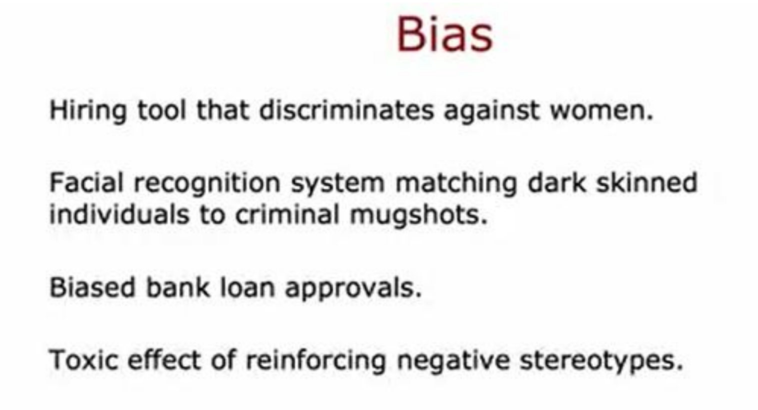
Fairness, impartiality and ethics

Machine learning algorithms today are affecting billions of people. You've heard me mention ethics in other videos before. And I hope that if you're building a machine learning system that affects people that you give some thought to making sure that your system is reasonably fair, reasonably free from bias. And that you're taking an ethical approach to your application. Let's take a look at some issues related to fairness, bias and ethics.

Unfortunately in the history of machine learning that happened a few systems, some widely publicized, that turned out to exhibit a completely unacceptable level of bias. For example, there was a hiring tool that was once shown to discriminate against women.

The company that built the system stopped using it, but one wishes that the system had never been rolled out in the first place. Or there was also well documented example of face recognition systems that match dark skinned individuals to criminal mug shots much more often than lighter skinned individuals. And clearly this is not acceptable and we should get better as a community at just not building and deploying systems with a problem like this.



Bias

Hiring tool that discriminates against women.

Facial recognition system matching dark skinned individuals to criminal mugshots.

Biased bank loan approvals.

Toxic effect of reinforcing negative stereotypes.

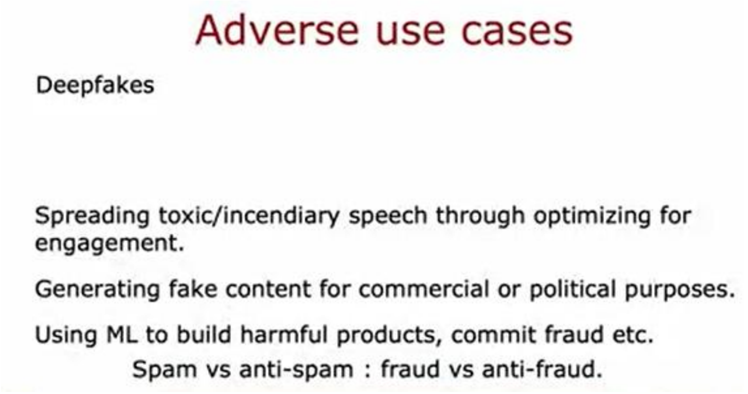
In the first place, there happens systems that gave bank loan approvals in a way that was biased and discriminated against subgroups. And we also really like learning algorithms to not have the toxic effect of reinforcing negative stereotypes. For example, I have a daughter and if she searches online for certain professions and doesn't see anyone that looks like her, I would hate for that to discourage her from taking on certain professions.

In addition to the issues of bias and fair treatment of individuals, there have also been adverse use cases or negative use cases of machine learning algorithms.

But I have killed the project just on ethical grounds because I think that even though the financial case will sound, I felt that it makes the world worse off and I just don't ever want to be involved in a project like that. Ethics is a very complicated and very rich subject that humanity has studied for at least a few 1000 years.

When aI became more widespread, I actually went and read up multiple books on philosophy and multiple books on ethics because I was naively hoping it turned out to come up with if only there's a checklist of five things we could do and so as we do these five things then we can be ethical, but I failed And I don't think anyone has ever managed to come up with a simple checklist of things to do to give that level of concrete guidance about how to be ethical.

So what I hope to share with you instead is not a checklist because I wasn't even come up with one with just some general guidance and some suggestions for how to make sure the work is less bias more fair and more ethical. And I hope that some of these guidance, some which would be relatively general will help you with your work as well.



Adverse use cases

Deepfakes

Spreading toxic/incendiary speech through optimizing for engagement.

Generating fake content for commercial or political purposes.

Using ML to build harmful products, commit fraud etc.

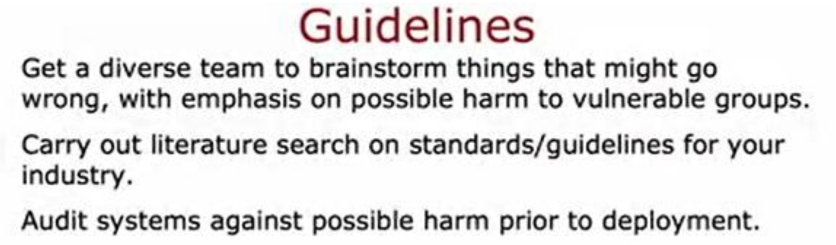
Spam vs anti-spam : fraud vs anti-fraud.

So here are some suggestions for making your work more fair, less biased and more ethical when before deploying a system that could create harm. I will usually try to assemble a diverse team to brainstorm possible things that might go wrong with an emphasis on possible harm.

Two vulnerable groups I found many times in my life that having a more diverse team and by diverse I mean, diversity on multiple dimensions ranging from gender to ethnicity to culture, to many other traits. I found that having more diverse teams actually causes a team collectively to be better at coming up with ideas about things that might go wrong and it increases the odds that will recognize the problem and fix it before rolling out the system and having that cause harm to some particular group. In addition to having a diverse team carrying out brainstorming.

I have also found it useful to carry out a literature search on any standards or guidelines for your industry or particular application area, for example, in the financial industry, there are starting to be established standards for what it means to be a system. So they want that decides who to approve loans to, what it means for a system like that to be reasonably fair and free from bias and those standards that still emerging in different sectors could inform your work depending on what you're working on.

After identifying possible problems. I found it useful to then audit the system against this identified dimensions of possible home.



Guidelines

Get a diverse team to brainstorm things that might go wrong, with emphasis on possible harm to vulnerable groups.

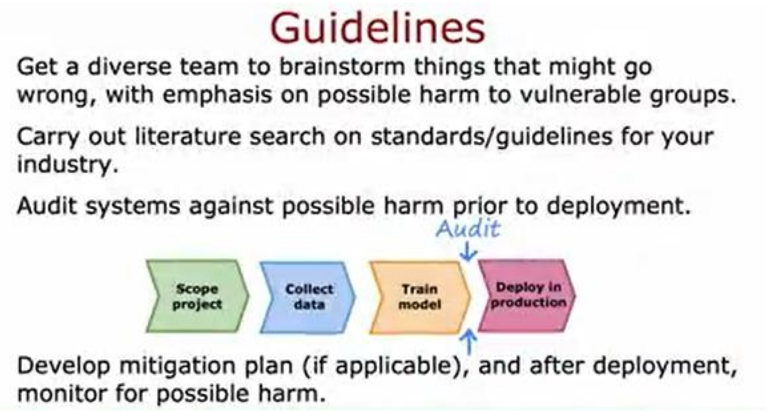
Carry out literature search on standards/guidelines for your industry.

Audit systems against possible harm prior to deployment.

Prior to deployment, you saw in the last video, the full cycle of machine learning project. And one key step that's often a crucial line of defense against deploying something problematic is after you've trained the model. But before you deployed in production, if the team has brainstormed, then it may be biased against certain subgroups such as certain genders or certain ethnicities.

You can then order the system to measure the performance to see if it really is bias against certain genders or ethnicities or other subgroups and to make sure that any problems are identified and fixed. Prior to deployment. Finally, I found it useful to develop a mitigation plan if applicable. And one simple mitigation plan would be to roll back to the earlier system that we knew was reasonably fair.

And then even after deployment to continue to monitor harm so that you can then trigger a mitigation plan and act quickly in case there is a problem that needs to be addressed. For example, all of the self driving car teams prior to rolling out self driving cars on the road had developed mitigation plans for what to do in case the car ever gets involved in an accident so that if the car was ever in an accident, there was already a mitigation plan that they could execute immediately rather than have a car got into an accident and then only scramble after the fact to figure out what to do.



Guidelines

Get a diverse team to brainstorm things that might go wrong, with emphasis on possible harm to vulnerable groups.

Carry out literature search on standards/guidelines for your industry.

Audit systems against possible harm prior to deployment.

Develop mitigation plan (if applicable), and after deployment, monitor for possible harm.

I've worked on many machine learning systems and let me tell you the issues of ethics, fairness and bias issues we should take seriously. It's not something to brush off. It's not something to take lightly. Now of course, there's some projects with more serious ethical implications than others. For example, if I'm building a neural network to decide how long to roast my coffee beans, clearly, the ethical implications of that seems significantly less than if, say you're building a system to decide what loans.

Bank loans are approved, which if it's bias can cause significant harm. But I hope that all of us collectively working in machine learning can keep on getting better debate these issues. Spot problems, fix them before they cause harm so that we collectively can avoid some of the mistakes that the machine learning world had made before because this stuff matters and the systems we built can affect a lot of people.

And so that's it on the process of developing a machine learning system and congratulations on getting to the end of this week's required videos. I have just two more optional videos this week for you on addressing skewed data sets and that means data sets where the ratio of positive To negative examples is very far from 50, 50. And it turns out that some special techniques are needed to address machine learning applications like that. So I hope to see you in the next video optional video on how to handle skewed data sets.