AI Ethics

Warning: values are inherently subjective. Some of what we will discuss today is either my personal opinion or it is intended to be provocative, to make you think.

While many of these topics can be emotional, it’s important to expose yourselves to them from time to time, because they affect you on a daily basis.

How can a machine learning model be ‘good’?

What types of problems does it solve?

- Solve medical problems - “solve cancer”
- Education: helps learn better
- Accelerate plant growth, in agriculture
- Discover new science!
- Maximize utility: benefit everyone
- Security and safety: monitoring everyone.
- Improve sales and maximize profits for your company:: MORE MONEY
- Reduce congestion in traffic
- Entertainment suggestions. Generate movies. based on user preference.
- Less effort in repetitive tasks.
- Solve agricultural and pests problems.
- Restoring art and historical archives.
- Translation
- Maximize performance in sports
- Finance: manage markets
- Natural disaster management

How can a machine learning model be ‘bad’?

What types of problems does it create?

- Emotional and intellectual dependence
- Kill creativity and intelligence
- Bias towards certain groups of people
- Efficient destruction: war, drones.
- Try to please everyone
- Privacy: make it easier to monitor people
- Propaganda
- Replacing humans: jobs, relationships
- Environmental impact
- Overfitting
- Hard to trust anything online: deep fake, fake news, fake pictures.

Why build AI systems?

Why build any technology?

- Who are we building for? For ‘us’ or for ‘others’?
  - Profit
  - Fame
  - Control

The World is a Complex System

Solving real-world problems isn't actually 'possible'

- Good solutions to problems can have unintended consequences: the world is chaotic.
  - Example: automating a factory with AI. What are the consequences?
- My hypothesis: just like 'all models are wrong but some are useful':

All technologies are destructive, but some are beneficial

- In a complex world, the line between destruction and utility is drawn by context —a line that's often blurred and rarely static.
- Real solutions look very different from the ones you get an "A" on. (ARGD 060)

“All animals are equal,
but some are more equal than others”

Social systems are hierarchical

- **Who** controls advanced technology? And what do they use it for?
- Who will you work for? And what type of technology will you develop?
- Do you have to “choose your evil”?
- Technologies are often used by the powerful to control the weak, but they can also provide opportunities for the weak to become powerful.
- Is there a trade-off between power and autonomy?

The Alignment Problem

What if the values of an AI agent are not aligned with ours?

- Who ‘us’?
- Different people have different values? Who decides what are the values an AI should adopt?
- Assuming we agree on the values, can we constrain an autonomous ‘super-intelligence’ to follow our rules?

The Environmental Problem

What's the carbon footprint of machine learning models?

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

https://arxiv.org/pdf/1906.02243

The ‘Replacement’ Problem

If AI can do our job, what do we do?

- Universal Basic Income?
- Argument: technologies don’t replace us, they require us to use them to solve the same tasks faster and more effectively. They only replace those who don’t know how to use them.
- Counter-Argument: this time it’s different. AI is better than us at everything.
- Are AI and capitalism in its current form compatible? Does productivity have to be the means by which we justify our survival and comfort?

The Fairness and Bias Problem

AI is trained on biased data, and will make biased decisions

- Since most data online comes from the privileged part of society that has access to information outlets (internet, news, etc.), a model trained on that data will embed biases from that subset of society; often against the less privileged part of society.

Where do we go from here?

My 3 takeaways

- Keep questioning assumptions in perpetual discomfort of not having a ground truth. That’s true for both ethical and technical aspects.

- Find **real problems** of **real people** in **real need** and solve for their **context**. It doesn’t have to be cutting edge technology. Keep it simple and effective. Don’t settle on the first solution you find. If you find a solution, there’s probably a better one, keep looking. Build, look, re-build, improve, question, re-build, look,…

- Remember that we live in a society: we build for others. It’s good to know who you’re building for. In an interconnected world, that’s often impossible to do; but it doesn’t mean that we should give up on cultivating that awareness.