Full Self-Driving Skynet, and Other Artificial Intelligence Myths

The realities of decision making with deep machine learning models

Brad Flaugher

November 28, 2022

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 a work in progress, I hope it helps demystify the world of deep learning as I understand it.

Humans won't be able to control superintelligent AI, talk about that here[2]

Talk about Bostrom and GPAI here, and Erdi's answer to that. [3] [4]

Talk about the alignment problem and Ethical freakouts about AI. Talk about the big 3 from [5] [6]

Funding and startups, everybody is doing it, I'm trying to make sense of it

Brad Flaugher

Contents

List of Figures

List of Tables

List of Listings

Playing chess in 1997 | 1

1.1 My baby computer

If you woke up as a computer science student in 1984, maybe you grabbed the latest textbook from Dr Elaine Rich on artificial intelligence[7] [1] you would have certain ideas about how to teach computers how to do things.

Like young parents [2] maybe you thought the best way to teach your (digital) offspring how to learn was to start with the basics, the way you did as a kid. Say dada! Then let's teach you what a verb and a noun are, then maybe we will be able to translate english to french one day, right? In between lessons let me teach you how to play chess, but to do that let me teach you how the pieces move, and then let me beat you a couple hundred times before bedtime.

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

1: the book is now in its third edition and unlikely to be updated as Dr. Rich as retired. sad.

2: including your author

1.2 say "Hello, World!", little computer

Noam Chomsky [3] says that lanugage evolves over time, but may have an underlying universal structure or something [4]. For us English speakers, or especially 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 [5]. 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.

Folks have tried this and it's been tough going for them, notable projects include...

3: wikipedia.org

4: find a citation, dude

5: for a weekly dose of language nerdery go here https://www.economist.com/johnson

1.3 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.[6] 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.

6: 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.

```
def get_days(month):
    if month in ["Jan"]:
        return 31
    elif month in ["Feb"]
        if leap_year:
            return 29
        else:
            return 28
    elif month in ["Nov"]:
        return 30
    else:
        throw exception "I don't understand"
```

You can see how annoying programming this stuff is alreday, what about numerical months and other shit...?

1.4 Deep Blue brutally defeats Gary Kasparov

This was a win, but it did so by brute force, this is cute and a fine way to solve a chess problem, most of life is not like chess... so most problems are ill-suited to be explicitly coded in this manner, we need something else.

1.5 Meanwhile at the Post Office

[7]

7: check out Yann LeCun domonstrating a convolutional neural network hereyoutube.com

1.6 Artificial Intelligence vs Programmer Intelligence vs Data Intelligence

TODO talk about this book [8]

[8]: Douthat (2022), *Can We Resist the Age of the Algorithm?*

2 Self-driving with statistics

Figure 2.1: The Mona Lisa. https://commons.wikimedia.org/wiki/File:Mona_Lisa,_by_Leonardo_da_Vinci,_from_C2RMF_retouched.jpg

Hardware got amazing, we gave up teaching the way we teach ourselves and let the data do the work

We leveraged huge statistical models to regress our way to success

We used building blocks of regression and neurons to train huge models

These models are statistical and deterministic, but ultimately chaotic black boxes..

TODO talk about these books [9] [10] [11] [12]

A Shakespearean comedy of numbers

3

It's all numbers man!!!!

3.1 Knowlede Representation, or not

This is some text and a link to Hey if you want to site something on the side use[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 beloive any kind of deep learning network will achieve the goal of AGI 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." [13]

```
cd myproject
docker run tensorflow
#profit!
```

`tex.stackexchange.org` for help.

[2]: Andreu et al. (2021), *Humans won't be able to control a superintelligent AI, according to a study*

Derivative artworks of the future? 4

GPT-3, BERT and Bloom

Link some cool shit here, Draw Owl!

Who owns this shit anyway? Copilot and FSF plus lawsuits

Prove it, asshole!

I can make a "model" that behaves like a database. Just memorizes shit

5 The data is the hardest part

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

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?

"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 [14]

[14]: McCullough (1992), *Truman*

Representation, "fixing the training set" [5], or the Impossibility of Fairness from a model.

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

TODO talk about these books [15] [16] [6] [5]

"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 singer 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." [13]

```
cd myproject
docker run tensorflow
#profit!
```

`tex.stackexchange.org` for help.

The police and Big Tech are profiling me!

6

Classification is everywhere, it's also very useful. Just get over it.

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

```
cd myproject
docker run tensorflow
#profit!
```

tex.stackexchange.org for help.

The useful chaos of spaghetti code 7

Interacting Layers of Statistical Understanding Useful Chaos

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

You can't understand ML layers, but they are useful nonetheless.

8 Self-stabilizing concept drift

Many models need to constantly be retrained

Does progress slow down because we keep reusing the work of the past to generate our work?

My horse drives itself, thanks!

9

Can Models bring Incremental or Revolutionary Change?

Respond directly to Jon Krohn's TED talk about monkeys being dumber than us... what about construction equipmenth that's stronger than us, or racism/eugenics people that are dumber than us [17]

[17]: (2022), *Jon Krohn*

"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 [18]

[18]: Keynes et al. (2012), *The Collected Writings of John Maynard Keynes (Volume 5)*

Who is affected the most?

What should individuals do?

What should governments do?

What should businesses do?

Skynet: did you try unplugging it? 10

Appendix

A

????

Let's say we want to build an ensemble model to analyze poetry, put a haiku into craiyon's online shit, then we categorize the resulting photo. [2]

[2]: Andreu et al. (2021), *Humans won't be able to control a superintelligent AI, according to a study*

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] Abraham Andreu and Qayyah Moynihan. 'Humans won't be able to control a superintelligent AI, according to a study'. In: *Business Insider* (Sept. 24, 2021). (Visited on 09/24/2021) (cited on pages iii, 4, 13).

[3] Péter Érdi. *Ranking: The Unwritten Rules of the Social Game We All Play*. Oxford, England, UK: Oxford University Press, Oct. 2019 (cited on page iii).

[4] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 1st. USA: Oxford University Press, Inc., 2014 (cited on page iii).

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

[6] Reid Blackman. *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI*. Harvard Business Review Press, July 2022 (cited on pages iii, 6).

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

[8] Ross Douthat. 'Can We Resist the Age of the Algorithm?' In: *The New York Times* (July 30, 2022). (Visited on 07/30/2022) (cited on page 2).

[9] MacAskill2022. 'The Case for Longtermism'. In: *The New York Times* (Aug. 5, 2022). (Visited on 08/05/2021) (cited on page 3).

[10] Cade Metz. 'The Long Road to Driverless Trucks'. In: *N.Y. Times* (Sept. 2022) (cited on page 3).

[11] Cade Metz. 'Stuck on the Streets of San Francisco in a Driverless Car'. In: *N.Y. Times* (Sept. 2022) (cited on page 3).

[12] Caglar Aytekin. 'Neural Networks are Decision Trees'. In: (2022). DOI: 10.48550/ARXIV.2210.05189 (cited on page 3).

[13] Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Basic Books, 2022 (cited on pages 4, 6).

[14] David McCullough. *Truman*. Riverside, NJ, USA: Simon & Schuster, June 1992 (cited on page 6).

[15] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown, Sept. 2017 (cited on page 6).

[16] Caroline Criado Perez. *Invisible Women: Data Bias in a World Designed for Men*. New York, NY, USA: Abrams Press, Mar. 2019 (cited on page 6).

[17] *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 10).

[18] 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 10).

Notation

The next list describes several symbols that will be later used within the body of the document.

c	Speed of light in a vacuum inertial frame
h	Planck constant

Alphabetical Index