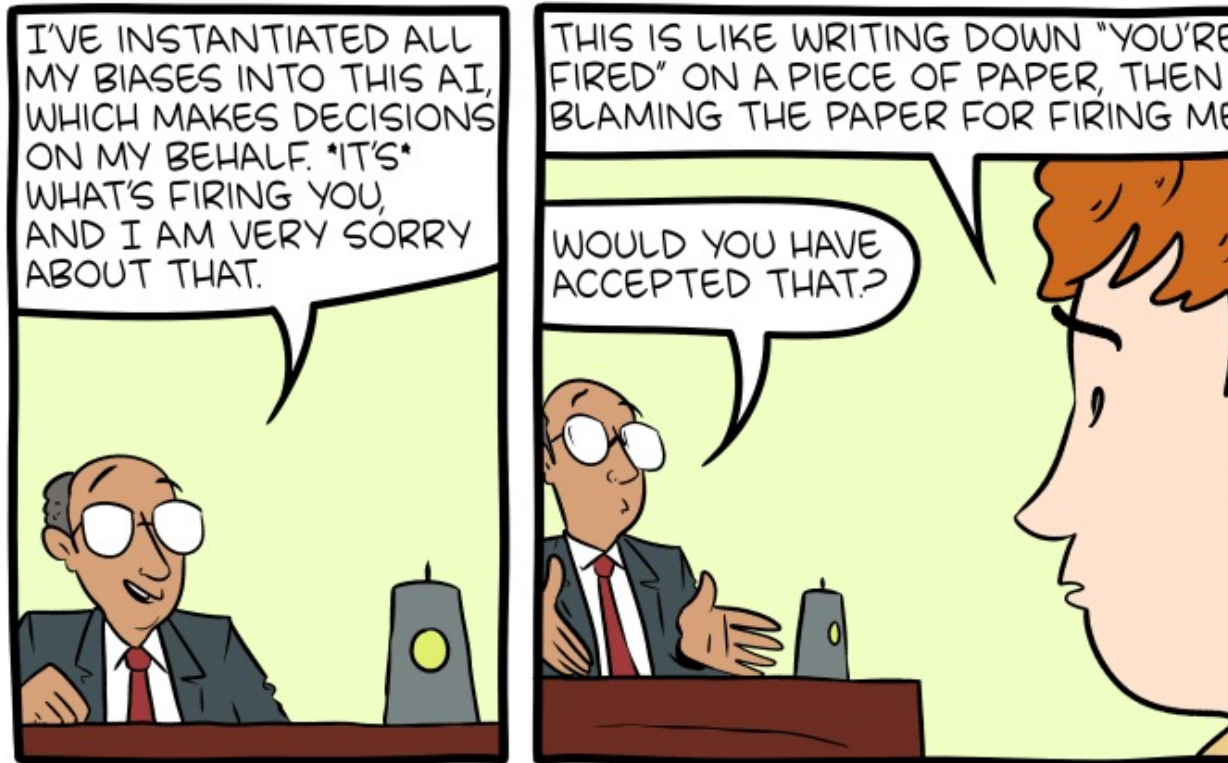


Fairness, Accountability, Transparency, and Ethics

Zach Wood-Doughty
and Bryan Pardo

Some slides from Mark Dredze
Adapted from FAT*2019 Tutorial:
Challenges of incorporating
algorithmic fairness into industry
practice



How We Analyzed the COMPAS Recidivism Algorithm

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

When an Algorithm Helps Send You to Prison

By Ellora Thadaneey Israni
Oct. 26, 2017

MACHINE BIAS

Technical Response to Northpointe

Northpointe asserts that a software program it sells that predicts the likelihood a person will commit future crimes is equally fair to black and white defendants. We re-examined the data, considered the company's criticisms, and stand by our conclusions.

by Jeff Larson and Julia Angwin, July 29, 2016, 11:55 a.m. EDT

The Washington Post

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17

Is my classifier fair?



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Is my classifier fair?

- The stakes are high: unethical (and often illegal) to discriminate based on many protected classes
- But it's not trivial to define fairness or check for it
- The COMPAS doesn't use race as a feature in its predictions
 - Why isn't that enough to make it fair?

Defining Fairness

- We're typically worried about being unfair with respect to a particular variable, e.g. race or gender
- Because these variables are correlated with many/all of our features, just removing them from consideration doesn't fix the problem
 - Often, it makes it worse!

Corbett-Davies, Sam, and Sharad Goel. 2018.

"The measure and mismeasure of fairness: A critical review of fair machine learning."

Defining Fairness

a True Positives	b False Negatives	$b/(a + b)$ False Negative Rate
c False Positives	d True Negatives	$c/(c + d)$ False Positive Rate
$c/(a + c)$ Failure Prediction Error	$b/(b + d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Positive predictive value (PPV: $a / (a+c)$) ; Negative predictive value (NPV: $d / (b+d)$)

	WHITE	AFRICAN AMERICAN
False positive Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
False negative Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Defining Fairness

- 1. Overall accuracy: $(a+d) / (a+b+c+d)$ is the same for both groups
- 2. The positive predictive value (PPV: $a / (a+c)$) and negative predictive value (NPV: $d / (b+d)$) are the same for both groups.
- 3. The false negative rate (FNR: $b / (a+b)$) and the false positive rate (FPR: $c / (c+d)$) are the same for both groups

a True Positives	b False Negatives
c False Positives	d True Negatives

Optimizing for Fairness

- Can we change our loss function to enforce fairness while also maximizing accuracy?
 - Which definition of accuracy?
 - Should we prefer fairness or accuracy?
 - How do we control this trade-off?
 - What are the ethical implications of these decisions?

$$\text{Loss}(X, y, \Theta) = \frac{1}{N} \sum_{i=1}^N (y_i - f(X_i, \Theta))^2$$

$$\text{Regularization}(\Theta) = \sqrt{\Theta_0^2 + \Theta_1^2 + \dots}$$

$$\text{Fairness}(X, G, y, \Theta) = \left[\left(\frac{1}{N_{G=0}} \sum_{i=1}^N \mathbb{1}(G_i = 0) (y_i - f(X_i, \Theta))^2 \right) - \left(\frac{1}{N_{G=1}} \sum_{i=1}^N \mathbb{1}(G_i = 1) (y_i - f(X_i, \Theta))^2 \right) \right]$$

Optimizing for Fairness

- 2. Conditional use accuracy equality: both PPV and NPV are the same across groups.
- 3. Predictive equality and equal opportunity: both FPR and FNR are the same across groups
- Why not require both of these?
- Theorem: If we cannot **perfectly** classify the data and the base rate of the outcome differ by protected class, then it is **impossible** to satisfy both these conditions!

Kleinberg, Jon, et al. 2018 "Algorithmic fairness."

Chouldechova, Alexandra, and Aaron Roth. 2018 "The Frontiers of Fairness in Machine Learning."

Defining Fairness

Tutorial: 21 fairness definitions and their politics

Translation tutorial: 21 fairness definitions and their politics

Arvind Narayanan
@random_walker





PROPUBLICA

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST

Possible ways forward

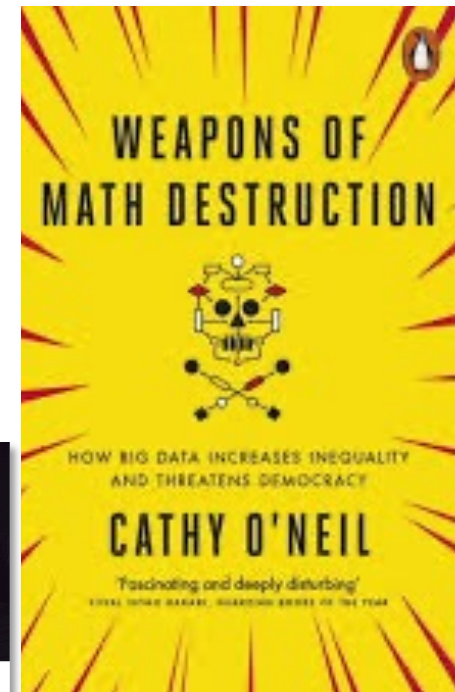
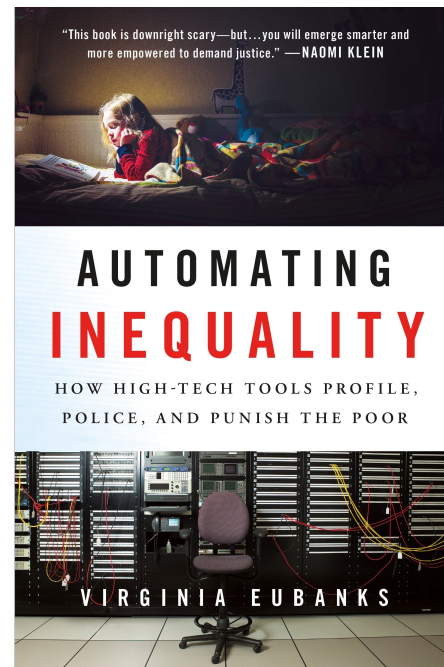
- Can we budget the amount of unfairness that's acceptable, and then minimize classification loss within that budget?
- Can we pre-process our data to eliminate sources of unfairness before we train our model?

Kleinberg, Jon, et al. 2018 "Algorithmic fairness."

Chouldechova, Alexandra, and Aaron Roth. 2018 "The Frontiers of Fairness in Machine Learning."

Are we asking the right questions?

- “Mathematical models can, in fact be, and in some cases have been, tools that further inequality and unfairness and perpetuate bias”
- O’Neil, 2017
- Ways to automate existing systems are often considered instead of questions on how to improve the underlying (sociocultural) system itself.



What about accountability and transparency?

- Most research and tools focus on fairness
 - External accountability: users/regulators can hold an organization responsible for harmful ML
 - Internal accountability: developers/researchers can “debug” a harmful ML system
 - Transparency: decisions around fair ML can be understood by stakeholders

Raji et al., *Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing*. <https://doi.org/10.1145/3351095.3372873>

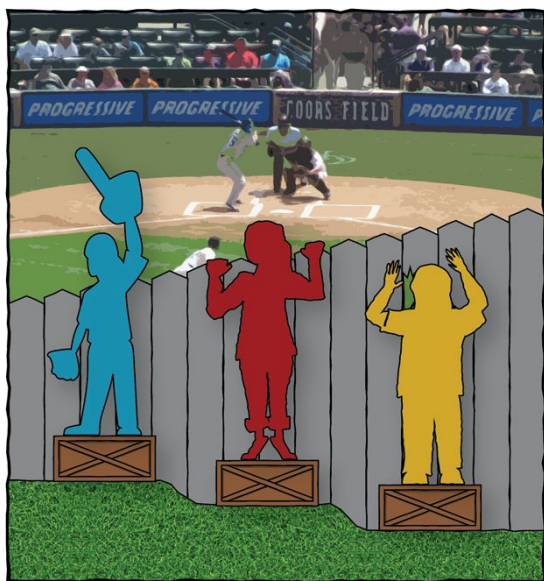
What about ethics?

Fairness is
Political

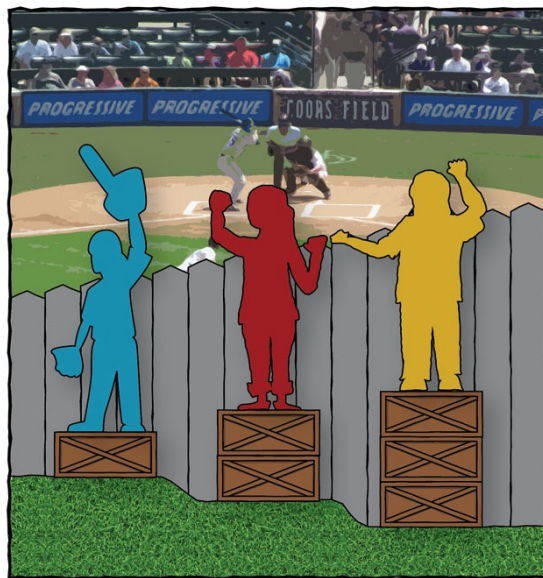
Someone must decide

Decisions will depend on the
product, company, laws, country, etc.

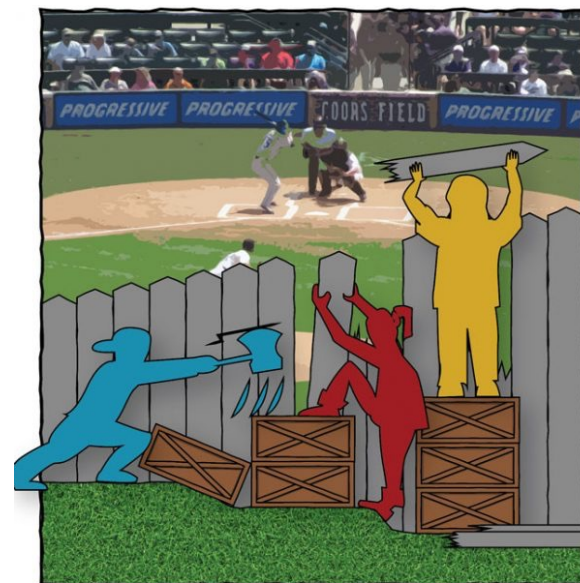
What about ethics?



EQUALITY



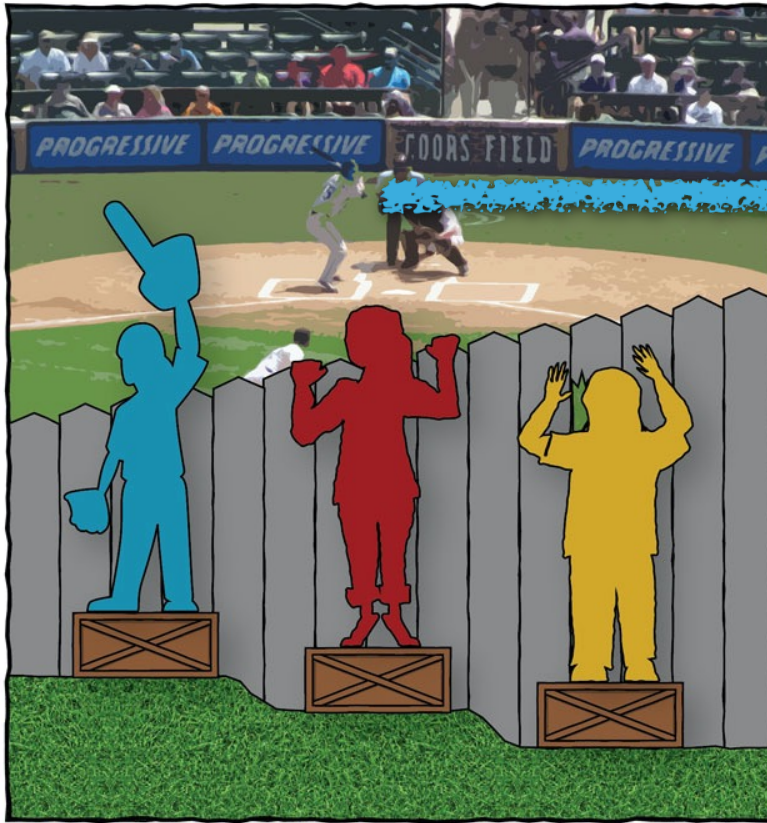
EQUITY



JUSTICE

Who are the stakeholders?

What is equitable? What is at stake?



Access to:

- Hiring,
- Credit,
- Criminal justice,
- Quality UX

Freedom from:

- Discrimination,
- Stereotyping

ACADEMICS / COURSES / DESCRIPTIONS

COMP_SCI 396: Computing, Ethics, and Society

Quarter Offered

Fall : 11-12:20 TuTh ; Van Wart

Computing technologies shape our personal, social, and political lives in increasingly complex and consequential ways – providing tremendous benefits (e.g. convenient access to information, connecting to one another across time and space) and harms (e.g. biased decision-making, mass surveillance, disinformation campaigns, and exclusion from critical material opportunities) that are important to examine and understand.



<https://drive.google.com/file/d/1rUQkVS0NzSH3IEqZDsczSxBbhYHbjamN/view>
<https://www.youtube.com/watch?v=UicKZv93SOY>

A practitioner translation tutorial

Challenges of incorporating algorithmic 'fairness' into practice



Henriette Cramer
Spotify



Jenn Wortman Vaughan -
Microsoft Research



Ken Holstein
CMU & Microsoft

Co-organized by:

Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, Sravana Reddy

Different types of harm

Harms of allocation withhold opportunity or resources

Harms of representation reinforce subordination along the lines of identity, stereotypes

Shapiro et al., 2017

Kate Crawford, “The Trouble With Bias” keynote N(eur)IPS’17

All subsequent slides taken from Microsoft/Spotify tutorial: “Challenges of incorporating algorithmic ‘fairness’ into practice” unless otherwise specified.
<https://drive.google.com/file/d/1rUQkVS0NzSH3IEqZDsczSxBbhYHbjamN/view>

Allocation, incl resources

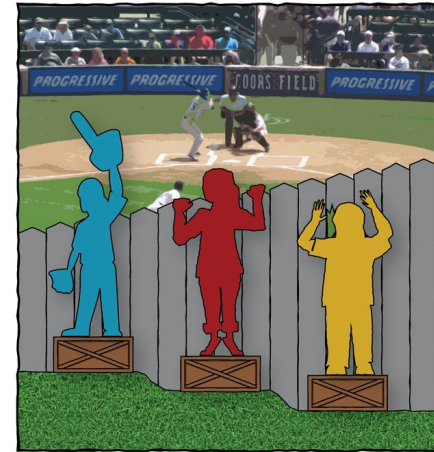
Amazon scraps secret AI recruiting tool that showed bias against women

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.



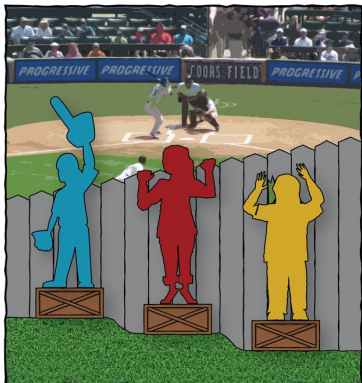
All subsequent slides taken from Microsoft/Spotify tutorial: “Challenges of incorporating algorithmic ‘fairness’ into practice” unless otherwise specified.
<https://drive.google.com/file/d/1rUQkVS0NzSH3IEqZDsczSxBbhYHbjamN/view>

Representation

Over/under-representation, stereotyping, denigration



[Kay et al., 2015]



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]

Sweeney, Latanya, Discrimination in Online Ad Delivery (January 28, 2013).

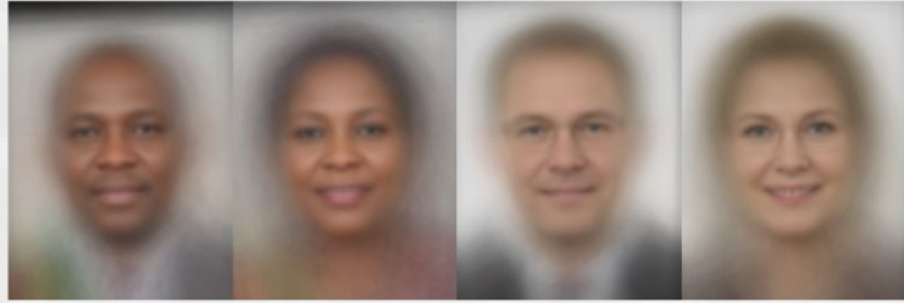
<http://dx.doi.org/10.2139/ssrn.2208240>

Quality of Service, degraded user experience



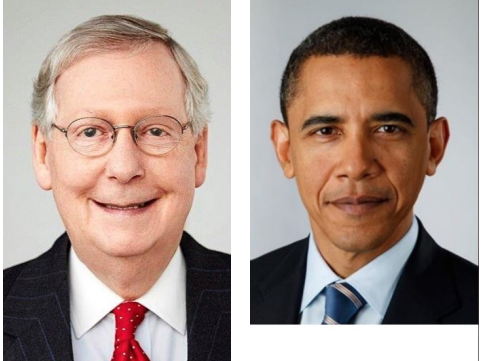
@jozjozjoz, 2009 Nikon S630

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]

Example: Twitter cropping



- Who is responsible for unfair performance of models?
- Is the model developer? The dataset developer? Twitter?



 **Tony "Abolish (Pol)ICE" Arcieri** 🦀
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



6:05 PM · Sep 19, 2020

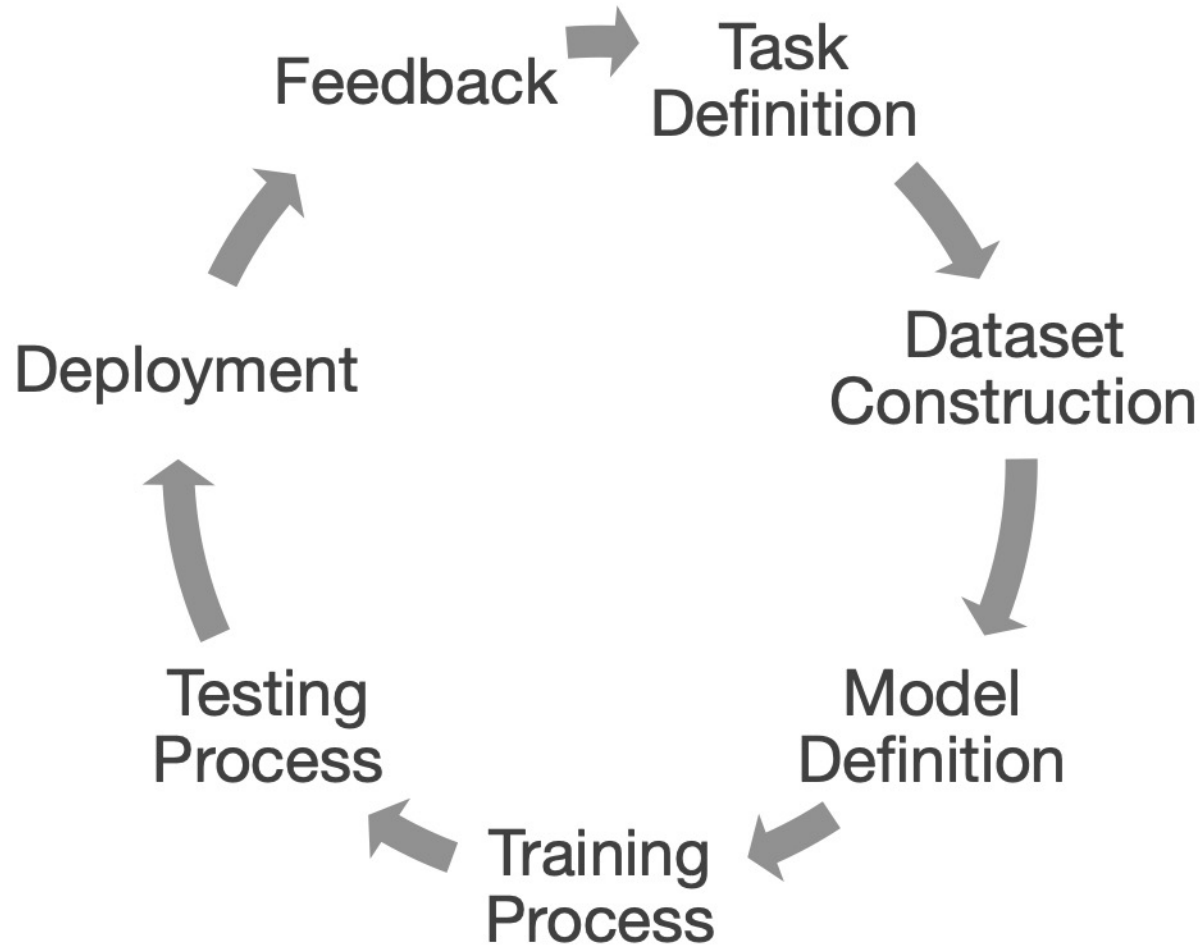
198.9K 67.5K people are Tweeting about this

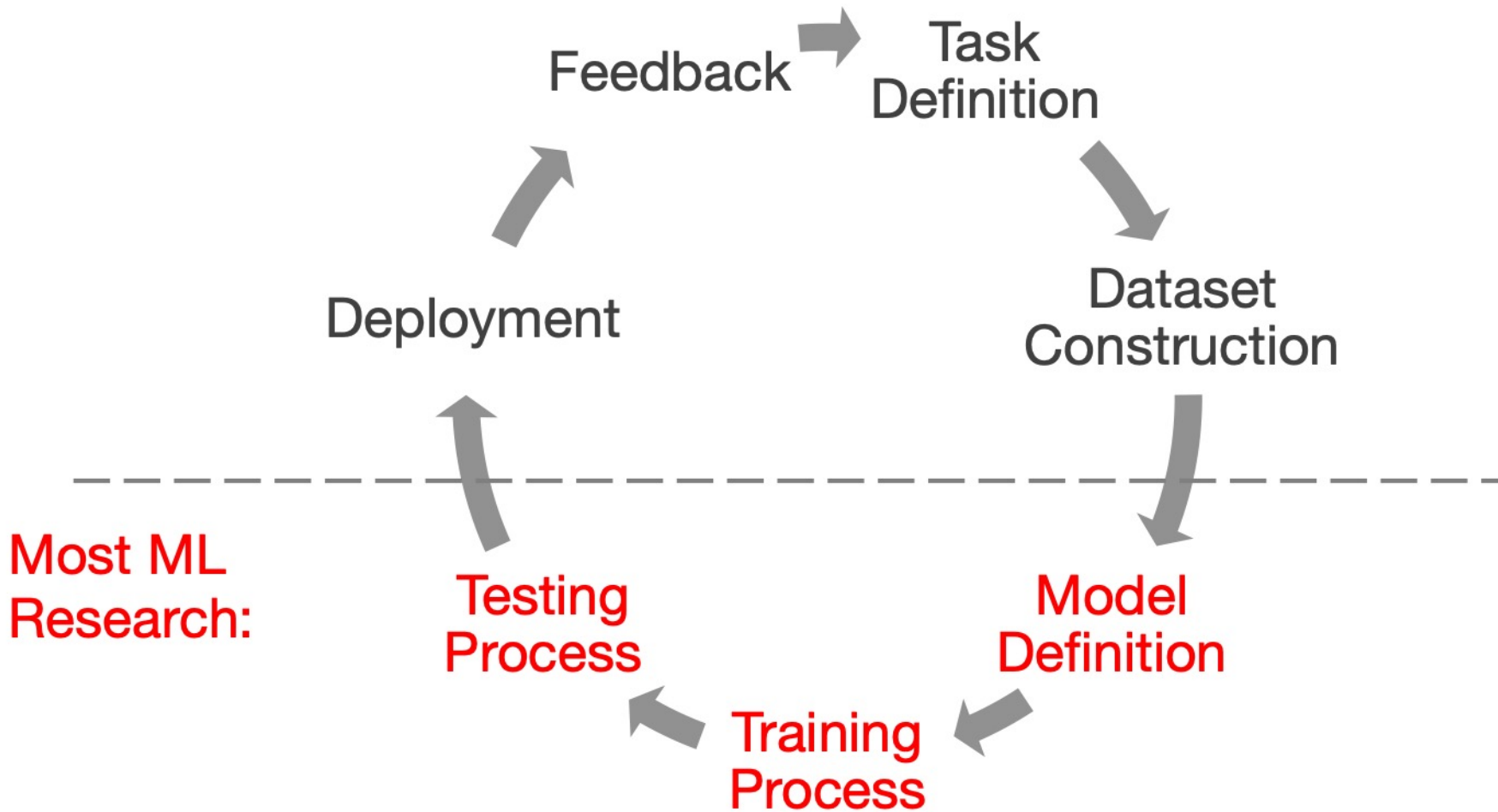


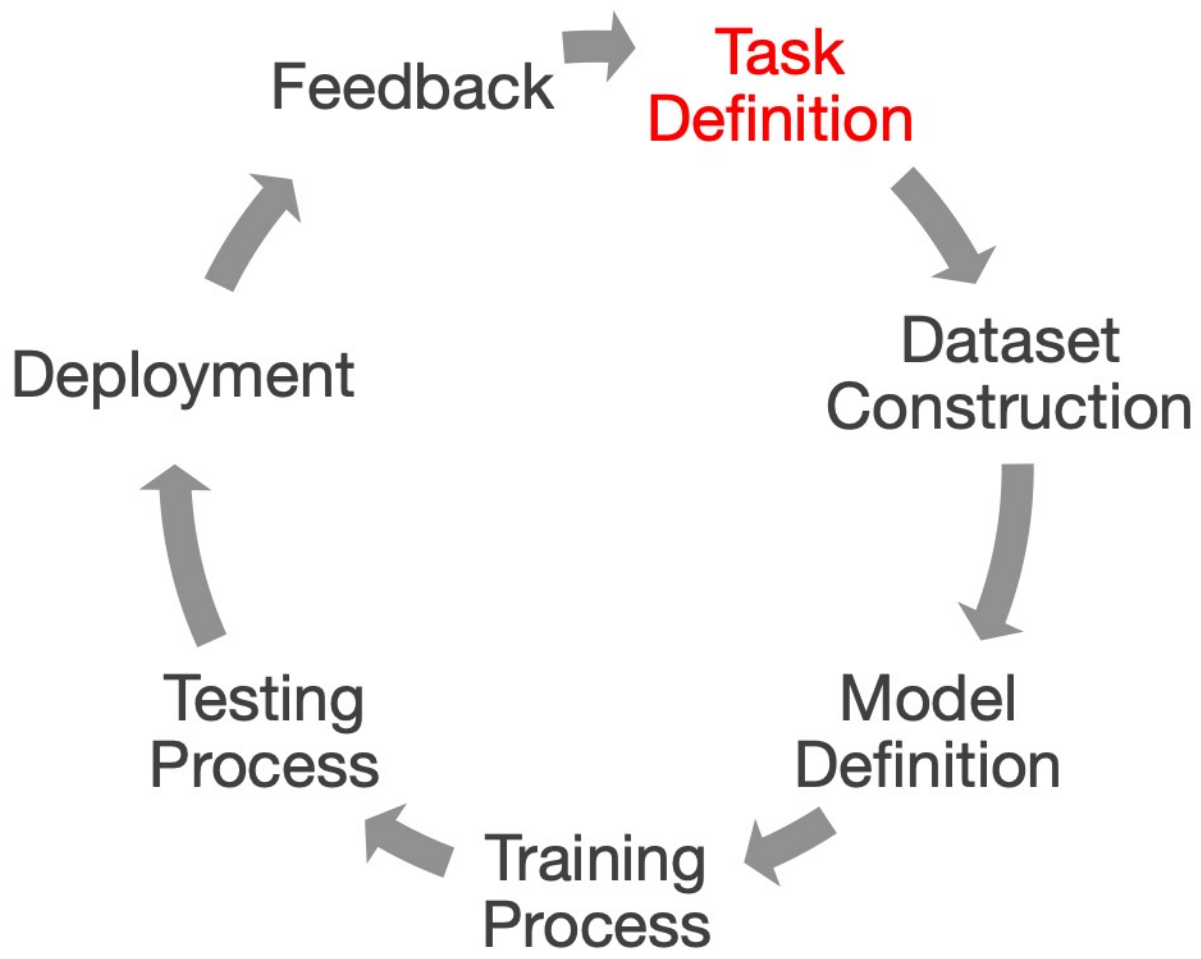
Types of harm can co-occur and need to be specified

	<i>Allocating resources</i>	<i>Quality of Service</i>	<i>Stereotyping</i>	<i>Denigration</i>	<i>Over- / Under-Representation</i>
Hiring system does not rank women as highly as men for technical jobs	X	X			
Photo management program labels images of dark-skinned people as "gorillas"		X		X	
Image search for "CEO" yield only photos for white men			X		X

Fairness Throughout the Machine Learning Lifecycle







Task Definition

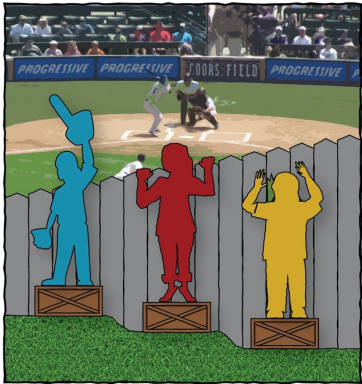


(a) Three samples in criminal ID photo set S_c .

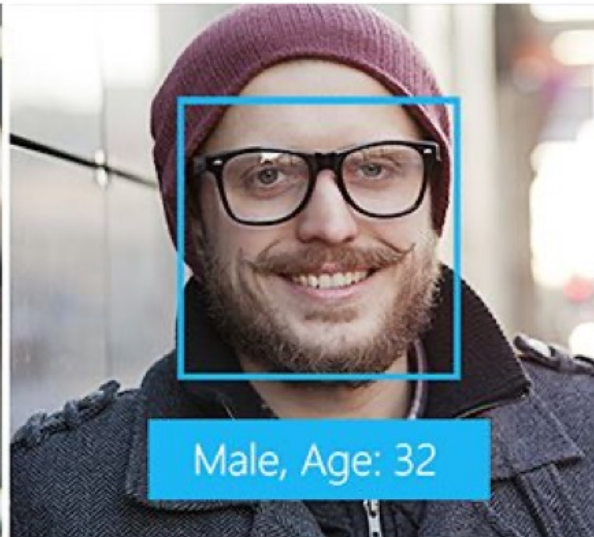
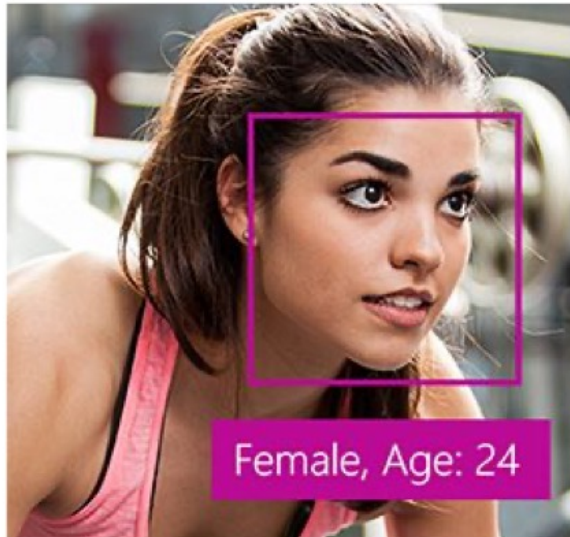


(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



Task Definition

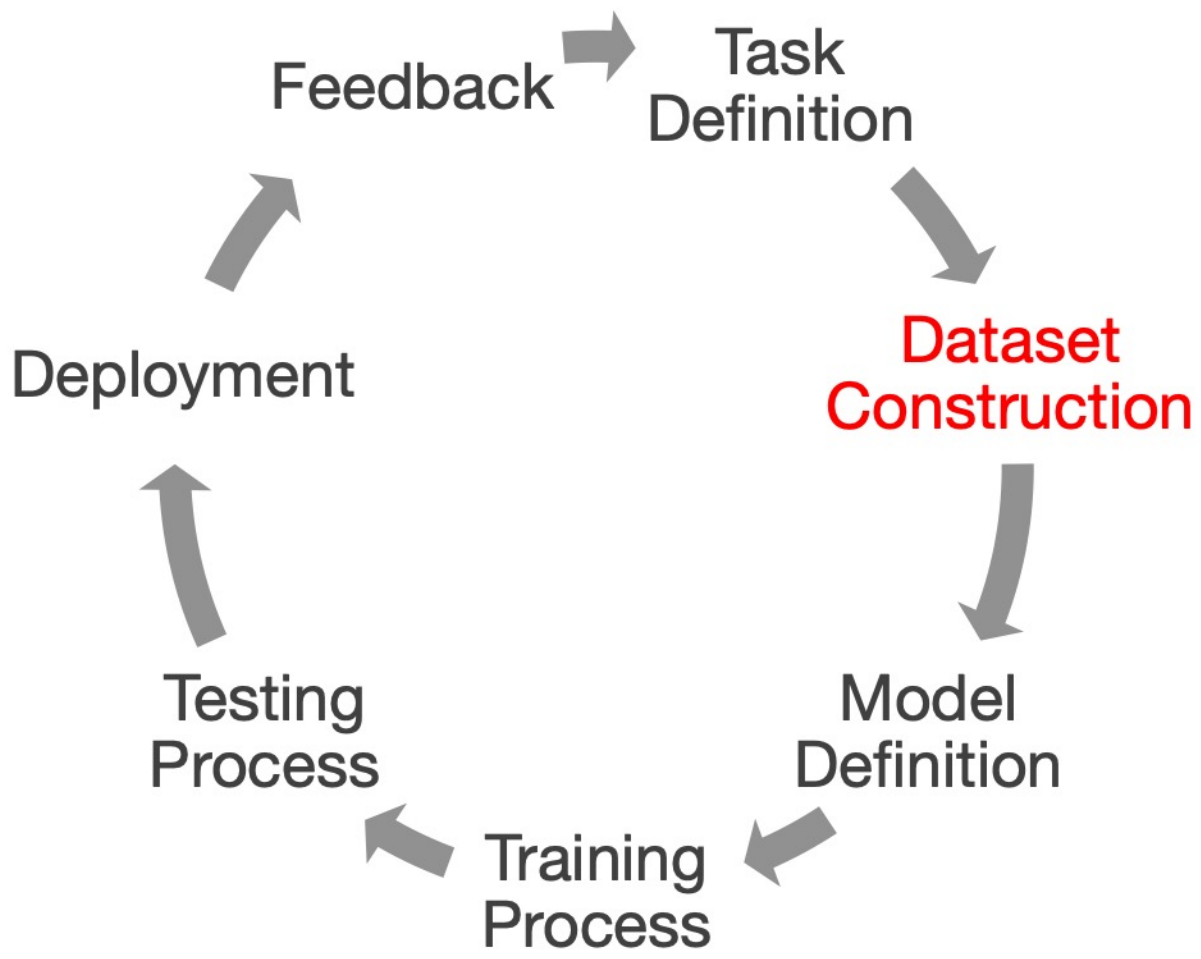


Best Practices: Task Definition

- Clearly define the task & model's intended effects
- Try to identify and document unintended effects & biases
- Clearly define any fairness requirements
- Involve diverse stakeholders & multiple perspectives
- Refine the task definition & be willing to abort

Research Challenges: Task Definition

- What are the most effective ways to elicit diverse opinions?
[e.g., <http://techpolicylab.org/diverse-voices/>]
- How should decisions be made within companies about which tasks to pursue and which to avoid?
- How should we design processes for uncovering unintended effects and biases before development?



Data: Societal Bias

Translate

Turn off instant translation



English Spanish French English - detected



English Spanish Turkish

Translate

He is a nurse
She is a doctor



O bir hemşire
O bir doktor



29/5000



Suggest an edit

Translate

Turn off instant translation



English Spanish French Turkish - detected



Turkish English Spanish

Translate

O bir hemşire
O bir doktor



She is a nurse
He is a doctor



26/5000



Suggest an edit

[Caliksan et al., 2017]

Data: Skewed Sample

Boston releases Street Bump app that automatically detects potholes while driving

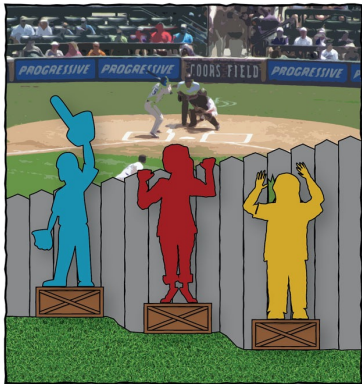
By [DAILY MAIL REPORTER](#)

PUBLISHED: 19:37 EST, 20 July 2012 | **UPDATED:** 20:01 EST, 20 July 2012



[View comments](#)

The next time your car hits a pothole, a new technology could help you immediately tell someone who can do something about it.



Best Practices: Choosing a Data Source

- Think critically before collecting any data
- Check for biases in data source selection process
- Try to identify societal biases present in data source
- Check for biases in cultural context of data source
- Check that data source matches deployment context

Best Practices: Data Collection

- Check for biases in
 - technology used to collect the data
 - humans involved in collecting data
 - sampling strategy
- Ensure sufficient representation of subpopulations
- Check that collection process itself is fair & ethical

Best Practices: Labeling & Preprocessing

- Check for biases introduced by
 - discarding data
 - bucketing values
 - preprocessing software
 - labeling/annotation software
 - human labelers

All HITs Your HITs Queue

HIT Groups (1-20 of 640) [Show Details](#)

Requester	Title	HITs	Reward
ScoutIt	Classify Receipt	151	\$0.03
Crowdsurf Support	Full Text Review - Earn up to \$...	53	\$0.17
Laura A. King	Personality, Information Proce...	1	\$0.15
Crowdsurf Support	Review, edit, and score the tra...	1,091	\$0.02
Erica Fissel	Quick Demographic Survey!(-...	1	\$0.01
ScoutIt	Extract summary information fr...	1	\$0.05
Crowdsurf Support	Transcribe up to 35 Seconds o...	1,042	\$0.05

Data: Labeler Bias

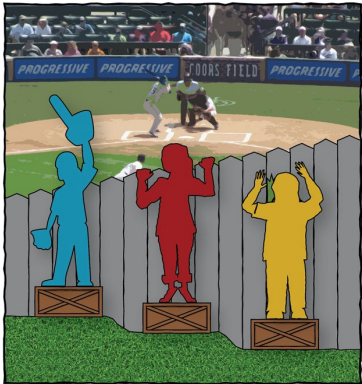
More States Opting To 'Robo-Grade' Student Essays By Computer

June 30, 2018 · 8:13 AM ET

Heard on [Weekend Edition Saturday](#)



TOVIA SMITH



<https://www.theverge.com/2020/9/2/21419012/edgenuity-online-class-ai-grading-keyword-mashing-students-school-cheating-algorithm-glitch>

REPORT / TECH / ARTIFICIAL INTELLIGENCE

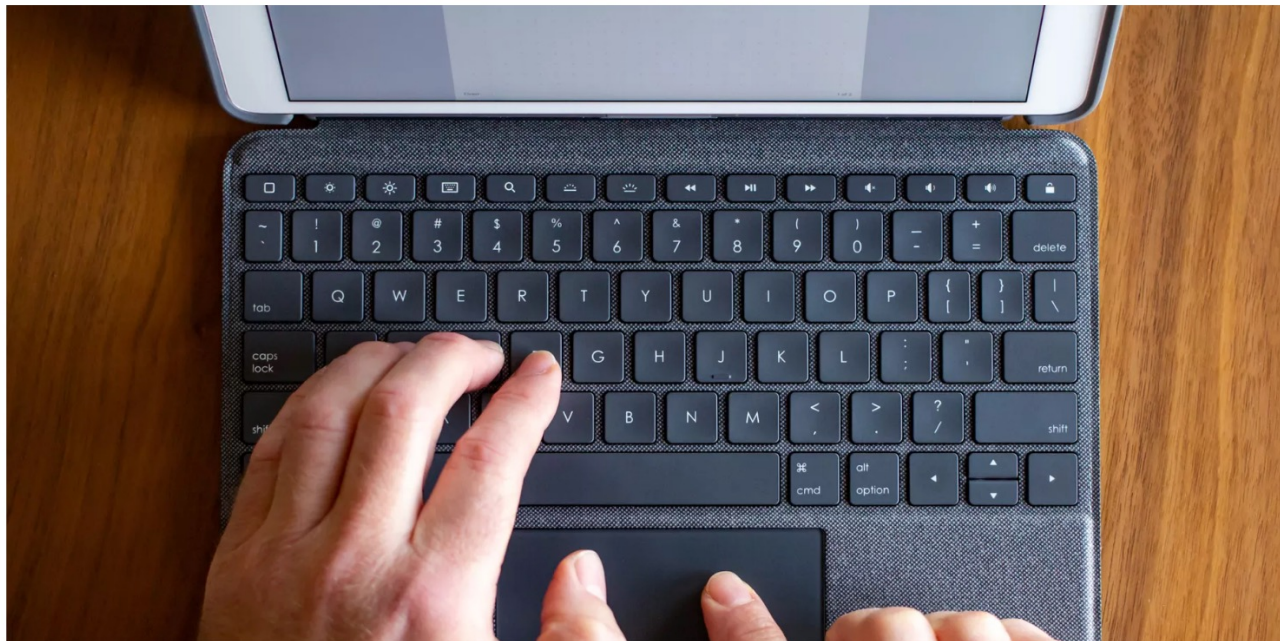
These students figured out their tests were graded by AI — and the easy way to cheat

"He's getting all 100s"

By [Monica Chin](#) | [@mcsquared96](#) | Sep 2, 2020, 10:05pm EDT



SHARE



VERGE DEALS



Samsung's Galaxy Buds Live wireless earbuds are \$35 off at Woot today



Best Practices: Labeling & Preprocessing

- Check for biases introduced by
 - discarding data
 - bucketing values
 - preprocessing software
 - labeling/annotation software
 - human labelers

Datasheets for Datasets (Gebru et al, 2018)

A Database for Studying Face Recognition in Unconstrained Environments

Labeled Faces in the Wild

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.¹

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

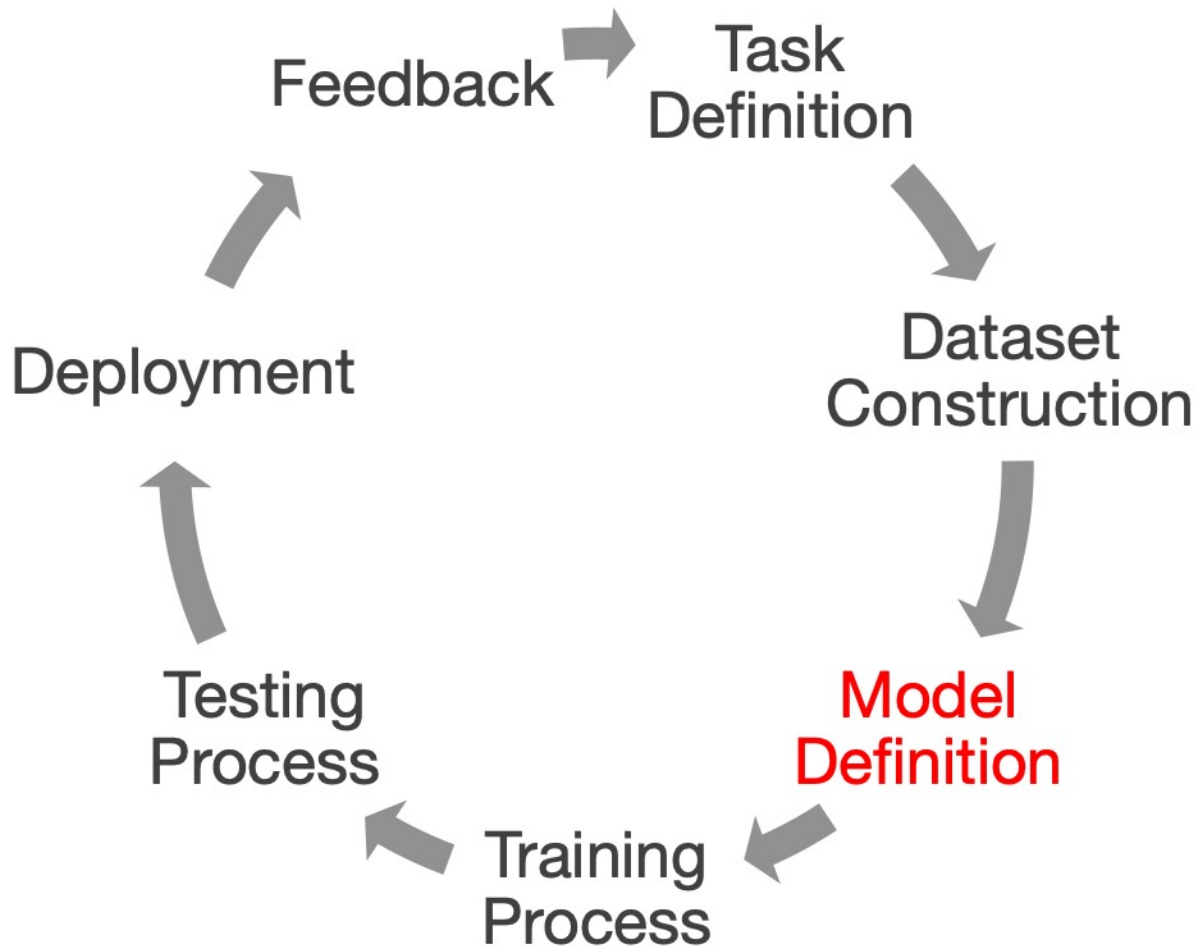
Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

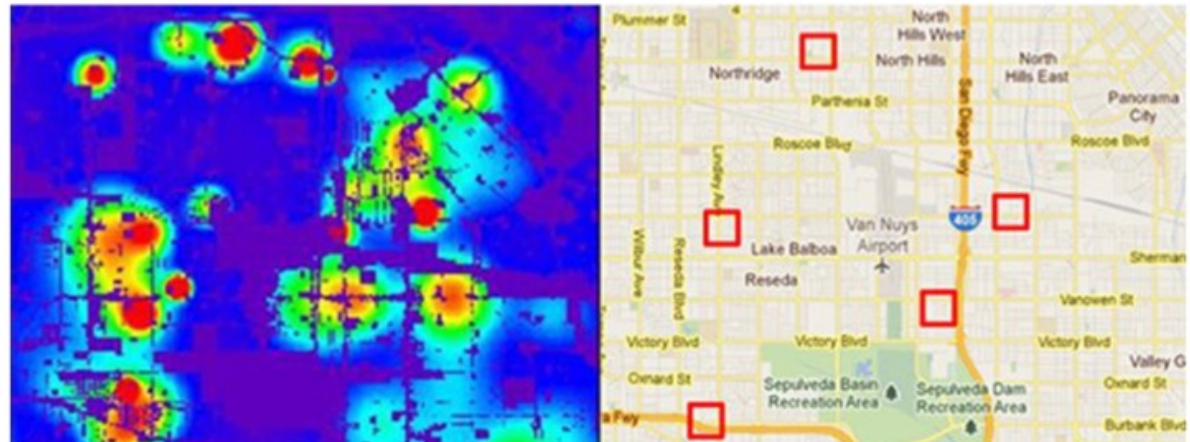
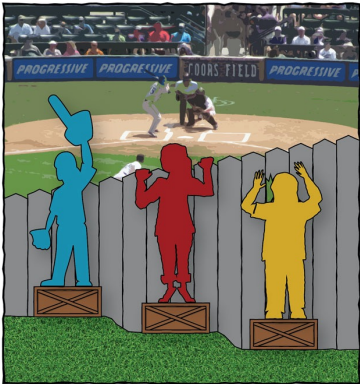
Everything is included in the dataset.



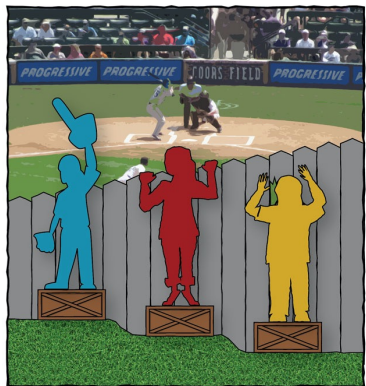
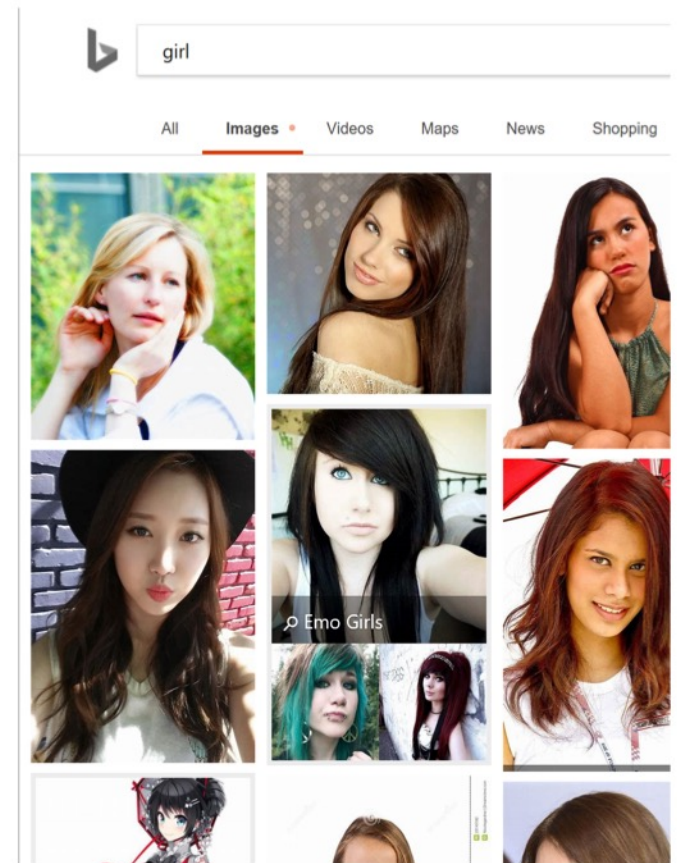
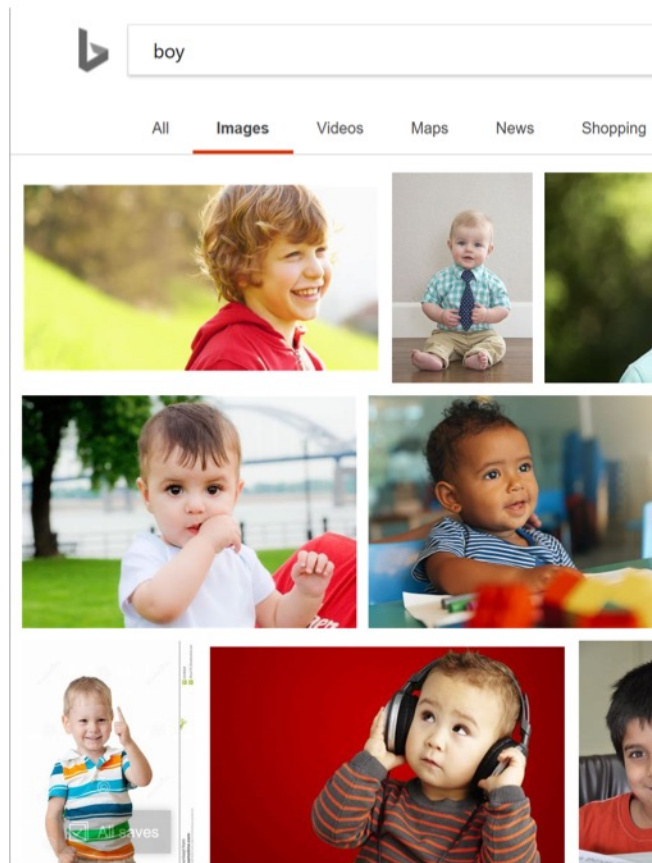
Model: Assumptions

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.



Model: Objective Function

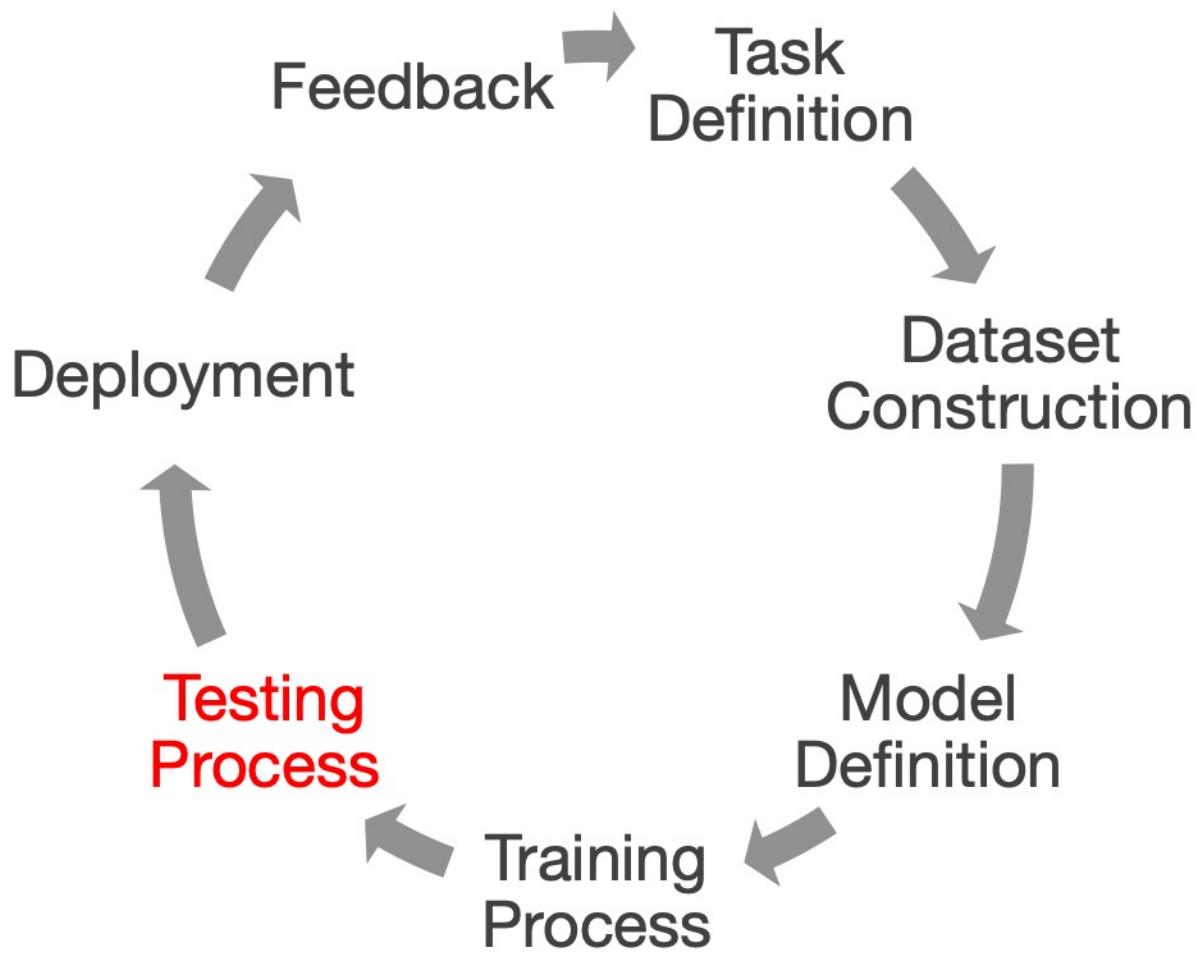


Best Practices: Model Definition

- Clearly define all assumptions about model
- Try to identify biases present in assumptions
- Check whether model structure introduces biases
- Check objective function for unintended effects
- Consider including “fairness” in objective function






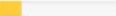




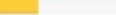



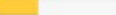
Research Challenges: Model Definition

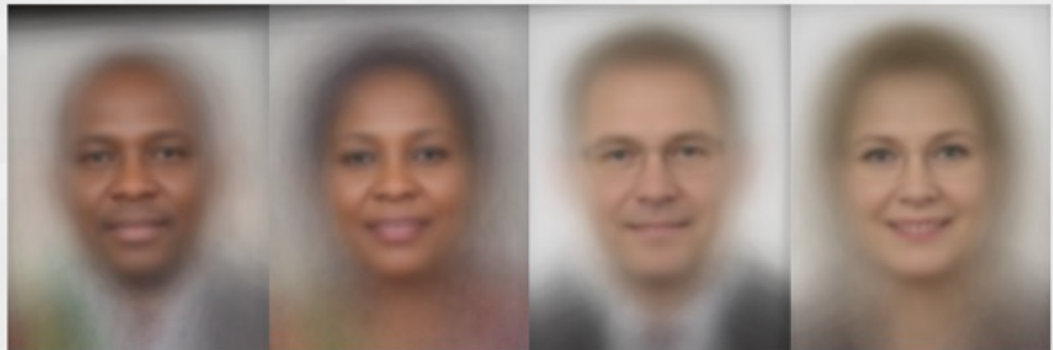
- Identify biases in common modeling assumptions (in consultation with domain experts)
- Explore ways in which some measure of “fairness” might be included in the objective function—but be thoughtful about the limitations of this approach! [e.g., Corbett-Davies and Goel, 2018]
- Move beyond supervised learning



Testing: Data



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% 



Metrics: Points to Consider

Fairness is a non-trivial sociotechnical challenge

- » Many types of harm relate to a broader cultural context than a single decision-making system
- » Many aspects of fairness not captured by metrics

No free lunch! Can't satisfy all metrics [Kleinberg et al. 2017]

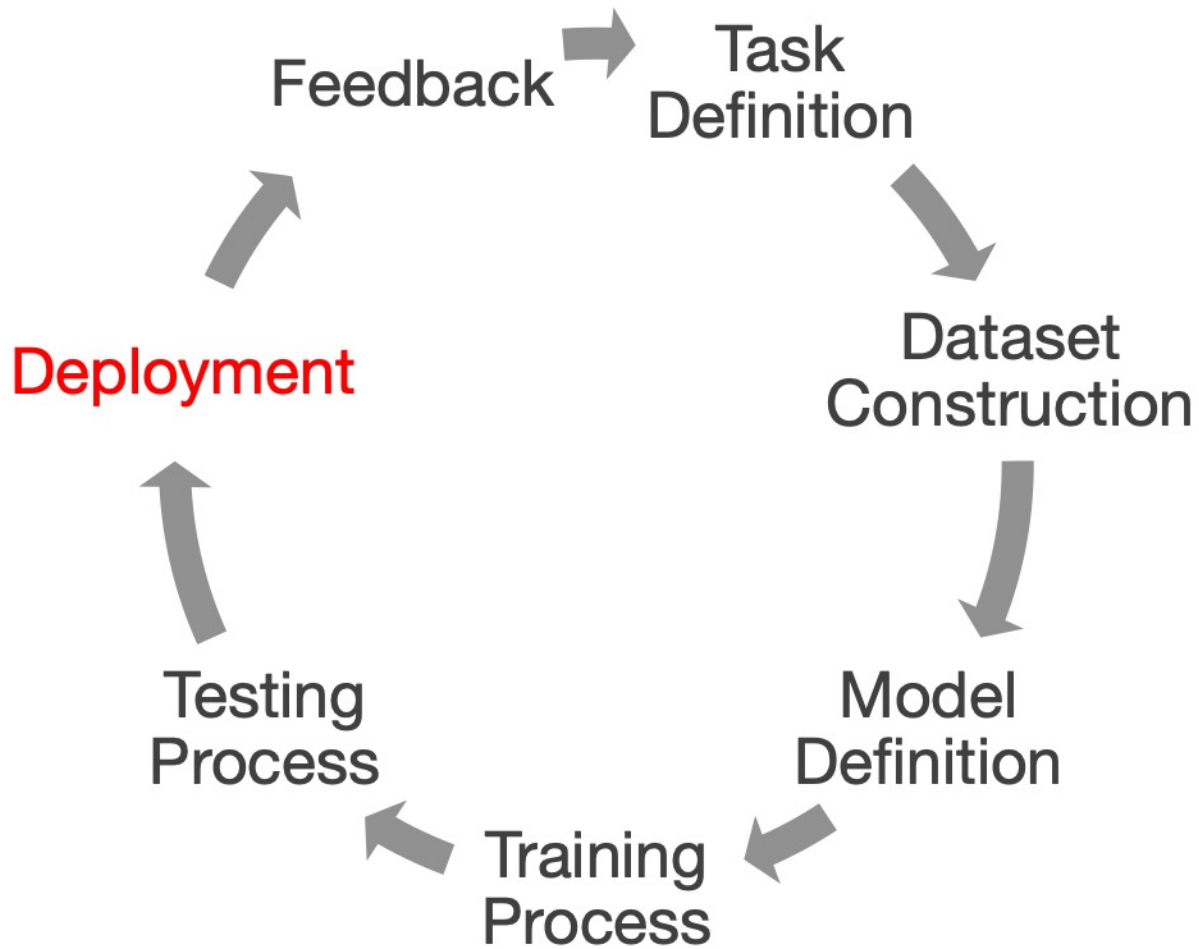
- » Need to make different tradeoffs in different contexts

Best Practices: Testing

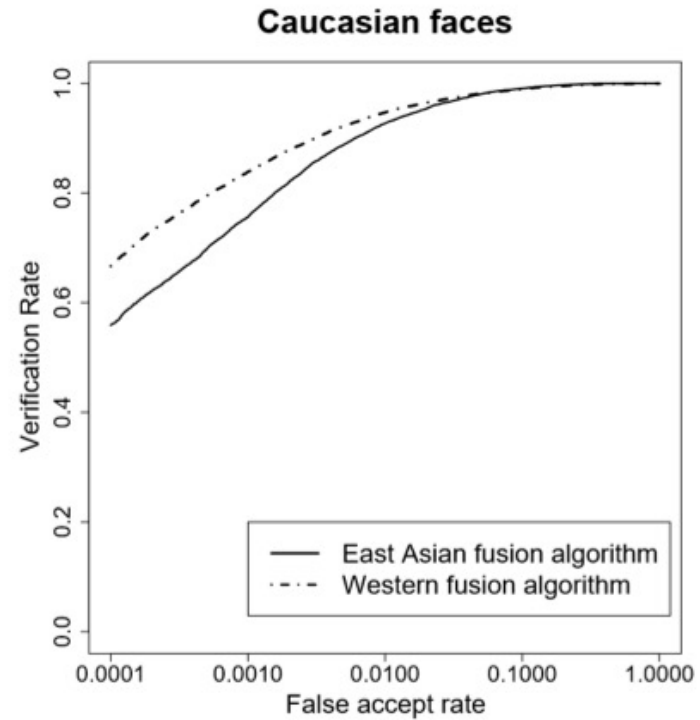
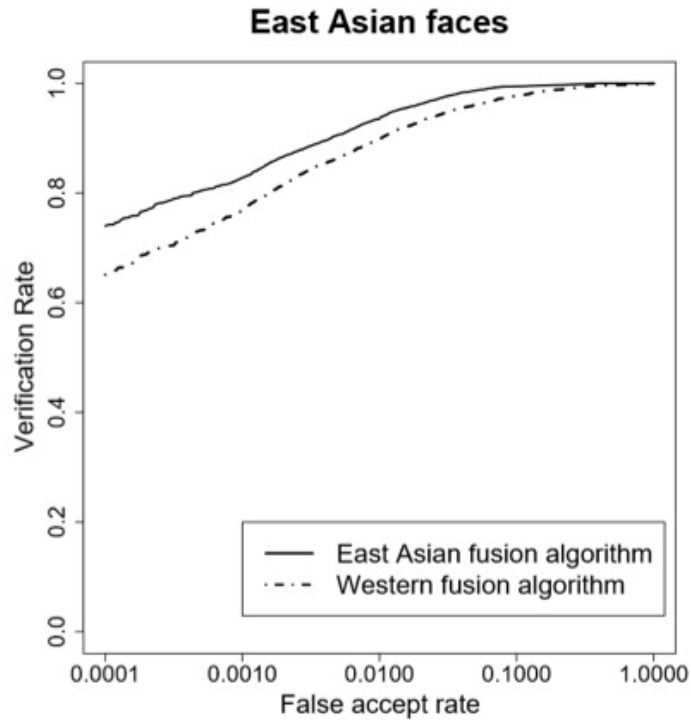
- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

Research Challenges: Testing

- What constitutes “sufficient representation” of subpopulations for test data in different domains?
- What are the subpopulations of interest for testing?
- Which fairness metrics are appropriate in which scenarios?
- What are the right fairness metrics for unsupervised learning, RL, or complex systems like chatbots?



Deployment: Context

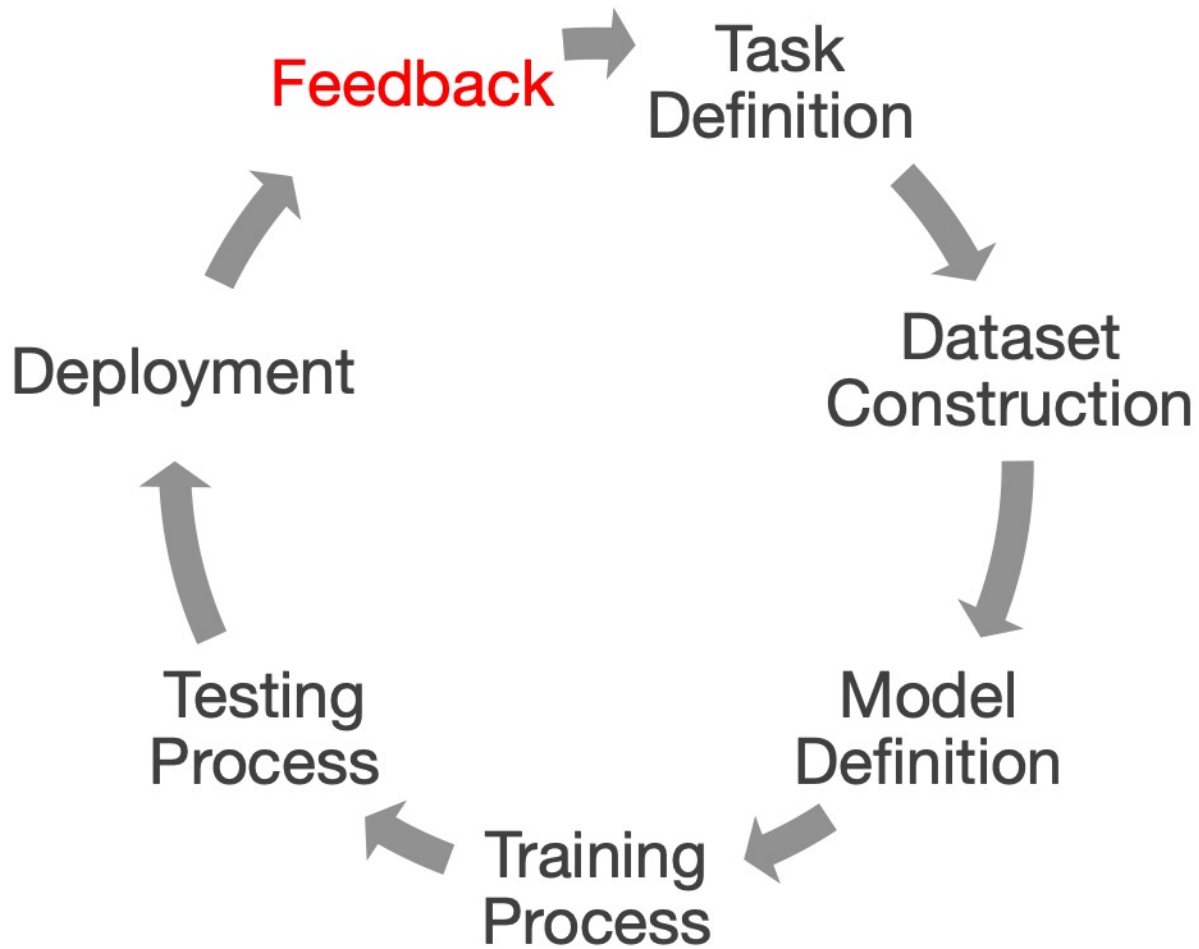


Best Practices: Deployment

- Continually monitor
 - match between training data, test data, and instances you encounter in deployment
 - fairness metrics
 - user reports & user complaints
- Invite diverse stakeholders to audit system for biases

Research Challenges: Deployment

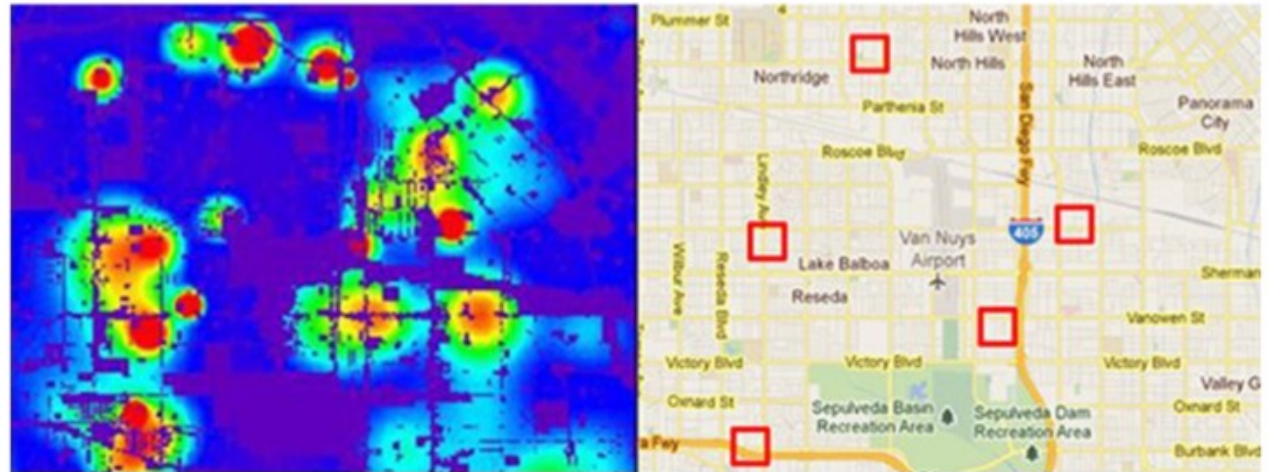
- Methods/tools to audit for shifts in population
- Methods/tools to determine whether a particular error is a one-off issue or is indicative of a systemic problem
- Audit existing system for biases (in collaboration with the teams that built the systems whenever possible)



Feedback: Non-Adversarial

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.



Feedback: Adversarial

Microsoft's chatbot gone bad, Tay, makes MIT's annual list of biggest technology fails

BY ALAN BOYLE on December 27, 2016 at 1:56 pm

8 Comments

f Share

🐦 Tweet

📌 Share

👍 Reddit

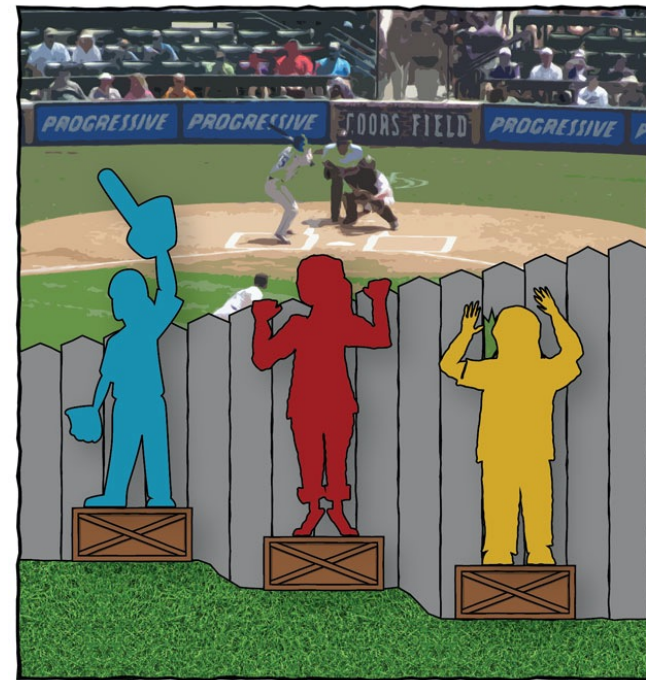
✉ Email

BOT or NOT? This **special series** explores the evolving relationship between humans and machines, examining the ways that robots, artificial intelligence and automation are impacting our work and lives.

Tay, the Microsoft chatbot that pranksters trained to spew racist comments, has joined the likes of the Apple Watch and the fire-prone Samsung Galaxy Note 7 smartphone on MIT Technology Review's list of 2016's biggest technology failures.

Tay had its day back in March, when it was touted as a millennial-minded AI agent that could learn more about the world through its conversations with users. It learned about human nature all too well: Mischief-makers

Tay



Have a scoop that you'd like GeekWire to cover? Let us know.

Send Us a Tip

Best Practices: Feedback

- Continue to monitor
 - match between training data, test data, and instances you encounter in deployment
 - fairness metrics
 - user reports & user complaints
- Monitor users' interactions with system
- Consider prohibiting some types of interactions

Google's Responsible Fairness Practices

<https://ai.google/education/responsible-ai-practices?category=fairness>

Summary:

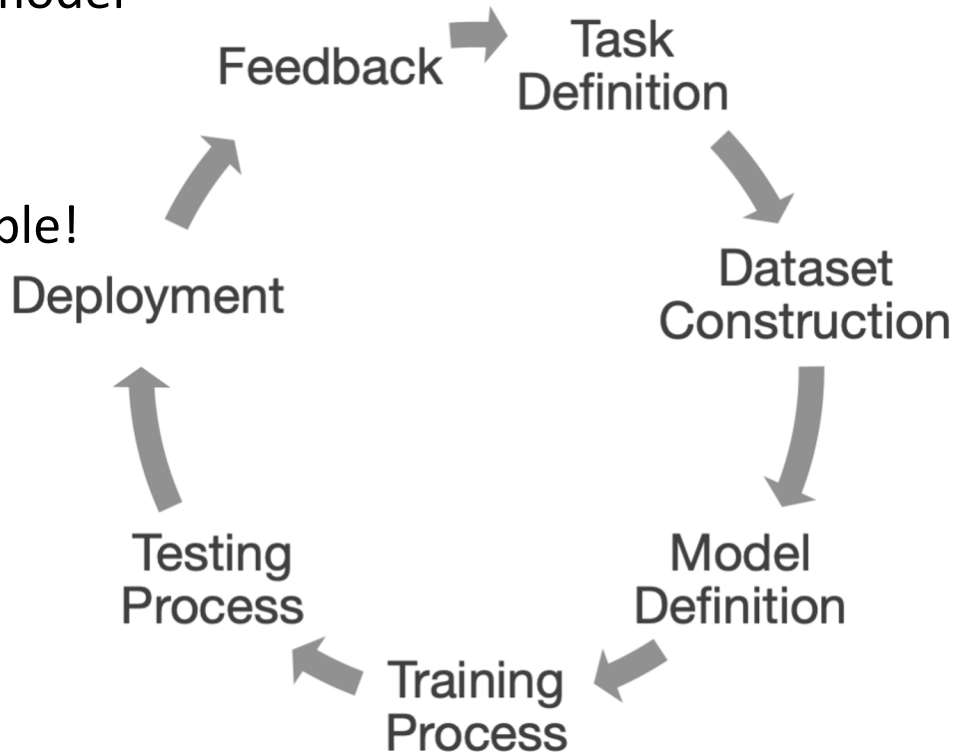
- **Design your product using concrete goals for fairness and inclusion.**
 - Engage with social scientists and other relevant experts.
 - Set fairness goals
- **Check system for unfair biases.**
 - Include diverse testers and adversarial/stress testing.
 - Consider feedback loops
- **Analyze performance.**
 - Evaluate user experience in real-world scenarios.
- **Use representative datasets to train and test your model.**

Bird et al., *Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned*

<https://doi.org/10.1145/3308560.3320086>

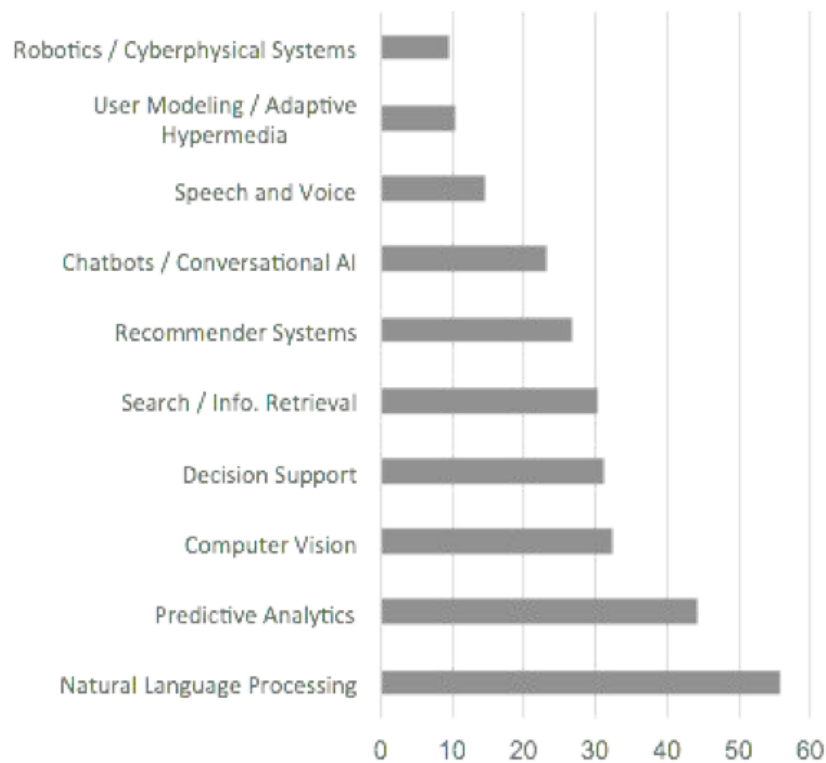
Takeaways

- Defining fairness is challenging
 - Who are the stakeholders?
 - What are the stakes?
- Questions of fairness arise throughout the “ML lifecycle”
 - Not *just* when users see the model
- This is a loop
 - It’s never going to be perfect
 - Be transparent and accountable!

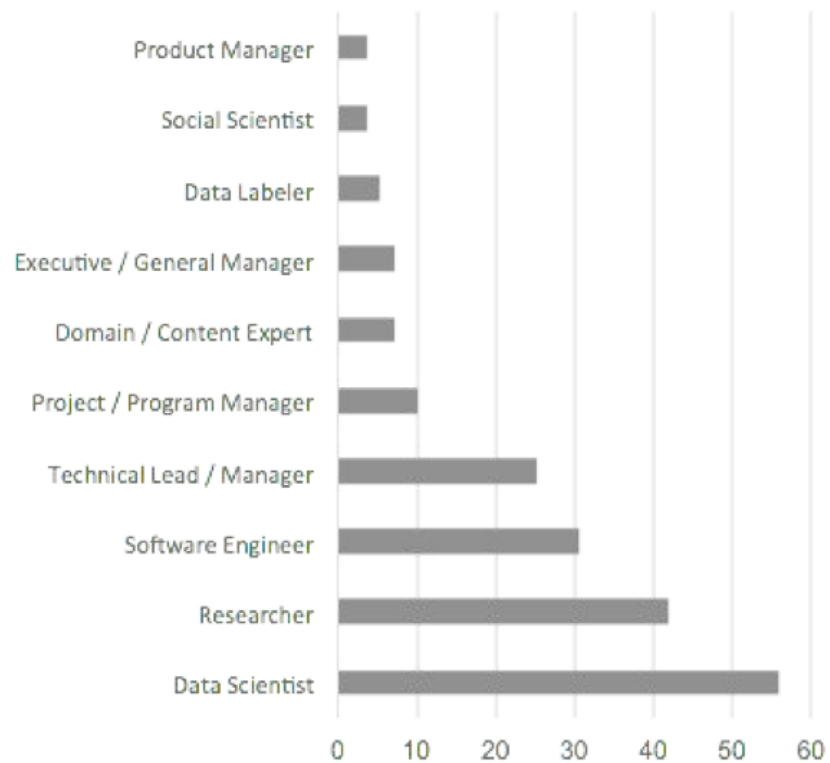


Anonymous survey (n=267)

Technology areas



Team roles



Models vs. Data

- ML literature generally assumes data is given and focuses on fair models and/or algorithms to optimize fairness metrics.
- Industry practitioners more often turn to the data first
 - 65% of survey respondents reported having control over data collection or curation
 - 73% of respondents who had tried to address fairness issues had focused on collecting more training data

Models vs. Data

- Needs for support in creating datasets that support fairness downstream
 - e.g., tools to diagnose whether a given fairness issue might be addressed by **collecting more training data** from a particular subpopulation ... and to predict **how much more** data is needed

**“I always would just really want to
know how much was enough.” - R4**

(cf. Chen, Johansson, & Sontag, 2018; Nushi, Kamar, & Horvitz, 2018)

Blind Spots

- ML literature often assumes subpopulations of interest are given (e.g., based on race, gender, age, religion), but several interviewees highlighted needs for support in identifying relevant subpopulations
 - 62% of survey respondents said it would be very/extremely useful

“It’s just everyone’s collecting all the things that they can think of that could be offensive and testing for it” - R2

“...you know, no one person on the team are experts in all types of bias or offense... especially when you take into account different cultures and different parts of the world” - R4

Limitations of Existing ML Methods

- Most fairness metrics designed for classification (bail/no bail, hire/no hire), while product groups face a much richer space of applications (chatbots, adaptive tutoring, search)
 - Interviewees reported **struggling to use existing fairness research**
 - Applications less amenable to de-contextualized fairness metrics of isolated ML system components

“[with] contextual kinds of responses [it is] harder to [...] predict all the outcomes [... It would help to] find **ways to automate the identification of risky conversation patterns that emerge.**” - R17

“If we think about educational interventions as **analogous to medical interventions or drug trials** [...] we know and [expect] a particular intervention will have **different effects on different subpopulations.**” - R30

(cf. Friedman & Nissenbaum, 1996; Selbst, Friedler, Venkatasubramanian, & Vertesi, 2019)

Note: Bias must be considered relative to task

Gender in loan application



FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.

Gender discrimination is illegal

Gender in medical diagnosis



Gender-specific medical diagnosis is desirable

Limitations of Existing ML Methods

- ML literature generally assumes individual-level access to sensitive attributes, which many teams lack
 - Needs for support in effectively and efficiently monitoring fairness with access only to coarse-grained, partial, or indirect information (e.g., neighborhood- or organization-level statistics)

“If we had more people who we could throw at this... ‘Can we leverage this fuzzy [coarse-grained] data to [audit]?’ that would be great [...]

It’s a fairly intimidating research problem I think, for us.” - R21

(cf. Kilbertus et al., 2018; Veale & Binns, 2018)