w7_2

Sun, Oct 30, 2022 10:03AM 50:31

SUMMARY KEYWORDS

fairness, bias, model, data, people, hiring, population, resume, female, subgroup, women, gender, problem, ai, group, metrics, means, based, design, criteria

05:54

Amen. Okay, right. Now

06:09

let's get a deeper dive into buyer. So I have to, I have to acknowledge that I still have plenty of slides. So I'm going to skip quite a bit in order to finish in time, because I under estimate I normally deliver two lectures on this at least I will, I will just do the, you know, the over coverage. But in deep dive, so

06:36

we all know the term that data is the new oil for AI, right.

06:45

But all the data is the new oil or a lot of AI and EDM systems. And we already discussed before that this is often where the source of bias would actually come from, where do buyers come from. So I want to actually go through them more deeper about the types of biases in AI and machine learning system. So in in previous half of the lecture, I have actually mentioned a couple of examples of cognitive biases, which are actually coming more from how humans interact with the system, or the world. Now, in terms of bias in machine learning. The next few slides really are derived from this paper from Robbie at all, which is a survey of fairness and bias in machine learning. So I mentioned about the list of cognitive biases more than 100 of them, in fact, of that bias and machine learning there. In this paper long alone, there are more than 23 types of biases, I wanted to discuss some, I may only cover some of them. So one of the common type of bias is representation bias. So we did actually discuss this a little bit when we discussed the gender the beauty contest, or the gender shades. So representation bias is, is basically how the data is defined, and how is how they actually sample from the population. Now, deep learning models have started, you know, and became a boom since 2012. Because of ImageNet. So a couple of things. There's a couple of reasons, of course, the the power of GPU, the power computation has increased. But also there's, there's this huge trove of model of images. They're actually already widely annotated, so that it contains millions of images. And let's see where they actually come from. Actually, it's sorry, it comes usually from the US population. So 45% comes from us. So you'll see a lot of objects that are predominant in the US population, but may not be in other other population, as you can see here. So representation bias occur here, because when you build these datasets, it only will then will only work well for certain types. So another example, if you build a model, for example, based on smartphone data. If let's say data you want to build a data to design The transportation system in US based on how people move in the city, and let's do it, let's sample it from mobile data geolocation or location sharing. So if you have locations, Sherry and I both in your phone for some apps, and you'll be able to use that, for example, to design and optimize ride share, or Uber or public transport schedules. Unfortunately, there'll be representation by is there because individuals over 65 is all in us, even to date, if you look at Pew Internet Research metrics, they're usually underrepresented. Underrepresented, so because the they may not use either the latest smartphone that has to look accurate location sharing, and therefore they could actually be discriminated against. And this actually, in fact, the same case in regional Australia. Believe it or not, we have more than 20% of our population in regional Australia that do not have access to the internet. So if we actually designing, I mean, they only have access potentially, to satellite phones, but they're terribly expensive. If you're designing for regional Australia, of example, you really have to think where and how you're going to collect the data to be able to perform a research, they actually the more representative of their population.

11:48

Another type of bias that is quite common is historical bias or a dimension about the CEO. So historical bias actually can occur because of the state of the world of which the data was generated. And this is actually an accessing bias in history. So for example, when you think about a president of United States, before Barack Obama, what you have in your mind, when I mentioned President of United States, you only think about white people. And even and until to date, you won't be able to sing off women, everyone's men. So if that's basically the same the world in which we are in, so as of 2020, for example, if you do a search image search on SEO, the thing is, this is actually historical bias like Krypton because as I turned 20, only 7% of fortune 500 CEOs are women. That's very low. But it's interesting also that research shows that female companies with female CEOs CFOs are generally more profitable than companies with men in the same position

13:10

so what do they suggest

13:13

it looks like women are held to a much higher hiring standards to mandate scrutinize they're really you know investigated before hiring decision made for women to be placed on that kind of positions. Now we're going to use AI to make hiring process more equitable. How do we fix this should we even use AI to make hiring process more equitable?

13:46

Any comment? No, yes. Our friend here Yes.

13:58

I mean, I think it would make it like four biased Okay, great. I mean, given our own, the percentage of women are like, if the AI picks up on that it's done.

14:19

What Yeah, so So you're saying that yeah, we'll just pick up on what's the view of the world. Any other opinion yes. Are you designing that? That issue? Yes, Tom. Okay.

14:55

So when you say I, right, so there's a couple of the frictions that A you might have less than you dislike a technology or discuss in your in your assignment is an AI tool that I mean house can be you know how CV or resumes are ranked in hiring obviously you're for example, oh, that's uncommon here Rachel is removed Jen gender, racial information or CV, anonymize a CV will that solve the problem? If the race racial information removed Yes, hmm.

15:54

The problem goes so much deeper than just like whether your name is male or female. It's like what school you went to can tell us about it. Like who you got your resume template off, can tell you something about you know who your friends are. And maybe if you've got an offer woman, and there's like different resume templates for men and women, just because of whatever reason, maybe it picks up on that. Maybe it picks up on your word choices. Like there's so many more things that just the gentleman said on the on the president?

16:23

In fact, yes, yes. Go ahead. I think

16:30

race and gender are generally written resumes anyway. Not really a matter of moving that.

16:38

Right. So removing race and gender very difficult as my friend offers, if you

16:43

want to do something, while up until very recently about I mean, like 10 years ago, it will be a bit longer it was normal and accepted in Germany, that on your resume, you will miss your parents, occupations along with a photograph of yourself.

17:01

Once you look like until you want your

17:05

parents to give you the standard, if you're somebody that you would just expect them to, you're not right, you're still expected to do it. But

17:15

I think it's very uncommon now. Especially if they say like you shouldn't remove your photo. Yeah, and things like that. So

17:25

good. Progress.

17:29

Thank you, my friend. So I've got a visitor here you did. Visiting patients should have been from Germany. So this is a very difficult problem. And in fact, Amazon has that one in one company a couple years ago, because they've just got so many applicants, I think we just don't have the time in the world to sift through all the resumes. And let me just, you know, build automated tools to actually just do you know, the feel free and the ranking and then we can you know, the, the hiring committee will just be able to look at the top rank resumes. But even as you say, friend here, say, Okay, let's just analyze remove, or the racial information, all that and it's still a very difficult thing to do. But even the way apparently the way male and female write cover letters are quite different. Males tend to be very confident, and they write the achievement. And often sometimes if this is not my does not a stereotype. But this was actually the study, often over inflate the achievements. whereas females tend to undersell whatever cheap.

18:50

And this is quite common. Yes,

18:53

as a similar statistic that's researched and said, there's a list of criteria for job application, men will feel comfortable applying for that job. If there's something like 50% of the applicable criteria, then eight, and for women, it's closed at like 85 or 90% of the criteria. So even before you get to the point, we're doing iria, the criteria you buy.

19:12

That is true. That's true this way, um, you know, often, you have another comment, based on his

19:19

solution, we got two ways we could either make more or less of them. The second was money. Right.

19:30

So selection criteria. That's right. Often, if we pack more, we want this technical skills, everyone. If it's, for me hiring a postdoc, I said, I want students to have publishing this this. And men often can respond with Oh, I haven't auditioned them but I've published a list. I can say that they're similar. We may we'll take it no, I'm, I'm not eligible. And I need to make me more confident. Well, but that is basically the world we are in right now. So that is true. And that's why often when I make sure that it's gone past to my female colleagues to see whether this is actually going to be intimate.

20:26

Dating for the female to apply. There's another threat. Actually, it's

20:33

interesting. You'll see, there's research that says that those words put in your resume. And actually, it's a ton of women. So even though like the words that you use, even if it's the exact same criteria, changing those words, could change the balance of now,

20:49

that is really true. And therefore, this is very much a social problem than a technical problem. Yeah, so the issue of mitigating bias is very much a social technical problem. So you really need to have someone who, who's really good at understanding the social aspects of human factors. And this is why companies these days who actually want to invest in a aspect of fairness, transparency, D biasing the solution, they're actually not hiring just computer scientists or programmers. They're hiring designers, they're hiring social scientists actually understand how to unpack the social issues.

21:37

It is still a very difficult problem.

21:41

And by the way, that Amazon suppression my ranking tool is completely decommissioned, because it's completely unfair. Now, measurement bias, we talked about compass before mentioned bias happens when, when the the feature feature that was selected by the model often is measuring the wrong thing and is biased towards a certain certain groups subgroups. Say, in here, we again because this is black box, we can't really tell how compass work, but only based on research that was done on campus. And they tried to look at the data, it may also be to do with the skin color of the even if you remove the photos, even if you d anonymize it, you remove the photos, the proxy variables, such as the data taken from the survey that I mentioned before, whether they say you come from divorced families, or whether you have families and friends who have got offense before there can be cross parables, simply because you you're you're born into underprivileged society, or those who actually don't have basic access to support. It doesn't mean that you're more risky. But but this this system, in fact, highlighted that. And that was the issue the issue. So, these measurement bias can occur when this accuracy of the data vary between across groups, and when there's process variable that happen. Now. This is similar to conference thing now let's read this case study. So your local hospital uses a model to identify high risk patients before the death of serious conditions. Based on information like past diagnosis, medication, demographic data, the model uses this information to predict healthcare costs. The idea being that patients with higher costs likely correspond to high risk patients. Despite the final model, this specifically excludes race, it seems to this demonstrate racial discrimination. The algorithm is less likely to select eligible black patients, how can this be the case?

24:20

Now there was omission bias. Where's the omission bias? Potentially? Yeah, let's do this. Passing too much. Okay. So it could be the health care costs, if socially, maybe, you know, black people are not that high in like the, like the income brackets so to speak, maybe they're less prone to actually go into the hospital. And that could affect the history and the accuracy of the data because there's just not that much data for black patients. to slow your slaughter healthcare costs and costs was used as a proxy for race. And although race is excluded, but

25:16

this was actually became a proxy for race, right? Because using that healthcare costs it actually perfectly segregate it into the black and the white, because the black are have less access to health care and

25:35

have less to healthcare system this

25:40

does. Sorry, there's a risk. Okay, genetically predisposed? The answer is no black people are more likely to be working in jobs that they can't take time off from, they're more likely to be in communities where they don't have the ability to take time off and go to the doctor, which means that when they eventually do, they also got a client federal duties. But they actually do go to the doctor, it's worse, and they're hired more, but none of that has to do with the color of their skin. Because even like last year, there would be debates about like, oh, black people, well, not the color of their skin.

26:20

That's right. Thank you for that Tom. So So Tom was alerted to the fact that the occupation of black people actually became a proxy variable of accessibility to COVID. Another common type of bicycle aggregation bias. So this is when you use one size, when you train just one model across to fit all the kinds of population you have. Now, the problem is, each of these model might have each each of the subgroup might have different conditional distribution. So the probability of you know, X or a Y happening is depending on you know, your history and your profile data. Now, if you combined them all, distribution will skew. So for example, if you're building a life example, based on this data, Hispanics have higher rates of diabetes and diabetes related complete complications than non Hispanic whites. But if you're a built AI system, or machine learning system to monitor, diagnose and monitor diabetes, in all suburban populations,

27:50

it is important for you not to build a model one model to fit them all. It's just not going to work.

28:01

So, you either have to include ethnicity as a feature in the data. So, you should actually use it. But how do you use it you have to use it in a way that you measure the outcome for each of the subgroups, or you can also train a separate model for each subgroup. So you actually do segmentation in your population data and you train a separate model for each of the subgroup and then you look at and compare now, outputs are performance of that those models for the across different subgroups. Now, there's a really nice illustration, again, from Robbie's paper I rely and this is an example of Simpsons paradox. So, what do you see here is a hypothetical nutrition study then measure the outcome body mass index in in relationship to how much pasta do you take every day? Now, depending on how much path that you take, so, the y is the hammer how much the function of daily pasta calorie intake so, and the the X is is basically your, your BMI. So, now, the thing is, if you just use multivariate linear regression, you saying Oh, yeah, actually, you know, we combine all of them we just use regardless of whatever the age group the age, I mean the gender for the entire population, what you see on the left is for the entire population, um, and when the positive trends suggest that increase past assumption, so why increases it actually increases the A higher BMI. So unnecessary suggestion, you can see the positive correlation.

30:05

But if you actually cluster them, if you actually cluster them

30:12

based on age, gender and all that, so dependent, let's say this is based on gender, or based on age group, in inside disaggregating, this, you what you're seeing on the right is actually negative correlation. So the trend within each subgroup actually negative. So, what is interesting is increased Pathak assumption, in fact, is associated with lower BMI. How does it happen? There's an app on observables that happened here. They are products of the unknown, that there are groups of people, they just have a higher fitness, right. So no matter how, how much faster they take, they just have to have lower BMI. So it's nothing to do pass that it's more to do with the fitness level of the group. So this is why we call it aggregation bias. So remember, even so this is, again, part of how you treat your data. Evaluation bias. So if you want to compare your model, say, Oh, my model works a lot better than any other model out there. Machine learning AI, people love this, because they always try to be the state of the art benchmark, and try to be better. And this is what leads to a very competitive face recognition technology. A decade ago, but completely biased. Because they're comparing themselves to already a bias benchmark. There's a lot more, I won't even go through them, you can read the paper. And once it's once it's already deployed, there are also other issues, for example, reporting bias. So what is reporting bias? Say, you know, reviews, right? We, we tend to only recommend, right reviews, what when you say an accommodation, or you go to a restaurant, that's so terrible, you just couldn't eat the food that was served to you, you just have to actually document it and, and report it, or it was really good, that actually blew you away. So this reporting bias. So the things that are happening, and the ordinary, they don't get reported. Um, so that's why for example, if you have a sentiment analysis model to predict a book review, where there's positive negative, they won't be able to work really well. If the languages are more subtle, or, you know, they happen on about borderline. So biases can happen in all stages of a software deployment deployment, from the data generation, to selecting the population, to capturing the data set, or even generating those trying to even balance your data, even when you're actually deploying it. In real world, implication, why because sometimes the models is the train in a certain context and the point and not the context. And there'll be bias from the origin of the source when it's actually deployed. So, um, lots of things to read about, apart from Robbie, if you're interested. I really like this article by Alexandra Altano and her group from Microsoft research on social data, biases, methodological peoples and ethical boundaries. I don't have much time left I will talk really briefly about algorithmic fairness. So, we already touched on before that bias and fairness have huge high degree of overlap, but slightly different focus bias, talk about representation fairness, talk about the outcome. So, so, these people were one of the first in the world that that that basically indicates that you can you can capture some of the social constructs we believe of that what belief is fair into an algorithm. We can design a metric to measure how fairness across groups, for example, or across individuals, and whether we want to treat it more on the what you see is what you get, well view or we are all equal equal view. Now there, and why do we need to think about that, because

35:11

there are different types of decreased discrimination. So let's say in terms of hiring, it could be direct versus indirect. So we can make a completely indirect discrimination sighs these these job ads specifically only for male or female, you know, it could be very explicit as that or it could be it could be implicit could be indirect as well Tom site, the way it's actually formulated it, it actually seems to require man versus female counterparts. But it could be indirect here could also be to do with the data scattered so.

36:00

So in here, another example of indirect discrimination. So, Amazon rollout, same day delivery across select group of American cities. So genuine, this was a very genuine design, they didn't have a thing about bias or fairness. They said, Oh, we just first deploy it to a neighborhood with high number of current Amazon users. But predominantly non white neighbors were excluded, because they just don't use they're not many of the Amazon customers. So if Amazon look at the distribution, there's a long tail distribution, net long term distribution are non white. Now, unfortunately, only focus on the top k. Now. racial demographic data, we tend to correlate with the location. So this decision actually result in indirect discrimination.

37:10

There are different ways to metric

37:12

so many of them. I won't even be able to cover it in this lecture.

37:20

different metrics of these.

37:23

If you're interested, I can send you a couple more papers to read. And let's say you want to compute equalized odds or Equal Opportunity Democratic Party How would you compute this in your model? And how would you do it? How even come say, how i Who are you comparing? Is the individual? Is it group? Or is it subgroup? So individual means let's say they are 100 individuals in this lecture theatre. The output of the model has to be fair to every single individual here. Which means every single the upper model for me has to be compared against said, Tom unit and every single person as vice versa. So you will have this huge permutation here. In this class, there's a level of fairness, which is very expensive. And that's why people move away from individual fenders because it's very expensive to compute their steering group fare. So group fantasy is basically how do you define your groups, maybe based on gender or based on age group, depend on outcome? So, do we want to be fair to both male versus female or you want to be fair on older versus younger population, for example, or maybe there are multiple most groups or subgroups. So even maybe there are further subgroups.

38:53

Now, depending on the

38:58

the fairness metrics you choose, it might apply it can be pie for grid to measure fairness across food, such as for example, demographic parity. But there are some for example, famous true unawareness that only applies to individual.

39:20

Now, just giving you a quick example,

39:26

what is demographic parity? Demographic parity is basically the model is fair. If composition of people selected by model matches with the membership as a percentage on the app of the console, an example a nonprofit organization organized this conference right and there are 20 20,000 people sign up to the 10. local organizers right ML model to Slack 100 attendees who could potentially give interesting talk So the conference, so what is the demographic party looks like demographic party means? This because there's 50% of the attendees will be women. So this is based on group men, men versus women, then model will be designed in such a way that 50% of the selected speaker candidates are women. So that means 50 or 100 should be women. So this is demographic parity. Which means, for example, if the attendees not 5050, say the attendees are less than 30% are women, then at least demographic parity means that three out of the one 100.

40:43

Speakers should be women. So, that's record repairing. That is different to equal opportunity.

40:56

So, equal opportunity is to do if

41:01

the sensitivity of the model

41:05

I think it's much easier for me to explain with a chart. Another one is equal accuracy is to do with how, what is the output of the model how accurate it is a gate between male versus female, for example, should be equally accurate. If the model is 90% 98%. Accurate for man, it should be also accurate for female. Let me just give you before I talk about awareness, let's talk about this briefly.

41:41

Has any of you here? Not done machine learning course. Okay, most of you, maybe I don't know, I should cover this in the remaining time, maybe not.

41:55

But anyway, normally, you can compute this with your confusion matrix. Maybe I'll just skip that those because of time, said when we were thinking about designing advanced course of this right, so we can cover some of these in the advanced course. So a fan is quite nice.

42:22

Okay, the secrets though, will walk you through.

42:27

All right, um, group unawareness means okay, let's let's do this. Let's make what our friend said earlier suggest that, you know, let's make let's remove gender data to make the model fair to all the different gender groups, we can also remove information raise about age, but but there is this problem about fairness through unawareness. Because if we want to do that, we have to remove every single proxy in the data equals RT, which can be used to induce a postcode or zip code in US postal here. Because we do have suburbs that are predominantly Vietnamese, we do have suburbs that predominantly Italians have, we do have suburbs that predominantly are full of people who have retired, it's like a lot of little retirement villages. That means you really have to clean your data in such a way that you remove every single process variable that's possible. Is it possible is very difficult. So if you're interested to know more about this chapter with me later on, I can give you more details about examples of how do you ensure equal opportunity, for example, or equal accuracy or demographic power? Using confusion matrix, but I think I don't have time to discuss this. They are tools ready, I'm skipping about 10 slides, this confusion matrix. Now, there are some some tools that you can use because as well as there's one for example concretise and this quiz was used to actually measure fairness on competence output. Another one is AI fairness 360 by IBM. So you could then look at for example, output on statistical parity equal opportunity, push offs, equalized thoughts, all those things. So all the different fairness matters to choose. Based on that, based on these measures, there are different mitigation use, for example. So these are what we call pre processing. So you might want to look at how you pre process your data Add before you use it as an input to your model. And they're the these are some suggestions from Ai 360 from IBM, lots of them, but but in if we categorize them. So this come from our paper, there are three types of mitigation strategies, pre processing, are you doing it within a model when you're processing when you should definitely model so you can actually, because the models could learn a lot of different latent features and how they interact with each other. So the fairness can be done within a model as well. But the most popular method is post processing. So that's the one is widely used by IBM by fairland, Microsoft. And in this paper, we actually look at saying that, even if you try to mitigate unfairness, to processor post processing methods, which is like 70 80%, widely adopted by companies, and they're not robust, as soon as your data distribution changes, or actually deploy to a new environment, the fairness, the fairness representation will collapse.

46:10

So

46:13

fermo, Mr. A really good one, because it's completely open source, you can look at it as well. Just a quick example, you there's a dashboard here in Berlin, and you can use, you can try data, there's a you can look at the dummy data set. And then you can measure disparity in the performance of the prediction of the model, let's say assessing a credit rate rating, credit risk assessing.

46:42

For example,

46:46

a score, you know, whether someone should be awarded access to

46:52

select a school or something like that, for example.

46:57

A lot of these then can be used, they have some existing data and you can try it is now finally, I think it is still very difficult to define which fairness metrics to use. And this is why it's more a social problem than technical problem, because even if we can call them, we can, we can call this we can call it how things can be more fair. But it could be just fair to one stakeholder but not the others, you will you will discuss this in your assignment, right. Like if you pick famous, for example, what you view as an all.

47:53

Fairness on another stage, which is a fairness in optimization across different stakeholders, it becomes a multi criteria become a multi objective optimization. So, it's still a very difficult problem. That's why it's a very hard research area. And how do we even see, you know, synthesize what it means to be fair? Does it mean that we have to treat everyone equally. If we choose equal equality, then we can define a fairness metrics, for example, equalize or equalize odds, which I didn't have a chance to talk about. But there's a really nice feature here be quality for versus equity, equality, which will lead to a better world well, it really depends on the context as well. And the goal of the technology. And another open problem is how do we research on fairness, maybe maybe we should move away from the thinking of fairness, maybe the focus should be on fairness. There's a couple of things to read, if you want is also a very good nips guest lecture by Kate Crawford. There's also a really nice guest lecturer, nearly 2020 guest lecturer I forgot to input here. It's, it's adapted from the source of the scourge. So the keynote was very interactive are posted on Africa if you're interested to watch it. Anyway, all right, thank you. So that's it. And just quickly on your assignment, because there are a lot of different questions about stakeholders values, blah, blah, blah. If you're agreeable three, then you need to each of you need to represent our stakeholder. That means there will be at least three cycles if you will go for them as their poster holders, and each of you will pick a card that represents a value that you want to discuss. And, and everyone will every other stakeholder We'll discuss that as well. So that means there'll be metrics are three by three or four by four or two by two. Okay. And I've also given a more clarification on the scope of the technology and and how do you how do you describe your stakeholder or your user on it? If you have any other question, yeah, do let me know on it. All right. Thank you. See ya.