

Fairlearn:

Putting AI Fairness Research into Practice

Miroslav Dudík

Microsoft Research

Who am I?



Microsoft Research

FATE group.

Microsoft's Aether Committee

Fairness and Inclusiveness Working Group.

Fairlearn

An open-source, community-driven project
to help data scientist improve fairness of AI systems.

<https://www.microsoft.com/en-us/research/theme/fate/>

<https://www.microsoft.com/en-us/ai/our-approach>

<https://fairlearn.org/>

I'll talk about

How Fairlearn started as a technical project focused on algorithmic mitigation of fairness-related harms.

How it evolved into a community-driven project with socio-technical focus.

What we have learned along the way.

I'll talk about

How Fairlearn started as a technical project focused on algorithmic mitigation of fairness-related harms.

How it evolved into a community-driven project with socio-technical focus.

What we have learned along the way.

Note: There are many other open-source projects for AI fairness:
Aequitas, AIF360, FairML, FairTest, Themis ML, What-If Tool, ...

Lee and Singh (2021): The Landscape and Gaps in Open Source Fairness Toolkits. In CHI 2021.

Deng et al. (2022): Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In FAccT 2022.

Fairness-related harms

Negative impacts for groups of people, such as those defined in terms of race, gender, age or disability status.

See NeurIPS 2017 Keynote by K. Crawford: The Trouble with Bias, https://www.youtube.com/watch?v=fMym_BKWQzk

Allocation harms

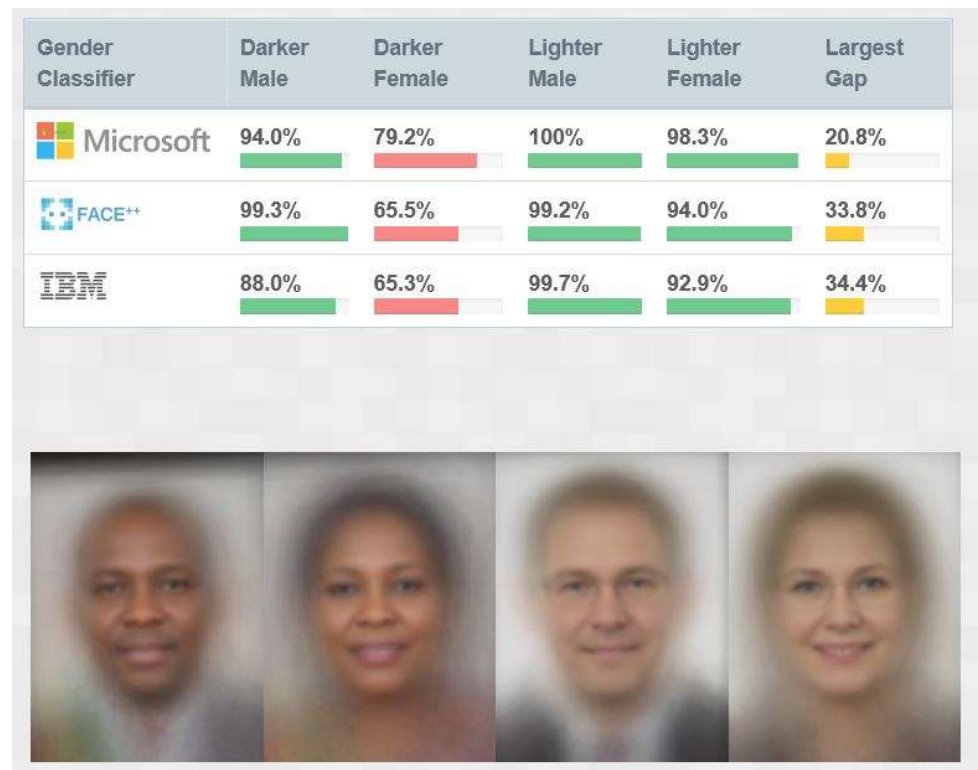
A system for **recommending patients** into high-risk care management programs is less likely to select Black patients than white patients of similar health.



Obermeyer et al. (2019): *Dissecting racial bias in an algorithm used to manage the health of populations*. Science 366 (6464).

Quality-of-service harms

Face recognition system works worse for women with darker skin.



Buolamwini and Gebru (2018): *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. In FAT* 2018.

Representation harms

Online ads more likely to suggest arrest records for names given primarily to Black babies.

Ads by Google

Latanya Sweeney, Arrested?

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

Latanya Sweeney

Public Records Found For: **Latanya Sweeney**. View Now.

www.publicrecords.com/

La Tanya

Search for La Tanya Look Up Fast Results now!

www.ask.com/La+Tanya

Sweeney (2013): *Discrimination in Online Ad Delivery*. CACM 56(5).

*Fairlearn started
with a research paper:*

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H.
A Reductions Approach to Fair Classification.
In ICML 2018.

But when we tried to
put this in practice...

But when we tried to
put this in practice...

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudík, M., and Wallach, H. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *CHI 2019*.

We found that:

Practitioners have more fundamental needs
before they can apply specific algorithms, such as needs for:

- fairness assessment
- access to domain-specific guides to fairness metrics and unfairness mitigation algorithms

Mitigating unfairness in binary classification

is a relatively rare use case, practitioners more often seek to solve:

- regression, ranking, complex ML domains (vision, speech, text, ...)

Holstein et al., Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *CHI 2019*.

We pivoted...

2018: Fairlearn starts as a Microsoft project to house Python code for the “reductions approach” paper.

2019: Focus shifts to assessment and visualization.

2020: Fairlearn whitepaper is published:

“We emphasize that prioritizing fairness in AI systems is a sociotechnical challenge. Because there are many complex sources of unfairness—some societal and some technical—it is not possible to fully “debias” a system or to guarantee fairness; the goal is to mitigate fairness-related harms as much as possible.”

Periodic Fairlearn community calls start.

Our community
held us accountable...

Our community held us accountable...

"I think that **de-abstracting would really help** make productive progress on what excellence in a **sociotechnical approach to ML** would be" - Community Member

<https://github.com/fairlearn/fairlearn/issues/413>

Our community held us accountable...

"I think that **de-abstracting would really help** make productive progress on what excellence in a **sociotechnical approach to ML** would be" - Community Member

"my impression at this point is that while folks on the team may think that **"fairness as a sociotechnical challenge"** is sort of an interesting aspiration, **right now there's not alignment for collaborating on more substantive work in that direction.** I'll keep following along, and feel free to @ me if other ways to collaborate come up down the line." - Community Member

Our community held us accountable...

"I think that **de-abstracting would really help** make productive progress on what excellence in a **sociotechnical approach to ML** would be" - Community Member

"my impression at this point is that while folks on the team may think that **"fairness as a sociotechnical challenge"** is sort of an interesting aspiration, **right now there's not alignment for collaborating on more substantive work in that direction.** I'll keep following along, and feel free to @ me if other ways to collaborate come up down the line." - Community Member

"what if Fairlearn was a place developers new to fairness came to get **help in learning and doing the work of approaching fairness as a sociotechnical challenge?**" - Community Member

Current vision for Fairlearn

Open-source toolkit with **metrics, algorithms, visualizations.**

Tasks **beyond binary classification.**

Importance of **education materials** (manuals, case studies, white papers).

Fairness as a **socio-technical** challenge.

Formulated requirements for our example notebooks

Excerpt from the contributor guide:

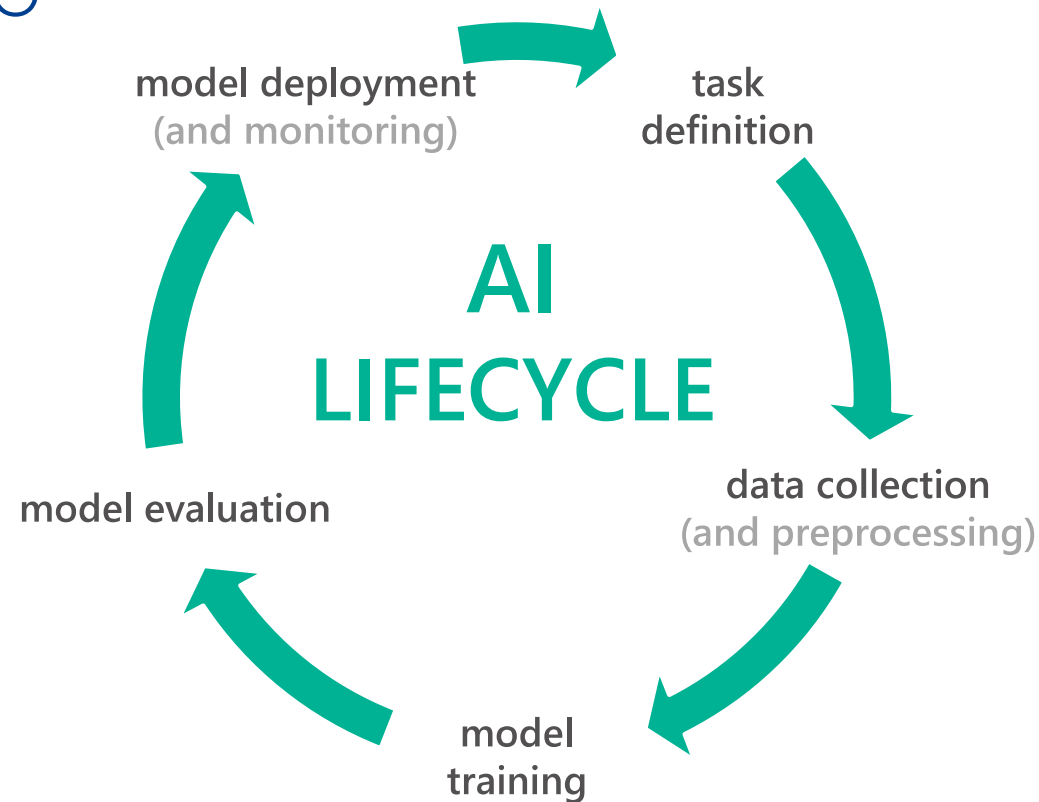
A good example notebook exhibits the following attributes:

1. **Deployment context:** Describes a real deployment context, not just a dataset.
2. **Real harms:** Focuses on real harms to real people. See [Blodget et al. \(2020\)](#).
3. **Sociotechnical:** Models the Fairlearn team's value that fairness is a sociotechnical challenge. Avoids abstraction traps. See [Selbst et al. \(2019\)](#).
4. **Substantiated:** Discusses trade-offs and compares alternatives. Describes why using particular Fairlearn functionalities makes sense.
5. **For developers:** Speaks the language of developers and data scientists. Considers real practitioner needs. Fits within the lifecycle of real practitioner work. See [Holstein et al \(2019\)](#), [Madaio et al. \(2020\)](#).

Please keep these in mind when creating, discussing, and critiquing examples.

Tutorial based on health-care recommendation scenario

Demonstrates how fairness harms can arise (and be mitigated) at any stage of the AI lifecycle.

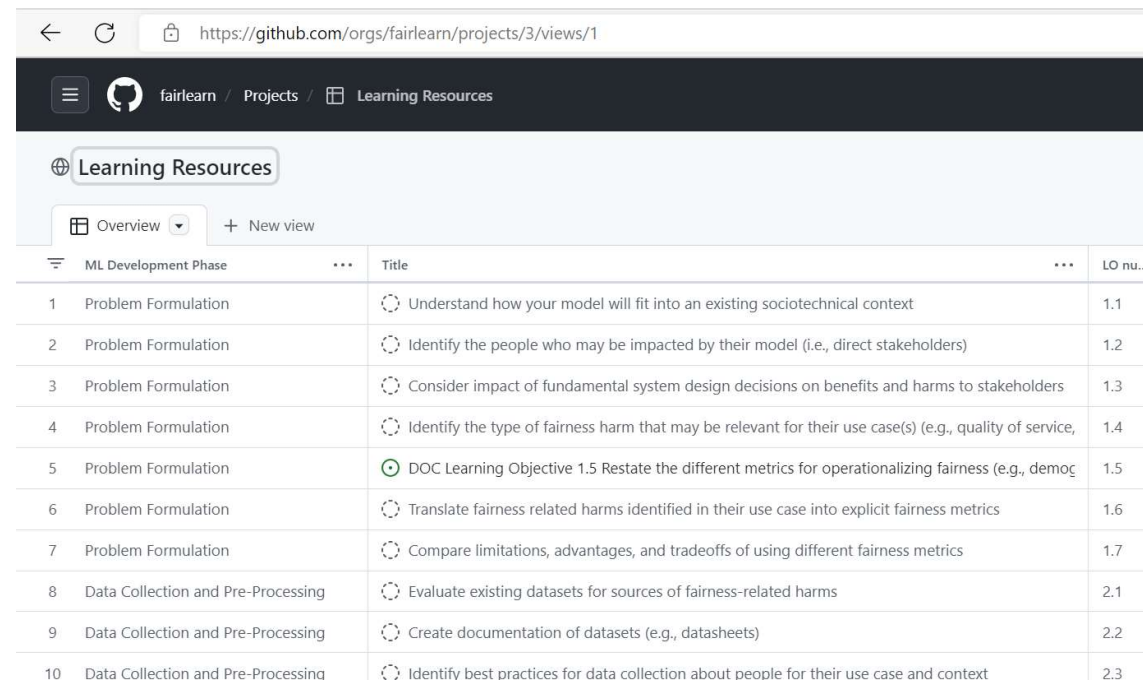


https://github.com/fairlearn/talks/tree/main/2021_scipy_tutorial

The tutorial draws motivation from [Obermeyer et al. \(2019\)](#), using the dataset developed by [Strack et al. \(2014\)](#).

List of concrete learning-resource needs

A working group identified and prioritized tasks to address gaps in educational materials.



The screenshot shows a GitHub repository page for 'fairlearn / Projects / Learning Resources'. The page displays a table with 10 rows of learning resources, organized by ML Development Phase. The table has columns for ML Development Phase, Title, and LO nu... (Learning Objective number). The resources are numbered 1 through 10, with phases ranging from Problem Formulation to Data Collection and Pre-Processing.

ML Development Phase	Title	LO nu...
1 Problem Formulation	Understand how your model will fit into an existing sociotechnical context	1.1
2 Problem Formulation	Identify the people who may be impacted by their model (i.e., direct stakeholders)	1.2
3 Problem Formulation	Consider impact of fundamental system design decisions on benefits and harms to stakeholders	1.3
4 Problem Formulation	Identify the type of fairness harm that may be relevant for their use case(s) (e.g., quality of service,	1.4
5 Problem Formulation	DOC Learning Objective 1.5 Restate the different metrics for operationalizing fairness (e.g., democ	1.5
6 Problem Formulation	Translate fairness related harms identified in their use case into explicit fairness metrics	1.6
7 Problem Formulation	Compare limitations, advantages, and tradeoffs of using different fairness metrics	1.7
8 Data Collection and Pre-Processing	Evaluate existing datasets for sources of fairness-related harms	2.1
9 Data Collection and Pre-Processing	Create documentation of datasets (e.g., datasheets)	2.2
10 Data Collection and Pre-Processing	Identify best practices for data collection about people for their use case and context	2.3

<https://github.com/orgs/fairlearn/projects/3/>

Lee and Singh (2021): The Landscape and Gaps in Open Source Fairness Toolkits. In CHI 2021.

Deng et al. (2022): Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In FAccT 2022.

What have we learned?

Running an open-source project is **a community building effort.**

To treat fairness as a socio-technical challenge requires a broad range of contributors with **diverse disciplinary backgrounds.**

Fairness issues often arise in early stages of AI lifecycle (e.g., task definition and data collection) and **non-algorithmic mitigation** strategies might be most effective.