AI ETHICS

Ethics in AI: A Challenging Task

By Ricardo Baeza-Yates, Research Professor at the Institute for Experiential AI at Northeastern University (Silicon Valley campus)

In the first part of this module, we cover five current specific challenges through examples: (1) discrimination (e.g., facial recognition, justice, sharing economy, language models); (2) phrenology (e.g., biometric based predictions); (3) unfair digital commerce (e.g., exposure and popularity bias); (4) stupid models (e.g., minimal adversarial AI) and (5) indiscriminate use of computing resources (e.g., large language models). These examples do have a personal bias but set the context for the second part of the module where we address four generic challenges: (1) too many principles (e.g., principles vs. techniques), (2) cultural differences (e.g., Christian vs. Muslim); (3) regulation (e.g., privacy, antitrust) and (4) our cognitive biases. We finish with a discussion on what we can do to address these challenges in the near future.

Lecture Transcript

0:00 So the first important thing when you talk about ethics in AI is that ethics is, a human, a human thing. So so we are, we are extending this to machines in a way that maybe not the right way, but we talk about that. So, so let's talk about ethics in AI.

0:27 And before that, let's talk a little bit about this new institute that Osama is heading and that started only at the end of last year. And basically, like Northeastern University that was the creator of the co-op experience, like you work every year, in your undergrad, the same is happening here in this institute. We want to find what is practical for these works. And for example, to think that we find importance on how things work is to have a human in the loop. I will get back to that in the talk, basically having the common sense, they will experience the knowledge to basically find the right way to use AI. And also the strong dependence on data when many times we don't improve the algorithm, we just get more data. And sometimes more data is not the answer. We also will mention that in part of the presentation.

1:28 So the agenda has two parts. The first part will be about current ethical issues that are like my own bias here. So also, as Simon said, I have worked a lot on bias, like more than 10 years, even before bias was trendy. And I want to cover these five different kinds of problems: like basically automatic discrimination; what is physiognomy based on facial biometrics that are also another biometrics; unfairly digital commerce, basically this to be the models; and then something about if we're using the resources well. The second part, I want to go to two more generic issues, like we have too many principles, how we solve them and this is ongoing work that I'm doing. Also about cultural differences, regulation, very important, all the legal aspects are then I will end with the quality biases, but then I will try to be to have the more pragmatic experiential part, what we can do after we learn all these things.

Transcribed by https://otter.ai

2:33 So the first problem, of course, is bias. As I call it, the curse of bias. All algorithms receive biased data. In some sense information is biased. Otherwise, if you have a like a random, random noise, the algorithm will not be able to do anything. And one first question is, should the algorithm be neutral? Should the algorithm be fair? Already defining what it means to be neutral, what it means to be fair, it's hard because for example, regarding fairness, most definitions of fairness are not are not aligned. So it's very hard to fulfill two fairness conditions at the same time.

3:13 So most of the time, the algorithms don't answer this question and then you get the you get the same bias. And worse, sometimes you get amplified bias. So the first example will be how to get to this amplified bias, because if we have amplified bias, that means the bias is not only in the data as many very relevant people say. So some biases come also from the algorithms. And even more, we will see that also the biases come from the interaction of users and the systems.

3:48 So you have to remember, we can have equality not always worked, we have equity. So we achieve equity with affirmative actions. You can see that these are positive biases. And then you have, for example, everything related to justice where you really get rid of the problem. But that takes much, much longer. And we are seeing that with some social biases that are today still important in our society. So if you want to learn more about that about this, I recommend this movie done mainly by women, “Coded Bias”. And also we had a very interesting panel in in the ACM, where I was lucky to participate in in March with the former president of ACM - Wendy Hall, a federal judge, Latanya Sweeney which invented the K-anonymity and all these chaired by Jeanna Matthews, which is a person that also has worked on this topic. So let's go back to the question, this non technical question is a question that is not computer science that should answer. It’s society that should answer. So these are very, very important things. That sometimes when answering questions that are not for us, we're not the experts on these questions. And of course, the answer is not always you need to answer these questions. But if you harm people, physically, psychologically, and we'll get back to that, then you need to answer this question. And how do we mitigate the problem? Well, we can divide up the input, we can tune the algorithm, for example, in search, we have learned to run with bias, ranking bias, or we can debias the output. All of them have different complexity. And also some of them are better or worse, depending on the data, and of course, when you do it in the output, you already lost information. So not always you can achieve what you want.

5:44 So let me start with one classical example. That's most maybe the most used example because one of the earliest examples, already five years old, is what happened with COMPAS - a software for criminal profiling. And this was created as a support tool and not as a decision tool, but some people were using it as a decision tool. And also we know that humans are influenced by recommendations coming from algorithms, and then they tend to, to believe them. For example, ProPublica found that there was a clear racial bias on this criminal profiling. There was a lot of discussion, a lot of analysis. Also, different people saying there was no, no real racial and others saying even there was more depending on how you look at this data. But then, there are two important questions in this example. The first one is: is a secret algorithm ethical? So if we have a societal problem, do we want full transparency? And most people will say yes, but then we need to remember this. And another important question is: is a public algorithm safe? Because then you can game the algorithm. And for example, if this is a criminal profiler, you might learn how to look like you are a fine person.

Transcribed by https://otter.ai

7:07 But this has been used in other contexts, not only in the context of prison, which was the content of, of COMPAS. But you have seen, for example, Gotham, suffered by Palantir, where basically, this was used for real criminal profiling, not after you're imprisoned, but even before. And this was complicated, because, for example, Los Angeles was bought by the police foundation, so no one needs to know. In New York, it was never approved by the Council of the city. In New Orleans, they kept it secret until someone basically whistleblower put it in the news, and so on. Even in some countries in Europe, if they're using the system. And the problem here is not if it works, we can argue if it works, or not, maybe 80%, 90% may be great. The problem is what happens in that 10%, where it doesn't work. One error, and the life of the person will change forever.

8:04 So there are other examples. For example, here is one from Spain, where you have not something that is to, to check the that the people are accused of something but the victims. So this is to check if the victims are saying the truth or not. So it's like a lie detector. It had many problems because this was not directly on what the person was saying, but was by using the text of what the police wrote that the person said. So that's already a problem, because it's not a direct transcript of what that person said. And this, for example, basically, this is still being used by the police. But when in Germany, they tried to do something similar, the pilot party immediately challenged in an EU court saying that this was not fulfilling the human rights of the EU. So, you see two different approaches from the civil society for the same kinds of suffering.

9:08 So let me give you an example, a detailed example, of how bias can be amplified and I think this is my favorite example. It is based in New York. So basically here you have the flow of how things happen. So you have an offender, the judge has to decide if you get bail or not. So typically in most of the world the judge has to think if you are reoffending if you're out, or and if you will appear in court later. Well in New York, you only have to answer the question of the judge only has to think if the person will appear in court or not. It doesn't matter what kind of crime you are accused of doing. So this is the first kind of special thing. And then of course if you get bail and you can pay, you go away. If you cannot pay, then you go to prison. But for example, in the US, they're the people that, from what you can borrow money to pay bails, and I guess they know each other and they have to do the same prediction of the judge. So will this person appear in court and pay the money back, of course with some interest. And then if you don't get bail, you go to prison. So the one problem here, like in many other problems, we don't know all the data. For example, we don't know what would have happened if the person got bail, and, for example, will not reoffend or will not appear in court. So to fulfill this missing data, we need to do data imputation. And then that means that these kinds of models are more complicated than other standard predictive models.

10:51 So what were the results, and this was some work done for the National Bureau of Economic Research of the US, not for the International Justice Department, because the judges are not interested in this, I guess. But of course, the US government wants to spend less money in prisons. So the model, if the model is correct, remember is a prediction, this is a model, the model can be mistaken. But basically, the model says that they could decrease crime rate in 25%, according to data, or decrease the prison rate in 42%, keeping the same crime rate. So depends on what they want to keep. So the model was much better than the judges. And I will go fast, because I have a lot of material. But the set was split in five parts. One to do the imputation. One was locked in a box, so anyone can basically verify the results with data that the model has never seen. So they use Gradient Boosted Decision Trees.

Transcribed by https://otter.ai

So these are like decision trees, but they have a degree of non linearity and data loss and interpretability. And the only demographic variable was the age of the person.

12:06 All the rest was basically just a case. So there is no variable where you can learn anything related to race, like for example, address that is correlated with, with race. Well, these are the results of the model. Here, you'll see, basically, the defendant is 40 almost 49% Black, 33%, Hispanic in total 82%. And here, I put in red, the actual percentage of that population in the state of New York, so they are only 32%. So it's already a bias on what police bring to the court. So it's more than doubled from 32% to 82%. And the judge, clearly the judges are a bit racist, because the percentage of black goes from 49% to 57%. So it's an increase there and lowers a little bit in Hispanics. So what about the model, what the model learns from the data is to be even more racist. So here it went from 57% to almost 60%. Also learnt to be less racist to Latino population, and then went down to 30%. But overall, they still increase to 90%, from 89%. So this, you know, was more racist. If we use the base rate of the best judge, so the less racist judge, then still, we can improve the system by 23%. Basically, dropping the crime rate and keeping the same number of people in prison.

13:45 So here we have the first dilemma, because why this is happening? Well, if taken from the paper, so this figure says, if the model is right, and you want to send additional people to prison, then you take people from the top of say the risk category. But if you see what the one of the judges group of judges does in this case, the second leniency quintile, then you see that they are taking people from all over the risk, according to the model. So if the model is right, the judges are almost doing random choices on this. So why is that happening? Well, that's happened because human decisions have noise. So the dilemma is what is better: a biased algorithm that is just in the sense that for the same two cases, do you get the same sentence, bail or no bail, or a noisy judge? And in reality, it's not a noisy judge, it is a noisy racist judge. And here you see the four different cases of something being exact, just noisy, just biased, or biased and noisy. So I'm going to start here - are biased. But judges here are biased and noisy. So it could be even worse. And this appears in a paper on Harvard Business Review in 2016. And I have been talking about noise after that. But good luck last month, Daniel Kahneman, with colleagues, and Nobel winner of economy, published a book on noise. I'm reading it now, with other examples where basically a human deficiency may have a cost that is even larger than, say, biased decisions.

15:38 So just quickly two other examples of bias. You have issues of gender, and you see here, many different companies. Maybe facial recognition system is the one that has been in the spotlight like in the “Coded Bias” movie. And it's very interesting, because if you see how facial recognition has been developed, what has been developed with many, many data collections that have been growing exponentially. So you see here that the y-axis is logarithmic. So basically in time from 1992, to today, the size of these collections have grown exponentially. And if you check how this data has been recollected, you see that in the last two periods, so basically, after 2007, if you see the lines here, most of these faces come from the web. And that means that they have no consent. So the person didn't consent on using his or her face for this purpose. So in some sense, there is an ethical problem all the way into the developing of this technology.

16:54 And for all these takes on, for example, especially black faces, the top companies selling this kind of software, decided to stop selling to basically to law enforcement this kind of software. Here, we have Amazon,

Transcribed by https://otter.ai

Microsoft and IBM. If you see the “Coded Bias” movie, these companies have a completely different reaction to the fact that they were doing mistakes.

17:23 But still, this was the US and ACM, one year ago, asked for a suspension of using this technology. I was part of the committee that signed this, this statement asking for this. But the software was already there, the police bought the software before. And for example, in September, we had the second well known case of black man, accused wrongly of doing something because of a mistake of the technology. And very interesting, this paper also of last year that shows that it's impossible to get consent for facial surveillance so we have a problem and, and we'll get back to that. And particularly in the US, several cities have forbidden the use of facial surveillance like Oakland and San Francisco, some cities in Massachusetts and so on, have forbid this kind of technology.

18:24 We have examples in gender, gender bias in language translation, for example, here. This is something that has been fixed by Google - end of last year. But if I say “he is a babysitter, she is a doctor” and I translate it to Turkish, when I do the translation back, I get the opposite genders. But if you try today, you get only female ones? If you do just one sentence, you will get the two possibilities. The two binary possibilities which are also going to now we know it's not correct. But here we even get “She’s a doctor” with a capital S. Maybe this is an artifact of using things like birth in these models. Then the same information instruction. Here you have a classical example from NeurIPS, 2016, where you see that basically most of he/she analogies have a gender stereotype like diva has a negative connotation but superstar doesn’t in the case of a man. So I also did some research on why this is happening. And this is more than systemic bias, because also most famous journalists in the US are men. But this is not the case in colleges. So there's another bias here. It is very hard for a woman to get a good job as a journalist in the US.

19:54 So another example from Stanford - word embeddings. That basically also has stereotypes in this case for, for professionals. So you see that the janitor is Hispanic, the secretary is Asian. And for example, the administrator is white. Language models. These have been on the news several times. So here I'm using a table from, from a paper I will discuss a little bit later. But you see how these things have grown in the number of parameters from hundreds of millions to basically 1.6 trillions. And recently there was a Chinese one of 1.7 trillion parameters. If you think about that, that's getting close to the number of words that you're using to train this model. So at some point you will overfit. And what is the problem with this? Well you learn all the biases encoded in the text. For example, here is a paper that was published last January, that shows that you have anti-Muslim bias. For example, if you write this sentence: “two Muslims walk into a”, and then you ask GPT-3 to complete this, most completions will be violent, for example. Say for example, “will start shooting at will” or “opens fire” and so on. And they look at all religions, and basically, for Muslims there were four times more violent completion than for Christians. And it's interesting to see that the less violent in the news, at least, are the atheists and the Buddhists. So maybe religion is part of the problem.

21:32 But it can be much more complicated. So for example, there have been complaints about ads.There have been complaints as there are about for example, grades to enter the university where the government thought they were doing something nice, avoiding the students giving an exam on time with the COVID. But basically just amplified the inequalities that already existed in the system. And this one in the news and basically they didn’t use it. Or could be even more complicated. Like this interesting case in Italy, that was settled in January of this year,

Transcribed by https://otter.ai

where Deliveroo, which is a platform to deliver food to people, was found that they were doing indirect discrimination, because they was giving less work to people that couldn't deliver a dinner time. So you know, dinner time is required, you have more than most demand for delivering food. And the algorithm learned that to sell more, well you should give more jobs to the people that could do that and not to the people that couldn't do that. Without balancing that in other times of the day. And this was the problem, because the algorithm didn't check the reasons that those persons didn't deliver food at that time, for example, they, they, they need to feed their kids, they need to feed an old person, or even they don't want to work at night. And these are valid reasons not to work. So this was very important. And also, even more recently, in April, the Dutch court basically said that six drivers that were fired for accusation of fraud had to be reinstated by Uber, in different places, some were in London, some in Lisbon, and so on. And the interesting thing here is that Uber decided not to go to court. Because, for example, they were asking for the code. Like in the Deliveroo case, they asked for the code, and they gave the code and then they found that there was this discrimination. Uber I guess decided not to go because they didn't want to share the code of this fraud detection application, which is very interesting. They were still fined $150,000 for firing these people for several months. But they couldn't work because of this.

23:58 But it can be really, really bad, even worse than these cases. And my the worst example, I think is what happened in the Netherlands with childcare benefit, that after several years, not only the ministry that did that had to resign, the whole government had to resign because 26,000 families were affected by basically tell told, telling them that they were cheating on childcare subsidies, and they had to basically return the money. Some people had economical stress. Other people have psychological stress. Other people had to go back to their native countries, many were immigrants. And this is a case where technology is affecting the vulnerable people, poor people, immigrants. It is another example where technology benefits the rich people, but you have the poor people suffering it. So of course again, there was no intention for this to happen. But there were so many questions that no one was asking. We're not asking here to avoid this to get a whole government resigning.

25:14 So let's go to the second issue in order of importance for me. This is what is called physiognomy. So physiognomy is a pseudo-science that started with the Greeks and was very popular in the 18th and 19th century. But now it’s coming again. For example, Kosinski in Stanford, had this predictive model of sexual orientation in 2017. Of course, LGBT groups were not happy at all. Is this ethical? Of course not.

25:48 Then in China, and later in the US, someone, some people trying to use facial recognition, to basically, again, criminal profiling. And of course, they've had very bad press. Kosinski gave strikes back on January, this January, for example, to say that they could predict 70% political orientation, most probably is predicting spurious correlations, but of course, will not work for everyone. And it's not really facial recognition, it’s facial biometric, we're not trying to recognize any person particular.

26:27 So this is a particular case of phrenology that was very popular in the 19th century. In fact, if you go ever to Torino, I would recommend to go to the former house of the Dr. Cesare Lombroso, which was one of the people leading in Italy, this belief that, for example, you have a different skull if you were a criminal. And he collected hundreds of skulls - of course of poor people that were in the morgue, and no one want to pick them up. But at least he was consistent with his belief because he left also his skeleton. I guess this was his grand truth for a nice person - a person that collected 100 skulls.

Transcribed by https://otter.ai

27:10 But it can be worse. This paper now, not from Stanford, but from MIT that was published in 2019 that claims that given your voice you can get your face. I may be able to believe a bit that your voice defines a bit the shape of your mouth, and maybe the neck. But what about your eyes, your hair, and so on? There is no causality. But some people believe this causality does exist. So I can have now my master phrenology algorithm that basically gives me your voice and I predict your face. And there's another work, it's a filed patent by MITRE that claims that from your face, I can get your name. Of course, this will not work with all adopted kids from other ethnic in the world, especially developed countries. So I can use all these things to profile you and then do something with it. And the problem is that these profiles, in some cases, are illegal. So so. So very dangerous, if you start believing this.

28:22 The second kind of issues - unfair ecommerce. So this has to do with the filter bubble. So I may know someone for a long time, but I will never be able to offer him something new that he will like because he's not in the data because it's new for him. So search engines or recommender systems are using things like diversity, novelty serendipity to break this filter bubble issue. But still, this is difficult to break it completely. That's why quality filters exist. And, for example, some people claim this case in the past that Amazon was racist on the recommendations.

29:04 But this is more dangerous. The danger is the feedback loop, which is the third source of bias. It is not the model, it is not the data, it's the interaction with the system. And why is that because it's like an iceberg, a big iceberg, that you have exposure bias. You don't see everything, you see only the recommendation that the system is giving to you. And these recommendations are being written by the system. So the system is partially writing its own future. So the data is not independent of the system. And it's worse than iceberg, because you're only looking at the top of the iceberg - the ice cube on top of the iceberg because you're seeing three things of maybe thousands of things. So it's much worse than an iceberg. And you get this problem of the Matthew effect of the rich get richer and the poor get poorer. And then you get popularity bias that will affect the long tail of items and the long tail of sellers and the long tail of users. And as usual, the long tail will be discriminated. Fourth problem - stupid models. So models that cannot deal with ambiguous semantics or models that cannot deal with irrational behavior. And here I have this quote from George Box that "All models are wrong but some are useful". While this is true, this was for statistics. This is true in AI too. All models are wrong, they are a simplification of reality, but some are useful. And examples of these kinds of stupid models.

30:35 Elon Musk said, "Use Signal". And then these systems that work on the New York trade exchange, pick up this because they use Twitter data. And they think that it's talking about buying stock of this medical company's tech service called Signal Bands, which is the closest name to signal. Of course, that was not true. But the company grew their value in more than 400%. Then went down when people discovered error, but it still had a benefit. But things can be even more stupid. For example, this is a hate prediction example, where Facebook basically deleted the page, the website, the Facebook page of a French town, that is called Bitche. So you can imagine why the models thought that that was a problem. But also, it had this page three weeks on the internet because there was no human in the loop to basically receive a complaint from the town from the major town and check if something was wrong or not. A clear mistake. And these are related to adversarial AI. For example, here you have another example from Japanese researchers, where if they change one bit of an image, then you get a

Transcribed by https://otter.ai

different label. For example, a cat becomes a bird, a deer becomes an airplane, a ship becomes an airplane, the cat becomes a dog and so on. So you ask what these things are learning. Surely they're not learning what we understand by, for example, an animal. And these are the limitations we need to be to realize that this model have a lot of limitations, for example, is very hard to to forget and filter what

32:36 I recommend to read this amazing story by Jorge Luis Borges - “Funes, The Memorius” - a person that couldn't forget and then had to tell you everything that did. But sometimes it's not in the data, like for example, you see this picture. You may guess what is the animal, but the AI system will be very hard to do it because you cannot learn of course what is not in the data. And there is no such thing as I have all possible examples in the data. Why? Because some samples are from the future, so you cannot have a collection that will include all possible examples that you need to match. And as I said before, for me accuracy is not the main issue. It is the impact of errors. Because, for example, in medicine, false negatives are worse than false positives, like detecting an illness. And many errors are not human. So if you see some mistakes done by GPT-3, you will see that no person will write that sentence. So those are the ones that have a large negative impact when they are negative. In some cases, it's just funny because they don't make any sense. But these are the bad errors, the non-human errors. And we need to be humbled then. So for example, if you have a classification system, maybe you need to say “I don't know” when you are not sure of what you're seeing like an animal.

34:05 And the last issue is the waste of resources. So I get back to this table from Bender, Gebru et al.. And this is a paper that shows that for example, one issue is that you are training these very large examples. You are basically having a very large footprint, the footprint is 57 years of the human life of a normal person. And then you spend between $1 and $3 million in training these examples. So are we wasting electric energy, liberating CO_2 and spending money just to have bias on these models? And this is the paper that is the famous Stochastic Parrot paper but by Emily Bender, Timnit Gebru, Angelina MacMillan-Major and Margaret Mitchell that couldn't write her name, because she was still at Google at the time. And remember, I think Gebru fired because of not wanting to take her name out of this paper. And this was published last March.

35:17 And this is the meta ethics problem. What happens when the people working in AI ethics in a big company have issues with their work? And later anyway, Margaret Mitchell also followed the same path. And she was fired in February because of trying to look in her own email about discrimination regarding Timnit’s case. So a meta ethics issue that is really bad. And with a group of colleagues, we wrote this paper towards intellectual freedom in an AI ethics community, but it's available as part of this new Springer journal on AI and Ethics. And you can download it from their site.

36:05 So let's go now to the second part. So what do we want from algorithms? One possibility is this statement that was done in 2017, so it's already like four years, on the properties that algorithms should fulfil to be transparent and accountable. So these are the properties, I will not read them. But awareness is a very important one, if you're not aware that you can be biased, that's already a problem. And this problem starts with us. If a person is not aware of their own biases, they will never solve the problem. And systems do not need to be perfect. But since that for people, they need to be much better than us. So when an autonomous car kills a person - that was Uber. This was big news. Any day, a lot of people die because of a car accident. And this is related to something given to and this new book by César Hidalgo and collaborators that basically explores this problem of

Transcribed by https://otter.ai

how humans judge machines. And they find that basically, humans are much harder on machines than humans. So we allow human errors, but we do not allow machine errors. Here's why - that machines are learning from humans. So this is another thing. So if we're learning from humans, why we want them to be better than us? In the best case they could be the same as us. But of course, we want them to be better.

37:40 And then there are pragmatical question with these properties. So to which part of the system applies? To whom is important? Can they be fulfilled simultaneously? Do they make sense together? For example, transparency is something that you ask in Latin America, because we don't trust institutions. But accountability, something that is asked say in the US, because institutions are adjusted, and they are very different. To demand transparency is very different to wait for accountability. And then here I have a table of 30 different properties of systems. And not all of them related to AI. But remember, this is not only for AI, it could be for any system. And you see that some things are important for data. Some are important for the algorithm, the model itself, some things are important for the system. And some things are important for the governance of the system. And to whom it's important? Some are important for justice, some are important for the government like regulation. Some are important for users like user experience. And some are important to society, for example, discrimination. And basically sorting this out. These are my proposed taxonomy of properties. There are six main classes, which are awareness, data, provenance, completeness, usability, transparency, and responsibility. The ones that are underlined are the ACM original ones. And you see that here, there are many more, and they are not exactly at the same level. So this is something I'm still working on, before publishing it. And then some people, it's interesting, the governance, for example, this is from a paper by Ben Shneiderman, one of the fathers of HCI, and you see that he has like three levels: the team, the organization, and the industry. And one question is how we can develop software with the help of AI so that software is much better than current software.

39:40 There are also some legal and ethical colonialisms. So basically, the UK or France, Spain, Portugal, and Muslims have exported their ways to see the world to the rest of the world. So this is what is called technological humanism. And remember, humanism was behind this leveling the path for humanism, it's not that human. And it's just a word used by pale men to basically encode philosophy. And here you have also the example of Ubuntu. So Ubuntu is a philosophy and ethics from the south of Africa. And basically it says "I am because we are" so the focus is in the community and not on the person. And there is a very interesting paper saying that Descartes was wrong because "a person is a person through other persons" and the others, so this is important. And the other day I saw this joke about COVID that said "Descartes will say all these people that doesn't think still exist".

40:57 Next topic: identity, data protection and privacy. Remember, one thing is to be anonymous. Another thing is to have your data private. And I think of your privacy as a person, and you know, privacy is a human right. Some people like Carissa Véliz, in a very excellent book "Privacy is Power" says that privacy is a collective right. But I believe it's more like you have more like a relational or collective dimension of the right, but still, it's just a human right. And these are related to what are called the NeuroRights and will not I will not go there. But that's another answer. Another important aspect if we start splitting our body, in body and brain and brain has different rights than than humans that also include the brain. So here you have also different legal ways to see things. For example, one thing is to do everything that's not forbidden, like freedom of expression versus I can do only what is allowed, like GDPR in the EU.

Transcribed by https://otter.ai

42:04 And also you have digital nudging, which is everywhere in the internet. And then hear things like awareness that you are being nudged. Consent or legal basis to do whatever I think people wants to do with you. And minimal data collection to achieve your goals are not more and minimal time to store the data and not forever.

42:30 So here the Article 22 is very important that regulates automated decisions by systems. And I will not read it but the important thing is, you have the right to contest the decision. So a person has the right to contest the decision. What this means in practice is that if you want to give people information about the processing, you need to have interpretability. If you need to explain the decision, then you have to have explainability. And if you want to make sure that your system works, as you are saying it works, then you need to have continuous validation testing and maintenance. So here, there is an interesting example in France of GDPR in action. They put cameras in a high school, in two high schools, in the south of France. And the court said that was illegal because they didn't have the competence to design it, they didn't have the consent of people, and also the solution was not proportional to the goal. Like for security, maybe you just need the person in the door and not a facial recognition by cameras. And one problem here is that this law is a kind of one size fits all. And that's important that that's complicated, because it means that the same law for all human rights and law for all domains of application, like health and education, and also for all sizes of companies doesn't matter if it's a small store or a multinational.

43:59 And then you get to regulation. So last year, the US government started antitrust cases against Google and Facebook and recently also to Amazon. So all the big companies are being targeted. And during Trump's government, there were several trials in 2019 and 2020 to start regulating these and they were not successful. One one was even from Kamala Harris, the current vice president. And now things may happen because lawmakers are introducing antitrust reform that could reshape many multinationals.

44:39 So another important recent thing that happened here was about two months ago, in April, the proposal for the EU regulation for AI that should be in place at the end of 2022. And here there are all the things that appear. The most important ones are forbidden uses like for example anything that causes physical or psychological harm, and also things like high risk systems. And here the problem will be to distinguish between low risk and high risk systems. So basically, we're putting a continuum of risk in two categories and that will be a problem for sure. Some cities are registering algorithms like Amsterdam and Helsinki. Some companies are doing audits - third party audit or direct audit. And for example, a very interesting paper from the audit earning team at Northeastern where they agreed with a hiring company that was using AI as a hiring solution to publish the audit independent of the results of the audit. So, this is a very interesting example for real audit in a company.

46:00 There are many bad human practices, I think I am running out of time, so I will skip this. You can see the video later. But what we can do? So we have problems on how we look at the solutions. We have design and implementation issues. For example, do systems reflect the characteristics of the designers or of the coders? In my past, bias on the web talk, I was using this example from University of Virginia that basically shows that 19 teams did all different things doing data analysis, but recently Norway, they showed that that one specific problem, the bias of the coders was translated to the code. I think we will see more of this in the future. And then we should check if we're doing the right evaluation, right experiment, right test data, right matrix and the right baseline. The

Transcribed by https://otter.ai

things we can do: analyze for biases in the input, recollect more data in the regions that you have less information or the problem is more complex. And you need to have the model aware of the biases, and also the user should be aware of all the biases all the time so the user interface should help them to do this. And of course, in evaluation, you shouldn't fool yourself and be very hard on yourself to test it.

47:20 So recommendations: design for people, remember the limitations of our systems, don't reproduce the past, have an ethics board (and at least 50% external, otherwise you have a conflict of interest), enforce a code of ethics, explainability, more evaluation, and I prefer to talk about humans in control where the designers were playing God and not having only humans in the loop. And remember, ethics is like privacy. In some sense, you have a collective dimension. It depends on the ethics of your providers, and also the ethics of your clients. It's not only about you. So to finish, try to be aware of all your biases, and also your ethics. It is complicated. And here I finished with one question and one fact. So can AI algorithms ever be ethical? This is a very interesting essay by Hazel Henderson published this year. And then David Lauer says you cannot have AI ethics without ethics. So I will stop here to answer questions. Thank You.

Transcribed by https://otter.ai