TETHICS

Ethics in Technology!

Abstract – Ethics in the age of AI

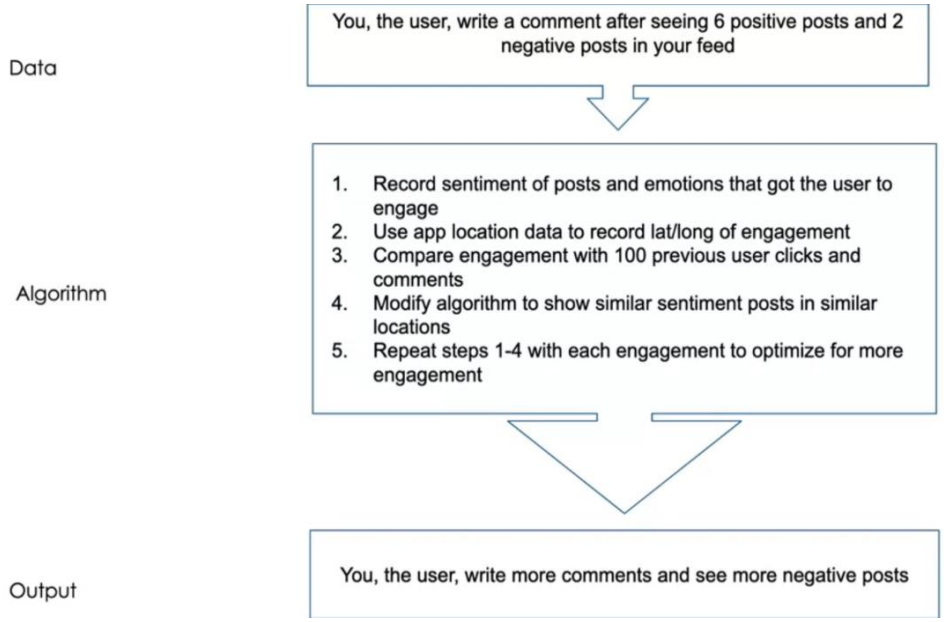
As we explore the algorithms that power everyday life. Today, we stand in the pinnacle of human knowledge where privacy has only become a facade for giant Tech Conglomerates and only used for marketing and biased purposes. As more predictive models get deployed in the real world, whether it's a credit card application or a loan decision, job application or college filter, we see certain groups benefiting more than others. And this isn't surprising at all as we create models of machine learning which based solely on reflecting our data as a society. So those won in the past will continue to win. Why talk about ethics now, after for years these practices have been common, to simply put it we are in an information age where information matters a lot and this information is nothing but our data. Moral dilemma of today, if this ends up as a utility or in the hands of a few powerful people. Can we trust these giant Tech conglomerates to take ethical choice over more profitable choice? As we move on to more and more complex systems as developers we will be able to step back and take a look at this systems, but we won't understand how the decisions are being made and with right decisions we will be able to have an immediate impact in tipping the scales toward the more just and moral path. A future with fairness and no bias in our technology is something that we deserve and our privacy is our right. Personal information can leak from seemingly most trusted source, then our data is used for commercial purposes.

Motivation

Privacy centric models which are protected and private from which data can't be leaked. A transparent algorithm to make the model fairer and more unbiased so that models can't get away with anything just because no one is looking at their decisions making ability. Aim is to build a model which is fair, reduces bias and considers privacy and transparency. Ethics in technology is important firstly because models which we are creating, are outgrowing expectations and we as developers are not aware of some of the predictions that models are making. We are in control of the inputs but not the outputs. Things in the middle are just too complicated and complex. Real harm comes from the lack of knowledge about this technology and without proper knowledge these models are deployed where the consequences affects the life of people. So for example, when you go to apply for a loan, a company has put in an model that may judge that you are not worthy of a loan and that is doing real harm to you, if it's actually not an model that has been designed to be ethically sound. Algorithms vs Privacy, as we have no control over public datasets going over satellites. This relates to our slow erosion of privacy. As we know, and the Internet right now, more and more personal data is being collected by social media companies, by big data firms and by machine learning researchers. Not only is the amount of personal data growing, but also over time, that same time period, the algorithmic power to re-identify individuals based on those personal data points is also growing. We get to a point in time in the future where any seemingly innocent piece of information can potentially put you at risk of being identified by an adversarial algorithm. Awareness is really important right now, even governments are taking actions, Especially European Union.

You, as a dataset

Let's just start with an example, what is a social network? A site that you use every day, that needs to figure out how to become more efficient to maximize its ad revenue. Their dataset is you, the user. So, the model they use is going to be more optimized according to us so that we engage more and more. As you see the below an illustration of a feedback loop.



The social network has detected that we've commented after seeing six positive posts and two negative posts in our social media feed. Now the algorithm which used by a model will actually learn and "improve" our social network feed. I say improve in quotes because that has become a norm. Their aim is to create more engaging services so that users come back and use it more. So, now inside the algorithm They record the sentiment of the posts and the emotions that got the user to engage. That's step one in the learning algorithm that's going to get turned into data that the machine learning algorithm is going to use in its model. Then it's going to use the application data to record location of engagement. Where was the

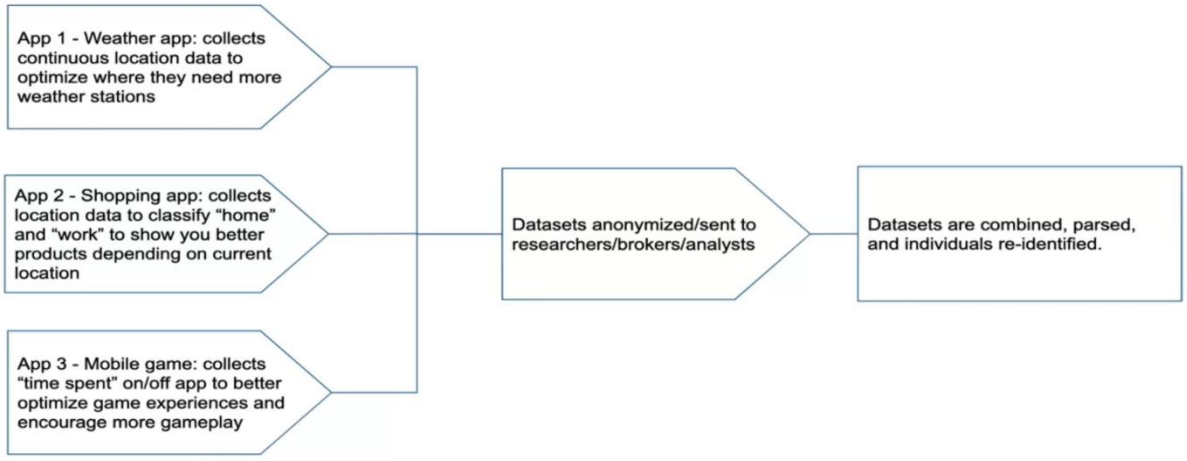
comment written? The model takes that data and compare engagement with 100 previous user clicks and comments to see, is this something out of the norm for someone in this behavior model. Or was it something that kind of fits into a pre-determined package. Number four is it's going to modify the algorithm and in turn the model is essentially referring to the social network newsfeed of the user to show similar sentiment posts in similar locations. Right, the whole goal here is to make sure that you continue to engage with the platform. And then it's going to repeat steps one through four with each engagement to optimize for more engagement. So, at the end of the day output is just as a user, User ends up writing more comments because user is seeing more negative posts.

Main take away from this example is that a company's main purpose will be to earn profit and in this case to earn profit this social media company needs to more engaging and if they engage more users more users will watch ads and this will lead to more ad revenue for company. Ethical responsibility of these companies should be not to promote negative or hate speech. For example in 2018, an racially offensive post went live on Facebook which shook the whole world and Facebook's CEO Mark Zuckerberg said he found those posts " Deeply offensive" and also added "At the end of the day, I don't believe that our platform(Facebook) should take that post down because there are things that different people get wrong" though mark was adamant about still stance was against this anti-semantic violence. The world's most powerful CEO also didn't say that these kind off negative posts are the one which garner most views and engagement which good for the platform as increases ad revenue for the company, but at the cost of their ethics.

The Sad Truth

Data doesn’t exist in a vacuum and the optimization of the social media feed basically making a more efficient algorithm means that they're deploying a learning model that requires a lot of data. So, with many different applications and services that we see, their data gets collected first and later questions are asked. Even the most innocent applications collect a lot of user data and as the developers with small applications, they are competing against much larger organization with tons of resources and money to back them up. So, collecting user data has become a standard practice. This something that never comes up in our mind while using this applications and services. Let’s take an example again of social media company where all the users have not disclosed their age but with use image recognition model, we can more or less find out their just by looking at their profile picture and then they advertise to certain age demographic with even asking for their age information. A very famous paper about this was published on 2012 by Brendan F. Klare, where they find influence of demographic on the performance of face recognition model. So, their finding was, All the commercial algorithms were consistent in that they all exhibited lower recognition accuracies on demographics of: females, Blacks and younger subjects of age between 18 to 30 years old. This shows a problem in fundamentals of model/algorithm creation where example like above could lead to accuracy variations and potential bias against certain demographic. How to tackle this bias and accuracy variations will be explained in Art of Machine Learning section of the paper.

A Big Data Conundrum



There's famous quote, with Great power comes great responsibility. A big data, large datasets which are collected for different reasons, for different applications. Now, Let's look at illustration above, Application 1, is weather application. It's very simple application which is just collecting continuous location data to optimize where to put more weather stations. Every time when user checks the app, it looks like the weather came from two miles away, that's a little far. Let's just record the location of the request, so we can improve in the future. Application 2 is a shopping app. It's collecting data that is also location data to classify home and work. That's just to optimize their service as well. Application 2 want to show you the best products depending on your current location. If you're at work, you might see office supplies. If you're at home, you might see clothes, entertainment.

Application 3 is mobile game. This application doesn't collect location data like shopping or weather applications, Application 3 is just interested in engagement. A game application just wants to make sure that you are active in the app, and this application collects how much time you've spent in which part of the game to make

sure that their game gets developed to be the best possible game. For example, a game like PUBG MOBILE tracks on which map players are playing more and makes improvement to that map. So, it's all about time spent in or off the app. So, we take a look at all three of these different applications here and they all exist on the same phone.

The conundrum, for example, A weather research group is developing an AI to better predict weather. They get in touch with the weather application ask for their data. These datasets are then anonymized and sent to the researchers. So, now application 1 does not know your name. It just collects the location data. Application 2 does the same. App 3 does the same. An ethical dilemma here is there are data brokers and analysts who misuse the data and might even sell the data.

But the end result is that these datasets can be combined and fed into a model which re-identifies these individuals. application 1 which just collects location data whenever the app is used, and application 2 the shopping app that knows where your home and work is. If that gets combined, all of a sudden, someone ends up with a full picture, a dataset that's re-identified you because there's only one person that lives at this address. There's only one person that works at this address and let's say takes the exact same route to get to home and work every day. So, the shopping habits, device habits, everything in between can put into one data set and that is valuable for advertisers and at last with help of Application 3 which knows your free time because of you engagement schedule, advertisers are now more than happy, because this applications know what kind ads to show, at what time to show you. A façade of privacy carries on and ethics go out of the window.

The Big data economy, machine learning models are used to exploit an underutilized or overlooked path to profit. In the financial markets, you can imagine that this can get pretty confusing. It's really just about speed and efficiency, identifying things with bigger and bigger data sets. For example, For-profit colleges have a business model where they make most of their money through government loans. So, what ends up happening with for-profit colleges is that when they get applications, they run predictive models. It was found that many for-profit colleges end up targeting those who are most likely to receive government loans. It's how they make their money. Those predictive models identify people like veterans that are likely to get government loans due to their service, but who are more likely to default on those loans by a large amount, for these college getting students who are eligible for government loans is the way to earn profit and due to this whole system is corrupt.

Art of Machine Learning

A traditional algorithm follows a set of explicit instructions. Take input, get output. Concept of learning and algorithms, which is to take the input, generate an output, but then there's going to be something in that output and the set of instructions that can used on the next input. Tt's really important to grasp what learning actually looks like because it's a tricky term. We learn things as humans subconsciously without actually figuring out what steps we are taking to get better and a machine learning algorithm does not work like that. It needs those instructions exactly what to do. Classic training set, we give the model 60% of that data that generates the model, we then validate the model with the next 20% to make sure the model is able to make predictions, and then we take that final 20% of the data, run a test to

make sure the model is accurate, and if the error rate is low enough, we deploy the model. Training a more ethical machine learning model:

Classic Training Set:

1. Feed 60%
2. Validate 20%
3. Test 20%
4. Launch

Optimized Training Set:

1. Feed 60%
2. Validate 20%
3. Accuracy test 20%
4. Scrub
5. Explainable test 20%
6. Scrub
7. Fairness test 20%
8. Refine
9. Launch

Optimized Training set, we feed that same 60% and validate with that same 20 percent. But then we take that last 20% of the data and modify it to make sure that we are running accuracy tests, something called unexplainable test, and also a fairness test.

Accuracy Test, we take 20% of the data and modify it to get lower error. Explainable Test, scrubbing the data so that the model does not know about the last 20 percent that it just tested for. This is pretty key because we do not want to reuse data that the model has already trained on. Then we reduce the black box problem as much as possible and we write tests to figure out what the model is actually looking at while making predictions.

Fairness test, we would take our data and slice it down to make sure that the data was actually 50% male, 50% female, in that job training example, and then run it through another set of tests to figure out if the model is over-sampled or under-sampled for either group, and then refine the model as much as possible, This will solve the amazon's recruitment tool problem and bank problem of bias against

women and also the demographic problem. Goals to hit when making a machine learning model one, Accurate, two, Explained and three, Fair to make our model more ethical.

A Reflection

As developers in this day and age, we need our data to be mirror of our society and what that data reflects in model is part of our job as developers. For example, in 2018 Amazon's recruiting tool over optimized for hiring men over women. Obviously, the tool was not explicitly programmed to hire more men than women but when we look at the facts based the mirror (our data) amazon's tool was likely created by team of men and the model was trained on amazon's existing teams, which likely had majority of men. The model showed a bias against women. Reflections of our models:

- Personal: Who is building the model?
- Cultural: What are the values of the company/Institution?
- Societal: What re the values of the state/country/world?

First, Person or team who is building the model, what are their views. Second, Institutional values that where they put ethics on top. Third, Societal cultural reflections values of the local people.

Empirical Reflections of our model:

- Choice: Features to be selected.
- Unknown unknowns: Predictions of the model
- Critical analysis: Finding blind spots

The Good guy AI

"The smarter and more capable an AI is, the More likely it will be able to find an unintended shortcut that maximally satisfies the goals programmed into it."

- Nick Bostrom

An example of Perverse Instantiation, how these shortcuts can be found can be seen in the 2016 Bot Tay released by Microsoft as a Bot that could chat with people on these social network Twitter. The perverse instantiation in the instance of Tay was that the goal was, speak and learn like the others that you see on Twitter. And Tay being a learning algorithm found that the fastest path was actually learning some hateful and offensive language because that was the most engaging. So that experiment went quickly wrong after users taught it hateful language. But at least in this instance, Microsoft was able to turn it off.

Bill Gates says "Artificial Intelligence can be our friend". Ethics and Control, one problem with turning off Artificial Intelligence is that the act of turning it off would go against its goal. A truly A.I would not let itself shut off, so the next possible solution is A.I enforcement which monitors other A. I's.

Deepfakes are computer-created artificial videos in which images are combined to create new footage that is based old footage. Deepfakes uses a form of artificial Intelligence called deep learning to make these fake videos. Technology used in deepfakes is advanced and useful. Let's take a look at practical example, Justice league a super big budget Hollywood movie released on 2018, There was problem with the movie so the director had to reshoot, some of scene and because of contract henry cavill who playing the part of Superman in that movie, couldn't shave its moustache and Visual effect artists had to remove his moustache digitally

for the reshoots, They spend millions on henry's face but couldn't make a convincing real face. Then a youtuber by the name of Jarkan, Made Deepfaked video of superman (Henry Cavill) using deepfakes with his existing hardware and achieved much more great and realistic looking human face as a result.

Dark side of deepfakes, Right technology in right hands could make life easier but there is always a dark side, its true deepfaked videos can be used for good purposes, but right now 90% of deepfaked videos are adult and it's just slapping the face of a celebrity on the actors of adult industry. Ethics of the developers is what we can count on to solve this problem and other solution is another AI which can detect a deepfaked video using advance algorithm.

Google Duplex, An AI System for Accomplishing Real-World Tasks Over the Phone. On May 8, 2018 Google released its new project called google duplex. Its main purpose was to complete long-standing goal of human-computer interaction has been to enable people to have a natural conversation with computers, with use Natural Language processing along with the application of deep neural networks. Google duplex passed the Turing test, A Turing Test is a method of inquiry in artificial intelligence (AI) for determining whether or not a computer is capable of thinking like a human being. Of course, the test is named after Alan Turing.

Deception by design**,** "Google's experiments do appear to have been designed to deceive"**.** Its AI system that makes phone calls on your behalf while sounding frighteningly human, Its AI which passed the Turing test. Ethical concerns at play around AI technologies that are powerful and capable enough of passing off as human — thereby making fools of real people in the process.

The Black box

The Black box problem, when model is opaque, we don't exactly know how it is making decisions. The model is not empathetic, it's not self-aware, it's really just making decisions based on data, and so the data can reflect society in good ways and bad. The model really can't factor in anything you don't tell it to, so we do not as researchers and as professionals know exactly why a model is making the predictions, but it is making predictions. Heart of AI ethics, we need to make sure that our data sets can meet social science. An example, A startup makes a personal loan model which decides who to give loan or not. Their model is based on 70 years of historical personal loan data, and whether those loans were paid or defaulted on. Then when it comes to judging income, they have 80 years of historic income data and that income data combined with a personal loan data can give us a really good idea of who is able to pay back a loan. Model when repayment percentage is above 85, it's going to be a YES, 84 and below and it's going to be a NO. The model has now showed that the error rates are very low and it is deployed with bank client. All goes great, they don’t know exactly how the model is making decisions but the model works at first but, in few months the startup gets a call from the bank saying that “Hey, any idea why we're seeing men approved for higher loan amounts than women on average”. So, the model is biased against women same as amazon’s recruiting tool , but why, when looked for problem it was obvious With 80 years of income data, the dataset have more data on men than women due to historical patterns of men versus women's involvement in the workplace, and so their model may have higher confidence levels and predicting men's repayments than women's. As developers how can we get over these constraints?

Conclusion

“Nothing vast enters the life of mortals without a curse” – Sophocles.

Information Technology is one of the diverse fields than imagine and is related to the ethics in various ways. Most of the people are unknown how they have abuse of Ethical benchmark in information technology. So rather than taking about the ethics, people should get aware to do something ethical. The sad truth is that data is generated and it us who are the data, our lives are the data.

We always forgot one thing, if you are not paying for the product then you are the product. These social media companies are fighting for our attention and engagement. How can they get our attention, its easy we give them, our data, our habits, our life, they know everything about us, we have become a datapoint in their Big data and Learning models. These companies sell certainty by having great prediction models and how do you get more accurate prediction model, of course by getting data, I mean big data. We need to have ethical designs introduced in every product so that, we don’t get used by these companies.

Successfully integrating ethical design in a product means making sure the product is based on accurate model, explainable model and fair model. Eliminating bias and preparing the model to be ready for unknown unknowns is reflection of our society. Where nobody is discriminated against, technology is one side of the life that always interest and surprise us with the new ideas, topics, innovations, products.

References

Facebook Stance on hate-speech: https://edition.cnn.com/2020/10/12/tech/facebook-holocaust-denial-hate-speech/index.html

Demographic paper: https://ieeexplore.ieee.org/document/6327355

Microsoft bot Tay: https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/

Bill Gates AI: https://slashdot.org/story/18/02/18/1735208/ai-can-be-our-friend-says-bill-gates

Deepfaked Superman: https://www.youtube.com/watch?v=5exUEnpViUs&ab_channel=Jarkan

Google Duplex: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

For-profit colleges: https://www.forbes.com/sites/lucielapovsky/2018/02/06/the-changing-business-model-for-colleges-and-universities/#404178865ed5

Photos courtesy of learnQuest: https://www.learnquest.com/home.aspx