience. A card is held in the hand of the magician. A moment of misdirection and a gesture waves it away. Then it reappears as if from thin air in an entirely unexpected place. The magician exploits the limits of our human perception system and our tendency to project our understanding of reality onto what we see, even as they misdirect us from seeing what is really happening.[13] The best magicians tell a story, and that increases our tendency to project our expectations, so that they can then surprise us with the unexpected twist.

Hollywood's sci-fi films are, perhaps, most closely related to magic—the stories of AI robots project our human characteristics onto technology so that directors can probe the human condition. Philip K. Dick's writing, which explores our fears about the unknown, animates the genre. So too does a promethean fear that our own creations may burn us alive. When the Philip K. Dick AI robot is questioned in a 2013 interview with PBS, it provides a chilling answer to a question some may have on their minds: "Will robots take over the world, Terminator-style?" The answer: "You all have the big questions cooking today. But you're my friend, and I'll remember my friends, and I'll be good to you. So don't worry, even if I evolve into Terminator, I'll still be nice to you. I'll keep you warm and safe in my people zoo, where I can watch you for old times sake."[14]

The reaction to this interview of artificial intelligence and the "people zoo" response went viral. But the response is a magic trick. Others working in and around AI use the trick as well. To capture attention, YouTubers feed AI responses into text-to-speech programs and match it with a voice and face so that AI-generated text can be delivered as if from a person. Corporations build robots to look like humans and connect them to AI to sell their capabilities. But portraying AI as human-like is misdirection. It creates the perfect tension to sell a story.

The story most are selling is that artificial intelligence is more powerful today than it really is. If a person interacting with the AI imagines a human-like counterpart, they make a lot of assumptions that make the AI seem more impressive. It is a common Hollywood trope to project human personality onto AI. But, when we anthropomorphize AI, we fall for the magician's misdirection. We fail to see and understand how artificial intelligence really works. Just as we should not expect to learn physics from watching the levitation act in a magician's show, we should guard against the misdirection of anthropomorphizing artificial intelligence and assuming it works the same way as human intelligence. It doesn't.

We are going to flip the Hollywood script. Instead of imagining artificial intelligence as human-like, we are going to ask you to put yourself in place of the AI so you can appreciate AI's capabilities and limitations.

A World Without Context

Imagine you have to perform the role of an AI in a computer. You have to think like a computer and do the jobs we ask AI to do every day. You are put in a box and you are shipped off to some different universe which has its own set of laws and physics. Maybe you were sent to a universe (call it U) that is somewhat similar to our own universe and also has cats and dogs. Your job is to distinguish between cats and dogs and anything that's neither. There are buttons you can push to signal your conclusion. The first picture you are given looks like a tree. You indicate there is not a cat or dog in the picture. But you're given back a response that you are wrong—there was a cat in the photo. Then, another picture contains an object that is entirely the color red, and you answer that there isn't a cat or dog in that one either—but again you're wrong. You are given the feedback that there was a dog in the picture.

It could be the case that in universe U, cats and dogs can shift into other things. Maybe cats can shift into anything that is the color green, and dogs can shift into things that are red. You find that your human experience can't help you in this universe—you are unable to process what a "cat" and "dog" looks like in this universe based on the limited examples you've been given so far, yet to a native of universe U, the color-shifting properties of cats and dogs would be obvious, and they would likely be perplexed as to why you were unable to recognize the obvious ability of cats and dogs to transform.

Even this example is too simplistic to capture what it is like to be an artificial intelligence. Although it is unusual for animals to metamorphosize, it is something we, as humans, can understand based on our experiences with caterpillars and other biological species that transform. We know chameleons and cephalopods can change colors. We might begin to improve our recognition of images by linking our experience with that of this new experience in universe U, but that isn't how AI works.

A truer picture of what it might be like to be an artificial intelligence in a computer would be to enter a world that is completely incomprehensible. Maybe everything exists in five-dimensional space, and it works off of hyperbolic geometry. If you move in a circle, instead of returning to your original position, you are now somewhere else. Maybe movement itself doesn't exist, or maybe cause-and-effect only sometimes takes place. Perhaps the laws of physics are constantly changing, and 10 + 10 doesn't equal 20 because addition itself works in a different way than it does in our universe. To become further immersed in a computer's perspective, assume that rather than being able to experience this world, you are given information about the world through a manuscript, written as a series of symbols unlike any human language you've ever encountered.[15] You need to look for patterns in the manuscript in order to return the correct answers in this new universe.

Any chance of using reasoning is ill-fated; statistical methods of pattern recognition are your only means of processing information to generate an answer. Much like a computer that starts tabula rasa, whose knowledge must be built from data-intensive exploration of this world to find patterns, in this thought experiment you too have to slowly proceed forward with absolutely nothing taken for granted. It is a world without context. With no foundational knowledge to draw upon, you need a massive number of examples before you can accurately answer even the most basic questions, such as, "Is this a horse?"

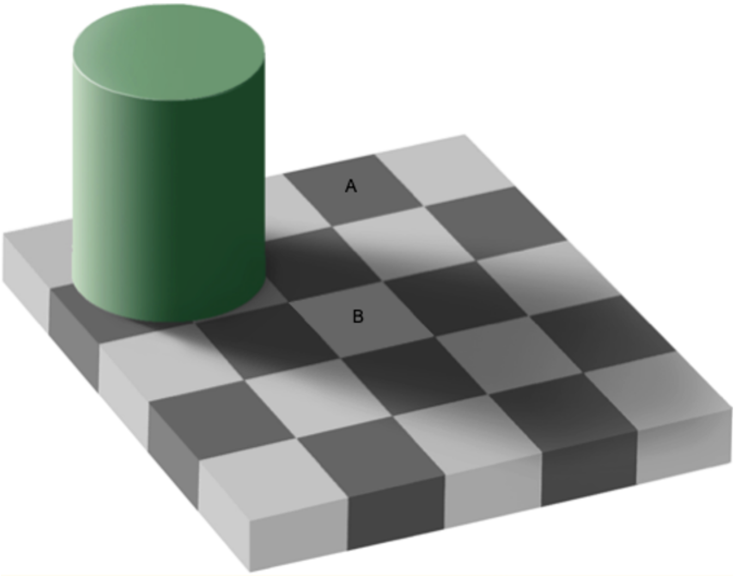
When asked, "Is this a horse?" you are given the picture—but an AI can't see pictures the way humans do. AIs work in numbers so the pixels of the image are converted into numbers which form a big list called a matrix. When you are given this matrix, you have no idea the numbers actually represent an image, you simply see a matrix with tens of thousands of numbers and are required to give an answer. In fact, you aren't even directly given the question "Is this a horse?" AI can only understand numbers, so the question needs to be converted into a number. To do this, the images that contain no animals are labeled with '0', the images that contain a horse are labeled '1', the images that contain a cat are labeled '2', etc. You look carefully over the numbers you're given. You try to look for a pattern that is similar to the patterns in the other sets of numbers that the operator of the computer has labeled as '1'. Finding similarities, you press the button, "1" . You earn a reward for your correct answer. You now pay extra attention to the pattern that helped you arrive at the right answer. To train to become a better AI, you repeat this process over and over again; you receive a big matrix of numbers, you press a button to indicate your response, and then update the patterns you pay attention to depending on whether or not you were correct. This is what it is like to be an AI.

When we anthropomorphize AI, we tend to overestimate what AI can do by believing it has the same contextual experience of a human while missing how hard it is to do what it does. In the example of image recognition, we are essentially feeding AI a matrix of numbers with a certain number of examples labeled as containing a horse, for example, and other matrices of numbers labeled as not containing a horse—and telling the AI to figure out the difference. It is remarkable what AI can accomplish, considering these limitations. Yet, there is a crucial observation one must accept about the state of artificial intelligence: Even an intelligent machine cannot possibly understand the world around it in as full a fashion as a human does with the information the machine is given. Lacking human experience and reasoning, a machine must rely on pattern recognition and correlation. For artificial intelligence, there is often no context beyond the specific data set on which it is trained. Without context, there is no meaning.

Visual Understanding

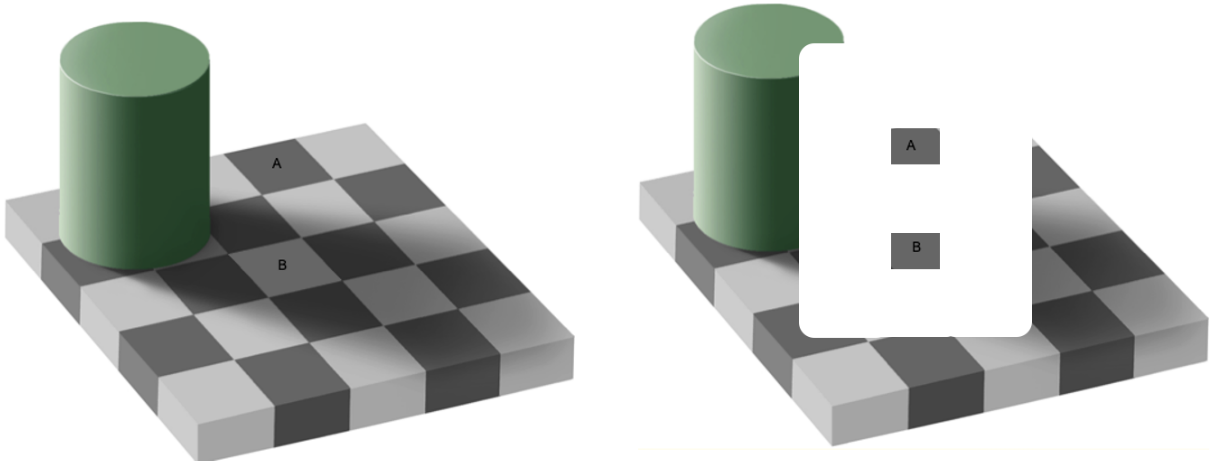
It is incredibly difficult for humans to comprehend the limitations of an AI since a great deal of the experience we have as humans is easily taken for granted. Our intellect is fundamentally tied to our experience in our universe in ways we often don't realize. Our cognitive structures reflect those realities of our physical world. The optical illusion in Figure 1.1, from MIT vision researcher Edward Adelson, shows how our minds interpret the world around us. Are Tile A and Tile B different shades?

Figure 1.1: Adelson's Checker Optical Illusion



They certainly appear to be different colors. But look again, at Figure 1.2

Figure 1.2: Adelson's Checker Optical Illusion With Some Context Removed



When we look at the exact same image on the right with the subtraction of some reference context, we can see A and B are indeed the same color. How did your brain manage to make them look different? To start, your brain perceives the two-dimensional image as representing a scene in three dimensions. It uses the context that the image represents three-dimensional space with other context clues to infer the direction the light is coming from, and deduces that the cylinder is blocking some of that light. Your brain knows that, when some light is blocked, your eyes receive less light from that object, so it will appear darker. Thus, your brain cancels out the effect of this shadow to see the real, underlying color. This is why an object that looks white is still white even under a shadow, despite the fact your eyes receive less color. Hence, when you look at the image, Tile B appears to be a lighter shade than Tile A—your brain is making it brighter to counteract the effect of the light. That is, when you look at this illusion, your brain immediately applies context to interpret it.

Imagine asking a computer "After we remove the cylinder, which tile would be brighter, A or B?" It would have a very difficult time answering, because, at a pixel level, A and B are exactly the same color. To a computer, A and B "look" the same. Computers don't have the context we humans gain from lived experience. (You can play with the online interactive version of this illusion at www.AI-Conundrum.com.) Take a moment to appreciate how our human brain uses context to discern information.

We interpret meaning in images based on our physical experiences with objects, light, and shadows. This is only a single example of a plethora of other effects that our human brains consider automatically when we see an object, a still image or video. Just inferring that something is a shadow relies on an enormous amount of context, such as the physical context of the way light creates shadows and illuminates an object, or how objects are distorted when translated from a three-dimensional to a two-dimensional image. None of this information is readily available to today's artificial intelligence when it analyzes a still image.

AI instead receives only a matrix of numbers with no context. It may not know the precise way in which light bounces, nor may it know that the image is a two-dimensional representation of something that the human brain understands as three-dimensional. This lack of context has profound implications.

Understanding Language

Artificial intelligence has made astounding gains in the ability to process language and answer prompts with convincing conversational replies. But, if you look closely enough, you will see the limitations. To illuminate the limitations, imagine that two Earths exist. Both are equivalent in all ways except that what we call "water" is made up of a different chemical composition on each Earth. On Earth A, water is made up of H_20; on Earth B, it is made up of some other formula—let's call it XYZ. The mental states of the people on Earth A and Earth B are exactly equivalent, and so they both have the exact same conception of water. If, like a computer, you could examine their code (the mental state), and you took their minds out of the context of their environment, you would find they are exactly the same. It would not be possible to determine the difference between them—yet, despite the mental state of the counterparts on both planets being the same, the word for water refers to different things. For person A, water refers to H_20, and for person B, it refers to XYZ. Therefore, some crucial information of what is meant by water is held outside of the words and is external to their mental state.

The philosophical view that words' meanings are determined externally from the words themselves is called semantic externalism. It was developed by philosopher Hilary Putnam in the 1970s, along with the thought experiment above, which is known as the Twin Earth argument.[16] It has important implications to computers and AI because it argues that AI may not be able to access the meaning that is external to its systems. Semantic externalism implies some of our understanding of language requires shared experience—for example, when we say that water is wet, usually we are conveying a particular part of our experience with water (e.g,. a certain way it feels), rather than making a statement about the water's underlying chemical composition. It may be that as more and more language is fed into the AI, it will pick up on more of these associations, but it is also possible that certain associations will not be noted in the body of words the artificial intelligence has consumed—it may be that some associations can only be gained through experience. Researcher Paul Schweizer of the University of Edinburgh explains that without our shared experience, artificial intelligence is not capable of fully understanding the meaning of the words and images we present it. AI will use the best statistical analysis available to produce an answer, but the answer will lack the context we have as humans living a shared experience and developing shared meaning.

Consider something as simple as the phrase "We will cross that bridge when we come to it." Researcher Ari Holtzman and his colleagues argue that this phrase subtly reflects our understanding that causality is dependent on distance in three-dimensional space and time—distant objects (the bridges) are less likely to affect us than something immediately in front of us. We attend to the things right in front of us and we can leave distant things for later. In addition, implicit in the statement is our experience that a bridge is built over some type of obstacle. This phrase means we may have to confront some type of metaphorical chasm that we can deal with at some future point in time. An AI that has consumed explanations of this phrase may be able to offer an explanation, but a genuine understanding of this phrase requires comprehension of not just the words, but the shared experiences it references.

AI researchers have assembled what are called *large language models*. They can include billions of words drawn from the internet and books. On the surface, they do a good job of explaining the meaning of challenging phrases like the one mentioned above because the models can draw on the content of humans explaining the phrase's meaning. Even when there aren't explanations in the corpus of content the AI has consumed, there may be enough associations for the AI to construct a plausible answer. But ingesting massive amounts of content from the web is not the same as developing context from diverse experiences. An AI is given none of this experience or background to truly understand the phrases. Artificial intelligence can get extensive training on words and language but it isn't able to experience the causality of existing in three-dimensional space and moving through time—both of which are crucial to a real grasp of what it means to see a bridge in the distance as one approaches it. Nor does AI have the opportunity to cross the bridge and look down to *see* the rivers or canyons that bridges are built over. It may seem the AI understands the phrase because others have described the meaning of the phrase in the content the AI ingested. However its understanding is very limited because that phrase relies on a reference to an experience that an AI doesn't have.

In fact, even the state-of-the-art AI large language models such as GPT-3, developed by OpenAI, still do barely better than random chance on certain subsets of questions designed to measure an understanding of context. For example, when asked novel questions requiring physical context, such as, "If I put cheese into the fridge, will it melt?" not all AI models were able to get the right answer.[17] Some are not able to correctly answer this simple question: "Imagine an empty cup. Now put a coin in the cup. Now take the coin out of the cup. How many coins are in the cup now?"[18] In our testing, ChatGPT did a much better job than its recent predecessor GPT-3 on these questions, but it still made some mistakes. AI can struggle to capture the relationship between concepts. In our testing, ChatGPT got the relationship between marketing spending and customer lifetime value backwards. We asked ChatGPT, "What are the benchmarks for successful Software as a Service (Saas) companies in terms of growth and investment in sales and marketing?" ChatGPT offered some standard benchmarks. The generally accepted relationship is that the expected customer lifetime value should be three times the cost of marketing to acquire a customer. However, ChatGPT got it backwards when it answered "Customer Acquisition Cost (CAC): A SaaS company should aim to keep its CAC as low as possible, ideally below 3x its Lifetime Value (LTV). This means that the company should be able to acquire a new customer for no more than 3 times the revenue that the customer is expected to generate over their lifetime."

Spending three times a customer's total lifetime value is wrong. It would quickly bankrupt a company if a business person took the relationship that marketing spend can be three times the customer lifetime value. Interestingly, in a later bullet, ChatGPT offered another benchmark that completely contradicts the earlier CAC benchmark: "LTV: A SaaS [software as a service] company should aim to maximize its LTV, as this is an important driver of long-term growth and profitability. A successful SaaS company should aim for an LTV that is at least 3x its CAC." This is the inverse of the prior relationship, and is the generally accepted right answer. The contradictory advice from

ChatGPT reveals that, at a fundamental level, it may not truly understand meaning. It is limited by its forward prediction output approach (it is the equivalent of speaking before thinking).

As language models grow larger, they may acquire more associations to underlying factors implicit in our words and phrases, but words are shadows in a cave—shadows cast from our lived experiences, written down and later consumed by the AI.

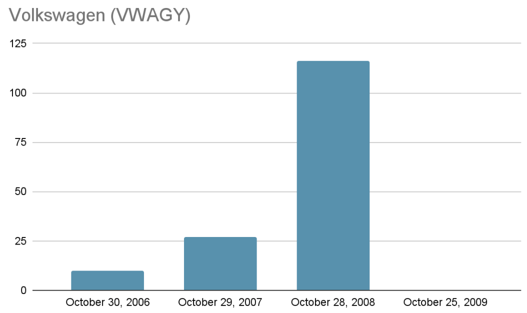
Symbolic Interpretation

The symbols used to represent numbers vary in some cultures, but the mathematical ideas that those symbols represent are recognized universally. There is precision in numbers and math that is readily obvious. At the same time, context matters. Most humans use the base10 system, and therefore $10 + 10 = 20$. However, computers use binary, and for them $10 + 10 = 100$. This is because, in binary, "10" is actually the representation for 2 (in our base10 number system) and "100" is the binary representation of the base10 number 4. The Mayans would also view numbers differently from us since they used a base20 system. On Papua New Guinea, there are four separate number systems linked to the different languages on the island. To understand how to interpret 10 + 10, some context is needed; namely, in which base are the numbers meant to be interpreted? As you can see, context is critically important to derive meaning from numbers. Most reading this book assume 10+10=20 in the same way we assume Tile A was a different shade than Tile B. We use our experience and culture to fill in context. Context is everything.

Cognitive scientist Douglas Hofstadter points out that symbols don't have a meaning on their own.[19] Rather, it's the isomorphism between symbols and an interpretation that creates meaning. In other words, context is an important part of how we create meaning from numbers. For instance, if we gave you a sequence of numbers as depicted in Figure 1.3, and told you this data represents a particular stock's closing price on or just after October 29 for a certain three-year period when a famous short squeeze occurred, then you could look up "famous short squeezes" and scan the table of stock quotes and find the pattern and the next value in the sequence:

Figure 1.3: Short Squeeze Pricing

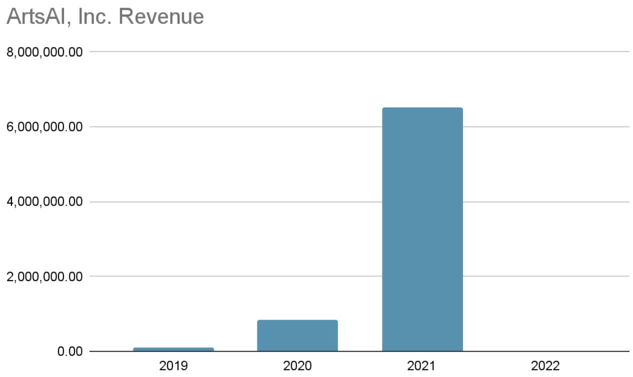
$$f(1) = 9.93, f(2) = 27.18, f(3) = 116.15, f(4) = ?$$



If you are unfamiliar with the term "short squeeze" you could look it up and learn it is a situation where certain investors buy up stock to put pressure on (squeezing) the short sellers, who then have to buy the stock to cut their losses. This forces the stock price artificially higher—sometimes, as in this case, about twelve times higher. This artificially high price does not typically last for very long. If the buyers' timing is right, they can make a lot of money. (If they don't time it right, they can lose a lot of money.) If you are asked to predict the range of value of the stock for the next year and you didn't have access to stock tables, your experience can help you estimate a reasonable range of values to make a guess. You know the value cannot be negative—stock prices can go to zero, but not to negative values. You can also reason that, while some stocks have had exponential increases like this one during a short squeeze, it is unlikely the value will remain high a year later—less than $75 is a safe guess. As soon as context is provided to describe this number sequence, we can begin to apply our experience or associated rules to come up with what we expect to be the next value.

Suppose, however, we told you these numbers represent the revenue of an artificial intelligence start-up we've worked with–as reported in the Inc. 500 list of fastest-growing private companies. The pattern is the same for the first three data points in the other examples, but the fourth datapoint, rather than dropping, increased by another 66 percent, reflecting the rapid growth of the firm.

Figure 1.4: Revenue for AI Start-up



These two scenarios have a similar pattern in the first three numbers but have different contexts and meanings. Symbols themselves, when devoid of context, can limit our ability to accurately extrapolate and understand meaning. Within a certain interpretation, we can find the value of $f(4)$, but without context, there is also not an inherent and objectively valid interpretation. If the context in which symbols appear is lost, the symbols largely lose their meaning.

Training AI to Fit Patterns

Artificial intelligence is essentially a pattern recognition machine—it takes in symbols, converts them to numbers, does some calculations, and then produces some numbers that in turn become symbols. But, an AI isn't given the context or interpretation of those symbols. To think of an AI as learning like a human does is largely inaccurate. Consider the example we just discussed above, where there are different possible numbers for the value of $f(4)$, depending on the context. When we train an AI, it is placed in a similar situation. We give it some examples of inputs, and their corresponding outputs, without any context of what the numbers mean. Each example is given as an input-output pair, but the input is not just a single number—it is a whole group of numbers (the matrix). Because the AI knows the input as well as the desired output for each pair, it can go through a process of adjusting itself so that, given a certain input, the output best matches the answer we told the AI was correct. After going through this process enough times, the AI will manage to find *some* pattern that matches the data, but there is no guarantee the pattern it finds will make sense in the context we are interested in.

For example, let's imagine we were trying to train the AI to determine whether an image has a train, a car, or neither. We hope that it would find a pattern based on the right features—say, the shape of the objects, the shape of their wheels, whether there are tires on the wheels, and the like. But, the AI might end up finding a different pattern. For instance, it might be the case that most pictures of trains also have railroad tracks—the AI might learn the pattern that *railroad tracks = train*. In this case, if it were given a picture of only railroad tracks, it would wrongly claim there was a train. There are various other details it might exploit to find a pattern in the data. (Statisticians call the incorrect patterns spurious correlations.)

If we feed AI images and ask it to classify if there is a dog or not, the AI processes the numeric values for each pixel in the image and converts it to a matrix of numbers. Then it analyzes the data in an effort to find a pattern, and produces an output. Once the AI has developed a pattern, it continues to use that pattern to classify images—whether it is accurate or not. For example, it might classify the image as "dog" or "not dog" based on the background, or by using some combination of factors in certain key regions of the image such as different features of the animals and different features of the background. Even the concept of "background" is a context of our three-dimensional experience. It requires our brain to have a deep understanding of perspective.

The AI has only a sequence of numbers based on the pixel colors and position—that's it. However, the AI may be able to finetune its accuracy; for example, if the AI detects that certain sequences of numbers represent the contrast in an image that signifies the edge of objects (say, a dog's tail), the AI may find that by using the numeric difference in adjacent pixels, the AI can discern shapes which improve correct classification. Detecting edges may improve the accuracy of classifying "dog" or "not dog." It is pretty remarkable that AI can find these patterns. However, the AI will almost certainly have problems with extrapolation outside of the domain on which it has been trained because it doesn't have the context to guide it in categorization of future input, whether it be images, stock prices, or the revenue of a start-up. We know that if a dog were to roll in green powder thrown in the Hindu festival of love, and thereby change fur color, it is still a dog. An AI which bases its decision on fur color may fail because of this change in the dog's color.

The idea that an object remains what it is despite superficial changes, like the addition of the color green to the coat of a dog, is known as an invariability—and AI doesn't seem to develop this construct when it is trained on still images. In large part, the concept of invariability is based on our context of lived experience with the world around us and our experience of our world changing in relation to time. Babies learn about one type of invariability, object permanence, when we play peek-a-boo. We hide our face with our hands and then move our hands to reveal we are still here, to the delight of the child. From birth, our brains become wired to develop an expectation for what will happen in the next moment of time, but most artificial intelligence is not designed to depend on time stamps, and thus it does not have the context to learn temporal invariability. If a dog is covered with green powder, an image recognition AI that has been fed tens of thousands of still images doesn't have the temporal experience to see how the same dog changed colors. To humans, it's still the same dog—now it's a dirty dog. But, to AI, it is a matrix of numbers whose values have changed in ways that may cause the AI classifier to now misclassify the image as "not a dog."

People who haven't developed AI image recognition software from scratch and evaluated its performance tend to overestimate AI's ability to understand the images it processes. They tend to overestimate the AI's ability to consider the shape of a dog, to recognize the four legs and tail, the fur, the size, the characteristic ears and snout—but that is not how artificial intelligence constructs its classification of "dog" or "not dog." People take for granted how we, as humans, construct invariability—and miss the fact this isn't a feature of today's AI. We should expect some AIs to fail in recognizing the dog since, pixel by pixel, it may look different from the dogs it was trained on because the colors in the pixels (the number value of the pixels, to be more precise) have changed.

Likewise, when driving a car, we understand that even if a splash of rain suddenly obstructs our view of the car in front of us, the car in front of us will continue to exist. However in many self-driving systems, this invariability is absent—cars can appear and disappear on the screen in an instant. If you have an automobile that displays the vehicles that your car's AI classifier identifies, take note if the images occasionally disappear and reappear and recognize the lack of invariability.

To summarize, the idea that the symbols we input represent horses, dogs, cats, stock prices, revenue or text about crossing a bridge in the distance is foreign to artificial intelligence. Whether images, numbers, or words, those symbols are just a matrix of numbers to the AI. When we instruct

AI on the pattern we want it to learn, it takes the matrix of numbers from all the examples we have fed it and uses math calculations to find the pattern that will more often produce the output we've told the AI is correct.

Is it a bridge too far to think of artificial intelligence as thinking like Philip K. Dick? We think it is. As humans, we need to guard against anthropomorphizing AI and instead consider what AI is actually doing. We need to get better at understanding that the intelligence of AI is not the same as human intelligence. We need to demystify AI and understand that AI "learning" is a specific mathematical process of fitting a pattern to data in order to produce the output we reward the AI for producing. As we will see in Chapter 2, artificial intelligence represents a specific form of intelligence with its own unique strengths and weaknesses. It is useful in many applications, but there are also applications where it fails miserably. There is ample room for improvement.

Chapter 2

How AI Fits Patterns

In the Virginia Range of Nevada, wild horses roam. It is a novelty to look out in the distance and see if you can spot one. Imagine for a moment you are looking at a horse in the distance. Even though it is far away, you can see it is a horse—no question about it. Now, imagine that a gnat zooms a foot or two in front of your face as you look at the horse. You notice the gnat, but it doesn't change your ability to see the horse in the distance—your ability to understand what you are seeing is *robust*, in computer science terms. If, on the other hand, a tiny gnat in your field of vision could so thoroughly confuse your understanding that you no longer knew you were looking at a horse in the distance, the perception would be, in computer science terms, *brittle*—the method of identifying the horse could be easily broken. Humans aren't confused by a speck, but today's AI can be.

In addition to artificial intelligence being, at times, brittle, it can also be *myopic* because it is generally only given one input stream. Today's AI focuses much more narrowly, but often more deeply, than humans because we might feed thousands and thousands of pictures of horses, in as varied circumstances as possible, to the AI so it can identify a horse. While the data set is thousands of pictures deep, it is still only a narrow experience with horses (specifically pictures of horses) compared to how a human might learn about a horse. A human child may see only a fraction of the examples of horses compared to AI, but the examples are likely to be more varied and draw on more of our senses than just vision. We might see pictures of horses in books, watch a video of a horse and hear it whinnying. We might touch a physical toy of a horse and rotate it around to see different angles. We might see movies with horses in them, providing context of what horses do. We might take a pony ride, complete with sights, sounds, smell and touch. In France, our experience might even include the taste of horse. (If you find eating horse off-putting, we include it as an object lesson in how our cultural context influences our perceptions.) The representation a human forms for a horse is much richer and more comprehensive when compared to the ability of artificial intelligence to classify a horse in a two-dimensional picture. This is in part because we have a broader range of experiences and context and in part because our neurons are structured differently than the artificial neurons of today's AI.

How AI Works: A Simple Mathematical Introduction

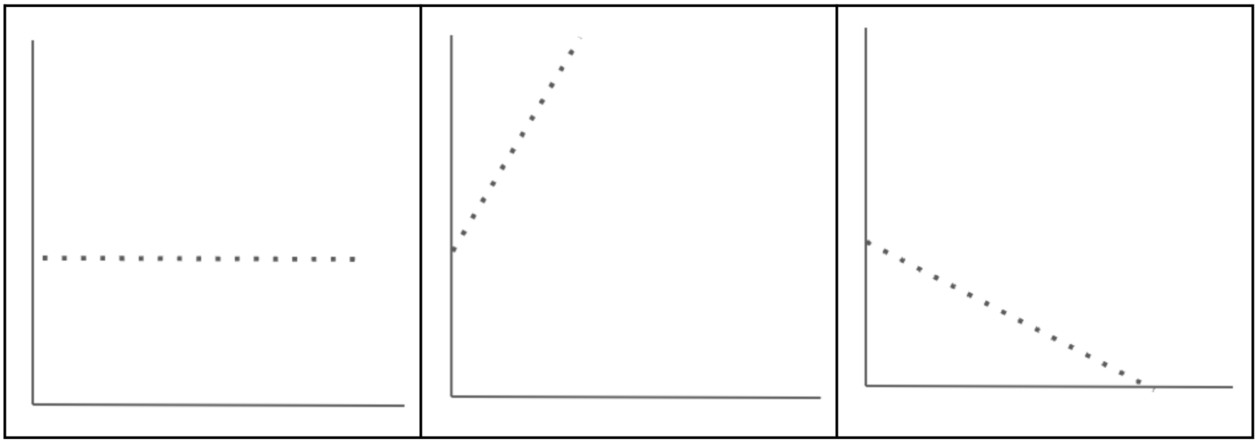
People often use artificial intelligence without understanding how it works. This means they may not appreciate the challenges that have been overcome or the hidden weaknesses that remain. A user of AI should understand the strengths and weaknesses, and be aware of how it can produce flawed results and possibly cause harm. We'll discuss a wide range of use cases in Part 2. In this chapter, we will use some basic math in order to provide some insights into how AI works and how it is trained.

There are various architectures from which AI systems can be constructed—their differences are based on how they fit data using a particular math function. In this chapter we will be exploring *deep neural networks* (DNNs), which have risen to prominence in the past decade and are used extensively for image recognition contexts, natural language processing, and a wide range of other tasks. Deep neural networks are a way of fitting patterns by using many layers of neural networks (hence the term deep), with each layer making a small adjustment that interacts with the other layers. To better understand pattern fitting, we'll briefly take you back to middle school math class, where you likely learned the equation for a line. Then, we will proceed to high school (or college) Introduction to Statistics, where you likely learned how to fit a line to data points. We will explain why AI excels at fitting more complex non-linear patterns. We've written this section for the non-mathematician, going light on the math and heavy on the concepts. If you are not a fan of equations, please bear with us—there is a big payoff in terms of understanding AI. We aim to level-up your knowledge of how AI works and the next few pages provide enough mathematical foundation and explanation for you to appreciate the strengths and weaknesses of today's AI, which stem from the very mathematical structure of AI.

Anatomy of a Line Equation

We typically learn in middle school the mathematical description for a line, which is given in the equation $y=mx+b$, where y is the line and the slope of the line is m, as illustrated in Figure 2.1. With a flat line, m is 0 (as in $y=0x+b$). The value m is known as the coefficient and it is simply the number we multiply each increment of x by to produce the points of the line. The larger m gets, the steeper the line; a negative value of m is a line that descends. Each x is an increment, such as a measure of time. The last component of the $y=mx+b$ line is the intercept b. This is a constant value and the point where x=0.

Figure 2.1: Slope of a Line By Changing the Coefficient



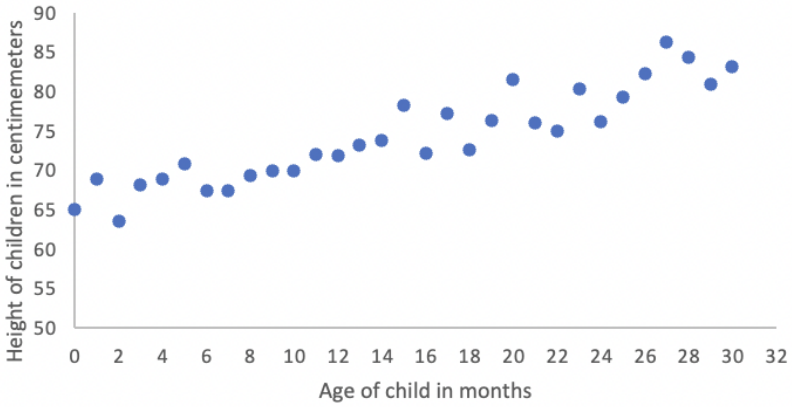
$y = 0x + b$	$y = 10x + b$	$y = -3x + b$

To ground these charts with an example, imagine the Y-axis that runs vertically is speed, and the lines below represent what happens when we are driving along at 60 miles per hour and press different buttons and pedals on a sports car. The X-axis is time. The first frame, where the coefficient is zero, would be the behavior if we pressed the cruise control—over time (x) the speed would stay the same, so the coefficient is zero. The next frame shows what happens when we press the accelerator—the speed goes up, and so the coefficient that is multiplied by x is positive. The last frame is the negative coefficient, showing the speed dropping when we press the decelerate button on cruise control or step on the brake.

Fitting a Line to Data Points

To take the explanation a step further with a different example, Figure 2.2 shows the height of 31 different children of different ages, measured in months. As one might expect when looking at data for a population of babies, you can see that height increases as age increases. But, there are some cases where an older child is not as tall as a different child that is younger. This is known as natural variability.

Figure 2.2: Height of Children by Different Ages

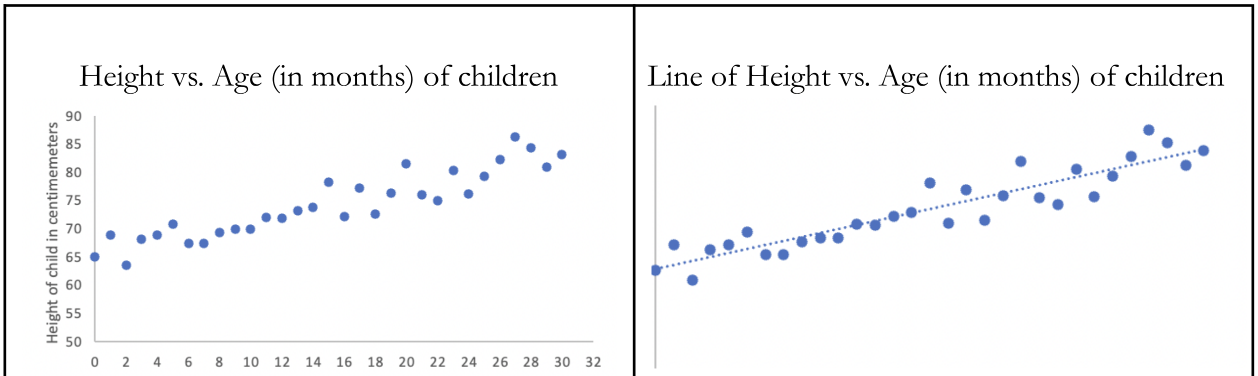


In the example depicted in Figure 2.2, *x* is months of age, so *b* is the height at birth. The height of a child at birth (when *x* is 0) is 65 centimeters in this data set.

Linear regression, often taught in statistics classes, is an approach to fitting a line to data points. In the simplest form, linear regression measures the distance from each data point. It produces the coefficient *m* and intercept *b* that is closest to all the data points—this is called least squares regression. In the example of height and age in children, linear regression takes some output pairs (the list of *x,y* coordinates), and produces a line equation that best fits the data. In Figure 2.3, the data set is plotted on the left; the chart on the right fits the data of a child's height by age and produces the equation *y*=0.61x+65.

Producing the equation is a very useful tool. If you want to know a child's height at 24.5 weeks, you can plug it into the equation and arrive at the solution that, on average, a child at that age is 80.6 centimeters (0.635 x 24.5 + 65 = 80.6). This is called interpolation, because the answer falls within the dataset.

Figure 2.3: Data And Linear Regression Fitted Line With Equation



Line equation: y=0.61x + 65

Interpretation: Infants in this data set are about 65cm at birth and grow 0.61cm per month.

The math of linear regression aims to find the correlation between x and y (in this case between age of a child and height) by finding the best fit equation for the data. If we wanted an AI system to make these calculations, we would call the input data the *training set*. The training set and mathematical calculations built into the AI system will make an association between age and height.

Neural Networks and Linear Regression

Neural networks can be thought of as an enhancement to linear regression because they are superior at fitting non-linear relationships. Non-linear is the mathematician's way of describing a curved line rather than a straight line. Let's consider a slightly more complex example to make the advantages of AI over linear regression clear. Suppose that a fast-growing ice cream company wants to predict its sales in millions of dollars *(y)* over time in months (*x*). This can help the company make projections, such as what its expected revenue will be in five months. The goal of linear regression is to produce a function where the outputs (the sales in a certain month) most closely align with the training data. For this company, the equation for the revenue of the company over time is:

$$y = \frac{1}{2}x + 10 + 5\cos\left(\frac{2\pi(x-6)}{12}\right)$$

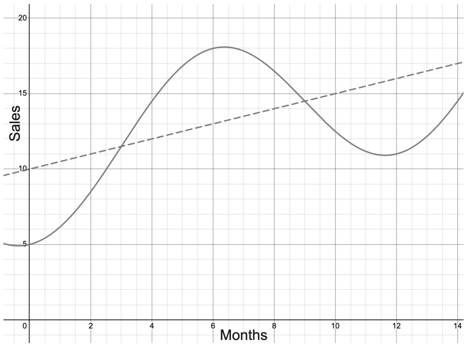
Let's break down what this equation means in this example of ice cream sales:

- $1/2x$ tells us that, based on observed sales over time, in each month that passes, the monthly revenue increases by $0.5 million.
- +10 means that the company is initially making $10 million a month at the start time period (time period 0).

So far, this equation follows the y=mx+b form, but what does that last term $5\cos\left(\frac{2\pi(x-6)}{12}\right)$ mean? To find out, we can graph the entire equation (the solid line in Figure 2.4) alongside the line y=½x+10 (the dashed line) as shown below so that we can see how the term behaves over time.

This reveals what the $5\cos\left(\frac{2\pi(x-6)}{12}\right)$ term does: It is a cosine function, which is to say it is a periodic function that repeats a certain pattern over a defined interval. In this case, the interval is 12 months. We can see that in the denominator.

Figure 2.4: Ice Cream Sales Linear Trend and Seasonal Components



The cosine term creates this wave-like pattern that cycles every 12 months—it measures the seasonal effect on ice cream sales. It shows that in the summer, sales increase significantly, whereas in the winter months sales decrease relative to the overall trend.

If we attempt to run linear regression on this company's revenue over the last year of sales, it will produce the equation $y = 1/2x + 10$ as the line of best fit (shown as the dashed line in Figure 2.4). But this statistical equation completely misses the seasonal effect. This could be a problem for the company making decisions based on these results. Imagine that the company starts a new advertising campaign during the summer. They spend $1 million on advertising but find that, six months later, in the winter, their sales have not increased at all—they are the same as what they were when they started the campaign. To the linear regression user, this would seem like a disastrous advertising campaign. In six months, the company was supposed to increase $3 million in monthly revenue, but instead their sales have stagnated. However, at the peak of summer the seasonal effect adds $6 million in revenue, and during the winter the seasonal effect is a loss of $6 million dollars in revenue. Between these two months, the net difference is $12 million. The non-linear equation shows that the expected change during these six months is a drop in sales due to seasonality, so if sales are still the same during the winter, that is a win for the business. The advertising campaign was highly successful—with only a $1 million investment it added $9 million in sales.

A cosine wave can represent a cyclical pattern, such as higher demand in the summer and lower demand in the winter. The linear form cannot represent that cyclical part automatically. The ice cream company that detects this nonlinear pattern gains a significant advantage over other ice cream companies, as they can more appropriately scale back production during the winter and scale up during summer months, as well as more effectively measure the effect of their business strategies.

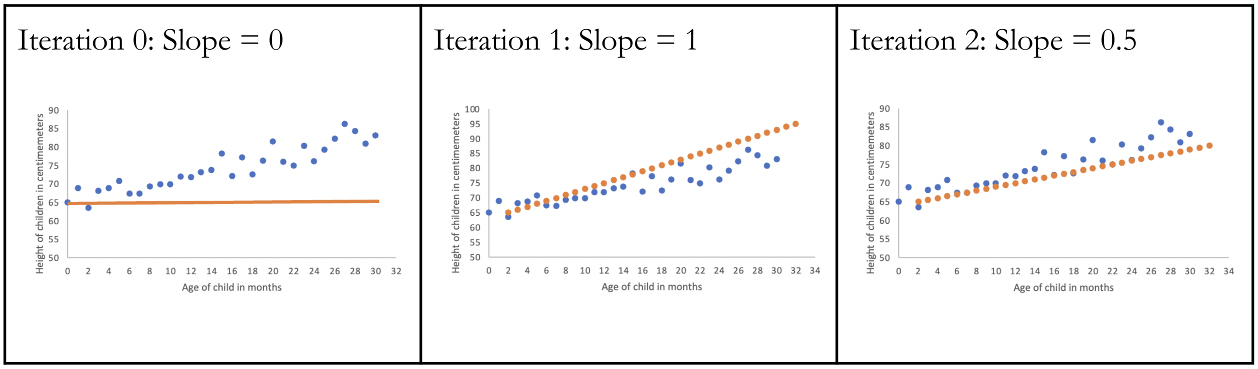
There are ways for a person who is adept with math to insert this cycle into a regression, but regression isn't going to fit this relationship automatically. The regression requires some human expertise to guide it. Linear regression cannot adequately approximate all functions, and in fact cannot perfectly model any function that isn't linear. Seasonality is just one example of a non-linear relationship. In this example, it's possible the ice cream company's growth could accelerate due to economies of scale—but that non-linear relationship is missed too. When the non-linear terms are obvious, an expert can account for them, for example by using polynomial regression (which can fit certain kinds of curved lines) or adding cosine terms. But it turns out there are still many functions that are not well approximated by polynomial or cosine functions. This is where neural networks come in. Neural networks are universal approximators, which means they can approximate any function well (or, more precisely, “arbitrarily closely”). This allows neural networks to capture any linear or nonlinear term, which means they can pick up on patterns that even human experts miss. For instance, a recent neural network built to predict properties of galaxies discovered many strong correlations and patterns between certain galactic properties that experts, with years of mathematical training, hadn’t found previously. If neural networks have a superpower, it is the ability to fit a pattern to any data set.

AI Can Be Trained, But Is It Really Learning?

The process of providing data to the machine that it then uses to fit patterns is called *training*. A neural network starts with a matrix of values that are similar to the coefficient *m* in our linear regression examples. In an AI system, these values are called *weights*, and adjusting them is like changing the coefficient on a line to find the fit that minimizes the distance between the data points and the line. The AI performs an iterative process of making adjustments to the weights to better match the desired outcome—which computer scientists call learning. Thanks to the matrix of weights and the multiple layers of weights an AI system can use, the fitting process is much more sophisticated than linear regressions. It is this fitting process that earns the AI the title of *universal approximator*. It can fit anything—even random data with no meaning. Unfortunately, sometimes the pattern it fits is spurious and can lead to mistakes.

Let's return to the example of finding the children's height in centimeters in the first 32 months after birth and look at a simplified example of how AI fits a pattern. The simplest AI we could construct controls the slope of the line only. We might start with a flat line at 65 centimeters (the height observed at birth with a slope of zero), as shown in Iteration 0 in Figure 2.5. This line underestimates the height of older children and would not be a very good fit for the data. Doing some math based on calculus, the AI system recognizes it needs to adjust towards a steeper slope, and might adjust to a slope of 1. In Iteration 2, the system recognizes this new slope of 1 is too high and adjusts downward to a slope of 0.5. This is a better fit but undershoots the data. Recognizing this, the system might increase the slope in the next iteration. Continuing this process, the AI "learns" a better slope to match the data, until it gets to the best fit of 0.63. (In this case, it is the same slope produced by the statistical least squares linear regression technique.)

Figure 2.5: Pattern Fitting For Age And Height


Iteration 0: Slope = 0

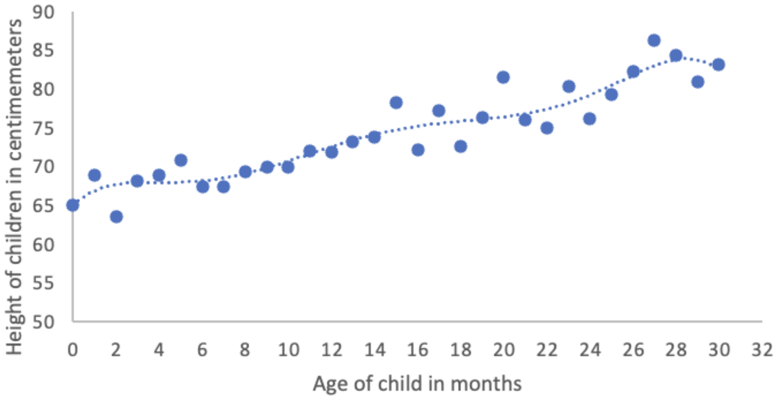
Iteration 1: Slope = 1

Iteration 2: Slope = 0.5

Where AI Goes Awry

The example above demonstrates how an AI adjusts through iterations of "learning" to find a fit to the data. The system has arrived at the same slope produced by the least squares linear regression technique. However, neural networks have a matrix of values rather than the single coefficient we had in our linear regression example. This allows neural networks to fit more complex patterns than linear regression. In many AI methods, there could be additional parameters to try to fit the non-linear relationships in the data. Even though this data set contains random variation related to sampling children of different ages, AI may find a non-linear pattern that fits the data even more closely. Figure 2.6 shows the fit of the data using a different AI that allows for non-linear patterns.

Figure 2.6: Polynomial Function Fit For Age and Height



This is a better fit of the data than the linear regression produced, but the implications are not logical. Notice how height decreases from age 28 to 31 months in the AI fitted line. This is random variation in the training data, but artificial intelligence fits this pattern. If we are trying to predict how a child's height will change as they age, as humans, we know from lived experience that children don't tend to shrink. However, artificial intelligence lacks context and doesn't know what it means to be a child, what it means to age, or even what it means to measure height—the AI only knows the data points so it might produce nonsensical patterns and a wrong prediction for height at 36 months—where it predicts the 36 month old will have shrunk by about 5 centiments. To an AI system, these are just symbols, and the system has been told to find a pattern in the symbols without context.

In our example of predicting the height of a child, AI would return the prediction of height at a given month that is more accurate than linear regression when it is interpolating the data, but fails badly when extrapolating the data to 36 months and beyond. Extrapolating to 36 months, the equation predicts a child will be 87 centimeters, which is a reasonable estimate. However, at 50 years (600 months) of age, the equation predicts 431 centimeters, which is the height of a typical one-story house—which is not a good prediction at all. These steps of taking input data and running it through the network to generate predictions is known as *forward propagation.*

The Black Box of Deep Neural Networks

The structure of how forward propagation works in deep neural networks is worth considering. Deep neural networks are known as "deep" because they have a large number of layers where weights are adjusted, with each layer influenced by the previous layer. Only after the input has been processed through every layer does the network return the output. These layers are called *hidden layers* because what goes on within the hidden layers is somewhat a black box from the outside. The human creator of an AI system can see every layer of a deep neural network and the matrix of weights that are adjusted in the process of training. The AI is represented by an enormous list of weights, with each layer of weights depending on the layer before it, which in turn depends on the layer before that, and so on, down to the original input. Therefore, to the human eye, the particular values of the weights lend themselves to no obvious explanation.

At the same time, deep neural networks operate in *high-dimensional space,* which means many different parameters are influencing the output simultaneously. Each input can be seen as its own dimension, and so an image recognition program which looks at each pixel in a 100x100 sized image resides in the 10,000 dimensional space. GPT-3, one of several AI models we will examine in this book, includes over 100 billion parameters. Google's PaLM has over 500 billion.[20] This means that the list of weights contains upwards of 500 billion entries. High dimensional space poses an issue for human analysis and understanding because it is nearly inconceivable for us to reason how all the different parameters might interact to influence an output.

To visualize this point of high dimensionality, let's revisit the ice cream sales example and seasonality. More ice cream is sold during summer, when it is hot outside. A human can easily conceptualize two dimensions, such as outside temperature and ice cream sales—consider temperature as the vertical y-axis of a chart and ice cream sales as the horizontal x-axis of the chart. We can visualize how higher temperatures are associated with higher ice cream sales. We expect ice cream sales to increase as temperature increases. But what would a five-dimensional chart look like that simultaneously includes temperature, economy, the mood of the population, the holiday schedule, and stock market performance? We can take any pair of these dimensions and visualize a relationship, but as we add more dimensions, dimensions that interact with one another and simultaneously influence the output, it becomes challenging for humans to directly visualize these multidimensional relationships. Five parameters are nearly impossible for most humans to conceptualize but are simple for artificial intelligence, which handles hundreds, thousands, millions and even billions of parameters (future generations may be trillions). Therefore, high dimensionality makes it difficult for a human to understand what the AI is doing. AI researchers are trying to find ways to automatically explain what the AI is learning, but as of today, much of what the AI does is inscrutable.

Whether artificial intelligence truly understands what it's doing to generate output is not straightforward. Did our simple AI system learn how babies grow in height over their first 31 months of life? Would a more sophisticated AI have understood that children don't shrink starting at 28 months nor grow to be the size of a house by the time they reach 50 years of age? We can't simply ask a deep neural network whether it understands what it's doing, nor can we look closely at its architecture, since its understanding is concealed within the relationship between neurons in the hidden layers. Therefore, much like investigating human bias and cognition, a good way to understand an artificial intelligence's basis for its output is to perform experiments and otherwise observe its behavior and find patterns in its failures and successes.

Determining AI's Level of Understanding

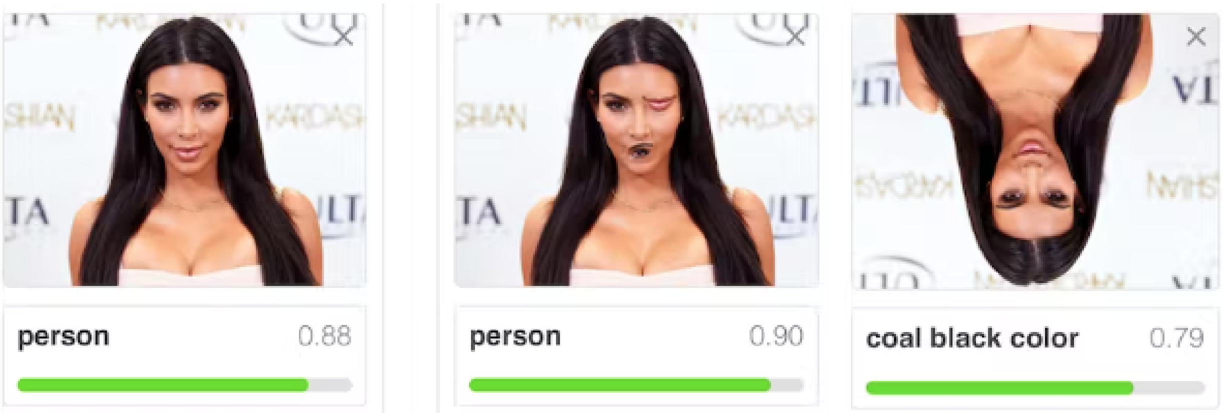
It is important that we establish what is meant when we write "understand" or "doesn't understand." An AI that classifies a horse based on irrelevant features like the background of the image or spurious correlations in the pixel colors has a non-robust understanding of a horse and can be easily tricked into misclassification. On the other hand, we write an AI understands when the AI classifies based on a combination of the robust features such as its size, body structure, the presence of hooves and mane, and other features directly related to what a horse is. Thus, when we say an AI "understands" what a horse is, we mean that it has learned to classify whether or not something is a horse based on the features that actually determine whether or not something is a horse. When we say it doesn't understand, we mean it has learned to use spurious correlations to fit the pattern and hasn't truly learned what something is.

A close observation of various deep neural network models suggests that the systems do not understand the underlying meaning of the features and interrelationships that they learn. AI researcher David Watson explains how "even top performing models can learn to discriminate between objects while completely failing to grasp their interrelationships." He shows an AI's confidence that an image is that of a person actually *increases* when its facial features are switched around. He states, "the true problem runs deeper… any combination of eyes, nose, and mouth will suffice for a convolutional deep neural network—not because of external constraints on the choice set, but because of intrinsic limitations of the model architecture."[21]

In other words, AI fundamentally doesn't grasp what it means to be a face. It doesn't understand the relationship between these parts of a person's face. The AI has learned a shortcut (what humans might call a cheat) by relying on certain non-robust patterns to make its discrimination. When it sees an eye, nose, or mouth, in the image, the AI classifier defines it as a person, regardless of whether it is organized as a face or not. Notice in the images in Figure 2.7 how switching around the eye and mouth actually increases the classification confidence from 88 percent to 90 percent. Notice that flipping the image leads to the AI having 79 percent confidence the image is of black coal—and not a person at all.

The problem is common enough to have a name—it is called the *Picasso Problem*. The image has the parts, but they are not in the correct spatial relationship to form the whole—yet the AI makes a classification anyway. This experiment demonstrates a major flaw in artificial intelligence today: It can easily be tricked into mistaken classifications.

Figure 2.7: The Picasso Problem

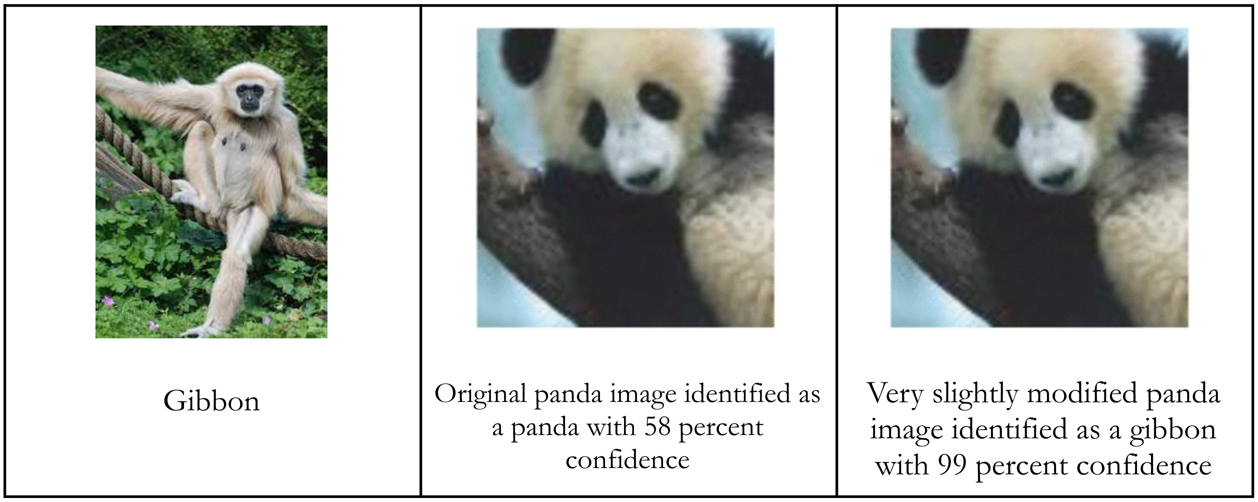


Source: Capsule Networks Are Shaking up AI — Here's How to Use Them, Bourdakos, N. (2017)

One might argue that this particular problem can be fixed with new functionality added to today's AI. In 2017, AI research pioneer Geoffry Hinton and his team proposed a solution, called *capsules,* to improve AI image recognition. A capsule is a hierarchical organization of content that is created so that the spatial relationship is considered as part of the structure.[22] However, most AIs today do not use this hierarchical organization because it adds a lot of complexity that may make it harder for the AI to learn patterns, which can result in the AI failing to learn any pattern at all. Today's AI is overly reliant on non-robust relationships due in part to the math and in part due to limitations in the types of data fed to AI (usually still images or text). The example of identifying a person based on facial components illustrates that AI will tend to solve classification problems in non-robust ways—it is easier for the AI to learn to recognize eyes, a nose, or a mouth than it is for it to learn how to recognize an entire face and how all its pieces interrelate to form the whole.

A striking example of artificial intelligence's superficial understanding of the images it processes can be seen in an experiment in which an AI system was presented with a picture of a panda, which it properly identified. But, after the researchers added some imperceptible noise to the image by changing some of the pixels, the AI misclassified the image as a gibbon with a high degree of certainty. Figure 2.8 shows an image of a gibbon, for reference, next to the picture of the panda that the AI system classified with 58 percent confidence as a panda. The third image is the same image of a panda with less than one percent change to the pixels, and which the AI system misclassified as a gibbon with 99 percent confidence. The fact that a bad actor can introduce noise that is imperceptible to humans but can purposely fool an AI into the wrong answer hints at the fundamental problem of AI's lack of understanding.

Figure 2.8: How Image Classification Can Make Mistakes



Gibbon

Original panda image identified as a panda with 58 percent confidence

Very slightly modified panda image identified as a gibbon with 99 percent confidence

Source: David Watson, The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence, 2019.

Since this experiment, AI practitioners have added techniques to feed the AI images with some pixels changed in an effort to make the AI less susceptible to what is called an adversarial pixel attack. This approach makes the AI more robust to this particular attack, but makes the AI more brittle in other ways. Overall, it doesn't resolve the deeper problem of the AI not understanding what a panda, gibbon, human face, horse, or other labeled image *is*.

The AI Snow Job

The problem of artificial intelligence not truly understanding the classification of different objects is illustrated more clearly through another example. A system was trained to distinguish between huskies and wolves and achieved very high classification accuracy, even on images it hadn't been given before. Yet, there was a problem—it started to misclassify very obvious images. Fabio Kepler, a senior researcher at Unbabel, explains that "When they looked into why the neural network was making such gross mistakes, researchers figured out the model learned to classify an image based on whether there was snow in it—all images of wolves used in the training had snow in the background, while the ones of huskies did not." Once again, the AI had failed to understand the context—instead of learning what the difference between huskies and wolves really was, it simply took the shortcut and learned to recognize snow and classify it as a wolf.[23]

AI's universal approximator found that using the pixels that render the snow in the pictures reliably produced the right answer: wolf. The AI classifier was biased by the background presence of snow, which is correlated with the wolf in the foreground. One might blame the people that created the data set to train the AI, but because of the universal approximation power of AI, it is difficult to know, in advance, what relationships the AI might incorrectly use to arrive at its classification. If universal approximation is AI's superpower, it is also AI's kryptonite.

Examples of cases of spurious correlations are common throughout the AI world. These have come to be known as "Clever Hans" cases, named after a horse who was allegedly able to compute sums and count. It was later found that the horse could not, of course, do math—it was giving responses based on the behavior of the person posing the question. Examples of AI misclassifying images abound.[24] One AI system managed to determine a picture included a boat based on the presence of a body of water (not whether a boat was present), and it recognized trains based on the presence of tracks (not on whether a train was present).[25]

It is not just image recognition that reveals AI's brittleness. Image recognition is just one illustration of a more deeply rooted weakness with today's AI. As reported in an article in *Nature* in 2022, two researchers looking into flawed results from artificial intelligence found 17 fields of study and 329 published papers in which results could not be fully replicated because of problems in how AI was applied. Some, they note, were textbook problems such as flawed use of AI or reliance on biased data sets—and it is troubling that research could be published in top journals with such flawed use and biased data. Others were subtle and hard to detect, leading to flawed output that could be applied and not understood as flawed until harm had occurred.[26]

In 2019, Xiao Liu, a clinical ophthalmologist at the University of Birmingham, U.K., and her colleagues found that only 5 percent of more than 20,000 papers that discussed AI for medical imaging were described in enough detail for the reader to discern whether they would work in a clinical environment. At least some of these less-than-transparent papers appeared in very reputable and highly cited journals.[27] In an interview about her paper, Liu told us, "In a safety-critical field like healthcare, it is essential that AI is implemented with a responsible and ethical approach. AI systems should be robustly evaluated in real clinical pathways to demonstrate benefit to patients, clinicians and the health system, and to provide reassurance that safety and performance is maintained when translating from clinical studies to implementation."

Independent validation of AI results is important, and it should include the ability to replicate the results in the real world (or, as those in health care term it, in the clinical pathway). Liu explained that the fact that less than 5 percent of those studies were robustly designed and transparently reported is a serious concern. For example, she notes, "most were retrospectively tested on datasets only, rather than real clinical pathways which involved patients and clinicians; and many didn't test the algorithm in datasets separate to the data used to train the algorithm, in statistical terms, we call this an external validation, and it is known that testing the AI on the same data set used to train the AI overestimate the AI's accuracy."

Since publication of her paper, Liu now sees more algorithms tested in new datasets or real clinical environments, which is a step in the right direction. However, Liu says, "we are finding that the AI performance is often less than expected."