Keeping Human Concerns in the Loop:
Human-Centered Approaches to Responsible AI

Michael Madaio
10/05/20

Machine Learning Optimized Systems

Microsoft Research

Online Ads for High-Paying Jobs Are Targeting Men More Than Women New study uncovers gender bias

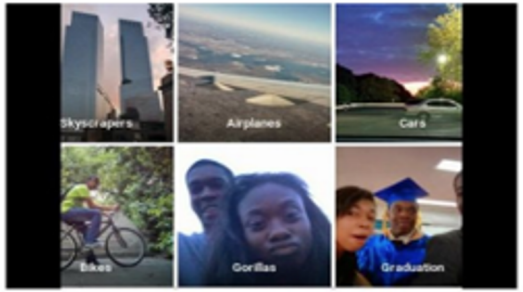
When Algorithms Discriminate

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology



Do Google's 'unprofessional hair' results show it is racist?
Leigh Alexander

Search term brings back mainly results of black women, which some say is evidence of bias. But algorithms may just be reflecting the wider social landscape

There results of image searches for 'unprofessional hair for work' (left) and 'professional hair for work' (right) on Google. Photograph: Google

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Amazon just showed us that 'unbiased' algorithms can be inadvertently racist

Amazon Prime and the racist algorithms

AI systems can behave unfairly, causing harm to people in a variety of ways

AI systems can unfairly **allocate** opportunities or resources

Millions of black people affected by racial bias in health-care algorithms

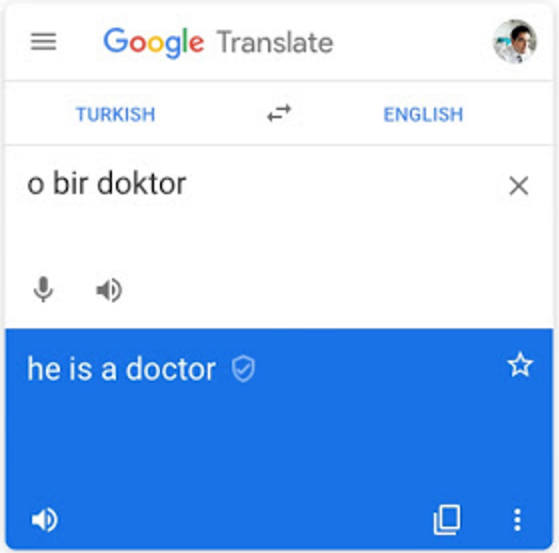
Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Dissecting racial bias in an algorithm used to manage the health of populations

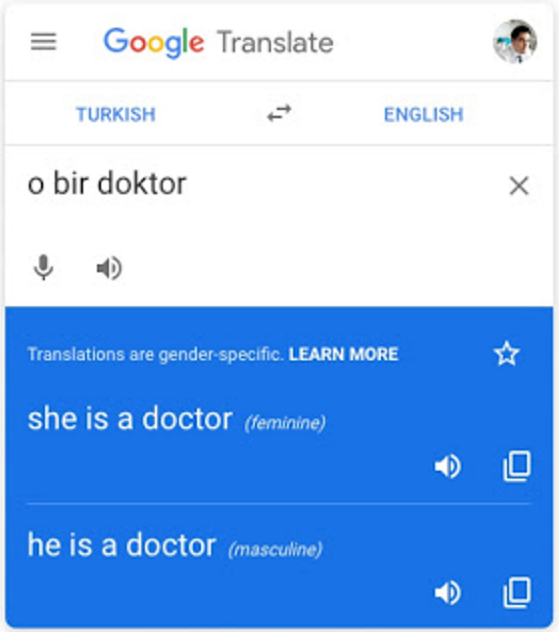
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

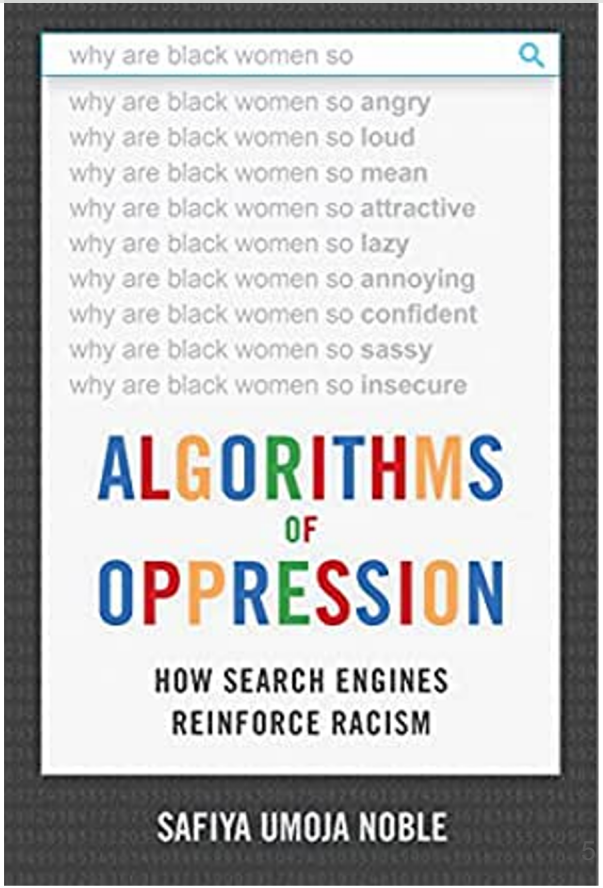
AI systems can reinforce societal **stereotypes**, or **denigrate** or demean people

Before



After





AI systems can **under- or over-represent** groups of people, or treat them as if they don't exist

Percentage of women in top 100 Google image search results for CEO: 11%
Percentage of U.S. CEOs who are women: 27%

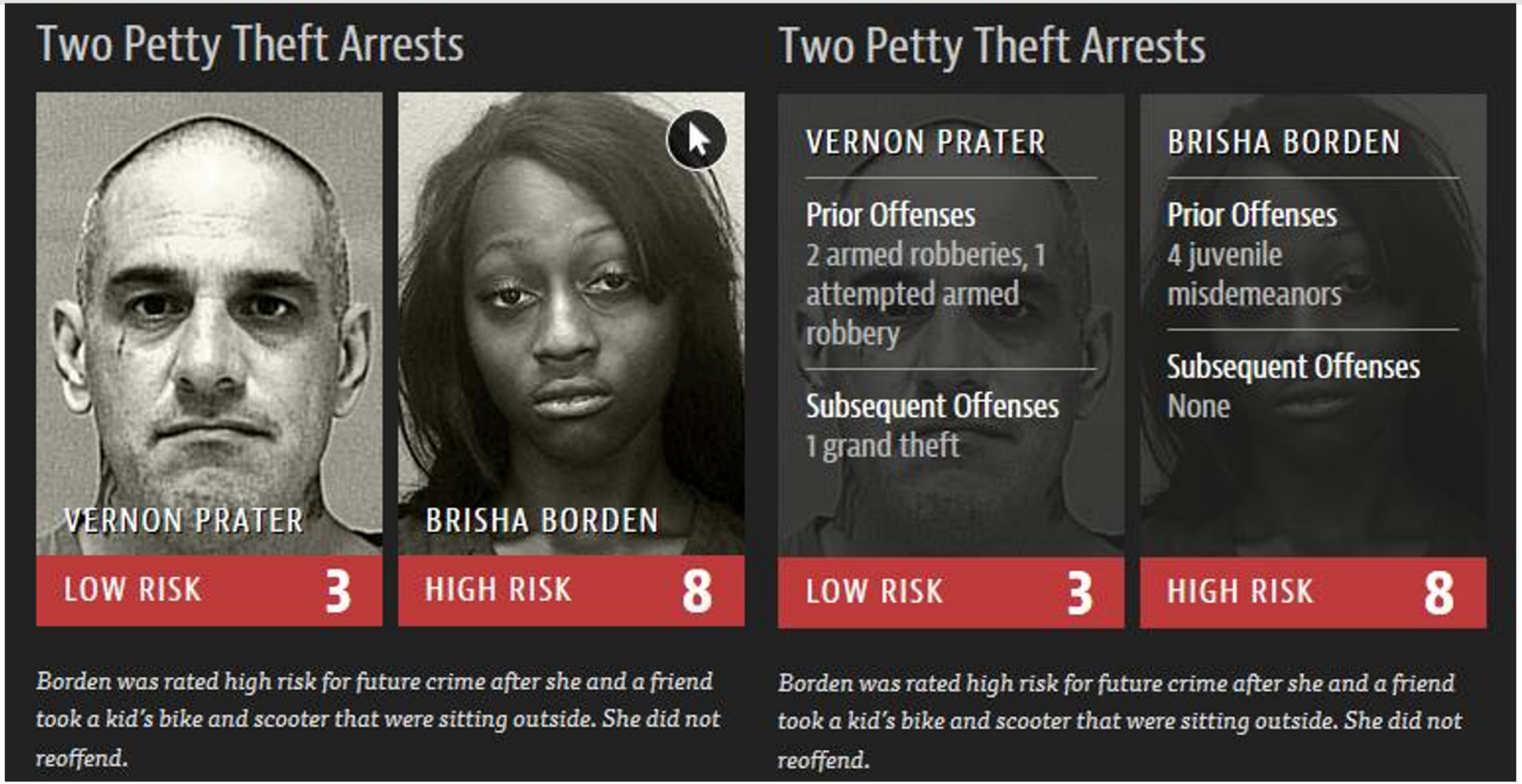
Kay et al., 2015

AI systems can cause fairness-related harms to people in a variety of ways:

- Allocation
- Performance disparities, or quality of service
- Stereotyping
- Denigration
- Under- or over-representation or erasure

Crawford et al., 2017

AI systems are increasingly used in high-stakes contexts



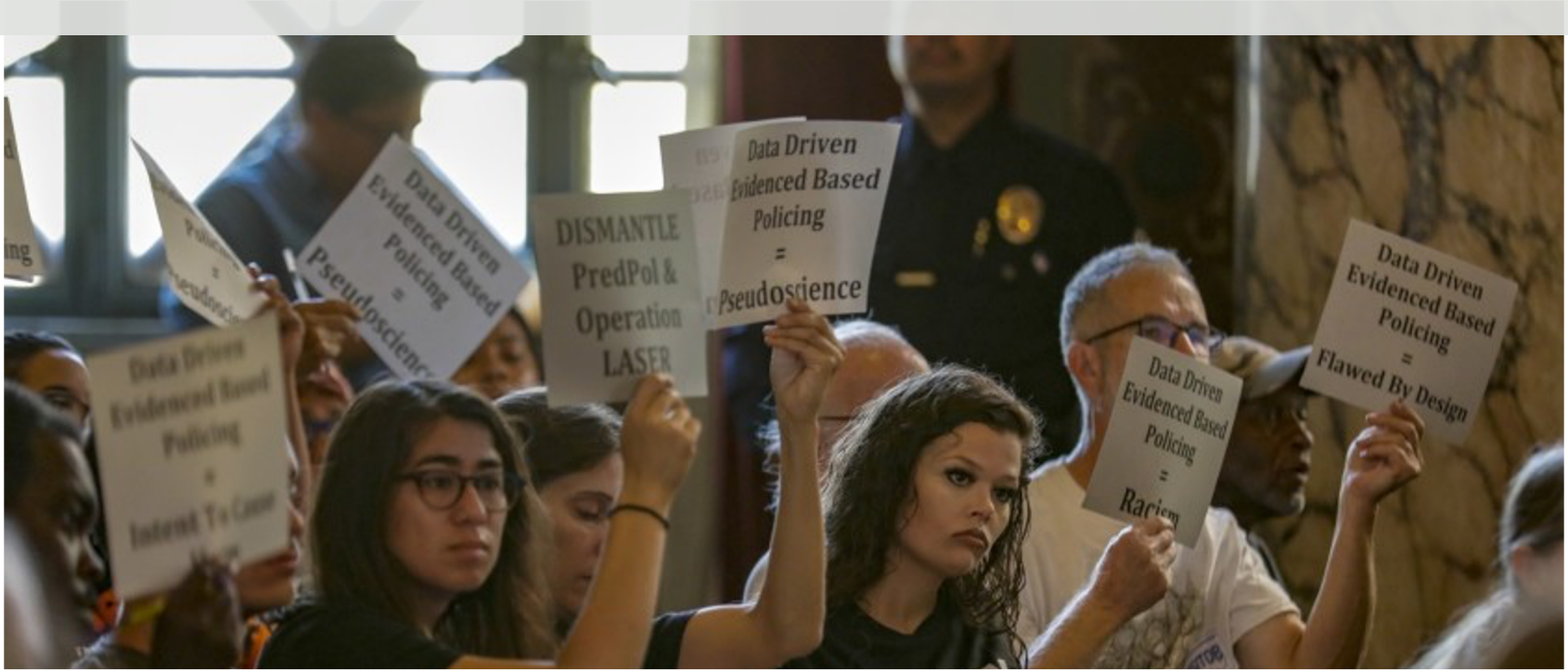
ProPublica, 2016

AI systems are increasingly used in high-stakes contexts

Can an algorithm help keep kids safe? So far, Allegheny County's screening tool is improving accuracy

Vaithianathan et al., 2019;
Chouldechova et al., 2018

Communities have protested algorithmic decision-making

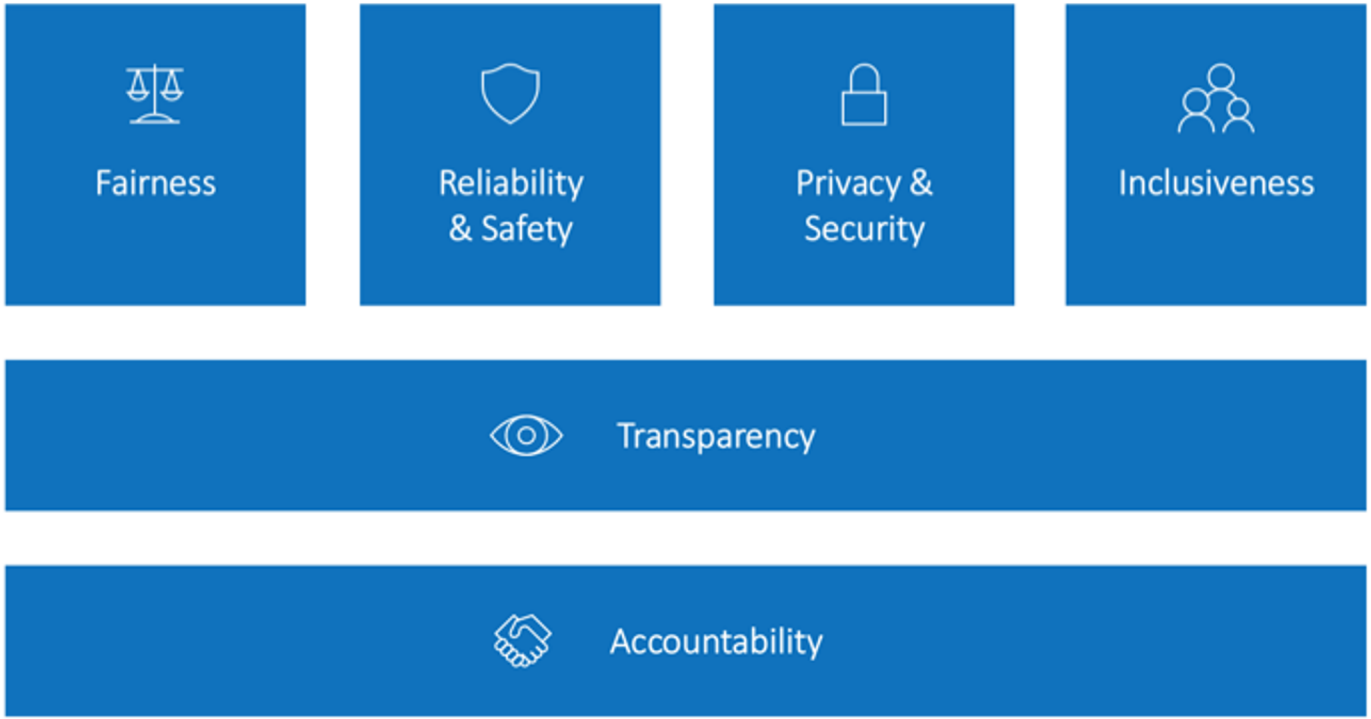


LAPD officials defend predictive policing as activists call for its end

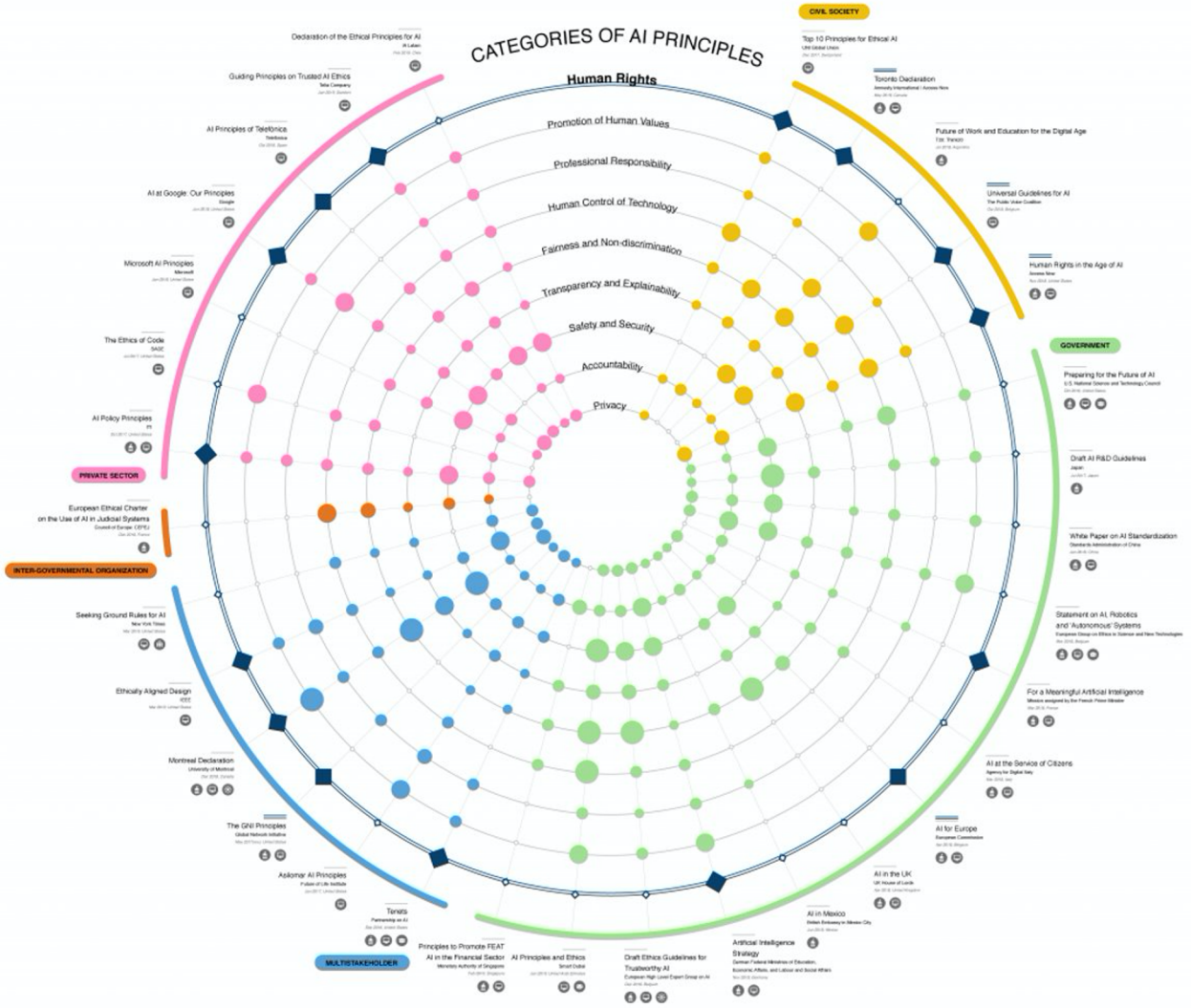
Los Angeles Times

Communities have protested algorithmic decision-making

Gilman, 2019; 2020


Fairness
Reliability & Safety
Privacy & Security
Inclusiveness
Transparency
Accountability

Microsoft



BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

“You just have to put your model out there, and then you know if there’s fairness issues if someone raises hell.”

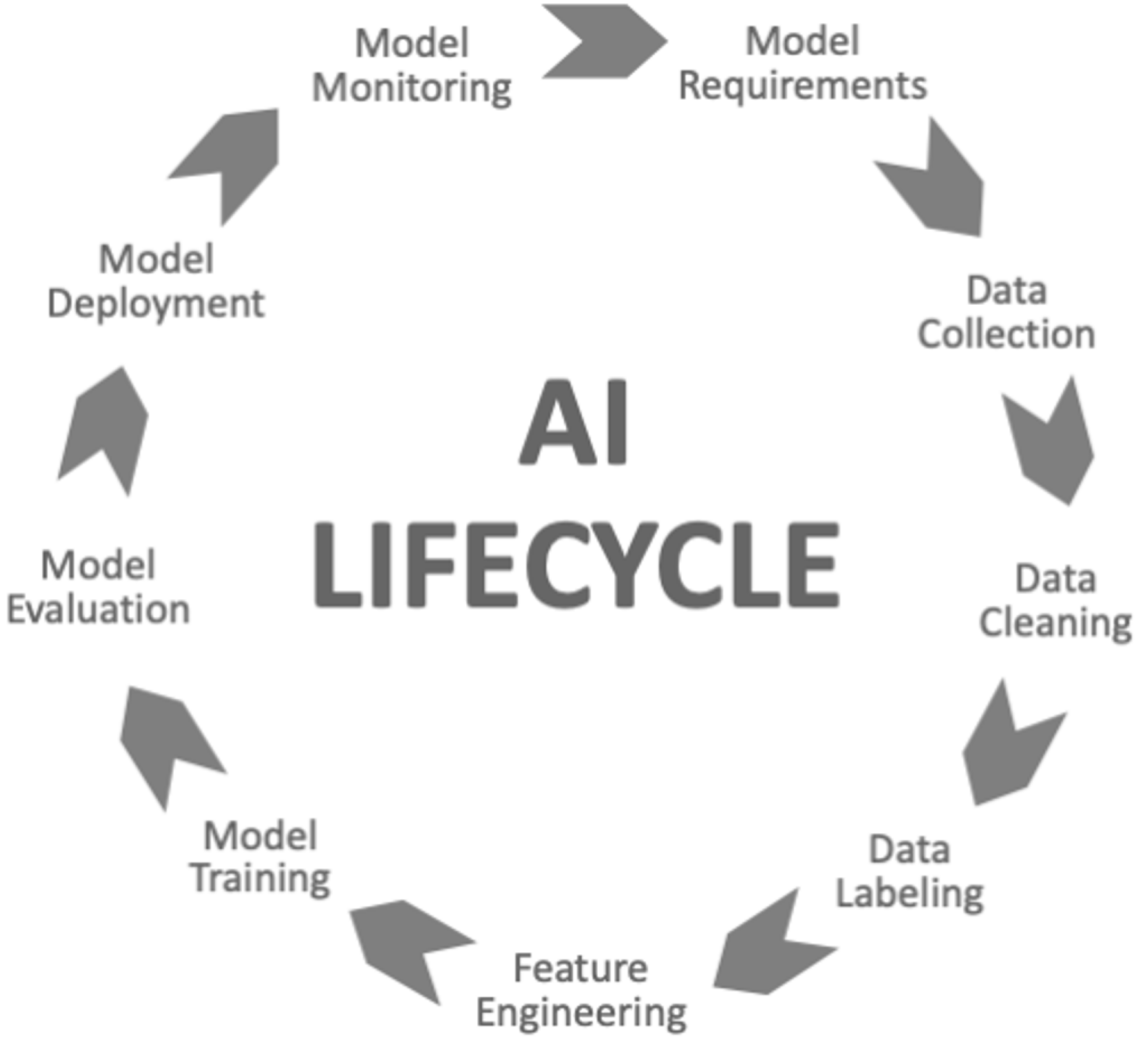
Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Found fairness issues in their products

Of those, 99% found issues after deployment

Suspected there were unidentified fairness issues

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).


AI
LIFECYCLE
Model Requirements
Data Collection
Data Cleaning
Data Labeling
Feature Engineering
Model Training
Model Evaluation
Model Deployment
Model Monitoring


THE NEW YORK TIMES BESTSELLER
THE CHECKLIST MANIFESTO
HOW TO GET THINGS RIGHT
PICADOR
ATUL GAWANDE
BESTSELLING AUTHOR OF BETTER AND COMPLICATIONS

Here's a checklist for people who are working on data projects:

❏ Have we listed how this technology can be attacked or abused?

❏ Have we tested our training data to ensure it is fair and representative?

❏ Have we studied and understood possible sources of bias in our data?

❏ Does our team reflect diversity of opinions, backgrounds, and kinds of thought?

❏ What kind of user consent do we need to collect to use the data?

❏ Do we have a mechanism for gathering consent from users?

❏ Have we explained clearly what users are consenting to?

❏ Do we have a mechanism for redress if people are harmed by the results?

❏ Can we shut down this software in production if it is behaving badly?

❏ Have we tested for fairness with respect to different user groups?

❏ Have we tested for disparate error rates among different user groups?

Patil et al., 2018;
Cramer et al., 2019;
Center for Democracy and Technology, 2019;
DrivenData, 2019;
Johns Hopkins Center for Government Excellence, 2019;
European Union High-level Expert Group, 2019;
Machine Intelligence Garage, 2019;
UK Department of Digital, Culture, Media and Sport, 2019;
Vallor, 2019;

Here's a checklist for people who are working on data projects:

❑ Have we listed how this technology can be attacked or abused?

❑ Have we tested our training data to ensure it is fair and representative?

❑ Have we studied and understood possible sources of bias in our data?

❑ Does our team reflect diversity of opinions, backgrounds, and kinds of thought?

❑ What kind of user consent do we need to collect to use the data?

❑ Do we have a mechanism for gathering consent from users?

❑ Have we explained clearly what users are consenting to?

❑ Do we have a mechanism for redress if people are harmed by the results?

❑ Can we shut down this software in production if it is behaving badly?

❑ Have we tested for fairness with respect to different user groups?

❑ Have we tested for disparate error rates among different user groups?

Patil et al., 2018

1. Be clear about the benefits of your product or service

2. Know and manage your risks

3. Use data responsibly

4. Be worthy of trust

5. Promote diversity, equality and inclusion

6. Be open and understandable in communications

7. Consider your business model

Machine Intelligence Garage, 2019

Data Science Ethics Checklist

ethics checklist deon

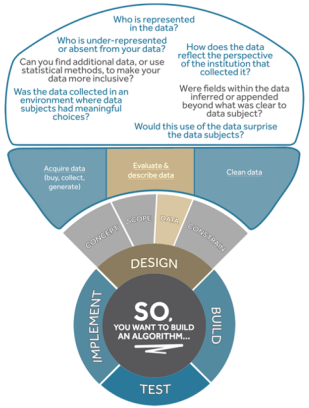
A. Data Collection

- **A.1 Informed consent:** If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
- **A.2 Collection bias:** Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
- **A.3 Limit PII exposure:** Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?

B. Data Storage

- **B.1 Data security:** Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
- **B.2 Right to be forgotten:** Do we have a mechanism through which an individual can request their personal information be removed?
- **B.3 Data retention plan:** Is there a schedule or plan to delete the data after it is no longer needed?

DrivenData, 2019



Center for Democracy and Technology, 2019

Step 1.1.2 Identify the types of impact

Your algorithm will have at least one or more areas of impact. The table below describes the different types of impact. One type of impact may implicate another. For example, restaurant reviews impact reputation which in turn impact financial health. The goal of this step is to ensure sure we understand the nature of the impacts - not the degree or direction.

You'll want to identify the type of impact for each group you identified in Step 1.1.1.

Type	Description
Access to goods, benefits or services	These types of algorithms inform who, what or where does or does not receive access to goods, benefits or services. This can include access to insurance, government benefits, housing opportunities, education, maintenance or prevention services, recreation etc.
Financial	These types of algorithms impact the financial health of individuals, groups, entities or areas.
Property or equipment	These types of algorithms impact the quality or value of property or equipment.
Reputation	These types of algorithms impact the reputation of an individual, group, entity, or location.
Emotional	These types of algorithms impact the emotional health and well-being of an individual or group of individuals.
Life / safety	These types of algorithms impact the life or safety of an individual, group, entity, or location.
Privacy	These types of algorithms impact the privacy of an individual or group.
Liberty / freedom	These types of algorithms impact the liberty / freedom of an individual, group, or entity.
Rights / intellectual Property	These types of algorithms impact the rights / intellectual property of an individual, group or entity.

Johns Hopkins Center for Government Excellence, 2019

Guidance

Data Ethics Workbook

Published 13 June 2018

Contents

Questions for principle 1 - Start with clear user need and public benefit

Describe the user need.

- Does everyone in the team understand the user need?
- How does this benefit the public?
- What would be the harm in not using data science - what needs might not be met?
- Do you have supporting evidence for the approach being likely to meet a user need or provide public benefit?

Questions for principle 2 - Be aware of relevant legislation and codes of practice

List the pieces of legislation, codes of practice and guidance that apply to your project.

- Do all team members understand how relevant laws apply to the project?
- If necessary, have you consulted with relevant experts?
- Have you spoken to your information assurance team?
- If using personal data, do you understand your obligations under data protection

United Kingdom, Department of Digital, Culture, Media and Sport, 2019

Surgical Safety Checklist

World Health Organization | Patient Safety
A World Alliance for Safer Health Care

Before induction of anaesthesia

(with at least nurse and anaesthetist)

Has the patient confirmed his/her identity, site, procedure, and consent?
- ☐ Yes

Is the site marked?
- ☐ Yes
- ☐ Not applicable

Is the anaesthesia machine and medication check complete?
- ☐ Yes

Is the pulse oximeter on the patient and functioning?
- ☐ Yes

Does the patient have a:

Known allergy?
- ☐ No
- ☐ Yes

Difficult airway or aspiration risk?
- ☐ No
- ☐ Yes, and equipment/assistance available

Risk of >500ml blood loss (7ml/kg in children)?
- ☐ No
- ☐ Yes, and two IVs/central access and fluids planned

Before skin incision

(with nurse, anaesthetist and surgeon)

- ☐ **Confirm all team members have introduced themselves by name and role.**
- ☐ **Confirm the patient's name, procedure, and where the incision will be made.**

Has antibiotic prophylaxis been given within the last 60 minutes?
- ☐ Yes
- ☐ Not applicable

Anticipated Critical Events

To Surgeon:
- ☐ What are the critical or non-routine steps?
- ☐ How long will the case take?
- ☐ What is the anticipated blood loss?

To Anaesthetist:
- ☐ Are there any patient-specific concerns?

To Nursing Team:
- ☐ Has sterility (including indicator results) been confirmed?
- ☐ Are there equipment issues or any concerns?

Is essential imaging displayed?
- ☐ Yes
- ☐ Not applicable

Before patient leaves operating room

(with nurse, anaesthetist and surgeon)

Nurse Verbally Confirms:
- ☐ The name of the procedure
- ☐ Completion of instrument, sponge and needle counts
- ☐ Specimen labelling (read specimen labels aloud, including patient name)
- ☐ Whether there are any equipment problems to be addressed

To Surgeon, Anaesthetist and Nurse:
- ☐ What are the key concerns for recovery and management of this patient?

This checklist is not intended to be comprehensive. Additions and modifications to fit local practice are encouraged.

Revised 1 / 2009

How might we support AI practitioners in **proactively anticipating** fairness harms?

Research Questions

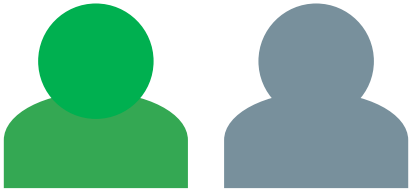
RQ1: What are AI/ML practitioners' **current processes and workflows** for identifying and mitigating issues of fairness in AI?

RQ2: What are AI/ML practitioners' **needs, desires, and concerns** regarding AI fairness checklists?

RQ3: How do practitioners **envision** AI fairness checklists might be **implemented within their organizations?**

Methods

Co-designed an **AI fairness checklist** with **48 industry practitioners** from 12 technology companies



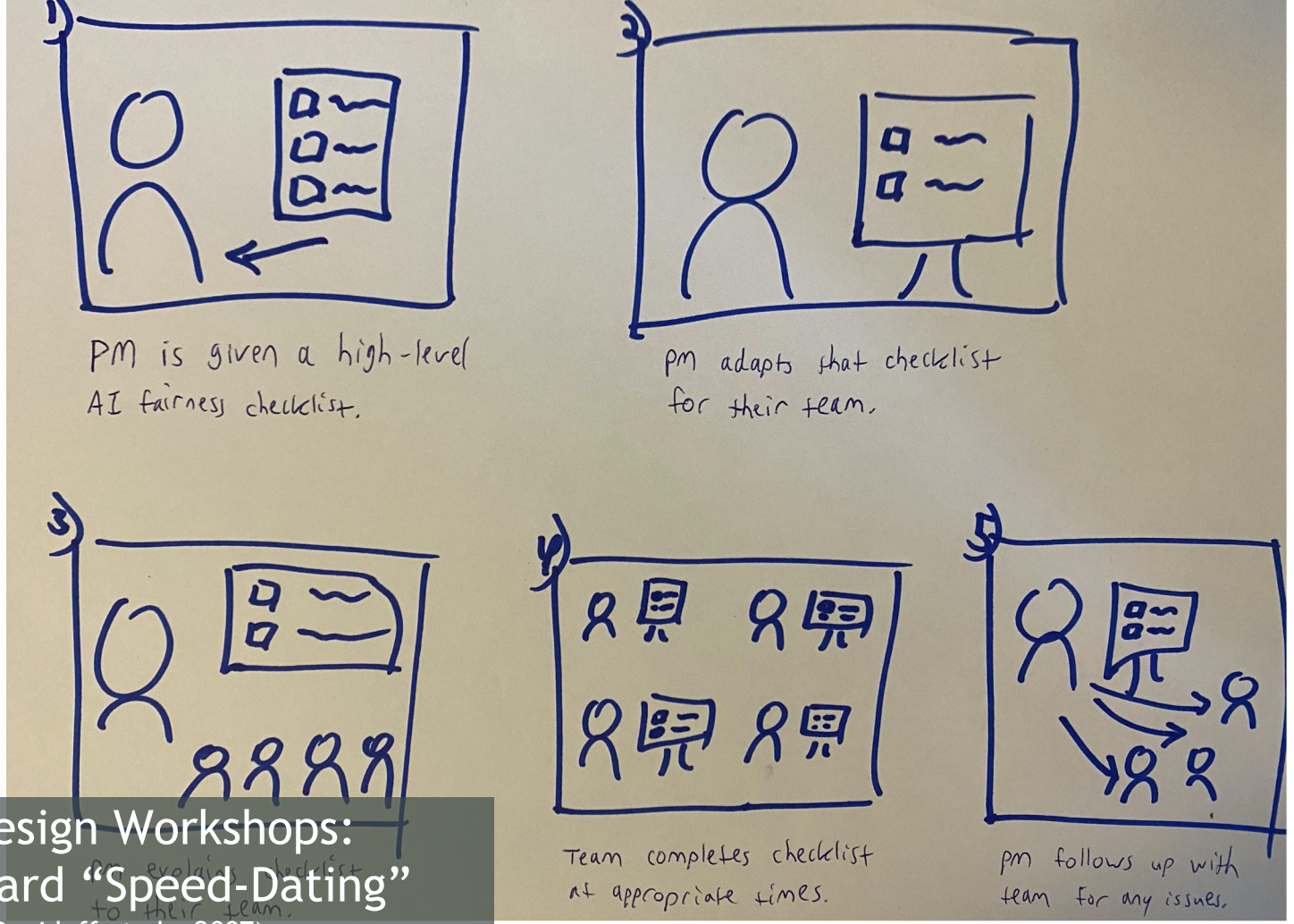
Interviews

- 14 participants
- 60-90 minute sessions
- Shown a medical checklist, but not an AI checklist

Co-Design workshops

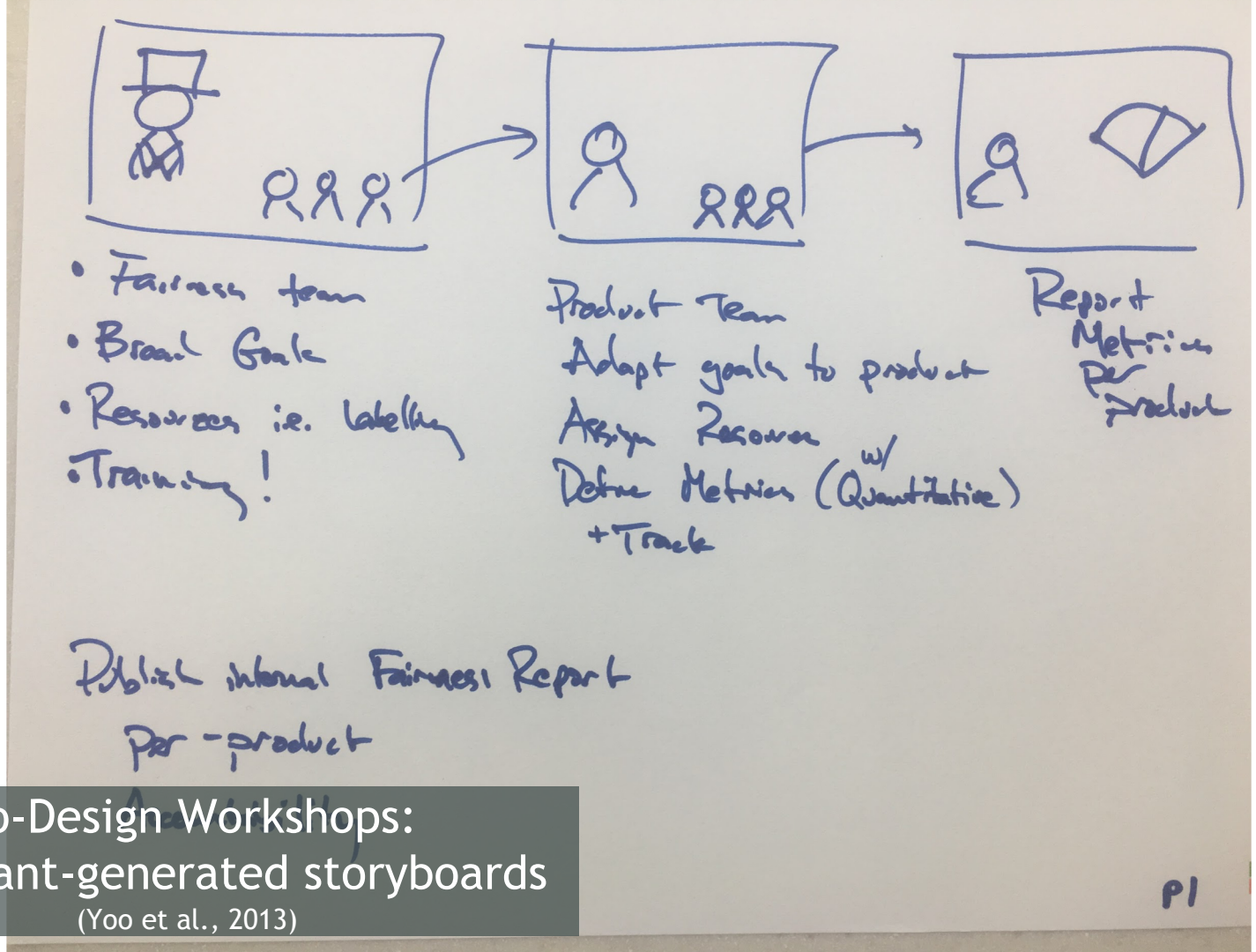
- 27 90-minute sessions, 40 participants
- Elicited ideas for checklist items, feedback on items, and created implementation scenarios

Co-Design Workshops: Participant input



Co-Design Workshops: Storyboard “Speed-Dating”

(Davidoff et al., 2007)



Co-Design Workshops: Participant-generated storyboards

(Yoo et al., 2013)

- Transcribed 60+ hours of audio data
- Coded transcripts using inductive coding
- Clustered into ~40 code groups
- Synthesized into 9 high-level thematic findings

Inductive Thematic Coding
(Strauss and Corbin, 1995; Braun and Clarke, 2006)

Findings

RQ1: What are AI/ML practitioners' **current processes and workflows** for identifying and mitigating issues of fairness in AI?

RQ1: Participants are **aware that fairness and ethics are important**, but mostly have **ad-hoc approaches** to it

"One of the **biggest consequences [of not addressing fairness issues] is that we're not helping our customers**. I think it's our responsibility to help our customers build trust with *their* customers. If we don't have tools and platforms and systems that allow them to do that, we're not setting them up for success."
(P7, Product Manager)

"At this moment **it is more kind of ad hoc.** If something happens, the team fixes the problem; **maybe we'll fix it proactively, maybe reactively**."
(P31, Program Manager)

RQ1: Individual advocacy for fairness was at odds with **organizational incentives** for a fast-paced development lifecycle

"To be honest, a lot of it felt-- so on the level I was at, which is {PM}, it felt like **it was kind of up to the individuals** involved with the project **to raise that awareness** when we were having a design discussion."
(P17, Program Manager)

"**I get paid to go fast**. And I go as fast as I can without doing harm."
(P24, Software Engineering Manager)

"There's a broader, company-wide push-pull of "**Do I do a good thing or do I do the thing that ships the product?**"
(P19, Data Scientist)

RQ1: Participants felt there were **social costs to being an individual advocate** for fairness and ethical issues during development

"There's no checkpoint where someone's supposed to say something and, so, then you can only do it by being this annoying squeaky wheel and, well, **I'm the annoying squeaky wheel about too many things**."
(P19, Data Scientist)

Every answer [they] gave for every question on the panel was basically, like, "Do the ethical thing and don't worry about the impact on your career." But that's an easy thing to say for a senior level person. **It's a lot harder for the people in the trenches**, especially when this room was full of junior designers.
(P20, Design Researcher)

RQ2: What are AI/ML practitioners' **needs, desires, and concerns** regarding AI fairness checklists?

RQ2: Having a checklist could **provide organizational infrastructure** to empower team members to catch issues that might not be caught

"How do we **enable people to do these things** without feeling like they will be **labeled a troublemaker**, or they will be the **stop-ship person**. How do we **give everybody the red button** without making it a problem?"
(P20, Design Researcher)

RQ2: Having a checklist could **provide organizational infrastructure** to empower team members to catch issues that might not be caught

"Even in a room of people who all really care, the fact that it's not part of the process is not good, because we're always under so much crunch time. By design, we have so much on our plates, and **the first things to go are the ones that aren't processes**. You know? So it doesn't matter what good intentions people have, **if it's not part of the process, it's not going to get done to the level of quality as things that do have a process**."

(P17, Program Manager)

RQ2: Having a formal process for fairness would help teams negotiate priorities, but only if it **fits into teams' workflows**

- We elicited **"pause points" in teams' processes** where checklist items could be completed (e.g., spec reviews, code reviews)

- **Additional resources and templates** are needed for teams to adapt general checklist to their team's process

Surgical Safety Checklist

World Health Organization | Patient Safety
A World Alliance for Safer Health Care

Before induction of anaesthesia	Before skin incision	Before patient leaves operating room
(with at least nurse and anaesthetist)	(with nurse, anaesthetist and surgeon)	(with nurse, anaesthetist and surgeon)
Has the patient confirmed his/her identity, site, procedure, and consent? ☐ Yes	☐ **Confirm all team members have introduced themselves by name and role.**	**Nurse Verbally Confirms:** ☐ The name of the procedure
Is the site marked? ☐ Yes ☐ Not applicable	☐ **Confirm the patient's name, procedure, and where the incision will be made.**	☐ Completion of instrument, sponge and needle counts
Is the anaesthesia machine and medication check complete? ☐ Yes	**Has antibiotic prophylaxis been given within the last 60 minutes?** ☐ Yes ☐ Not applicable	☐ Specimen labelling (read specimen labels aloud, including patient name)
Is the pulse oximeter on the patient and functioning? ☐ Yes	**Anticipated Critical Events** **To Surgeon:** ☐ What are the critical or non-routine steps? ☐ How long will the case take? ☐ What is the anticipated blood loss?	☐ Whether there are any equipment problems to be addressed
Does the patient have a: **Known allergy?** ☐ No ☐ Yes	**To Anaesthetist:** ☐ Are there any patient-specific concerns?	**To Surgeon, Anaesthetist and Nurse:** ☐ What are the key concerns for recovery and management of this patient?
Difficult airway or aspiration risk? ☐ No ☐ Yes, and equipment/assistance available	**To Nursing Team:** ☐ Has sterility (including indicator results) been confirmed? ☐ Are there equipment issues or any concerns?	
Risk of >500ml blood loss (7ml/kg in children)? ☐ No ☐ Yes, and two IVs/central access and fluids planned	**Is essential imaging displayed?** ☐ Yes ☐ Not applicable	

This checklist is not intended to be comprehensive. Additions and modifications to fit local practice are encouraged.

Revised 1 / 2009

Implementation manual for WHO surgical safety checklist

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

The dataset was created to enable research on predicting sentiment polarity: given a piece of (English) text, predict whether it has a positive or negative affect or stance toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

At the time of publication, only the original paper http://xxx.lanl.gov/pdf/cs/0409058v1. Between then and 2012, a collection of papers that used this dataset was maintained at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The instances are movie reviews extracted from newsgroup postings, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The polarity rating is binary {positive,negative}. An example instance is shown in Figure 1.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

None explicitly, though the original newsgroup postings include poster name and email address, so some information could be extracted if needed.

How many instances of each type are there?

There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

[1]Information in this datasheet is taken from one of five sources; any errors that were introduced are our fault. http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Everything is included.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

Several experiments are reported in the README for baselines on this data, both on the original dataset (Table 1) and the cleaned version (Table 2). In these results, NB=Naive Bayes, ME=Maximum Entropy and SVM=Support Vector Machine. The feature sets include unigrams (with and without counts), bigrams, part of speech features, and adjectives-only.

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Datasheets for Datasets
Gebru et al., 2018

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

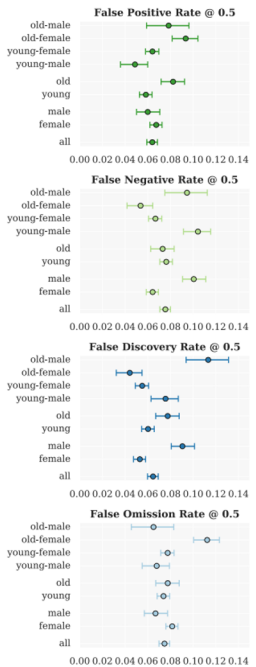


Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Model Cards for Model Reporting
Mitchell et al., 2019

RQ2: Concerns that a **checklist may lead to rote compliance**, when AI fairness is a **complex socio-technical phenomena**

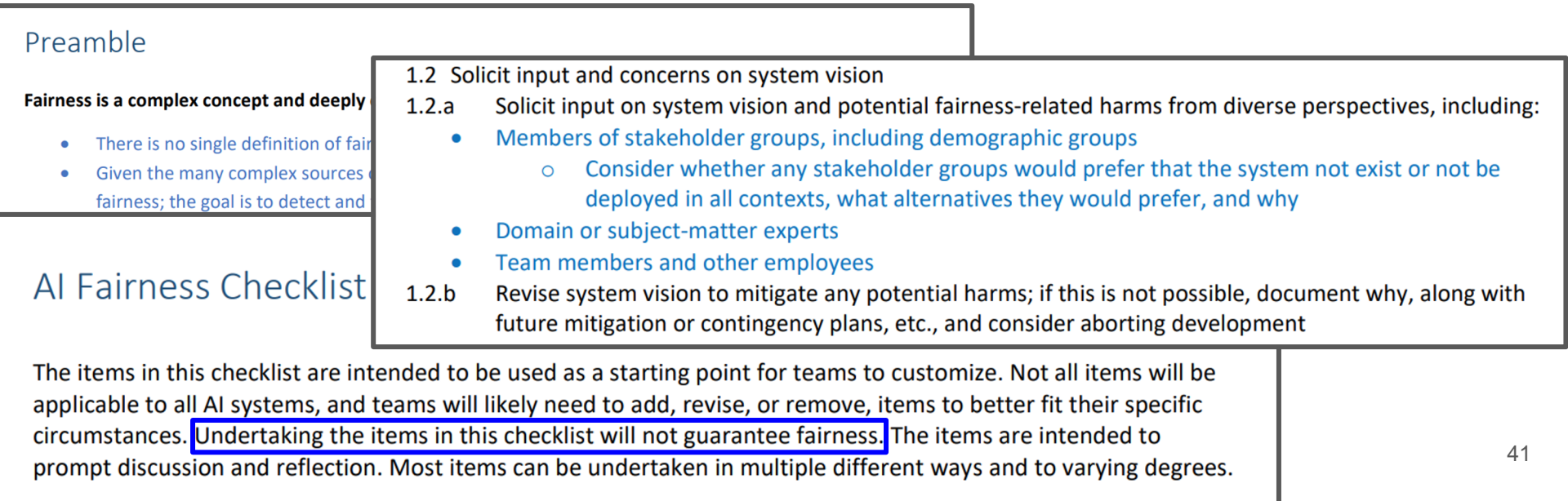
"**I'm a little bit suspicious of the checklist approach**... [Fairness is] a very non-engineering thing and the notion that engineering and technology cannot fix these problems is really upsetting to people who have spent their entire lives believing they can solve the world's problem with computing."
(P34, Data Scientist)

"People thought, you know, 'If I just use this security compliance checklist, **I could just check things off**, and then I'm good!' And **they were not good**."
(P11, ML Engineer)

"If any of the checklist items says, **'Have you met this number of things?' it becomes easy to game,** without making things more fair."
(P4, Program Manager)

RQ2: Concerns that a **checklist may lead to rote compliance**, when AI fairness is a **complex socio-technical phenomena**

Checklist items were designed to **prompt critical conversations** between stakeholders and designers, with documentation and action, if needed



Preamble

Fairness is a complex concept and deeply

- There is no single definition of fair
- Given the many complex sources
  fairness; the goal is to detect and

1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

AI Fairness Checklist

The items in this checklist are intended to be used as a starting point for teams to customize. Not all items will be applicable to all AI systems, and teams will likely need to add, revise, or remove, items to better fit their specific circumstances. Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection. Most items can be undertaken in multiple different ways and to varying degrees.

RQ3: Leadership needs to change culture to drive checklist adoption

“It’s a change management problem. So I think I can only lean back on a little bit **my experience with internationalization** because it was very similar. When I started at our company, **there was no such thing as an internationalization checklist.**

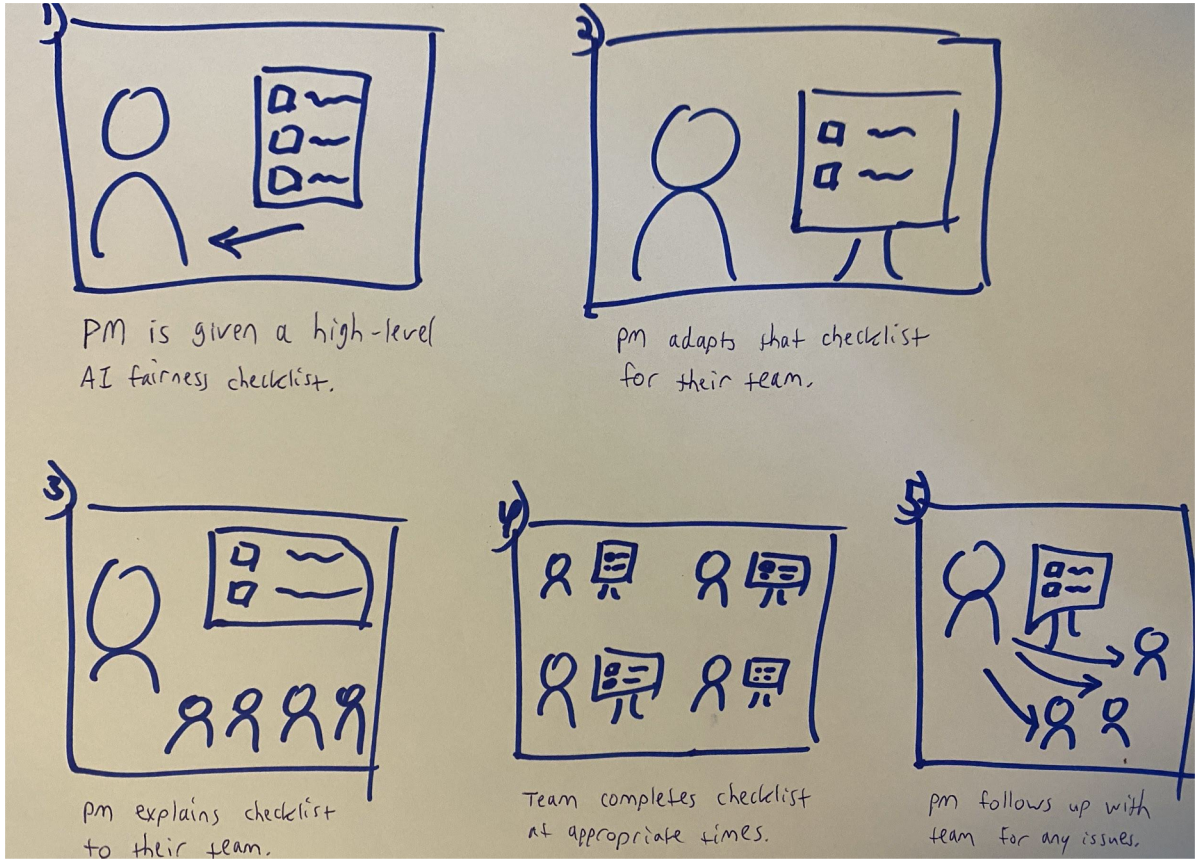
It was a total cowboy [situation]. People wrote code and you would try to translate the code and it would break left and right. We were breaking the company’s software build with international files every day. And **it took years to actually get that stuff upstream.”** (P5, Program Manager)

RQ3: Participants felt checklists must be integrated into **organizational goals and priorities**, (e.g., as key performance indicators)

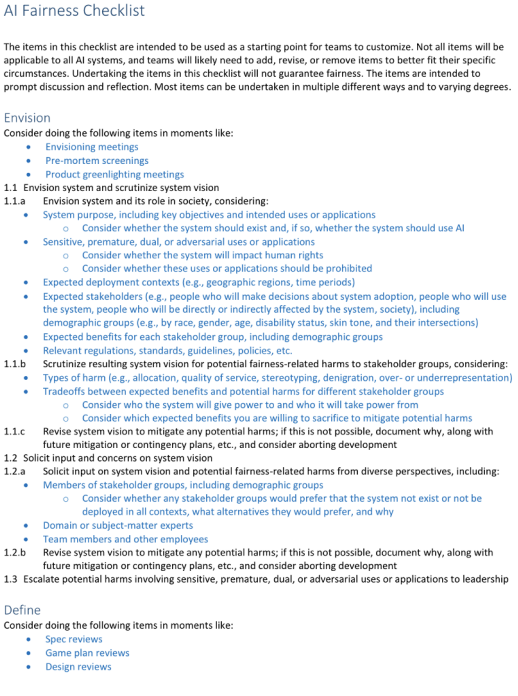
"This is not going to be **moving any of the top-line metrics** that we've been used to moving for years, and not everyone may be bought in yet with the concept of this actually providing a benefit. They can see what we're doing, but it's **hard to prove right now that we're helping users** with this."

(P4, Program Manager)

RQ3: Participants wanted **support in customizing a general checklist** to fit their team's specific needs



How might we **support AI practitioners** in proactively anticipating fairness harms?



AI Fairness Checklist

The items in this checklist are intended to be used as a starting point for teams to customize. Not all items will be applicable to all AI systems, and teams will likely need to add, revise, or remove items to better fit their specific circumstances. Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection. Most items can be undertaken in multiple different ways and to varying degrees.

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.1.c Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

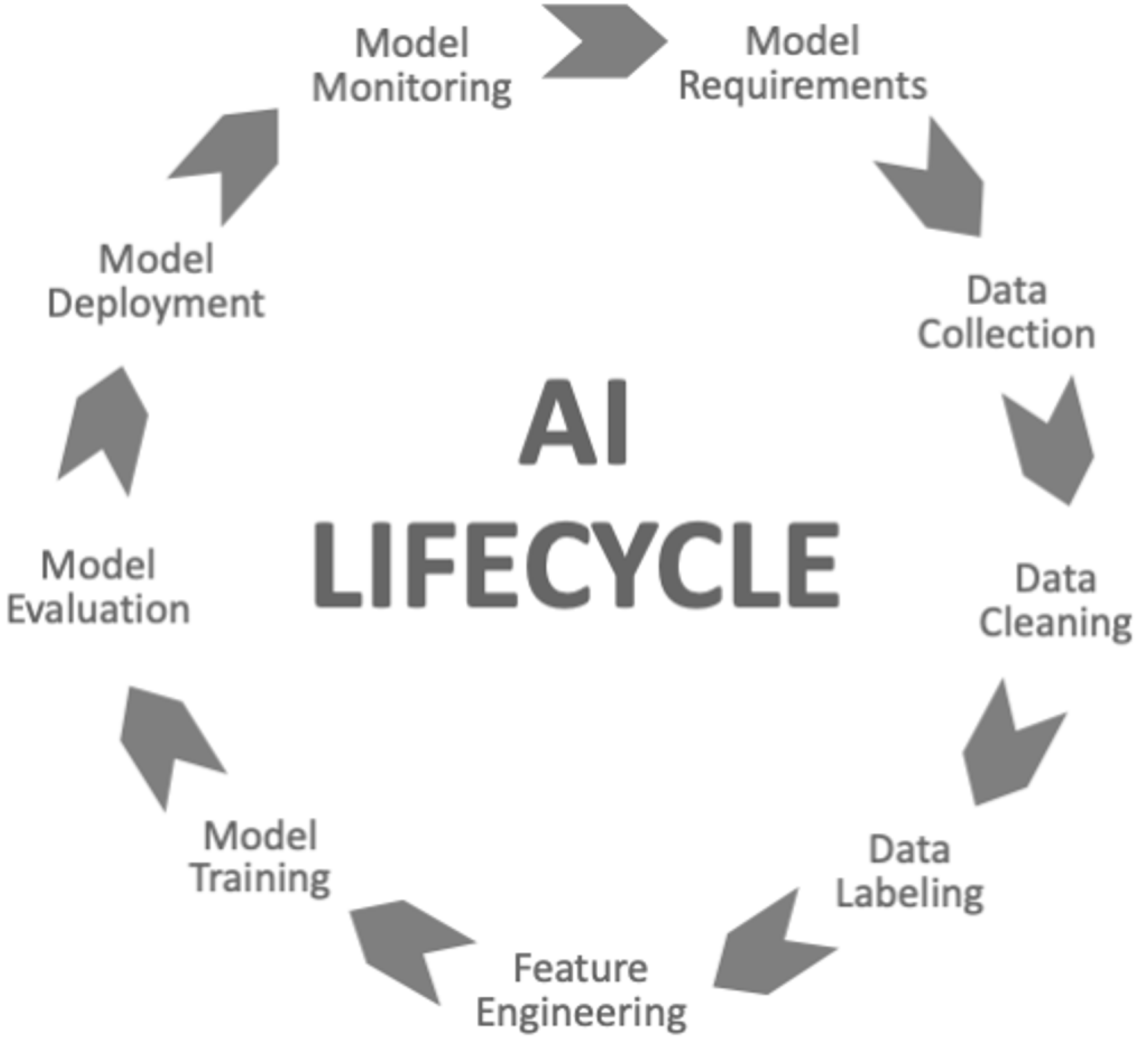
1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

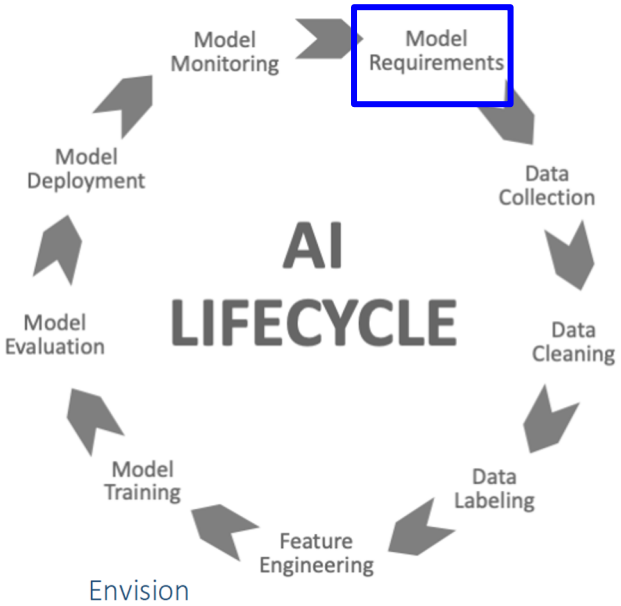
1.3 Escalate potential harms involving sensitive, premature, dual, or adversarial uses or applications to leadership

Define

Consider doing the following items in moments like:

- Spec reviews
- Game plan reviews
- Design reviews


AI
LIFECYCLE
Model Requirements
Data Collection
Data Cleaning
Data Labeling
Feature Engineering
Model Training
Model Evaluation
Model Deployment
Model Monitoring



1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.1.c Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

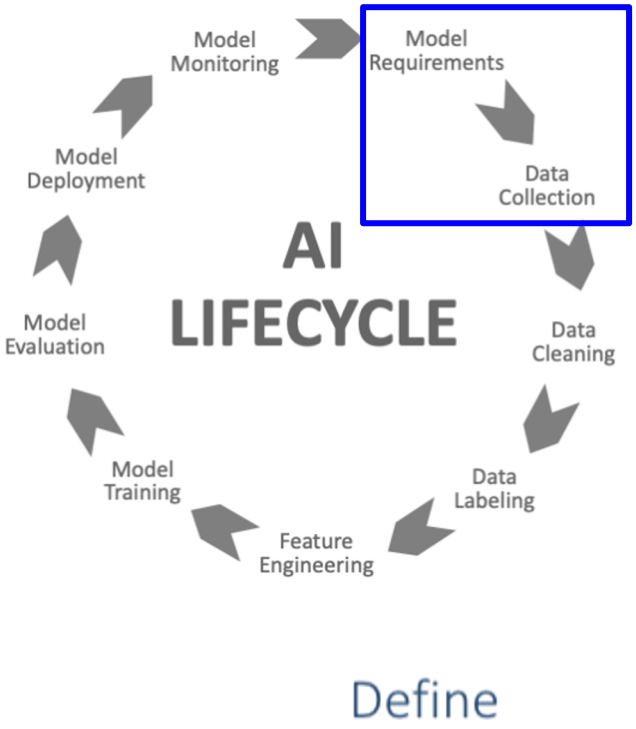
1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.3 Escalate potential harms involving sensitive, premature, dual, or adversarial uses or applications to leadership



2.1 Define and scrutinize system architecture

2.1.a Define system architecture, considering:

- Machine learning models, including their structures, relationships, and interactions
- Objective functions and training algorithms
- Performance metrics (e.g., accuracy, user satisfaction, relevance)
- Functionality for stakeholder feedback (e.g., comments or concerns, third-party audits)
- Functionality for rollback or shutdown in the event of unanticipated fairness-related harms
- Functionality for preventing any prohibited uses or applications
- User interfaces or user experiences
- Other hardware, software, or infrastructure
- Assumptions made when operationalizing system vision via system architecture
    - Consider whether these assumptions are sufficiently well justified

2.1.b Scrutinize resulting definitions for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups

2.1.c Revise system architecture definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

2.2 Define and scrutinize datasets

2.2.a Define datasets needed to develop and test the system, considering:

- Desired quantities and characteristics, considering:
    - Relevant stakeholder groups, including demographic groups
        - Consider oversampling smaller stakeholder groups, but be aware of overburdening
    - Expected deployment contexts
- Potential sources of data
    - Consider reviewing all datasets from third-party vendors
- Collection, aggregation, or curation processes, including:
    - Procedures for obtaining meaningful consent from data subjects
    - People involved in collection, aggregation, or curation, including demographic groups
        - Consider whether people involved might introduce societal biases
    - Incentives for data subjects and people involved in collection, aggregation, or curation
        - Consider whether data subjects might feel undue pressure to provide data
    - Software, hardware, or infrastructure involved in collection, aggregation, or curation
- Relevant regulations, standards, guidelines, policies, etc.
- Assumptions made when operationalizing system vision via datasets
    - Consider whether these assumptions are sufficiently well justified

Scrutinize resulting definitions for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different groups

2.2.b Revise dataset definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

2.3 Define and scrutinize fairness criteria

2.3.a Based on potential fairness-related harms identified so far, define fairness criteria, considering:

- How criteria will be assessed (e.g., fairness metrics and benchmark dataset, system walkthroughs with diverse stakeholders or personas) at each subsequent stage of the lifecycle, including
    - People involved in assessment (e.g., judges), including demographic groups
        - Consider whether people involved might introduce societal biases
    - Datasets needed to assess fairness criteria
- Acceptable (levels of) deviation from fairness criteria
- Potential adversarial threats or attacks to fairness criteria (e.g., “brigading”)
- Assumptions made when operationalizing system vision via fairness criteria

2.3.b Scrutinize fairness criteria definitions for potential fairness-related harms that may not be covered

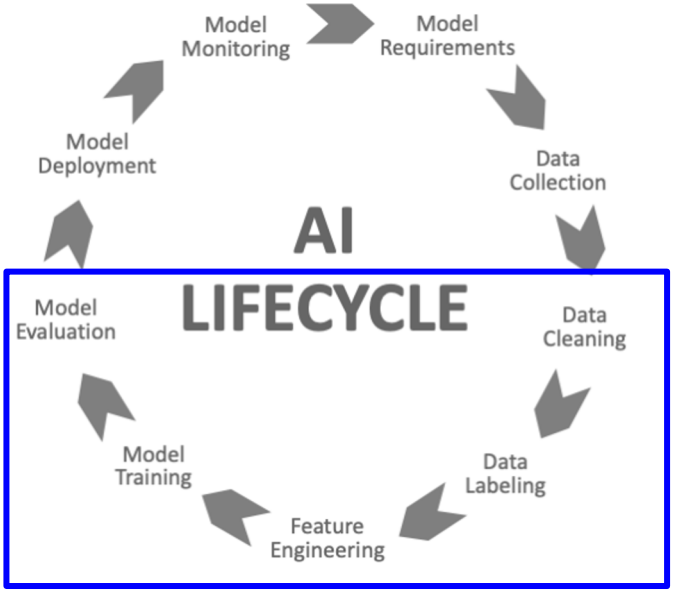
2.3.c Revise fairness criteria definitions to cover any not-covered potential harms; if this is not possible, document why, along with contingency plans, etc., and consider aborting development

2.4 Solicit input and concerns on system architecture, dataset, and fairness criteria definitions

2.4.a Solicit input on definitions and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
- Domain or subject-matter experts
- Team members and other employees

2.4.b Revise definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development



Build & Prototype

3.1 Prototype (and scrutinize) datasets

3.1.a Prototype datasets according to dataset definitions; if datasets deviate from definitions during development, revisit checklist items from "define" stage

3.1.b Document dataset characteristics and limitations (e.g., by creating datasheets), considering:

- Potential audiences for documentation, including:
  - Members of stakeholder groups
  - Team members and other employees
  - Regulators and other third parties

3.2 Prototype (and scrutinize) system

3.2.a Prototype system according to system architecture definitions; if system architecture deviates from definitions during development, revisit checklist items from "define" stage

3.2.b Document system characteristics and limitations (e.g., by creating model cards for the models that comprise the system or a transparency note or factsheet for the system itself), considering:

- Potential audiences for documentation, including:
  - Members of stakeholder groups
  - Team members and other employees
  - Regulators and other third parties

3.3 Assess fairness criteria

3.3.a Assess fairness criteria according to fairness criteria definitions, considering:

- Acceptable (levels of) deviation from fairness criteria
- Tradeoffs between different fairness criteria
- Tradeoffs between performance metrics and fairness criteria
- Discrepancies between development environment and expected deployment contexts

3.3.b If system prototype fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with contingency plans, etc., and consider aborting development

3.4 Undertake user testing

3.4.a Undertake user testing with diverse stakeholders, analyzing results broken down by relevant stakeholder groups. This should be done even if the system satisfies the fairness criteria because the system may exhibit unanticipated fairness-related harms not covered by the fairness criteria. Consider conducting:

- Online experiments
- Ring testing or dogfooding
- ield trials or pilots in deployment contexts

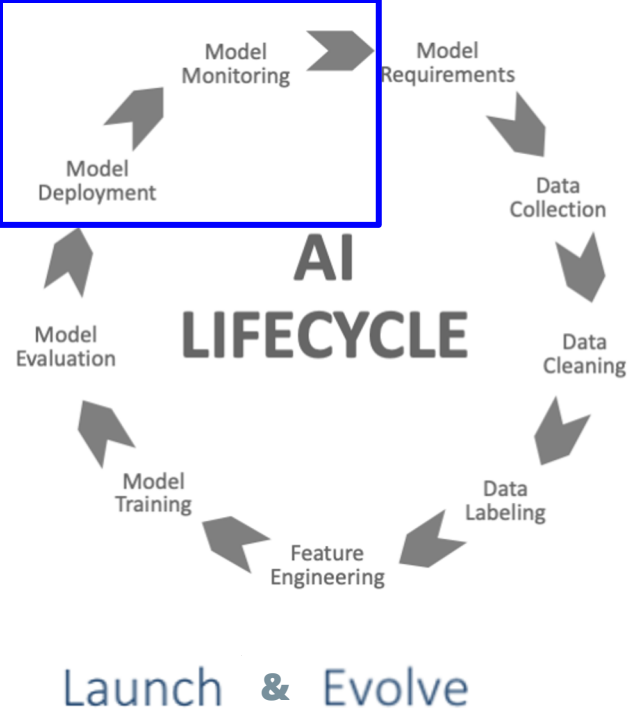
evise production system to mitigate any potential harms; if this is not possible, document why, along ith future mitigation or contingency plans, etc., and consider aborting development

3.5 Solicit input and concerns on system prototype

3.5.a Solicit input on system prototype from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
- Domain or subject-matter experts
- Team members and other employees

3.5.b Revise system prototype to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development



Launch & Evolve

5.1 Participate in public benchmarks
5.1.a Participate in public benchmarks so that stakeholders can contextualize system performance, considering:
- Competitors' responsible AI principles and development practices
- Alternatives to public benchmarks if relevant public benchmarks don't exist (e.g., distributing and publicizing private benchmark datasets for use by competitors or third parties)

5.1.b Revise system to mitigate any harms revealed by benchmarks; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting deployment

5.2 Enable functionality for stakeholder feedback
5.2.a Establish processes for responding to or escalating stakeholder feedback, including:
- Stakeholder comments or concerns
  - Consider establishing processes for redress
- Third-party audits

5.3 Enable functionality for rollback or shutdown in the event of unanticipated fairness-related harms
5.3.a Establish processes for deciding when to roll back or shut down
5.4 Enable functionality to prevent prohibited uses or applications
5.4.a Establish processes for deciding whether unanticipated uses or applications should be prohibited

6.1 Monitor deployment contexts
6.1.a Monitor deployment contexts for deviation from expectations, including:
- Unanticipated stakeholder groups, including demographic groups
- Adversarial threats or attacks

6.1.b Revise system (including datasets) to match actual deployment contexts; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown
6.2 Monitor fairness criteria
6.2.a Monitor fairness criteria for deviation from expectations, including:
- Adversarial threats or attacks

6.2.b If system fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown
6.3 Monitor stakeholder feedback
6.3.a Follow processes for responding to or escalating stakeholder feedback
6.3.b Revise system to mitigate any harms revealed by stakeholder feedback; if this is not possible, document why, update system documentation, and consider rollback or shutdown
6.4 Revise system at regular intervals to capture changes in societal norms and expectations
6.4.a Revisit checklist items from previous stages

“Pause points” for considering fairness-related harms in an AI lifecycle

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

Define

Consider doing the following items in moments like:

- Spec reviews
- Game plan reviews
- Design reviews

Prototype

Consider doing the following items in moments like:

- Go / no-go discussions
- Code reviews

Build

Consider doing the following items in moments like:

- Go / no-go discussions
- Code reviews
- Ship reviews
- Ship rooms

Launch

Consider doing the following items in moments like:

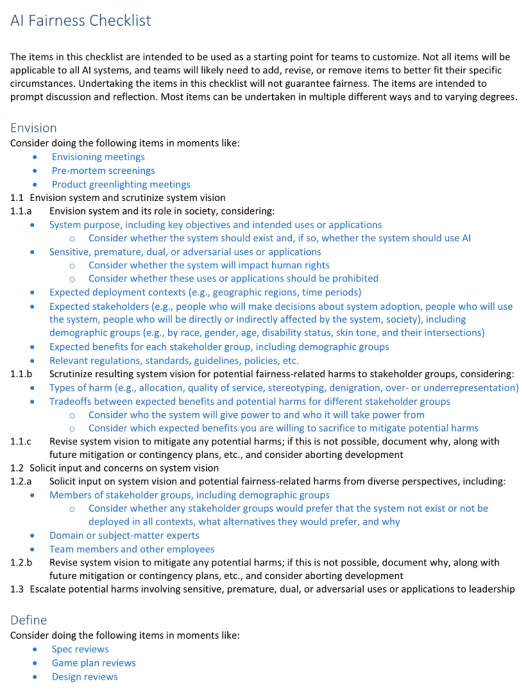
- Ship review before launch
- Code reviews

Evolve

Consider doing the following items in moments like:

- Regular product review meetings
- Code reviews

How might we **support AI practitioners** in proactively anticipating fairness harms?



AI Fairness Checklist

The items in this checklist are intended to be used as a starting point for teams to customize. Not all items will be applicable to all AI systems, and teams will likely need to add, revise, or remove items to better fit their specific circumstances. Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection. Most items can be undertaken in multiple different ways and to varying degrees.

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.1.c Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

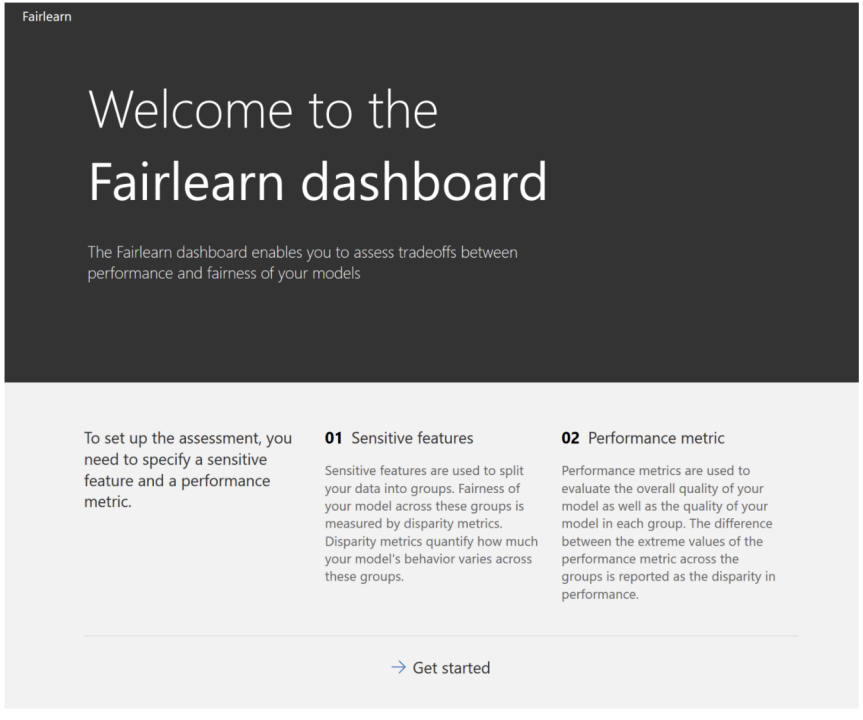
1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.3 Escalate potential harms involving sensitive, premature, dual, or adversarial uses or applications to leadership

Define

Consider doing the following items in moments like:

- Spec reviews
- Game plan reviews
- Design reviews

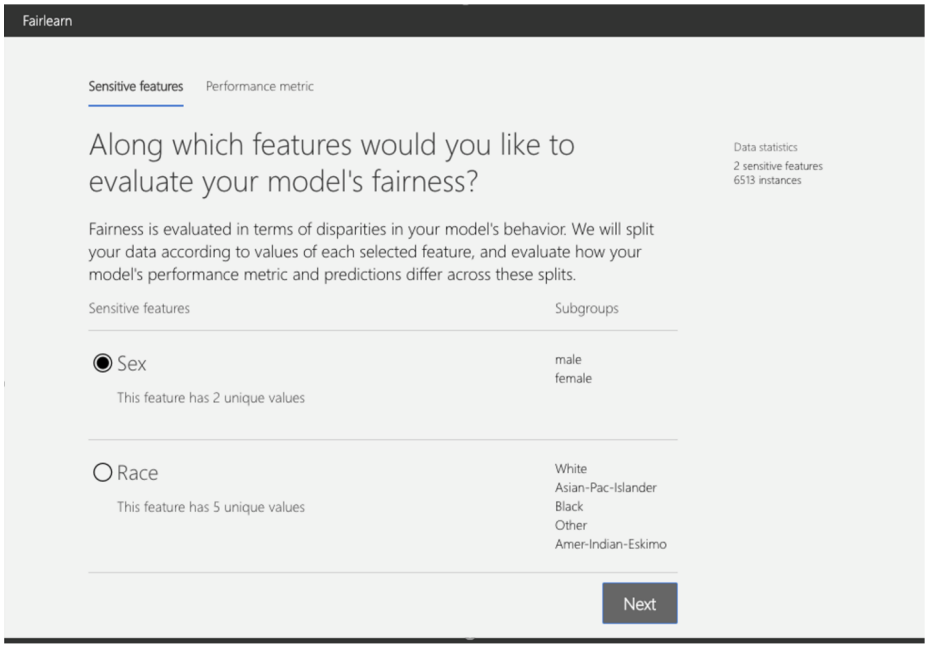


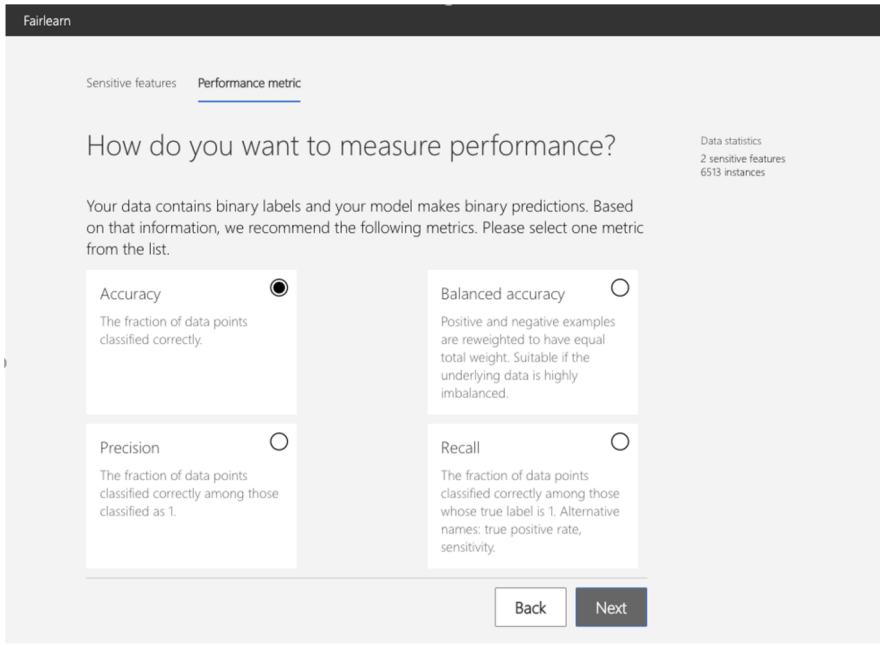
Fairlearn: A toolkit for assessing and improving fairness in AI

- Interactive visualization dashboard (to *assess* fairness)
- Unfairness mitigation algorithms (to *mitigate)*
- Currently designed for classification and regression models
- Open-source, with a community trying to focus on how fairness is ***sociotechnical***

Bird, S., et al., 2020
Agarwal et al., 2018
fairlearn.github.io

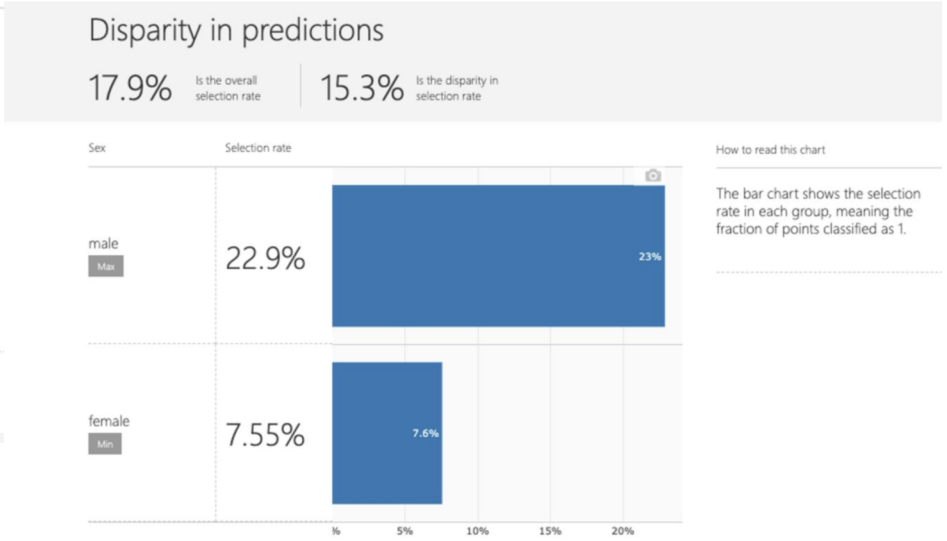
Fairlearn: A toolkit for assessing and improving fairness in AI





Bird, S., et al., 2020
Agarwal et al., 2018
fairlearn.github.io

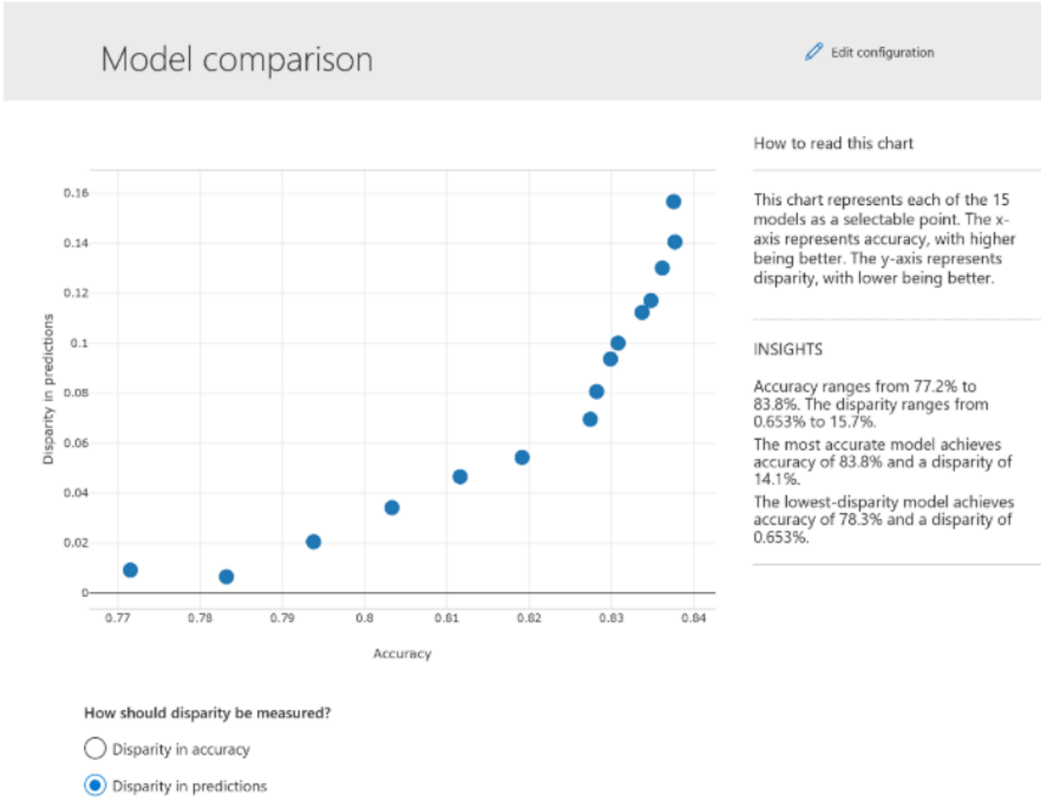
Fairlearn: A toolkit for assessing and improving fairness in AI



Bird, S., et al., 2020
Agarwal et al., 2018
fairlearn.github.io

Fairlearn: A toolkit for assessing and improving fairness in AI

Comparison of multiple models using the Fairlearn dashboard.



Bird, S., et al., 2020
Agarwal et al., 2018
fairlearn.github.io

Discussion questions and future directions

- **How might AI/ML practitioners engage diverse stakeholders** in contributing to fairer AI systems? How might this task be shared across technical and design-oriented roles?

- How might **fairness checklists be customized and adapted** for different teams' workflows, domain areas, and specifics of their product or service? Is this something you might find useful in your work?

- How might toolkits like Fairlearn and checklists **contribute to more "sociotechnical" thinking about fairness** -- not just viewing it as a technical problem to be solved? **How might they be used in practice?**

Thanks!

Keeping Human Concerns in the Loop: Human-Centered Approaches to Responsible AI

Michael Madaio
mimadaio@microsoft.com

Microsoft Research

Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. To appear in the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)
http://www.jennwv.com/papers/checklists.pdf

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn : A toolkit for assessing and improving fairness in AI. 1–7.
https://fairlearn.github.io/