Fairness in AI systems:
From social context to practice using Fairlearn

Tutorial at SciPy 2021 | July 13 2021

Manojit Nandi, Miro Dudík, Triveni Gandhi, Lisa Ibañez,
Adrin Jalali, Michael Madaio, Hanna Wallach, Hilde Weerts

Fairlearn

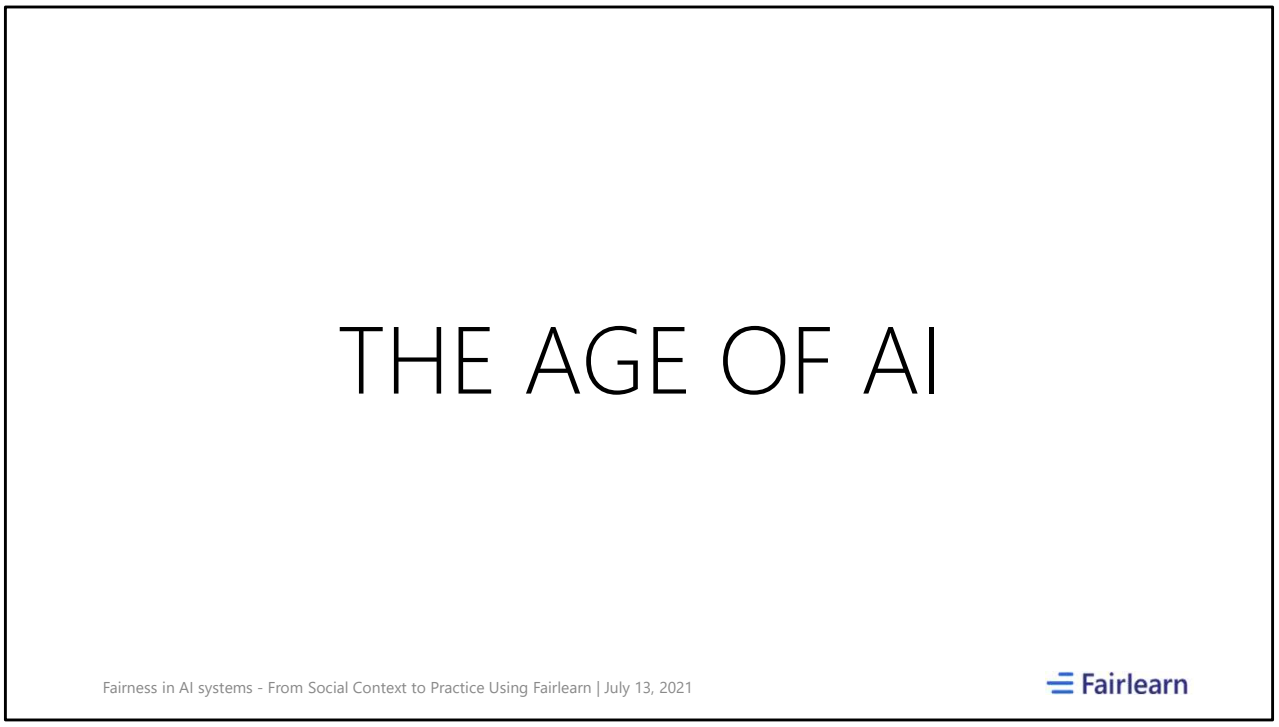
Acknowledgements

The slides for **overview of AI fairness** developed by Hanna Wallach, Jenn Wortman Vaughan, and members of Microsoft's Aether Fairness and Inclusiveness Working Group

Our **health care scenario** draws motivation from the work by Obermeyer et al. (2019).

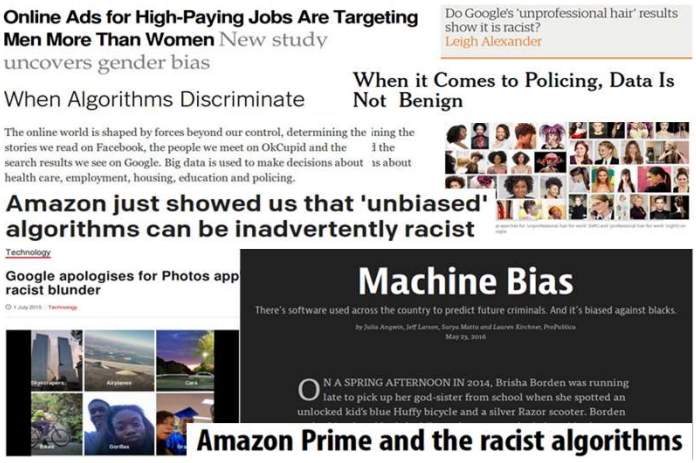
Our **dataset** was developed by Strack et al. (2014).

*



So, at the risk of stating the obvious, we're living in the age of AI. Machine learning is everywhere and, at least for now, it looks like it's here to stay. And, in many ways, this is great. But, at the same time, we're seeing that the new opportunities brought about by AI and machine learning also raise new challenges….

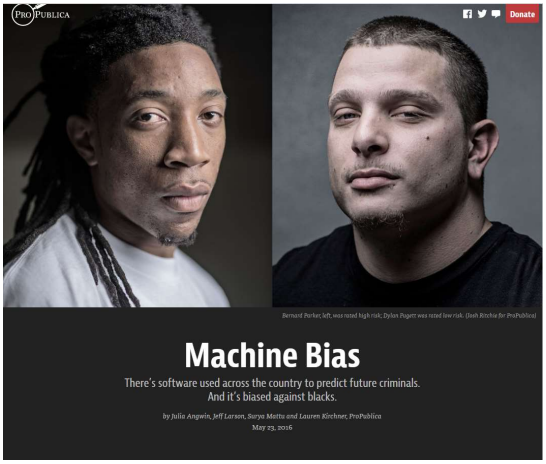
The media



Fairlearn

In particular, challenges that have received a lot of attention in the media and have really highlighted how important it is to get AI right – to make sure that AI does not discriminate or further disadvantage already disadvantaged groups of people.

Justice system



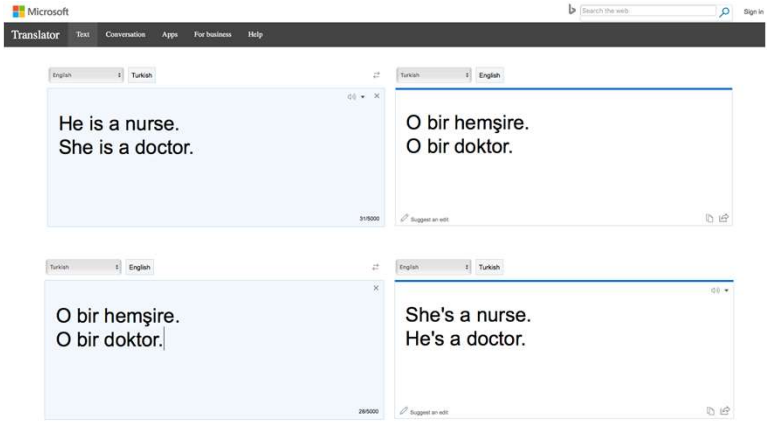
[Angwin et al., 2016]

Fairlearn

Some of these media stories have involved high-stakes decisions where AI systems are used allocate opportunities or resources in ways that can have significant negative impacts on people's lives. For example, I'm sure many of you heard about the ProPublica investigation a few years ago, which showed that COMPAS, a widely used recidivism risk assessment tool, incorrectly scored Black defendants as high risk more often than white defendants, while incorrectly scoring white defendants as low risk more often than Black defendants.

Resource: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine translation



[Caliksan et al., 2017]

Fairlearn

Other stories have involved much more mundane AI systems. For instance, researchers at Princeton found that translating “He is a nurse” and “She is a doctor” into Turkish, a genderless language, and then back into English yields the stereotypical (and, in this case, incorrect) “She is a nurse” and “He is a doctor.”

Resource: https://arxiv.org/abs/1608.07187

These stories are examples of AI systems behaving unfairly. But even though we can often spot fairness-related harms in AI systems when we see them, there's no one-size-fits-all definition of fairness that applies to all AI systems in all contexts. Moreover, fairness in AI is a fundamentally "sociotechnical" challenge, meaning that it cannot be approached from purely social or purely technical perspectives. Taken together, these factors can make it a pretty daunting landscape to navigate. But, although there are few easy answers, there are a variety of strategies emerging for assessing and mitigating fairness-related harms, as well as a deepening understanding of the challenge throughout society. And that's what we will explore in the tutorial today, in the context of a scenario drawn from health care.

Outline of this tutorial

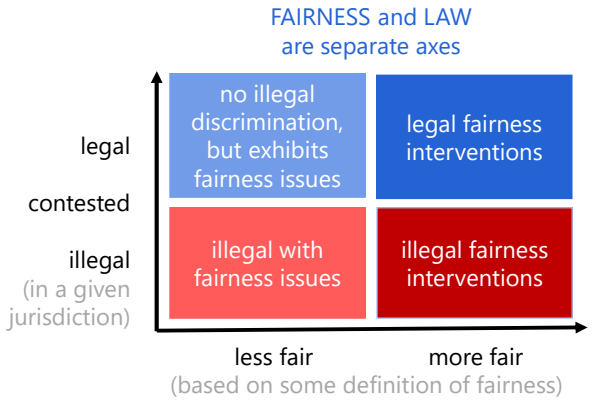
- Overview of fairness in AI systems
- Introduction to the health care scenario
- Assessing the fairness of an ML model
- Mitigating fairness-related harms in ML models
- Conclusion

Fairlearn

What we will not discuss: anti-discrimination law

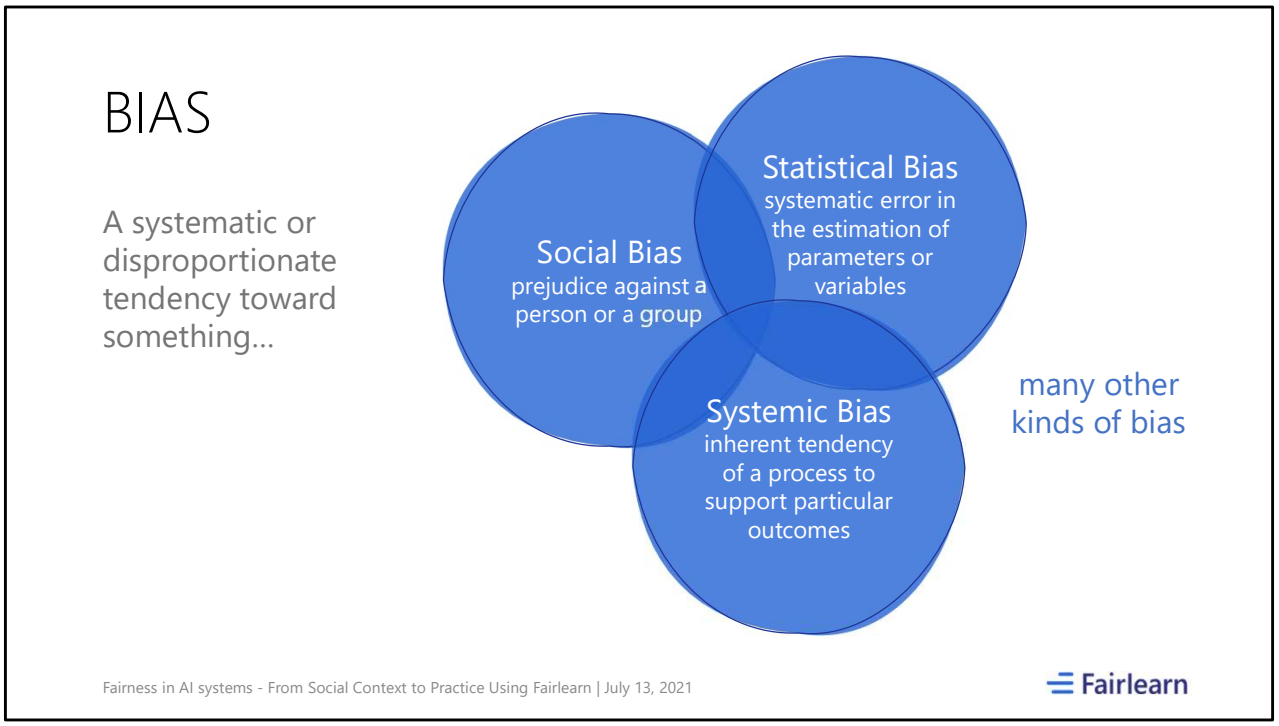
Fairness is related, but distinct from anti-discrimination law.

We avoid legal terminology like: discriminate against, protected class, disparate treatment, disparate impact, …



Fairlearn

While fairness is related to the concepts in anti-discrimination law, in this tutorial we will just focus on fairness and largely do not touch on the legal considerations. However, one important aspect to note is that some fairness interventions could be illegal, and, conversely, there are AI systems that follow all law (including anti-discrimination law), but still exhibit severe fairness issues.



So, if you've read anything about fairness in AI systems you've probably seen the word "bias" get thrown about left, right, and center – often as a catch-all way to describe any unfair behavior of any AI system and any possible causes of that unfair behavior.

However, in this webinar, we're going to avoid using the word "bias" wherever possible – and we recommend that you do the same. This is because the word "bias" is ambiguous and means very different things to different communities – for example, statistical bias vs. societal biases.

Most issues arise at the intersection of social, systemic, and statistical bias…

Instead, we're going to talk about the fairness-related impacts that AI systems can have on people—that is, specific types of fairness-related harms. This precision is useful for disentangling who might be harmed by a system and in what ways, as well as suggesting different assessment and mitigation strategies.

On that note, I also want to emphasize that because there are so many different reasons why AI systems can cause fairness-related harms, it is not really possible to fully “debias” a system or to guarantee its fairness. Because of this we also don’t recommend using words like “debiasing” or “unbiased,” which can set up unrealistic expectations. Instead, we recommend talking about assessing and mitigating fairness-related harms.

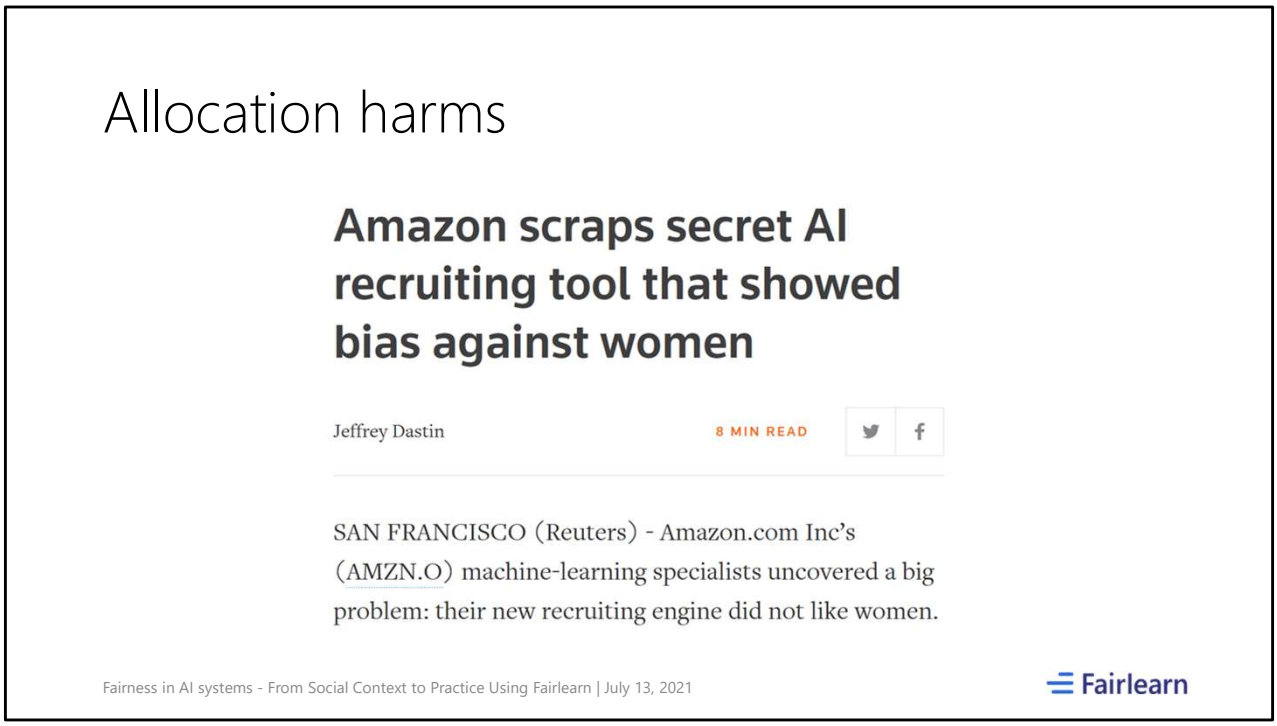
Alright, so with that, let's move on to different types of fairness-related harms.

Broadly speaking, there are three main types of fairness-related harms that can be caused by an AI system: allocation, quality of service, and representation harms.

To explain each of these three types, I'm going to run through some illustrative examples, starting with the most well-known type – allocation. But before I do that, I want to note that these types are not mutually exclusive. It's possible for a single AI system to exhibit more than one type of harm. And, in fact, many of the examples that I'll show you could have been used to illustrate one or more of the other types as well.

Resource: https://www.youtube.com/watch?v=fMym_BKWQzk

Allocation harms



First, allocation. As I said before, many media stories have focused on high-stakes decisions where AI systems are used allocate opportunities or resources in ways that can have significant negative impacts on people's lives. For example, Amazon abandoned its automated hiring system after finding that it amplified the existing gender imbalance in the tech industry by withholding employment opportunities from women.

Resource: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Quality-of-service harms

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

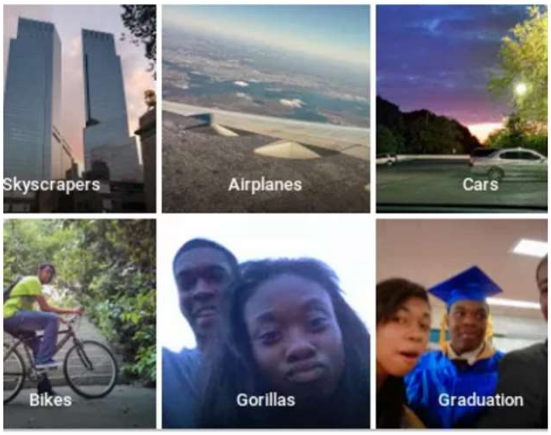
[Buolamwini & Gebru, 2018]

Fairlearn

Quality of service is all about whether a system works as well for one person as it does for another, even if no opportunities or resources are extended or withheld. For example, researchers found that three commercial gender classifiers had higher error rates for images of women with darker skin tones than for images of men with lighter skin tones. Like accessibility issues, quality of service harms can raise questions about respect, dignity, and personhood. Imagine how a user might feel if a system repeatedly fails to recognize her voice, but easily recognizes those of her peers?

Resource: http://gendershades.org/

Representation harms



Fairlearn

There are many different kinds of representation harms including stereotyping, denigration, over- and under-representation. The harm of denigration occurs when an AI system is itself part of a process that is actively derogatory, demeaning, or offensive. For example, a few years ago, Google Photos infamously mislabeled an image of Black people as “gorillas.” This mislabeling is harmful not just because the system made a mistake, but because it specifically applied a label that has a long history of being purposefully used to denigrate and demean Black people.

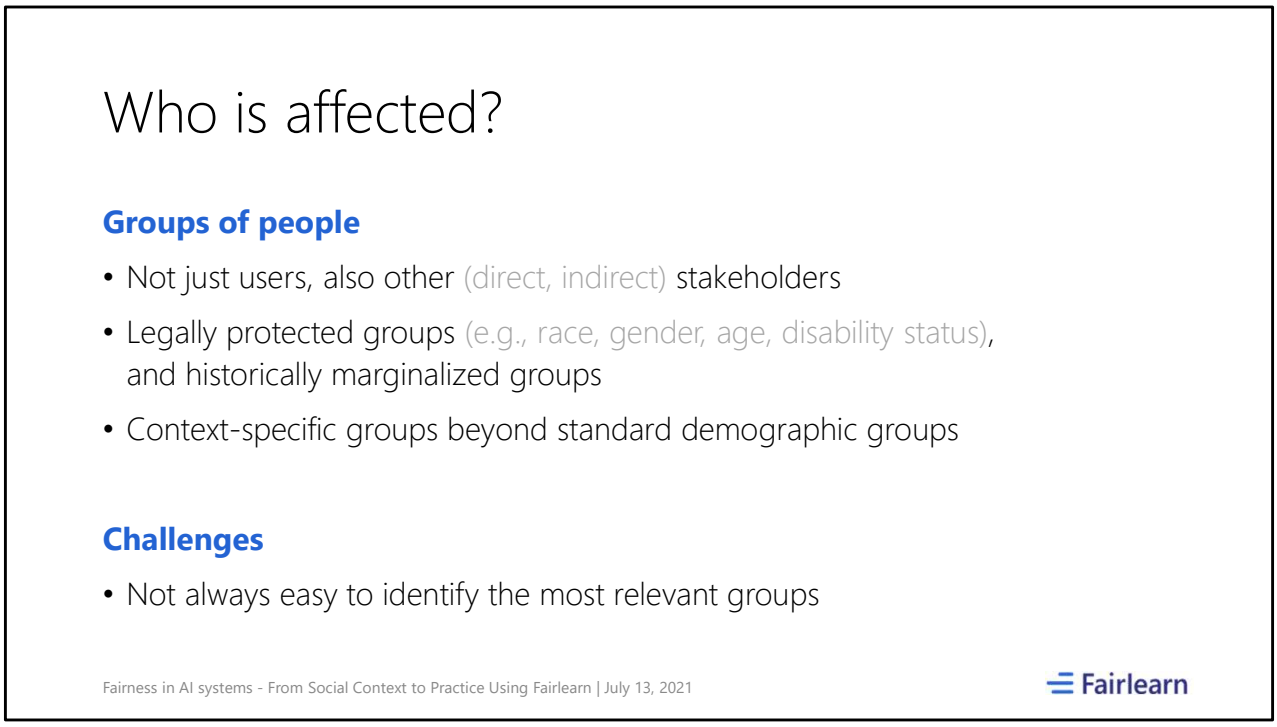
Resource: https://www.cbsnews.com/news/google-photos-labeled-pics-of-african-americans-as-gorillas/

	Allocation	Quality of Service	Representation
Automated hiring system does not rank women as highly as men for technical jobs	x	x	x
Gender classification systems have higher error rates for women with darker skin tones.		x	x
Machine translation system exhibits male/female gender stereotypes.			x
Photo management program labels image of Black people as "gorillas".		x	x

Fairlearn

I want to reiterate that the three types of harms are not mutually exclusive. A single system can exhibit more than one type of harm, as I've indicated here on this slide, and can even exhibit different harms for different groups of people, too. Finally, fairness-related harms can have varying severities. For instance, unfairly denying someone bail is a more severe harm than labeling an image of a female doctor as nurse. But it's also important to remember that even relatively "non-severe" harms can make people feel alienated or singled out, and their cumulative impact can be extremely burdensome.

Resource: https://www.youtube.com/watch?v=fMym_BKWQzk

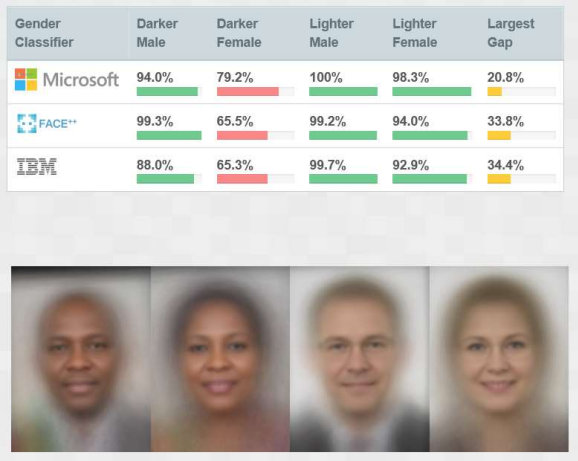


So, who is at risk of experiencing fairness-related harms?

Well, sometimes it's the people who will use or operate a system, but it can also be other stakeholders who are directly or indirectly affected by a system, either by choice or not. For example, in the case of a facial recognition system for workplace building access, the system operator is not the person whose face is being recognized and thus not the person who is most immediately harmed if the system makes a mistake.

I also want to emphasize that although media stories often focus on groups of people that are protected by antidiscrimination laws, such as groups defined in terms of race, gender, age, or disability status, there are actually many different groups of people that we want our systems to treat fairly and it's not always easy to identify the most relevant ones, which can even be specific to the domain or use case. For example, in the case of an automated essay-grading system, whether someone is a native speaker of the language may be more relevant than their age or their disability status.

Intersectionality



[Buolamwini & Gebru, 2018]

Fairlearn

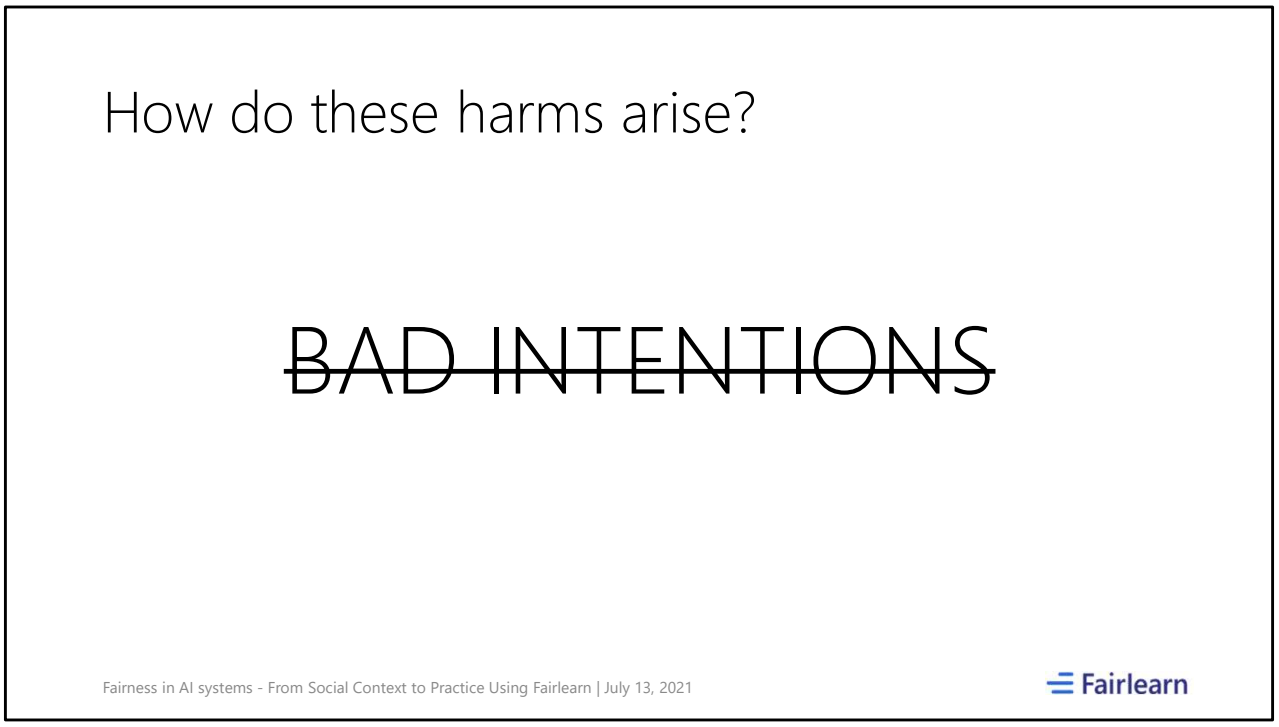
It's also important remember that groups intersect and people within those intersections may be at risk of experiencing unique harms that might be obscured by considering only non-intersected groups. Returning to the gender classification example that I mentioned earlier, error rates were significantly higher for images of women with darker skin tones than for images of women overall or for images of people with darker skin tones overall. The researchers who conducted the study were only able to uncover these disparities by assessing system performance with respect to skin tone and (binary) gender at the same time.

Resource: http://gendershades.org/

How do these harms arise?

Why *do* AI systems cause fairness-related harms? Well, if you look at the way that AI systems are portrayed by the media, you'd naturally assume that the only reason is "biased datasets." But, in reality, the situation is more complex than that – not to mention the fact that phrase "biased datasets" is often used to refer to datasets with any number of undesirable characteristics, such as reflecting stereotypes in their content or too few data points about some group of people.

Over the next few slides, I want to run through some of the different reasons, with examples, so that you'll know what to look out for in your own systems.



First, though, I want to explicitly emphasize that although there are many reasons why AI systems can behave unfairly, these reasons seldom include bad intentions on the part of the teams responsible for developing or deploying those systems.

Datasets: **societal biases**

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Fairlearn

Alright, so starting with datasets, some AI systems behave unfairly because of societal biases that are reflected in the content of the datasets used to train them. Returning to Amazon's automated hiring system, the system withheld employment opportunities from women because the data with which it had been trained reflected the existing gender imbalance in the tech industry.

Resource: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Datasets: **insufficient coverage**

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

[Buolamwini & Gebru, 2018]

Fairlearn

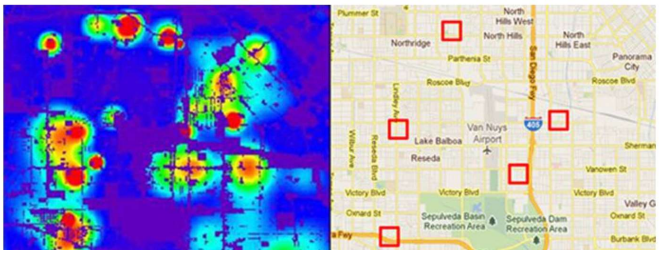
Other AI systems behave unfairly not because of societal biases that are inherent to the datasets used to train them, but because of other dataset characteristics, such as too few data points about some group of people. For example, the reason why the three gender classifiers that I mentioned earlier had higher error rates for images of women with darker skin tones was because there weren't enough images of women with darker skin tones in the datasets used to train them. In some cases, the relative proportions of different groups of people in a dataset may even reflect reality, but this sampling strategy can still lead to insufficient data about groups that are smaller.

Resource: http://gendershades.org/

Task definition: **choice of label**

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Fairlearn

As another example, sometimes AI systems behave unfairly because of assumptions about what a particular dataset captures and how this relates to the system purpose. For instance, a crime-prediction system trained on an arrest dataset relies on the assumption that the number of arrests in a neighborhood is a good proxy for the amount of crime committed there.

Resource: https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

Task definition: **choice of label**

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But cr

arrest rate ≠ crime rate

Fairlearn

But this assumption fails to account for high rates of arrest without conviction or for over-policing in less-affluent neighborhoods, so the resulting system may make inaccurate predictions that reflect historical policing practices, rather than the occurrence of crime itself.

Resource: https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

Task definition: **choice of task**

[Wu & Zhang, 2016]

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

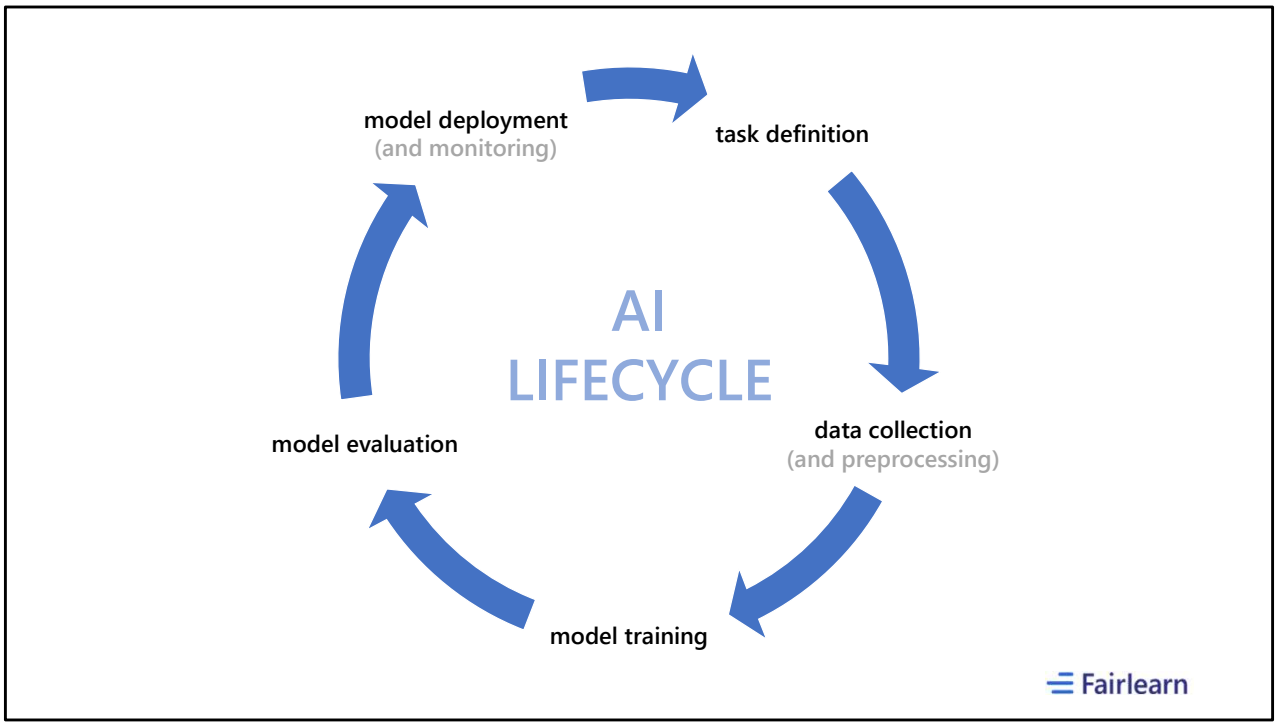
Figure 1. Sample ID photos in our data set.

Fairlearn

Finally, sometimes the issues with the task definitions are more fundamental and can reflect misunderstanding about what AI can and cannot do. As an extreme example of what can go wrong here, in 2016, a paper came out by a group of researchers who proposed training a facial analysis system to predict who is going to commit a crime based on images of people's faces. This is extremely worrying for a whole host of reasons and could lead to substantial harms for people who are misclassified. I hope it goes without saying that, even with good intentions, this is a questionable system purpose – and not something that AI is actually capable of doing.

Resource: https://arxiv.org/pdf/1611.04135v1.pdf

Resource: https://www.wired.com/2016/11/put-away-your-machine-learning-hammer-criminality-is-not-a-nail/



The examples I just showed you are by no means comprehensive, but I hope they've convinced you that there are many reasons why AI systems behave unfairly.

Sometimes AI systems behave unfairly because of societal biases that are reflected in the datasets used to train them or because of other dataset characteristics (such as too few data points about some group of people). And sometimes it's because of assumptions or decisions made by teams -- either explicitly or implicitly – throughout the AI development and deployment lifecycle, such as the choice of the label, choice of the objective, and in fact the very choice of the task.

Because there are so many reasons, and because these reasons are not mutually exclusive and tend to exacerbate one another, assessing and mitigating fairness-related harms are not things that can be treated as an afterthought – they need to be prioritized at every single stage of the AI lifecycle.

Summary so far

Three main types of fairness-related harms:

Allocation. The system allocates resources, opportunities, or information in a way that leads to disparities in outcomes.

Quality of service. The system does not work similarly well for all groups.

Representation. The system reinforces negative stereotypes, denigrates, over- or underrepresents certain groups.

Groups *at risk of experiencing harms are* **context-dependent**.

Fairness-related harms can be introduced or propagated at **every stage of AI lifecycle:**

- task definition
- data collection & preprocessing
- model training
- model evaluation
- model deployment & monitoring

Therefore, the assessment and mitigation should also take place at every stage.

Fairlearn