Full Self-Driving, Skynet and Other Artificial Intelligence Myths

Modern decision making with deep machine learning models

Brad Flaugher

January 9, 2023

We have now accumulated sufficient evidence to see that whatever language the central nervous system is using, it is characterized by less logical and arithmetical depth than what we are normally used to.
– John von Neumann [1]

Preface

This book is an attempt to organize my brain on paper. I have been thinking about decision making with computers, and the computer's relationship to the brain for too long now, and before I get dementia (I'm 35) I would like to put a stake in the ground and tell everyone where they can stick their thoughts about the future.

I am simultaneously very hopeful for the future and horrified by the way that people talk about it. I think many so-called "Data Scientists" are full of shit when they talk, and some of them know it but would like to keep their jobs. I think that many investors are snowed by fancy language and toy demos that they extrapolate into the future. Users of digital products are woefully ignorant about how the products work and how they fit into a larger ecosystem. I'm amused, annoyed and would like to clear things up for everyone, including myself.

This book contains strong beliefs, weakly held. If I say something stupid I'll eat my hat and update the book if I have time, or just trash the thing and say "please don't read this, consider it my diary from 2023". However, for now I think I am right and I think you should listen to me for a moment.

I am a self-employed programmer and financially secure, these are my credentials. I have no corporate master and I am not raising money to create the future of AI, so have a limited bias towards AI boosterism. I write programs close enough to the bleeding-edge that I can see the present as fast as almost anyone.

So let's go on this journey together, I'll try and explain to you (with words, code and examples) how modern AI works, and when you should be scared (self-driving cars without the appropriate infrastucture) or when you should not be scared (please shut up about Skynet becoming self-aware). I'll also try do case studies in as many arenas that I think are interesting. If you have notes for me you can reach me at brad@bradflaugher.com.

Please skim this book, please skip around and read some stories. I think the first three chapters are important for everyone to know. The rest of the stories can be chosen a la carte. Enjoy.

Brad Flaugher

Contents

Preface iii

Contents iv

1 History: From Explicit Pules to Big Data and Statistics 1
- 1.1 The AI Textbook in 1997 . . . 1
- 1.2 Does AI Need to Know Grammar to Translate? . . . 1
- 1.3 Explicit Rules and Codified Human Knowledge . . . 2
- 1.4 IBM Tries Every Possible Chess Move . . . 3
- 1.5 Statistical Analysis of Handwriting is the Way of the Future . . . 4
- 1.6 Less Programmer Intelligence and more Data Intelligence . . . 4
- 1.7 From Explicit Rules to a Black Box, and Beyond . . . 5

2 The Regression Theory of Everything 7
- 2.1 Let’s Avoid Knowledge Representation! . . . 7
- 2.2 A Simple Neural Network is also a Linear Regression . . . 7
- 2.3 Dummy Variables for Dummies; "It’s All Numbers, Man" . . . 7
- 2.4 Try That Again With 2,354,356 Parameters . . . 7
- 2.5 Multicolinearity and the End of Science . . . 7
- 2.6 Let’s Test Some Random Inputs! Feature Importance and Explainability . . . 7
- 2.7 The Universal Machine Learning Workflow . . . 7

3 Decision Making, Uncertainty and Chaos 8
- 3.1 Concept Drift . . . 8
- 3.2 Theories of Creativity . . . 8
- 3.3 From Spaghetti Code to Lasagna Neuron Layers . . . 8
- 3.4 Garbage In, Garbage Out . . . 8
- 3.5 Garbage In, New Perspective Out? . . . 8
- 3.6 The Impossibility of Fairness . . . 8
- 3.7 What Are We Prediciting Again? . . . 9
- 3.8 The Perverse Incentives of Data Scientists: Job Security by Obscurity . . . 9
- 3.9 Humans Love Computers . . . 9

4 Classification and Profiling: Studies in AdTech and Policing 10
- 4.1 Everything is Clasification . . . 10
- 4.2 Classification in Reverse; Generating the Stereotype . . . 10
- 4.3 Code Your Own Classifier in 5 Minutes! . . . 10
- 4.4 The End of Nuance . . . 10
- 4.5 If You Were Unable To See A Doctor, is it OK? . . . 10
- 4.6 Lazy Thinking . . . 10
- 4.7 Data Where There Is None . . . 10

5 Self-Driving With Statistics 11
- 5.1 Trolley Problems . . . 11
- 5.2 The Third Rail . . . 11
- 5.3 Purple Lights and Changes In Fashion . . . 11
- 5.4 Multicolinearity (Reprise) . . . 11
- 5.5 See You In Court, Asshole! . . . 11

6 Generating Art: Avante-Garde or Derivative? 12
- 6.1 Predictive Keyboards 12
- 6.2 GPT On A Napkin 12
- 6.3 Compress The World's Achievements Into One Database 12
- 6.4 Upscaling What Is Lost 12
- 6.5 Who Owns This Stuff Anyway? 12
- 6.6 Use Critical Thinking to Become A Critic 12

7 Revolutionary for Whom? 13
- 7.1 Self-Driving Horses 13
- 7.2 The Battle of the Assistants 13
- 7.3 Employees That Are Better Than You 13
- 7.4 Who is Helped the Most? 13
- 7.5 Who is Hurt the Most? 13
- 7.6 How to Respond 13
- 7.7 This Book is a Case Study 13

8 Unplugging Skynet 14
- 8.1 AI and Human Safety 14
- 8.2 Useful Incompatability 14
- 8.3 Checks and Balances 14
- 8.4 When You Can't Tell The Difference 14
- 8.5 Sucker Traps 14
- 8.6 Dead Inside 14

9 Errors and Omissions 15

Bibliography 16

History: From Explicit Pules to Big Data and Statistics 1

1.1 The AI Textbook in 1997 . . . 1
1.2 Does AI Need to Know Grammar to Translate? 1
1.3 Explicit Rules and Codified Human Knowledge 2
1.4 IBM Tries Every Possible Chess Move 3
1.5 Statistical Analysis of Handwriting is the Way of the Future 4
1.6 Less Programmer Intelligence and more Data Intelligence 4
1.7 From Explicit Rules to a Black Box, and Beyond 5

1.1 The AI Textbook in 1997

Dr. Elaine Rich's textbook on Artificial Intelligence, published in the 1980s, was a groundbreaking work that helped to establish many of the foundational concepts and techniques in the field of AI. However, the rapid advancements in AI over the past few decades have led to many of the chapters in this textbook becoming obsolete.

One of the main reasons for this is the prevalence of deep learning, big data, and large-scale statistical models in modern AI. These techniques have largely replaced the symbolic, rule-based approach to AI that was emphasized in the textbook, making many of the chapters on knowledge representation and expert systems less relevant.

Additionally, the explosion of data and the availability of powerful computing resources have made it possible to apply machine learning techniques at a scale that was previously unimaginable. This has led to the development of highly effective machine learning models that can handle complex tasks such as image and speech recognition with a high degree of accuracy, making many of the chapters on simpler machine learning techniques such as decision trees[1] and linear regression less relevant. [2] [2]

We'll discuss this history and a few examples from the "early days" of AI to help us understand where we are headed. We'll start with machine translation, then discuss chess and finally neural networks, which will be the focus of the rest of this book.

1.2 Does AI Need to Know Grammar to Translate?

Noam Chomsky is a linguist and philosopher who has made significant contributions to the field of linguistics with his theory of universal grammar. Chomsky believes that all human languages share a common underlying structure, and that this structure is innate to humans. He proposes that this innate structure is the result of a "language acquisition device" present in the human brain, which allows us to learn and produce language. Chomsky also argues that the structure of language is largely independent of its content, and that the ability to produce and understand language is a fundamental aspect of human nature. His theory has been influential in the field of linguistics and has sparked much debate and research on the nature of language and its relationship to the human mind.

For English speakers or anyone who has learned English as a second language you'll have many examples of special cases, irregular verbs, bad english and former street slang that became good and proper over time. For

[1] Although mathematically, Neural Networks are Decision Trees

[2]: Rich et al. (2009), *Artificial Intelligence*

[2] the book is now in its third edition and unlikely to be updated as Dr. Rich as retired utexas.edu

programmers this is a nightmare, how can we codify human knowledge in a timely fashion? If we tried to write the rules of the english language in code (which many have tried to do) the rules themselves might change before we were finished writing them.

Explicitly translating languages through code is a difficult task because it requires a thorough understanding of the grammar, vocabulary, and syntax of both languages, as well as the nuances and subtleties of their respective cultures[3]. Simply coding rules for how to translate words or phrases from one language to another is not sufficient, as there are often multiple valid translations for a given phrase depending on the context in which it is used.

A more effective approach to translation is to use statistical techniques that rely on a large corpus of translated data, such as Canadian laws[4]. This type of data-driven approach involves training a machine learning model on a large dataset of translations, allowing it to learn the patterns and relationships between the languages. The model can then use this knowledge to make educated translations of new phrases or sentences, taking into account the context in which they are used.

While this approach is not perfect, it has proven to be highly effective in machine translation and can produce accurate translations even for languages that are very different from each other. The use of a large dataset of translations also allows the model to learn from the mistakes and variations present in real-world translations, further improving its accuracy.

1.3 Explicit Rules and Codified Human Knowledge

When we "teach" a computer to perform a task by explicitly writing down all of the rules of that task, we are really codifying human understanding.[5] When we codify human understanding we write down every rule that we know explicitly. For small tasks we can do this with 100 percent accuracy, and only minor headache on the part of the sofware developer.

For example, let’s write a boring function to tell you the number of days for a given month.

```
def days_in_month(year, month):
  if month in [1, 3, 5, 7, 8, 10, 12]:
    return 31
  elif month in [4, 6, 9, 11]:
    return 30
  elif month == 2:
    if (year % 4 == 0 and year % 100 != 0) or year % 400 == 0:
      return 29
    else:
      return 28
  else:
    return "Invalid month"
```

[3] For programmers this is a nightmare, how can we codify human knowledge in a timely fashion? If we tried to write the rules of the English language in code (which many have tried to do) the rules themselves might change before we were finished writing them.

[4] They’re in French AND English, which is useful data that we can use to correllate phrases and transform English to French and vice-versa.

[5] Programming this way makes some software development totally boring, I almost switched my major in college to math after considering what a life would look like manually writing rules for handling "edge cases" for the rest of my natural life.

Writing code can be a tedious and repetitive task, especially when it comes to debugging and testing. It can be especially frustrating when you're working on a large project and you're trying to track down a specific bug that's causing the program to crash. Testing code can also be boring, as it often involves running the same tests over and over again to ensure that the code is working correctly.

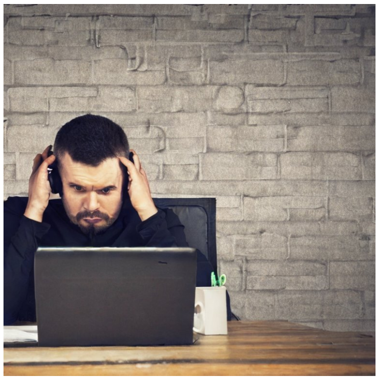

Figure 1.1: "a frustrated programmer writing boring rules on his computer" made with Stable Diffusion 2.1

Additionally, writing code can be boring because it requires a lot of concentration and focus. It can be easy to get lost in the details and lose track of time, especially if you're working on a complex problem. It can also be challenging to come up with creative solutions to problems, and it can be frustrating when your code doesn't work as expected.

While writing and testing code can be rewarding and fulfilling, it can also be a tedious and boring process. It requires a lot of patience, persistence, and attention to detail, and it can be easy to get frustrated and lose motivation. However, with practice and perseverance, it is possible to overcome these challenges and find enjoyment in the process of writing and testing code.

AI has traditionally operated by explicitly codifying human knowledge into machine-readable formats by doing the boring job of coding. This approach, which I'm calling "codified human knowledge" relies on humans to carefully structure and organize information in a way that can be understood by the AI system. The AI system then uses this structured knowledge to make decisions and perform tasks.

However, recent advances in AI have largely ignored the knowledge representation problem and instead have focused on using statistical techniques and neural networks to automatically learn patterns and relationships in data. This approach, known as "deep learning," involves training large neural networks on vast amounts of data, allowing the AI system to make educated classifications and transformations of data without explicit human guidance.

Deep learning has proven to be highly effective in a variety of applications, such as image and speech recognition, and has contributed to the rapid progress we have seen in AI in recent years. However, the reliance on large amounts of data and the lack of transparency in these models can make it difficult to understand how they are making decisions, which can be a concern in certain applications (hence the title of this book).

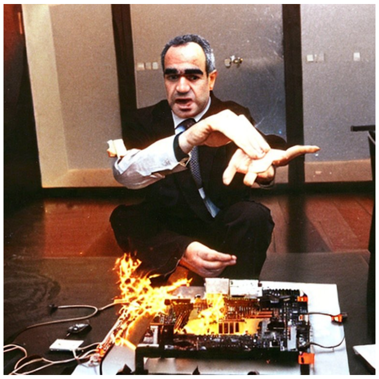

Figure 1.2: "Garry Kasparov setting a computer on fire" made with Stable Diffusion 2.1

1.4 IBM Tries Every Possible Chess Move

Deep Blue was a revolutionary computer developed by IBM that was specifically designed to play chess at the highest level. It was programmed with a vast database of chess knowledge and was able to analyze millions of positions per second.

Garry Kasparov was the reigning world chess champion at the time, and he was considered to be one of the greatest players in history. He had never lost a match to a computer before, and he was confident that he would be able to defeat Deep Blue.

However, things did not go as Kasparov had expected. Deep Blue was able to analyze the positions on the board with incredible speed and accuracy, and it was able to come up with highly sophisticated strategies that Kasparov had never seen before.

Despite Kasparov's best efforts, he was no match for the sheer brute force of Deep Blue's computational power. In the end, Deep Blue emerged victorious, defeating Kasparov in a historic match that changed the world of chess forever.

Deep Blue was a turning point in the development of AI, but Deep Blue's methods (namely calculating every possible outcome of a Chess game to determine the best move) was not suitable for many of the world's problems. It turns out that Chess is fun, but the world is not like chess.[6] The "real"[7] future of AI was being developed elsewhere, using statistics and a toy model of the brain to solve a very practical problem for banks.

1.5 Statistical Analysis of Handwriting is the Way of the Future

It was the early 1990s and Yann LeCun was a researcher at Bell Labs in New Jersey. At the time, the process of reading and processing checks was a tedious and time-consuming task that was done manually by bank employees. LeCun saw the potential for using artificial intelligence to automate this process, and he began experimenting with using convolutional neural networks (CNNs) to recognize patterns in images of checks.

At the time, CNNs were a relatively new type of neural network that had been developed in the 1980s for image recognition tasks. They were inspired by the structure of the human visual system, and were able to process images in a way that was similar to how the human brain does.

LeCun's work was groundbreaking, and he was able to achieve impressive results using CNNs to process checks. By 1993, he had developed a system that was able to read and process checks with a high degree of accuracy, significantly reducing the amount of time and effort that was required to process checks manually.[8]

LeCun's work on using CNNs for check processing was a major milestone in the field of artificial intelligence, and it laid the foundation for the development of many other applications of CNNs in the years that followed. Today, CNNs are widely used in a variety of applications, including facial recognition, image classification, and natural language processing. [9]

1.6 Less Programmer Intelligence and more Data Intelligence

I think it's useful to separate the knowledge in the AI problem-space into two groups. The data and the programmer together make the programs

[6] if you would like to read an example of the simplest chess engine that I could imagine that is written in a smililar "codified" way as Deep Blue check out sunfish.

[7] We might go back, and try again to more explicitly code everything up, and in some cases we still need to, but from the author's perspective, we live in a deep learning/neural network world.

[8] CNN digit OCR models are frequently featured in beginner training tutorials for deep learning libraries like PyTorch and Tensorflow, check one out on Github.

[9] check out Yann LeCun domonstrating a convolutional neural network in 1993 at youtube.com.

that we use every day, and for the rest of this book I'll try and separate the discussion of the smarts of each to help us better understand the world.[10]

Early AI relied heavily on a human programmer to design, write, and debug computer programs. Good progammers needed domain expertise, problem-solving skills, logical thinking, and the ability to learn and adapt to new programming languages and technologies.

We are now in the age of big data, and everyone knows that "data is gold". Statistical AI methods that are now most prevalent rely on extracting meaningful insights and knowledge from large datasets. This involves using statistical and analytical methods to discover patterns and trends in data, and using this information to inform business decisions or solve problems.

Figure 1.3: "a group of computer programmers striking outside of Microsoft's offices with placards saying 'rule-based programming is boring'" made with Dall-E 2

Throughout this book I'll discuss the interaction between programmer and data, and what can go wrong. Working with big data and statistics at a large scale has given AI tremendous abilitiy, but has made understanding and testing models infinitely more difficult. It is your authors belief that understanding the nuance of this interaction between programmer and data is essential to understadning modern AI.

1.7 From Explicit Rules to a Black Box, and Beyond

Artificial intelligence (AI) has come a long way since its inception, and the way that it makes decisions has changed significantly over time. In the early days of AI, explicit rules were used to tell the AI system what to do in certain situations. These rules were often written by humans and encoded into the system, and the AI would follow them to make decisions.

However, with the advent of deep learning, we have started to rely on a statistical understanding of the truth for AI to make decisions. Deep learning is a type of machine learning that involves training artificial neural networks on large datasets. These neural networks are able to learn patterns and relationships in the data, and can use this knowledge to make decisions.

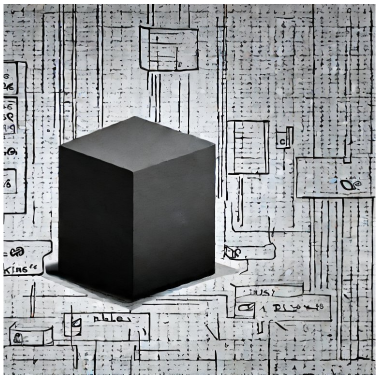

Figure 1.4: "from explicit rules, to a black box and beyond" made with Stable Diffusion 2.1

The use of deep learning has led to the development of powerful AI that is able to perform tasks that were previously thought to be impossible for a machine to handle. For example, deep learning has led to the development of AI systems that are able to recognize faces, translate languages, and even beat humans at complex games like chess and Go.

While deep learning has led to significant advances in AI, it has also made it harder to debug and understand how the AI system is making its decisions. With explicit rules, it was relatively easy to understand why the AI made a particular decision. However, with deep learning, it is often difficult to understand exactly how the AI arrived at its decision. This can make it challenging to troubleshoot problems with the AI system and to ensure that it is making decisions that are fair and unbiased.

In conclusion, AI has come a long way since its early days, and the way that it makes decisions has changed significantly over time. While explicit rules were once used to tell the AI what to do, we now rely on a statistical understanding of the truth for AI to make decisions. This has led to the development of powerful AI that is able to perform a wide range of tasks, but it has also made it harder to debug and understand how the AI is making its decisions.

The Regression Theory of Everything 2

"AI Scientists disagree as to whether these language networks posess true knowledge or are just mimicking humans by remembering the statistics of millions of words. I don't believe any kind of deep learning network will achieve the goal of AGI [Artificial General Intelligence] if the network doesn't model the world the way the brain does. Deep learning networks work well, but not because they solved the knowledge representation problem. They work well because they avoided it completely, relying on statistics and lots of data instead. How deep learning networks work is clever, their performance impressive, and they are commercially valuable. I am only pointing out that they don't possess knowledge and, therefore, are not on the path to having the ability of a five-year-old child." Jeff Hawkins, 2022 [3]

2.1 Let's Avoid Knowledge Representation! 7
2.2 A Simple Neural Network is also a Linear Regression . . . 7
2.3 Dummy Variables for Dummies; "It's All Numbers, Man" 7
2.4 Try That Again With 2,354,356 Parameters 7
2.5 Multicolinearity and the End of Science 7
2.6 Let's Test Some Random Inputs! Feature Importance and Explainability 7
2.7 The Universal Machine Learning Workflow 7

2.1 Let's Avoid Knowledge Representation!

2.2 A Simple Neural Network is also a Linear Regression

2.3 Dummy Variables for Dummies; "It's All Numbers, Man"

2.4 Try That Again With 2,354,356 Parameters

2.5 Multicolinearity and the End of Science

2.6 Let's Test Some Random Inputs! Feature Importance and Explainability

2.7 The Universal Machine Learning Workflow

Summarize and opine on The Universal Machine Learning Workflow [4]

[4]: Chollet (2022), *Deep learning with python, second edition*

Decision Making, Uncertainty and Chaos 3

"I hope for some sort of peace—but I fear that machines are ahead of morals by some centuries and when morals catch up there'll be no reason for any of it." Harry Truman, 1945 [5]

3.1 Concept Drift 8
3.2 Theories of Creativity 8
3.3 From Spaghetti Code to Lasagna Neuron Layers . . . 8
3.4 Garbage In, Garbage Out . . 8
3.5 Garbage In, New Perspective Out? 8
3.6 The Impossibility of Fairness 8
3.7 What Are We Prediciting Again? 9
3.8 The Perverse Incentives of Data Scientists: Job Security by Obscurity 9
3.9 Humans Love Computers . . 9

3.1 Concept Drift

3.2 Theories of Creativity

Is creativity combining exsiting things, or is it coming up with something new?

Does progress slow down becasue we heavily rely (more than usual) on the work of the past to generate future work?

3.3 From Spaghetti Code to Lasagna Neuron Layers

These layers are totally transparent, but you can't understand them because they're complicated :)

3.4 Garbage In, Garbage Out

You are essentially programming with data, so if your data sucks so will your prediction, you also really can't generalize, only correllate.

3.5 Garbage In, New Perspective Out?

What about cross-domain models, where maybe I train with a poetry dataset, and point my language model on nonfiction.... hmmm.

3.6 The Impossibility of Fairness

Representation, "fixing the training set" [6], or the Impossibility of Fairness from a model.

[6]: Christian (2020), *The Alignment Problem: Machine Learning and Human Values*

3.7 What Are We Prediciting Again?

are you predicting the right thing? Are you really predicting how valuable the company is or just whether it'll be the next meme stock?

3.8 The Perverse Incentives of Data Scientists: Job Security by Obscurity

Investors want predictions, Data Scientists don't have good data, but want to keep their jobs. Bad science rolls downhill to the user, and takes a long time to snuff out.

3.9 Humans Love Computers

Working together seems like a good idea, but how. Talk about 2 percent model control and model uptake.

Also talk about Kasparov's tournament. [7]

[7]: Mansharamani (2020), *Think for yourself: Restoring common sense in an age of experts and artificial intelligence*

Classification and Profiling: Studies in AdTech and Policing 4

4.1 Everything is Clasification . 10
4.2 Classification in Reverse; Generating the Stereotype . 10
4.3 Code Your Own Classifier in 5 Minutes! 10
4.4 The End of Nuance 10
4.5 If You Were Unable To See A Doctor, is it OK? 10
4.6 Lazy Thinking 10
4.7 Data Where There Is None . 10

4.1 Everything is Clasification

4.2 Classification in Reverse; Generating the Stereotype

4.3 Code Your Own Classifier in 5 Minutes!

Beginners tutorial

4.4 The End of Nuance

The unclassified?

4.5 If You Were Unable To See A Doctor, is it OK?

Increasing access of skin cancer detection.

4.6 Lazy Thinking

Online Advertising, Justice, Job Applications, Creditworthiness, Getting Insurance (Weapons of Math Destruction), Civic Life, /sideciteOneil2017 ; The Default Male, Invisible Women effects snow clearing schedules and drug discovery

4.7 Data Where There Is None

Talk about how generative tools can create data to be classified (like LAION).

Self-Driving With Statistics 5

5.1 Trolley Problems 11
5.2 The Third Rail 11
5.3 Purple Lights and Changes In Fashion 11
5.4 Multicolinearity (Reprise) . 11
5.5 See You In Court, Asshole! . 11

5.1 Trolley Problems

5.2 The Third Rail

5.3 Purple Lights and Changes In Fashion

5.4 Multicolinearity (Reprise)

5.5 See You In Court, Asshole!

Generating Art: Avante-Garde or Derivative? 6

6.1 Predictive Keyboards 12
6.2 GPT On A Napkin 12
6.3 Compress The World's Achievements Into One Database 12
6.4 Upscaling What Is Lost . . . 12
6.5 Who Owns This Stuff Anyway? 12
6.6 Use Critical Thinking to Become A Critic 12

6.1 Predictive Keyboards

6.2 GPT On A Napkin

Discuss this [1]

6.3 Compress The World's Achievements Into One Database

Lossy lookup of data, see how close you get to creating famous artworks or wikipedia pages?

Talk about AI upscaling and AI as lossy compression. we stored the entire corpus of human knowledge, but AI retrieves it in a silly way.

6.4 Upscaling What Is Lost

How to leverage this stuff.

6.5 Who Owns This Stuff Anyway?

Discuss FSF and Copilot, lawsuits.

6.6 Use Critical Thinking to Become A Critic

Talk about ChatGPT, deterministic vs probabalistic and Thomas Hobbes [2]

[1] GPT On A Napkin

[2] AI Homework

Revolutionary for Whom? 7

"The inhabitant of London could order by telephone, sipping his morning tea in bed, the various products of the whole earth – he could at the same time and by the same means adventure his wealth in the natural resources and new enterprise of any quarter of the world – he could secure forthwith, if he wished, cheap and comfortable means of transit to any country or climate without passport or other formality." John Maynard Keynes, 1920 [8]

7.1 Self-Driving Horses 13
7.2 The Battle of the Assistants . 13
7.3 Employees That Are Better Than You 13
7.4 Who is Helped the Most? . . 13
7.5 Who is Hurt the Most? 13
7.6 How to Respond 13
7.7 This Book is a Case Study . . 13

7.1 Self-Driving Horses

7.2 The Battle of the Assistants

Butler vs Indian Virtual Assistant vs Siri

7.3 Employees That Are Better Than You

Respond directly to Jon Krohn's TED talk about monkeys being dumber than us... what about construction equipmenth that's stronger than us physically, or racism/eugenics people that are dumber than us [9]

[9]: (2022), *Jon Krohn*

7.4 Who is Helped the Most?

7.5 Who is Hurt the Most?

7.6 How to Respond

7.7 This Book is a Case Study

Unplugging Skynet 8

"The second requirement of goal-misalignment risk is that an intelligent machine can commandeer the Earth's resources to pursue its goals, or in other ways prevent us from stopping it... We have similar concerns with humans. This is why no single person or entity can control the entire internet and why we require multiple people to launch a nuclear missile. Intelligent machines will not develop misaligned goals unless we go to great lengths to endow them with that ability. Even if they did, no machine can commandeer the world's resources unless we let it. We don't let a single human, or even a small number of humans, control the world's resources. We need to be similarly careful with machines." Jeff Hawkins, 2022 [3]

8.1 AI and Human Safety 14
8.2 Useful Incompatability . . . 14
8.3 Checks and Balances 14
8.4 When You Can't Tell The Difference 14
8.5 Sucker Traps 14
8.6 Dead Inside 14

8.1 AI and Human Safety

8.2 Useful Incompatability

It's a feature not a bug that no single comupter can control all others, AKA Security by Obscurity

8.3 Checks and Balances

8.4 When You Can't Tell The Difference

Talk about Taleb's aphorisms "Another definition of modernity: conversations can be more and more completely reconstructed with clips from other conversations taking place at the same time on the planet.", "You are alive in inverse proportion to the density of cliches in your writing."

8.5 Sucker Traps

"The sucker's trap is when you focus on what you know and what others don't know, rather than the reverse."

8.6 Dead Inside

"If you know, in the morning, what your day looks like with any precision you are a little bit dead - the more precision the more dead you are."

Errors and Omissions 9

Bibliography

Here are the references in citation order.

[1] John von Neumann and Ray Kurzweil. *The Computer and the Brain (The Silliman Memorial Lectures Series)*. New Haven, CT, USA: Yale University Press, Aug. 2012 (cited on page ii).

[2] Elaine Rich, Kevin Knight, and Shivashankar B. Nair. *Artificial Intelligence*. Tata McGraw-Hill, 2009 (cited on page 1).

[3] Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Basic Books, 2022 (cited on pages 7, 14).

[4] Franc ois Chollet. *Deep learning with python, second edition*. Manning Publications, 2022 (cited on page 7).

[5] David McCullough. *Truman*. Riverside, NJ, USA: Simon & Schuster, June 1992 (cited on page 8).

[6] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. New York, NY, USA: W. W. Norton & Company, Oct. 2020 (cited on page 8).

[7] Vikram Mansharamani. *Think for yourself: Restoring common sense in an age of experts and artificial intelligence*. Harvard Business Review Press, 2020 (cited on page 9).

[8] John Maynard Keynes, Elizabeth Johnson, and Donald Moggridge. *The Collected Writings of John Maynard Keynes (Volume 5)*. Cambridge, England, UK: Cambridge University Press, Dec. 2012 (cited on page 13).

[9] *Jon Krohn*. [Online; accessed 18. Oct. 2022]. Oct. 2022. URL: `https://www.jonkrohn.com/posts/2022/10/7/tedx-talk-how-neuroscience-inspires-ai-breakthroughs-that-will-change-the-world` (cited on page 13).