Data Science Ethics

These are my notes about data science ethics, an area I consider important especially when stu groundbreaking technologies like machine learning.

These notes are from the FastAI bonus chapter on ethics, specifically this YouTube Video: Ethic Data Science.

This notebook focuses on current ethical dilemmas surrounding data science - not future ethical problems. While technology has a done a lot of good for the world, it has and continues to cause significant harm.

What is Ethics?

> Ethics is the discipline dealing with what is good and bad; a set of moral principles.

Ethics is not fixed, like religion or law.

. It is a set of well-founded standards of right and wrong, prescribing what humans ought to d
. It is the study and development of one's ethical standards.

Let's have a look at two ethical philosophies and how they can be applied to our projects:

. **Consequentialism/utiliatrianism** - maximising good.

- Who will be directly and indirectly affected by the project?
- Will the effects in aggregate create more good than harm?
- Are we thinking about all types of harm/benefit?
  - psychological
  - political
  - environmental
  - moral
  - cognitive
  - emotional
  - institutional
  - cultural
- Do the risks of harm/benefit fall disproportionately on the least/most powerful in society?
- Have we considered **dual-use** - i.e., could the project also be used for harm?

. **Deontologicalism** - adhering to the 'right'.

- What rights of others and duties to others must we respect?
- How might the dignity and autonomy of each stakeholder be impacted by this project?
- What considerations of trust & justice are relevant to this project?
- Does this project involve any conflicting stakeholder rights? How can they be prioritised?

We can also look at ethics in AI through ethical lenses. It is not necessary to choose an e philosophy and live by it, but to instead consider our project in as many ways as possible.

- **The rights approach**: Which option best respects the rights of those who have a stake?
- **The justice approach**: Which option treats people equally?
- **The utilitarian approach**: Which option will produce the most good and do the least harm?
- **The common good approach**: Which option best serves the community as a whole, not jus members?
- **The virtue approach**: With option leads me to act as the sort of person I want to be?

Lenses and are deontological, while and and consequentialist.

It's also good to note that this discussion so far has been a very *western* view of ethics and mora Their are other worldviews to consider, and when implementing a project, it is important to consi cultural and ethical lenses of the people who have a stake.

Why is ethics in data science important?

Data collection has played a pivotal role in genocides, including the holocaust.

IBM used data science to decide whether people were Jewish and whether they should be execu They produced computer systems that were used in concentration camps and gas chambers - th machines required constant maintenance and an ongoing relationship between user and vendor

This is an important reminder how technology can be used for harm and why it is critical that eth considered when using technology.

How is speed/hypergrowth related to data ethics?

- Super-fast growth requries automation & reliance on metrics.
- Prioritising speed above all else doesn't leave time to reflect on ethics.
- Problems happen or surface on a large scale if the company grows too quickly.

Metrics

> Reliance on metrics is a fundamental challenge for AI.

Choosing appropriate metrics is very important when building an AI model, as deep learning is v effective at optimising metrics. While this is the strength of deep learning, it is also a fundamenta challenge, as inappropriate metrics can have a devastating impact.

Overemphasising metrics can lead to:

- manipulation
- gaming
- focus on short-term goals
- unexpected negative consequences

Let's have a look at an example, from the s when the UK started focusing intensely or numbers to improve performance in the healthcare system. This project was called "*What's mea what matters*", link to study here.

One of the metrics was around emergency department (ED) wait times. By **overemphasising** th metric, the following issues occured:

- Scheduled operations were cancelled to draft extra staff to ED.
- Patients were required to wait in queues of ambulances.
- Stretchers were turned into "beds" by putting them in hallways.
- There were big discrepancies in numbers reported by hospitals vs by patients.

This is an example of **gaming** occuring due to metrics. The healthcare system did not actually in anything it got worse, as the people in charge were manipulating processes to optimise a single

An essay grading software had similar issues in America. Metrics for grading an essay included length, vocabulary, spelling, subject-verb agreement - because these are metrics that are easy t measure. Therefore, it was not possible for the software to measure hard-to-quantify qualitities, l creativity.

As such, gibberish essays with sophisticated words scored best - an example of poorly chosen r not representative of what a **good** submission **actually** looks like.

Goodhart's Law is an important reminder why not to over-rely on metrics:

> "When a measure becomes a target, it ceases to be a good measure."

A metric is just a proxy for what you care about - and it turns out, its not so easy to measure **wha care about**.

Feedback Loops

Our online environments are susceptible to feedback loops, "when your model is controlling the round of data you get. The data that is returned quickly becomes flawed by the software itself." I referred to as **echo chambers**, particularly in social media.

For example, recommendation systems use watch time as a **proxy** for how interested we are in something. This leads to conspiracy content performing well, as it encourages it viewers to keep "uncovering" more "information". This was not an intended consequence when recommendation algorithms were originally built, but an unintended consequence now widely exploited.

Our online environments are designed to be addictive and content creators are always trying to metrics to improve their performance. This makes choosing appropriate metrics even harder.

Feedback loops are common with recommendation systems, as they return what the user likes.. what they are exposed to. It can reinforce and recommend damaging videos/articles/images etc. see this through Meta's role in the Rohingya genocide.

A good quote from James Grimmelman:

> "These platforms are structurally at war with themselves. The same characteristics that make outrageous & offensive content unaccpetable are what make it go viral in the first place."

In this way, disinformation is built into modern tech companies and into their business models.

Bias

Gender Bias

Commercial computer vision products perform significantly better on men and on white people a perform very poorly on women of colour. This research was conducted on several large commer products and they all showed this significant bias.

What is the source of this problem?

- Generally, unrepresentative datasets which were primarily built on white men. When the ber contains bias, this will be perpetuated on a larger scale with machine learning, as the algorit optimise to this biased dataset.
- Blackbox algorithms can be trained on many variables and cannot be analysed to check for
- Generally, bias in technology is sourced from bias in real-life - but, it has the potential to am especially if algorithms are trained to optimise biased metrics or benchmarks.

Historical Bias

Historical bias is:

> "a fundamental, structural issue with the first step of the data generation process and car exist even given perfect sampling and feature selection."

An example of this is with the COMPAS recidivism algorithm used in the US to predict whether s will re-offend to decide if they should pay bail. This algorithm was found to not only be supremely but also to be no more effective than guessing. It was upheld even after extensive research demonstrating its flaws.

Measurement Bias

Measurement bias is:

> when data collection methods systematically distort the true values of what is being measured.

An example of this is in this paper: Does Machine Learning Automate Moral Hazard and Error? T paper discusses an algorithm suggested to predict a person's risk of stroke to improve efficiency ED. What they found was that a number of irrelevant factors where most predictive of stroke, like "accidental injury" and "colonoscopy".

Why is this? The researchers hadn't measured the chance of stroke, but the chance someone ha symptoms, went to the doctor, got tests and recieved a diagnosis. And this is influenced by **MAN** factors then just the chance of stroke, including: race, class, gender, and health insurance.

Racial Bias

Humans are very biased, see these researched and peer reviewed examples of racial bias:

> "When doctors were shown patient histories and asked to make judgments about heart disease, they were much less likely to recommend cardiac catheterization (a helpful procedure) to black patients"

> "When whites and blacks were sent to bargain for a used car, blacks were offered initial prices roughly $ higher, and they received far smaller concessions."

If humans are biased, why does algorithmic bias matter?

. **Machine learning can amplify bias -** Bias in bios showed that the gender imbalance in me was amplified and made even worse when asking an algorithm to predict a person's job title
. **Algorithms are used differently than human decision makers -** people are more likely tc algorithms are objective, algorithms are more likely to implemented with no appeals process algorithms are often used at scale, and algorithmic systems are cheap.
. **Machine learning can create feedback loops**.
. **Technology is power. And that comes with responsibility**.

Disinformation

Disinformation is:

> false or misleading information that is deliberately created and spread to decieve people.

Disinformation can include so-called "fake news", where a single article or blog post is labelled a incorrect. However, on a larger scale, disinformation includes orchestrated campaigns of manipu

AI can be used to generate compelling but false information that can be dispelled at a large scal be subtle, even involving language models arguing with each other, where one slightly tak upper hand. False profiles for people can be generated, who appear to be reliable or professiona

In , it is estimated that nearly million out of million comments decidin neutrality laws in the US were fabricated, mostly by internet providers, source. This was prior to advent of generative AI, demonstrating the real impact of computer-generated content on our lav is expected to worsen unless legislative action is taken.

As humans, we've evolved to generate our opinions based on our in-group and to disagree with out-group. AI has the potential to amplify these differences and spread false and misleading info on a very large scale.

AI allows us to make forgery convincing, inexpensive, and automated. Solutions such as digital signatures have been suggested to address these concerns. It has also been suggested that **disinformation needs to be treated a cybersecurity problem.**

Diversity

Machine learning research is not very diverse (source), particularly for women and people of col

An important statistic from this article (from) is that % of women working in tech leaving, compared to % of men. Increasing the number of women learning coding and goi tech is not going to fix this problem.

It's important to have diversity on teams. The first female engineer at Quora implemented a 'bloc feature - something that otherwise would not have been implemented by her male colleagues. T important reminder why diverse experiences are valuable to a project, to a company, and to soci

What can we do as engineers?

- Vet the company you're joining for their ethics. We normally have lots of options and we can skill as leverage.
- While pressure from management might give us some leeway for unethical behaviour, it is i to be personally accountable for our actions and the harms we can cause.
- Talk to experts **and** people directly impacted by technology. Get feedback before and after r
- Ask yourself questions:
  - *Should we even be doing this?* Not everything that **can** be done **should** be done.
  - *What bias is in the data?* There will **always** be some level of bias in the data but it is im to understand how it is collected, etc.
  - *Can the code and data be audited?* Propietary black boxes can be very damaging and to trace, monitor, and evaluate.
  - *What are error rates for different sub-groups?* Are certain groups of data underperformi
  - *What is the accruacy of a simple rule-based alternative?* Have a baseline - if your comp model does not outperform the baseline, why are we even doing this?
  - *What processes are in place to handle appeals or mistakes?* There will be problems an important to have a robust system for them to be identified and rectified.
- Implement some tools:
  - **Ethical risk sweeping**: Implement regular ethical risk-sweeping - just as you would pe cybersecurity penetration testing (regardless of whether you find something), we should same with ethical risks. Assume you missed risks in initial development and reward pe spotting them.
  - **Expanding the ethical circle**: Whose interests, skills, experiences, and values have w assumed, rather than consulted? Who will be indirectly affected in significant ways? Wh use this product for an unexpected purpose?
  - **Think about the terrible people**: Who will want to abuse, steal, misinterpert, hack, de weaponise what we've built? What rewards/openings has our design inadvertently crea

- **Closing the loop**: Remember that ethical design is never finished. Identify feedback ch for ethical impact. Develop formal procedures and chains of responibility for ethical itera

Conclusion

While learning about machine learning and AI, I've been focusing on its applications for good - a there's no doubt that AI has the potential to drastically improve our world and to make society mo equal.

However, its important to look at its implications. Like all new technologies, AI has its drawbacks and misuse potential. This short ethics seminar was an important reminder that our powerful tec companies need to be help account, and that the financial incentive should be for good... not a ra the bottom. There needs to be significant legislative reformation to ensure AI moves the world fo

However, while ethics is a never-ending journey, I feel equipped with some knowledge and certa some awareness of the implications of the technology I'm studying. This will be an area of contin research for me, as I want to contribute my technological skills for good in the world.

And, while the motivations and actions of certain tech companies seems dire and bleak, innovati always been ahead of legislation. Its important as data scientists and engineers to advocate for e a personal, company and legislative level.

When cars were first invented, safety features were sparse as there was no financial incentive to them (even though they existed). With time, advocacy, and legislation, drastic changes have bee to improve car safety standards. And the common line used in the past was that "the people who cars are dangerous, not the cars themselves". **This is obviously misleading and irrelevant, ye enduring sentiment in innovation** It's important that we hold ourselves and our companies to a as we have a great responsibility to ensure technology is used for good.