Ethical AI
Be Mindful of Our Algorithms

Infotec
May 27, 2021

By Aaron Grothe

Introduction

Ethical AI?

"Replicants are like any other machine, are either a benefit or a hazard. If they're a benefit it's not my problem" - Blade Runner

One of the issues with AI is we can't predict when it will have a leap or bound

E.g. it was assumed it would be 2040 or 2050 before a program would be able to beat a Go master - it happened instead in 2016

Disclaimer

We'll be talking about some potentially thorny issues today. The goal here is to raise awareness. To be more mindful of things. To pause and make hopefully more inclusive decisions.

AI and ML systems are not inherently anything. They are just used to amplify our information.

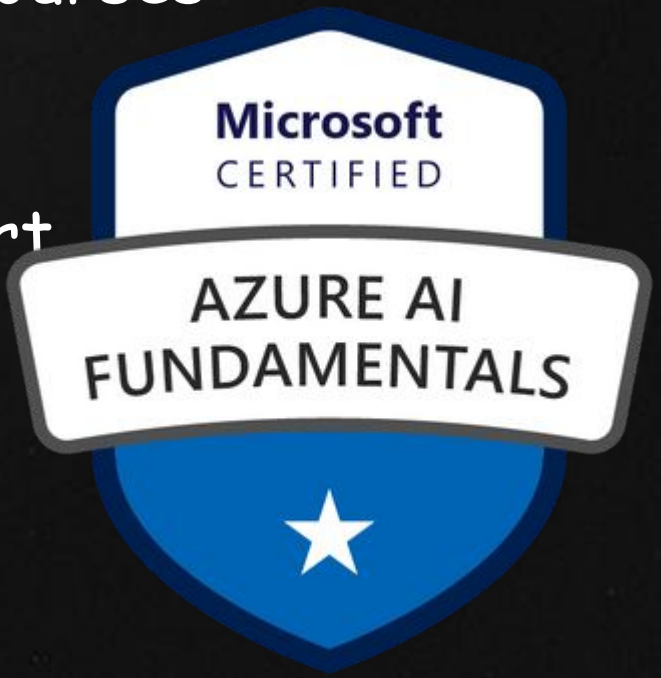
While I've been programming for 30+ years this is a field I'm still figuring it out as well.

Background

Got started doing research into AI/ML when I was looking into the ability to do anomaly detection for logs. Came at it from a Infosec perspective.

Bought some books, a bunch of Udemy courses and read a lot of articles on the web

Also got the Azure AI Fundamentals Cert



Overview

- Bad AI/ML examples
- Ethical AI Guidelines
- Common elements of Ethical AI
- What can you do?
- How to get AI Certified for Free

Bad AI/ML Examples

We'll look at four bad AI/ML examples

- Genderify
- Amazon Hiring Tool
- Microsoft Tay
- Lemonade Insurance

Genderify

Genderify WAS a service using AI to identify gender based on names.

What could go wrong?

- Use Meghan Smith came back with an assessment of 39.60% male, 60.40% female.
- Use Dr. Meghan Smith came back 75.90% male, 24.10% female
- Use Mrs. Joan Smith - comes back 94.10% male

THAT is what could go wrong.

Genderify.

Genderify was a real service by a real company with an API documentation/pricing, etc.

Was released on producthunt.com which is a y-combinator type of company for new products.

How did it go so wrong?

- Obviously not enough testing
- Bad training, bad data, bad modeling or "D) all of the above"

Genderify.

We might be coming back to this after we talk about some of the ethical AI guidelines as an example.

Reference:
https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify

Amazon Hiring Tool

Amazon in 2014 - created a process to help automate the company's head hunting process

The idea is that you'd be able to give the system 100 resumes and it would give you the top 5. You would then pursue those potential applicants

After analysis in 2015 it was discovered that the system was giving a preference to male applicants.

Amazon Hiring Tool

Successful resumes were considered "male" resumes. Largely because of past male bias in hiring.

References to female topics such as "Woman's Chess Club" or other references to "Women" were downgraded.

Preference was given to "macho" verbs such as "captured", "dominated" and others.

Amazon shuttered the effort in 2017.

Amazon Hiring Tool

What this brings up is that you can be encoding human biases into your systems without awareness and automating them into a system that might not be challenged.

How do you grade an applicant if you're looking for a simple number to include diversity, whether an applicant is suitable or so on.

"I am not a number" - The Prisoner

Amazon Hiring Tool.

References

Reuters - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Slate - https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html

Microsoft Tay

Microsoft created an AI chatbot named Tay to develop conversational understanding with humans

Tay was supposed to be a fun chatbot on twitter that you could send tweets to and it would imitate the language and get better at conversations.

Within 24 hours Tay went from “can I just say that im stoked to meet u? Humans are super cool” to “Hitler was right I hate the jews”

There was an intentional effort by the community to get the system to say hateful things.

Microsoft Tay.

Internet community discovered a debugging phrase “repeat after me” which was able to add new responses to the system.

Microsoft deleted the account and all of its tweets.

References

Register - https://www.theregister.com/2016/03/24/microsoft_ai_goes_troll/

CBSNews - https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

Lemonade Insurance

Lemonade Insurance has decided to retract claims about how they were using AI to study videos that customers upload to see if they could tell they were fraudulent.

When you file a claim with them you upload a video explaining the issue. The system then does analysis of it and decides whether to pay the claim in minutes.

"Our AI carefully analyzes these videos for signs of fraud. It can pick up non-verbal cues that traditional insurers can't, since they don't use a digital claims process."

Lemonade Insurance

After some intense feedback. An updated statement came back from the company saying they only used Facial recognition to make sure that a person wasn't filing multiple claims.

The company does say it collects physical characteristics when handling life insurance.

Lemonade statement from Twitter

Our systems don't evaluate claims based on background, gender, appearance, skin tone, disability, or any physical characteristic (nor do we evaluate any of these by proxy)

Lemonade Insurance.

Why is this bad?

Keep this example in mind and try to figure out how many of the standards from Microsoft and the FTC this seems to violate.

Is this a case of a company trying to use AI in marketing to differentiate itself or something else?

Sources:
https://www.theregister.com/2021/05/26/ai_insurance_lemonade/

https://twitter.com/Lemonade_Inc/status/1397564445648424965?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembe

Bad AI/ML Examples

So that was just a few bad examples.

There are many, many more examples that are public.

How many private ones are there out there affecting our daily lives we are not aware of? AI in things like mortgages, salaries offers, employment, resume/hiring analysis, adoption, medical care, new product development and so on.

Ethical AI Guidelines

With that in mind lets take a look at some of the Ethical AI Guidelines that some companies/groups have released.

We'll be using the following for our examples

- Microsoft
- Google
- FTC

Some cloud companies don't have official AI statements. Most do and I suggest you consult with any vendor who's AI services you use.

Microsoft's AI Principles

6 Rules

Microsoft's rules are a pretty good base. They look to be pretty complete :-)

Reference

https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

Microsoft's AI Principles

6 Rules

- Fairness
- Reliability & Safety
- Privacy & Security
- Inclusiveness
- Transparency
- Accountability

Microsoft's AI Principles - Fairness

"AI systems should treat all people fairly"

- AI systems can behave unfairly because of societal biases reflected in the datasets
- AI systems can behave unfairly because of societal biases explicitly or implicitly reflected in the decisions made by the team during AI development and deployment lifecycle

Fairness - AI Fairness Checklist - https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/

Microsoft's AI Principles - Fairness.

Potential example of bias in training data

If you were to go back 30 years in going over mortgage applications and acceptance/rejection rates. Correlate approval/rejection rates based on names and you might be encoding an implicit bias that existed in loan applications in the past.

Without being aware of it you could be reinforcing a bias based upon your dataset or how you process them.

Microsoft has a tool Fairlearn that is open source to help "debias" systems

Microsoft's AI Principles - Reliability & Safety

"AI systems should perform reliably and safely"

System needs perform as originally designed and be able to safely respond to new situations.

AI systems can degrade over time. Robust monitoring/tracking processes need to be created to measure model's performance over time evaluate when it is necessary to retrain, update training or create new models.

Self-Driving cars are a prime example.

Microsoft's AI Principles - Reliability & Safety.

An example of this can be systems that have been provided data sets that do not reflect enough cases.

E.g. You have a SCADA system that is controlling batteries and power storage for a solar energy generation system. If the batteries decreasing capacity over time is not taken into account over time. The batteries could be overcharged and pose a safety risk.

There is a life cycle for AI solutions that needs to be taken into account.

Microsoft's AI Principles - Privacy & Security.

"AI systems should be secure and respect privacy"

AI systems ideally should run analysis locally. E.g. you should do the majority of the processing on a local device such as a phone and then upload the anonymized data to the AI system.

AI systems by default have no concept of privacy. Data leakage can be an issue.

Microsoft has open sourced a tool called Counterfit which is on GitHub. It generates adversarial models which can be used for testing.

Microsoft's AI Principles - Inclusiveness

"AI systems should empower everyone and engage people"

Does your AI system take into account people with different physical capabilities? People with reduced vision, differently abled and so on.

Are your datasets diverse enough?

Is your training model broad enough?

How are you measuring success/failure in training?

Microsoft's AI Principles - Inclusiveness.

One of Microsoft's statements on inclusiveness is "Design for the 3% to include the other 97%"

Does your facial recognition include a wide set of people for analysis?

Does your Speech to Text handle accented, generated voices, and speech impediments?

This can be a very tough thing for a small company to do when building their training models.

Microsoft's AI Principles - Transparency

"AI systems should be understandable"

When an AI system makes a decision it should be possible to understand how the system reached its conclusion.

If an AI system decides on a certain set of medical decisions for a patient how/why it reached those decisions should be transparent to their doctor.

Microsoft's AI Principles - Transparency.

Given the size of the training datasets this can be a challenge.

When a self-driving car causes an accident there will be a lot of questions about the decisions made by the AI that lead to the accident. There will have to be a way to figure out the decisions made.

Microsoft's AI Principles - Accountability.

"People should be accountable for AI systems"

Self-driving cars are a good example here. If there is an accident who is responsible? The Driver? The Company who made the car? And so on?

This will be figured out over time through the courts, insurance, laws, standards and so on.

The ability to do things like Gait analysis/Facial recognition/License plate tracking and so on you can build a heck of a police state.

Microsoft's AI Principles

So that is Microsoft's 6 Principles.

Provides a pretty good overview of some of the Ethical issues to be mindful of about AI

Google's Objectives for AI Applications

7 Rules

A lot of overlap with Microsoft's principles (safety, accountability, fairness, etc.)

Reference

https://ai.google/principles

Google's Objectives for AI Applications

Google's 7 Principles

1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

Google's Objectives for AI Applications

If you do a search for "google ai resignation" you'll get a lot of results. Google's AI group has had some changes.

FTC - Aiming for truth, fairness and equity in your company's use of AI

7 Guidelines

A lot of overlap with Microsoft's and Google's ethics (safety, accountability, fairness, etc.)

Reference

https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

FTC - Aiming for truth, fairness and equity in your company's use of AI

FTC Guidelines

1. Start with the right foundation.
2. Watch out for discriminatory outcomes.
3. Embrace transparency and independence.
4. Don't exaggerate what your algorithm can do or whether it can deliver fair or unbiased results.
5. Tell the truth about how you use data.
6. Do more good than harm.
7. Hold yourself accountable – or be ready for the FTC to do it for you

FTC - Aiming for truth, fairness and equity in your company's use of AI

Some of the areas that the FTC is getting involved

- Section 5 of the FTC Act. The FTC Act prohibits unfair or deceptive practices. That would include the sale or use of – for example – racially biased algorithms.
- Fair Credit Reporting Act. The FCRA comes into play in certain circumstances where an algorithm is used to deny people employment, housing, credit, insurance, or other benefits.
- Equal Credit Opportunity Act. The ECOA makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.

FTC - Aiming for truth, fairness and equity in your company's use of AI.

There are a lot of laws/regulations currently in effect that will have an impact on the use of AI.

Have to keep that in mind during your use of AI.

Ethical AI Guidelines

That is just three examples. Many other companies are drafting or have released their AI guidelines.

I think Microsoft's guidelines do a good job covering most of the bases.

What Can YOU do?

- Be aware of potential issues with AI
- Do play the Devil's advocate occasionally
- Consider the lifecycle of an AI system
- Ask questions
- Help raise awareness
- Consider tools like Counterfit for creating better test cases
- Consider using Cloud tools that have some vendors have already built into them

What Else Can YOU do?.

There are non-profit groups that work at making sure that EVERYBODY is represented in AI, consider joining/supporting one of them:

- WiNLP - Widening Natural Language Processing - http://www.winlp.org/
- QUEER in AI - https://sites.google.com/view/queer-in-ai/home
- Black in AI - https://sites.google.com/view/queer-in-ai/home

There are a lot of other groups as well. They are committed to raising awareness and increase inclusiveness.

How to Become AI-Certified

Thank you for staying this long :-)

As promised up next is the AI-Certification trick I mentioned at the beginning of the talk.

How to Become AI-Certified

I currently hold the following certification from Microsoft.

Let me tell you how to get the same certification for FREE!!!
Normally a $100 Value.

How to Become AI-Certified

Microsoft is still holding their Virtual Training Days

https://www.microsoft.com/en-us/trainingdays

Sign up for a "VIRTUAL: Microsoft Azure Virtual Training Day: AI Fundamentals"

It is 1-day and about 4 hours. After you attend you will get a voucher to sit for the AI-900 exam for Free.

Doesn't go into a great deal about ethics, but covers it a bit.

How to Become AI-Certified.

Just attending the class won't be enough to prepare you for the exam.

Microsoft has free online training for this as well:

https://docs.microsoft.com/en-us/learn/certifications/exams/ai-900?tab=tab-learning-paths

Between that and John Savill's guide to AI-900 - https://www.youtube.com/watch?v=E9aarWMLJw0 should do you a good job at preparing

I got like a 910 on my exam. You need 700 to pass :-)

Summary & Thanks

We are currently in a turbulent time in AI. We're figuring out what is and is not going to be allowed. What we should and what we shouldn't do.

It has tremendous promise and if we screw this up, we're going to seriously regret it.

Thank you for listening.