Principles of AI Engineering
Chapter 7: Ethics, Fairness, and Transparency

Prof. Dr. Steffen Herbold

Credit:

Based on contents from Christian Kästner (https://github.com/ckaestne/seai)

Contents

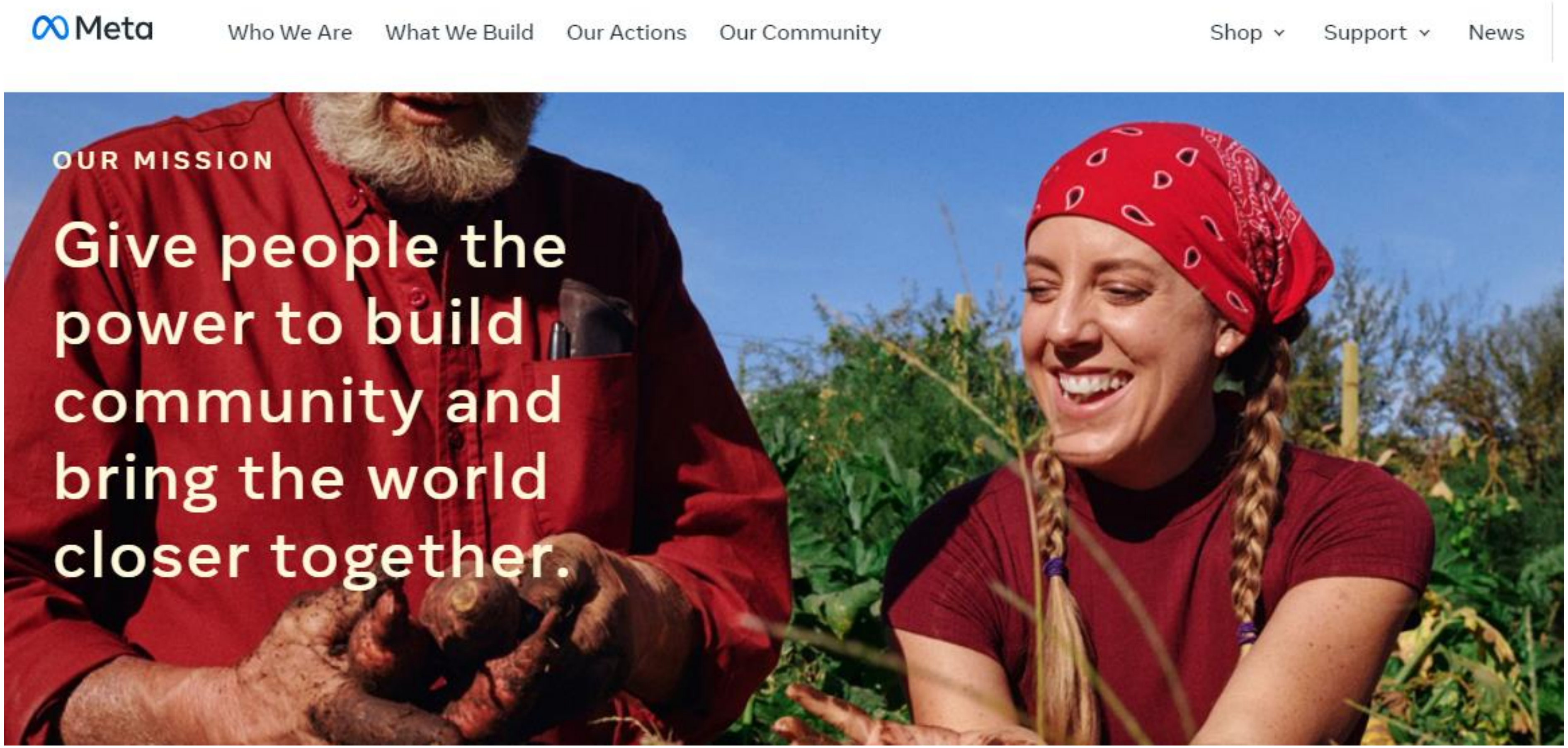
Ethics

Legal vs. Ethical

- Legal
  - In accordance to societal laws
  - Systematic body of rules governing society, defined by the government
  - Punishment for violation

- Ethical
  - Following moral principles of traditions, groups, or individuals
  - Branch of philosophy, science of a standard human conduct
  - Professional ethics are rules codified by a professional organization
    - E.g., ACM: https://www.acm.org/code-of-ethics
  - Not legally binding, (usually) no strict enforcement
  - High ethical standards may yield long-term benefits through image and staff locality

Example: Social Networks



Live exercise: What are the actual business objectives?

https://about.facebook.com/company-info

Social media business objectives

- Monetize interactions with social media
  - This is certainly legal and not, in itself, unethical
- How is monetization optimized?

User engagement!

https://commons.wikimedia.org/wiki/File:Social-media-g6e011fe56_1920.png

How to maximize user engagement

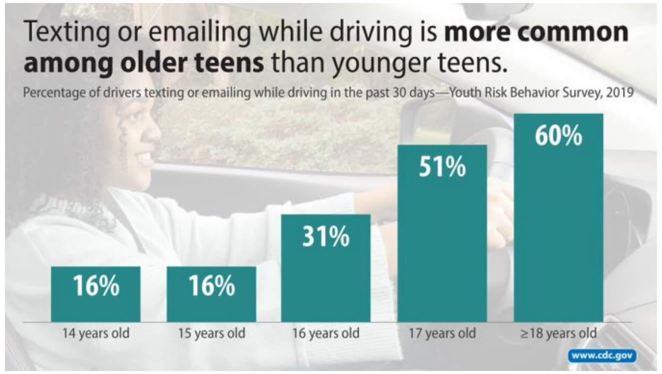
- Infinite scroll
  - Page automatically extended with more content before reaching the bottom
  - No natural stopping point, e.g., by clicking on next page
  - Encourages non-stop, continual use

- Personal recommendations
  - Suggest user content and news to increase engagement

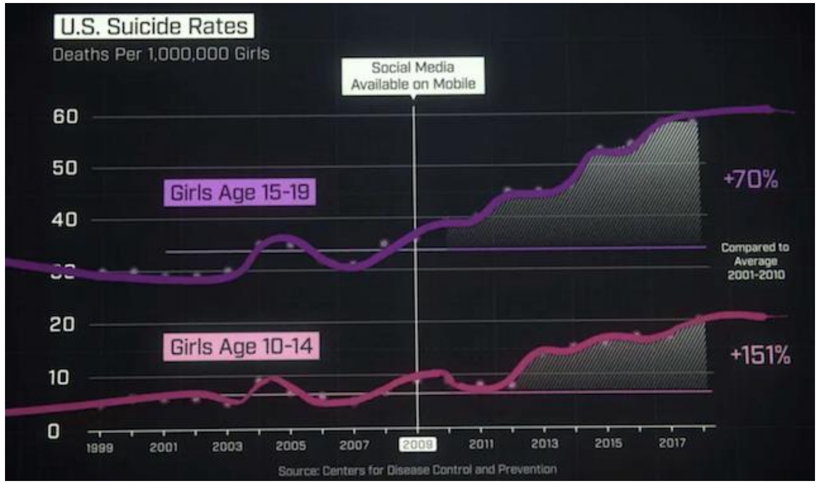
- Push notifications
  - Notify disengaged users to return to the app
  - New content, new social interactions (e.g., likes), …

Legal and even expected standard functions
(Missing this may lead to dissatisfied users!)

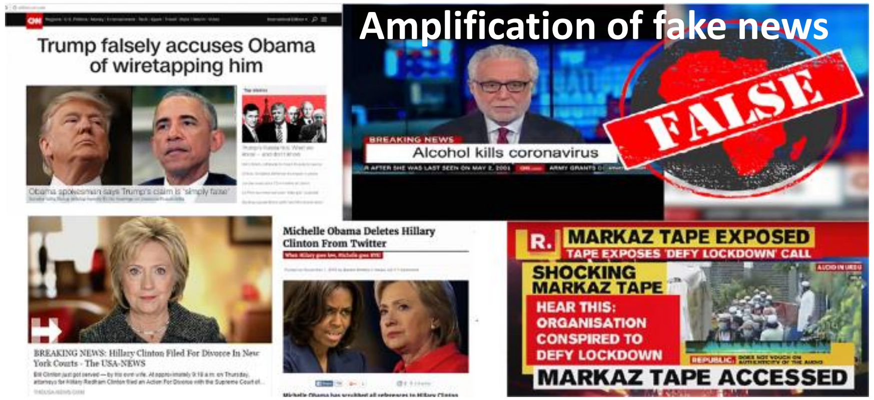
Negative side effects



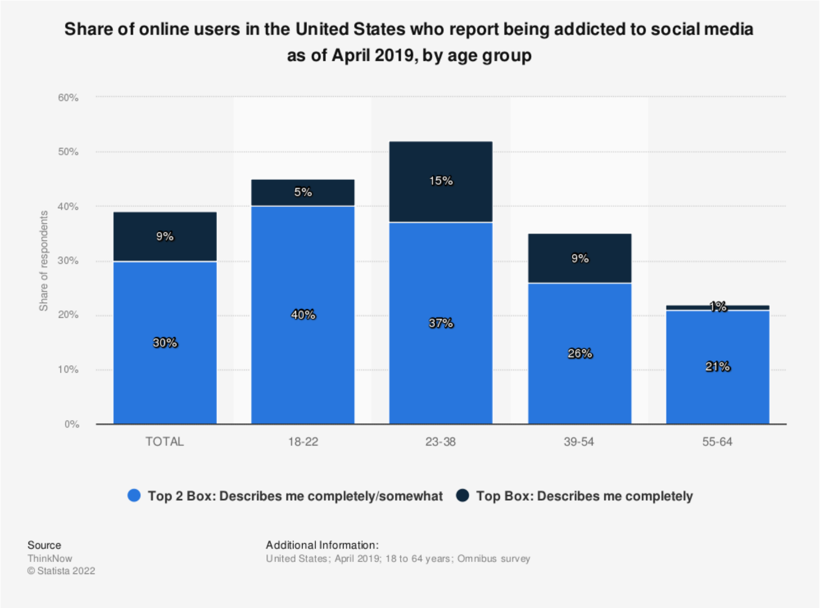
https://www.cdc.gov/mmwr/volumes/69/su/su6901a9.htm?s_cid=su6901a9_w



https://leftronic.com/blog/social-media-addiction-statistics/



Facebook / https://doi.org/10.1007/s11042-020-10183-2



So is this ethical?

Challenges

- Misalignment between organizational goals and societal values
  - Financial incentives often dominate other goals

- Insufficient amount of regulations
  - Little legal consequences for causing negative impact (there are exceptions!)
  - Poor understanding of social-technical systems by policy makers

Not new, but exacerbated by use of ML!

- Engineering challenges, both at system and ML level
  - Difficult to clearly define or measure ethical values
  - Difficult to predict possible usage contexts
  - Difficult to predict impact of feedback loops
  - Difficult to prevent malicious actors from abusing the system
  - Difficult to interpret output of ML and make ethical decision
  - …

Fairness

Definition

fair adjective

Save Word

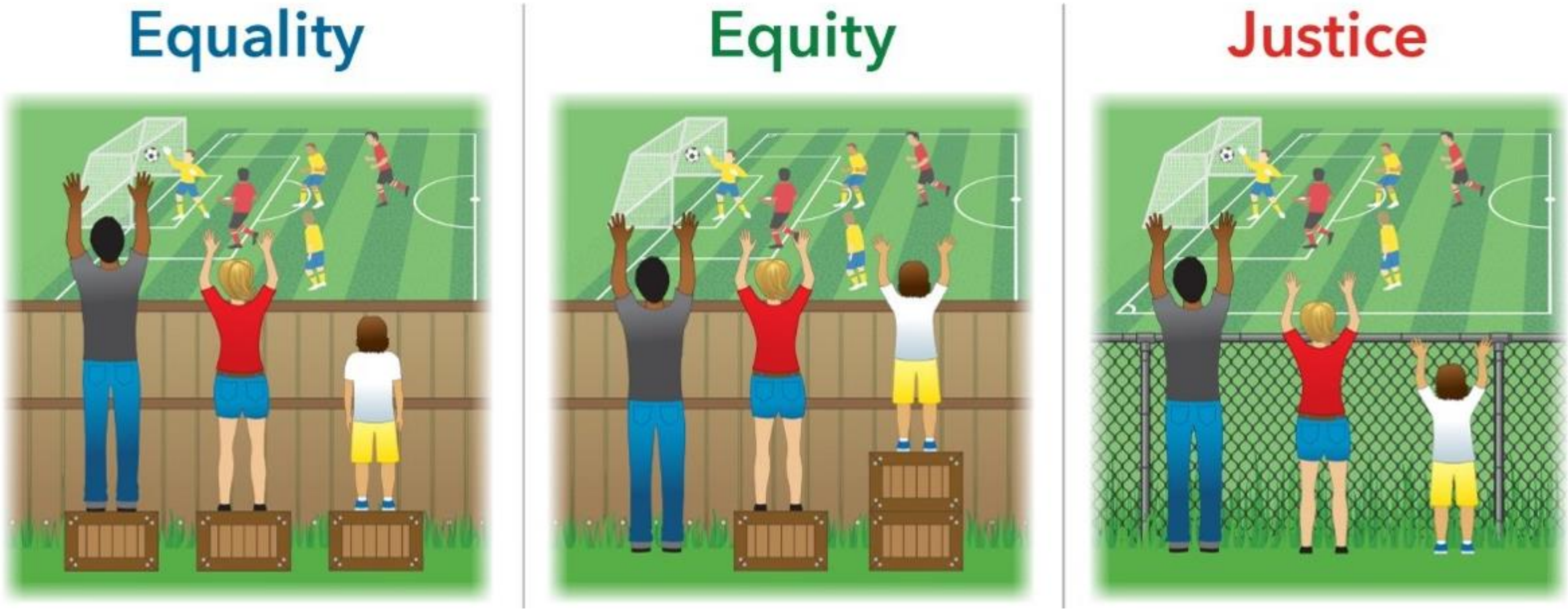
\ ˈfer \

Definition of *fair* (Entry 1 of 5)

1 **a** **:** marked by impartiality and honesty **:** free from self-interest, prejudice, or favoritism

https://www.merriam-webster.com/dictionary/fair

But which one is fair?!



Evenly distributed tools and assistance

Custom tools that identify and address inequality

Fixing the system to offer equal access to both tools and opportunities

… more on defining fairness later

Legal fairness protections

German Basic Law (Grundgesetz)

Article 3
[Equality before the law]

(1) All persons shall be equal before the law.

(2) Men and women shall have equal rights. The state shall promote the actual implementation of equal rights for women and men and take steps to eliminate disadvantages that now exist.

(3) No person shall be favoured or disfavoured because of sex, parentage, race, language, homeland and origin, faith or religious or political opinions. No person shall be disfavoured because of disability.

https://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html#p0026

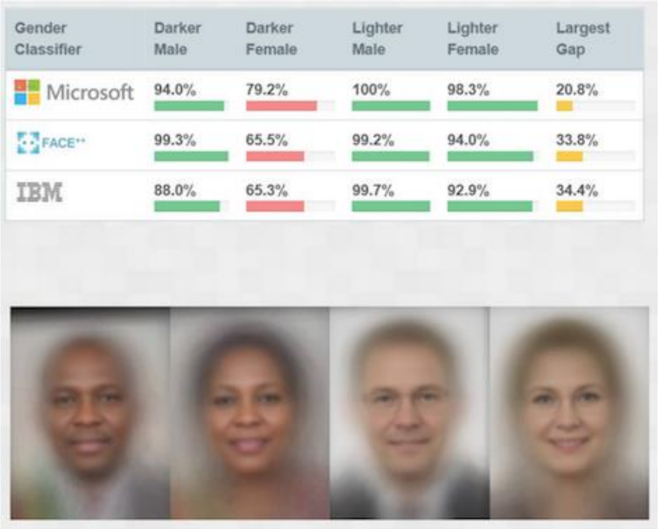
USA based on various laws (e.g., Civil Rights Act, Equal Pay Act):

Race, color, sex, religion, national origin, citizenship, age, pregnancy, familial status, disability status, veteran status, genetic information

Sometimes regulated by specific laws and agencies: credit scoring, education, employment, …

Types of harm on society due to unfairness

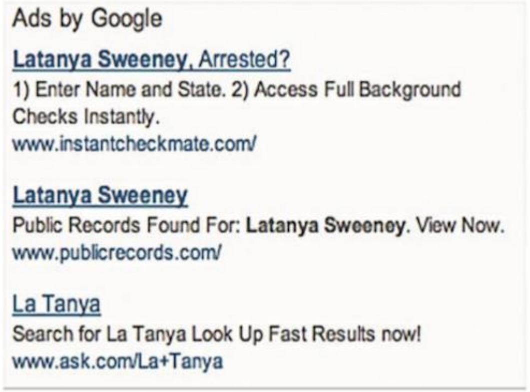
- Harms of allocation
  - Withhold opportunities or resources



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%

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Buolamwini & Gebru, ACM FAT* (2018).

- Harms of representation
  - Reinforce stereotypes, subordination along the lines of identity



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

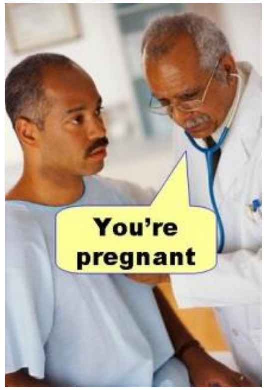
Identifying Harms

	Allocation of resources	Quality of Service	Stereotyping	Denigration	Over- / Under-Representation
Hiring system does not rank women as highly as men for technical jobs	x	x	x		x
Photo management program labels image of black people as “gorillas”		x		x	
Image searches for “CEO” yield only photos of white men on first page			x		x

Products can cause multiple harms ⟶ Identify harms while considering system objectives!

Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial (2019).

Not all discrimination is harmful



Medical diagnosis should take sex into account

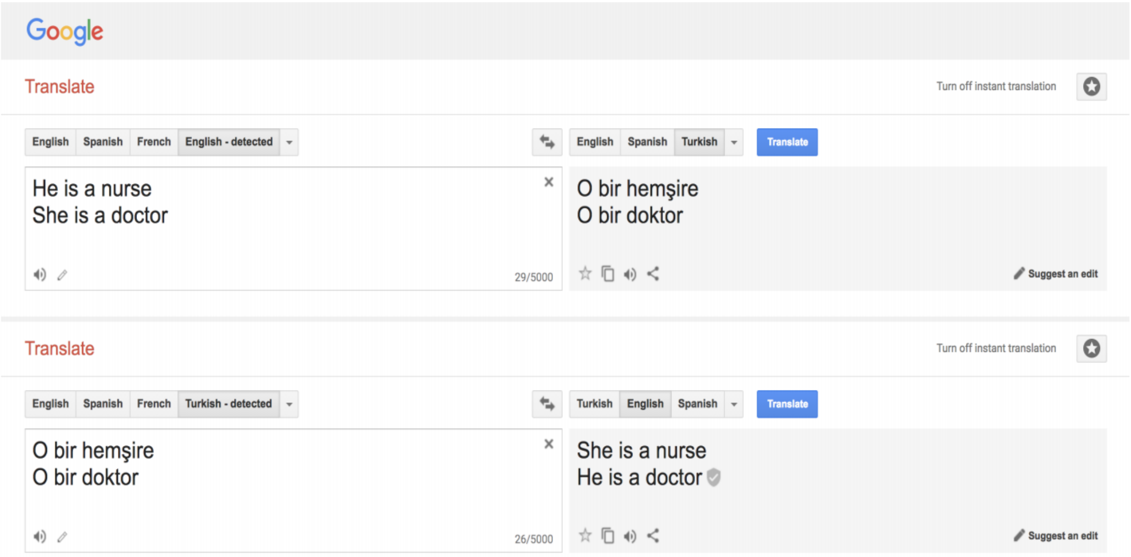
The problem is unjustified discrimination, i.e., by factors that should not matter

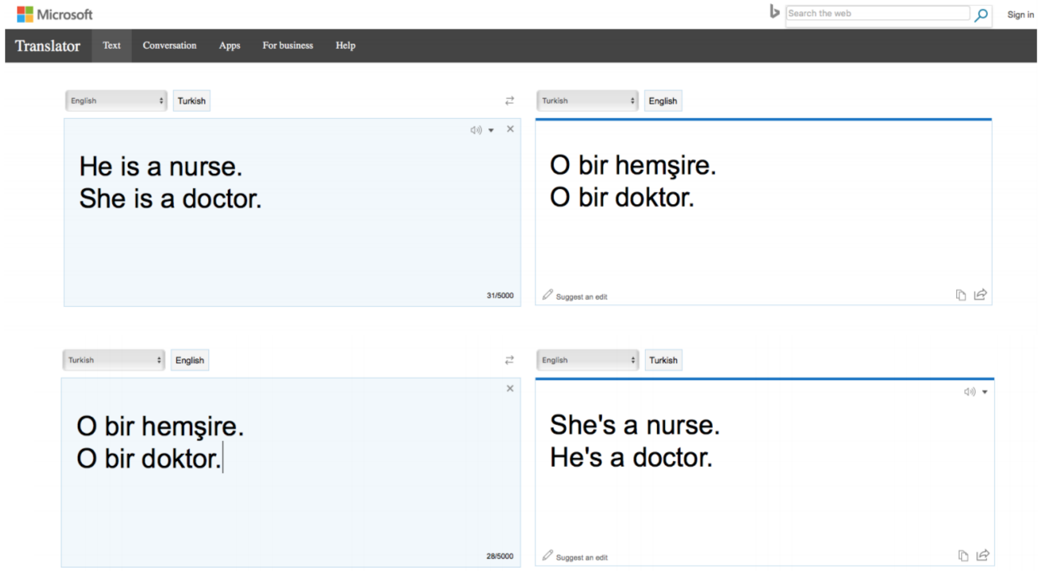
Discrimination is a domain-specific concept and must be understood in the context of the problem domain!

Live exercise: What are other examples where discrimination is not harmful?

Sources of bias

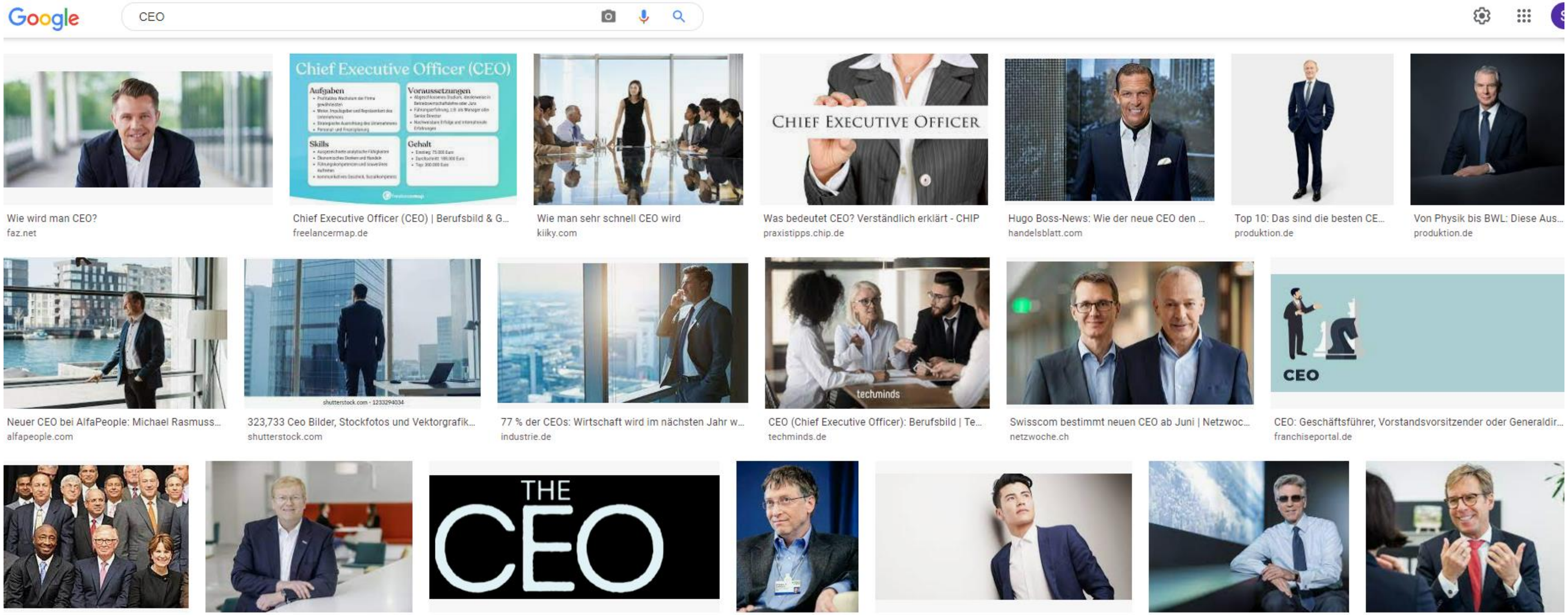
Bias is prevalent





Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science (2017).

Historical bias



Data reflects past biases, not intended outcomes!

Tainted examples

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

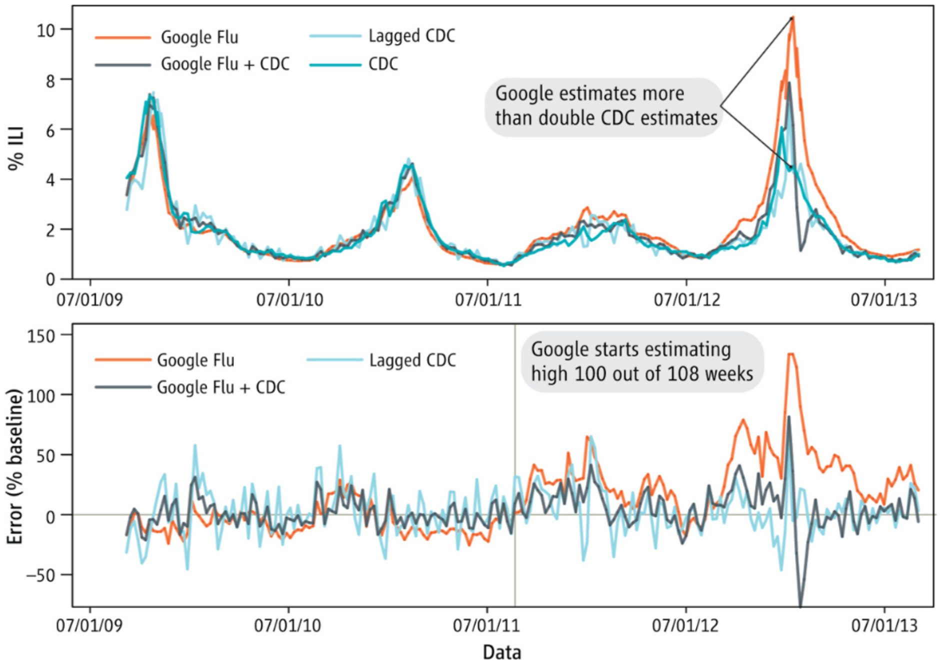
By Jeffrey Dastin 8 MIN READ

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

Data tainted by biased examples, e.g., past hiring decisions which were biased in favor of man

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

Skewed samples



GFT overestimation.GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. “Lagged CDC” incorporates 52-week seasonality variables with lagged CDC data. “Google Flu + CDC” combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Skewed samples may lead to overestimations

Flu model overestimates amount of flu, because it cannot distinguish between flu and flu-like illnesses

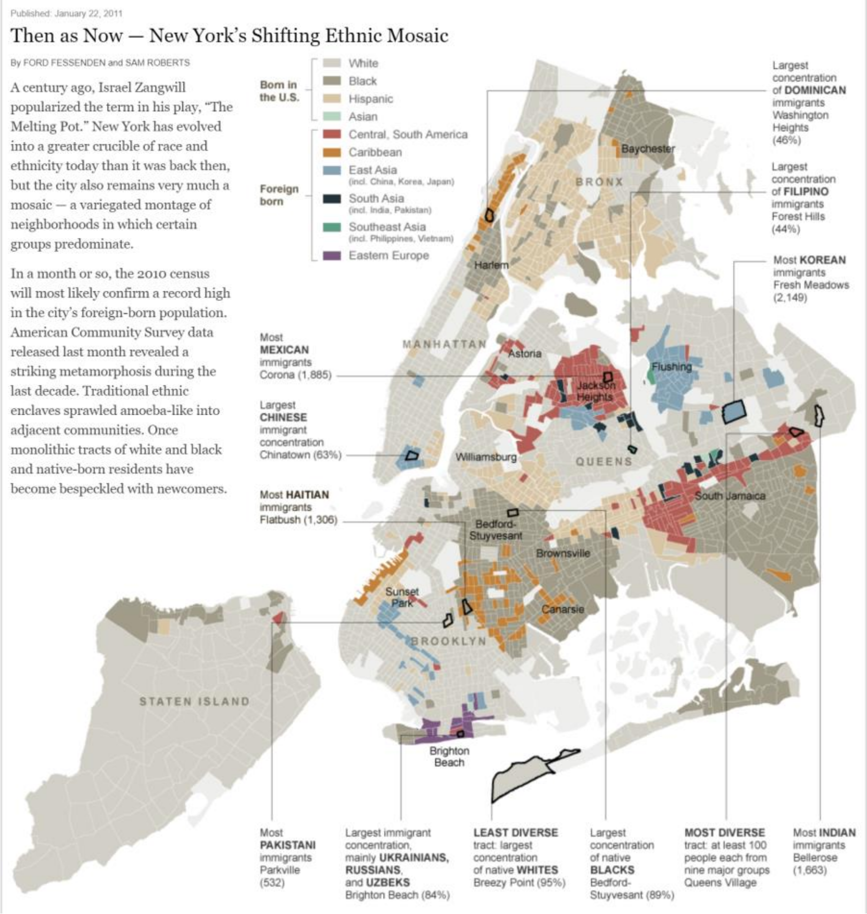
Lazer, David, et al. "The parable of Google Flu: traps in big data analysis." *Science* 343.6176 (2014): 1203-1205.

Limited features

- Features may not be equally reliable for all parts of the data
- Could lead to performance degradation for minorities
- Example:
  - Car insurance based on average accidents for age group and region
  - Penalizes uncommonly safe drivers

https://www.quoteinspector.com/images/car-insurance/car-insurance-coverage-drivers/

Proxy



Neighborhood as proxy for race

http://archive.nytimes.com/www.nytimes.com/interactive/2011/01/23/nyregion/20110123-nyc-ethnic-neighborhoods-map.html?_r=1

Sample size diversity

Used by Kodak to calibrate early color films. Very small sample without diversity.

Fairness definitions

Anti-classification

- Also known as *fairness through blindness*
- Concept
  - Ignore sensitive feature when making a decision
- Examples
  - Remove gender and race from credit scoring model
- Limitations
  - Sensitive attributes may be correlated with other features
  - Some ML tasks need sensitive attributes (e.g., medical diagnosis)

https://de.wikipedia.org/wiki/Rechtspflege#/media/Datei:Berner_Iustitia.jpg

Testing anti-classification

- Can be defined as invariant
  - $f(x[y = z]) = f(x[y = z'])$ for a classifier f where z, z' are arbitrary valid values for the protected feature y
  - Example: $f(x[gender = male]) = f(x[gender = female])$ to test anti-classification for gender

- Any inconsistency shows that the feature was used
  - This could also happen indirectly, through correlations!

- Reporting the rate of inconsistencies can help estimate the impact

$x[y = z]$ denotes that the feature y of the instance x is set to z without changing any other feature
$x.y$ denotes the value of feature y of the instance x

See for example: Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination." In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, pp. 498-510. 2017.

Group fairness

- Also called *independence* or *demographic parity*

- Can be defined as a probabilistic invariant
  - $P(f(x) = 1|y = z) = P(f(x) = 1|y = z')$ for a classifier f where z, z' are arbitrary valid values for the protected feature y
  - Example: $\mathrm{P}(creditscore = positive|gender = male) = P(creditscore = positive|gender = female)$ to test group fairness of gender

- Similar to anti-classification, but with probabilities
  - Results do not need to be the same for each instance, but no differences for the complete group

- Can be achieved without actually considering the objective
  - Could, e.g., assign positive credit rating randomly to match rate across groups

Testing group fairness

- Consider results on different slices (e.g., male, female) of test data or even production data
    - Alternatively: generate new test data according to the distribution of the protected classes

- Separately measure the performance
    - E.g., rate of positive predictions

- Define threshold of allowed deviations
    - If the measured performance deviates by more than ϵ between groups raise alarm

Separation

- Also called *equalized odds*
- For a classifier f with target function f^* (true value) where z, z' are arbitrary valid values for the protected feature y, the following two properties must hold:
  - False positive rate parity: $P(f(x) = 1 | f^*(x) = 0, y = z) = P(f(x) = 1 | f^*(x) = 0, y = z')$
  - False negative rate parity: $P(f(x) = 0 | f^*(x) = 1, y = z) = P(f(x) = 0 | f^*(x) = 1, y = z')$
- All groups have the same error rate for each class
- Example:
  - Both genders have the equal likelihood of being incorrectly denied a credit
  - Both genders have the equal likelihood of being incorrectly awarded a credit

Testing separation

- Consider results on different slices (e.g., male, female) of test data or even production data
  - Alternatively: generate new test data according to the distribution of the protected classes
- Separately measure the false positive and false negative rates
  - E.g., rate of positive predictions
- Define threshold of allowed deviations
  - If a measured rate deviates by more than ϵ between groups raise alarm
- Similar to testing group fairness, but with two specific criteria that both need to be fulfilled

Live exercise: Is this cancer classifier fair?

Overall Results

True positives (TPs): 16	False positives (FPs): 21
False negatives (FNs): 9	True negatives (TNs): 954

Male Patient Results

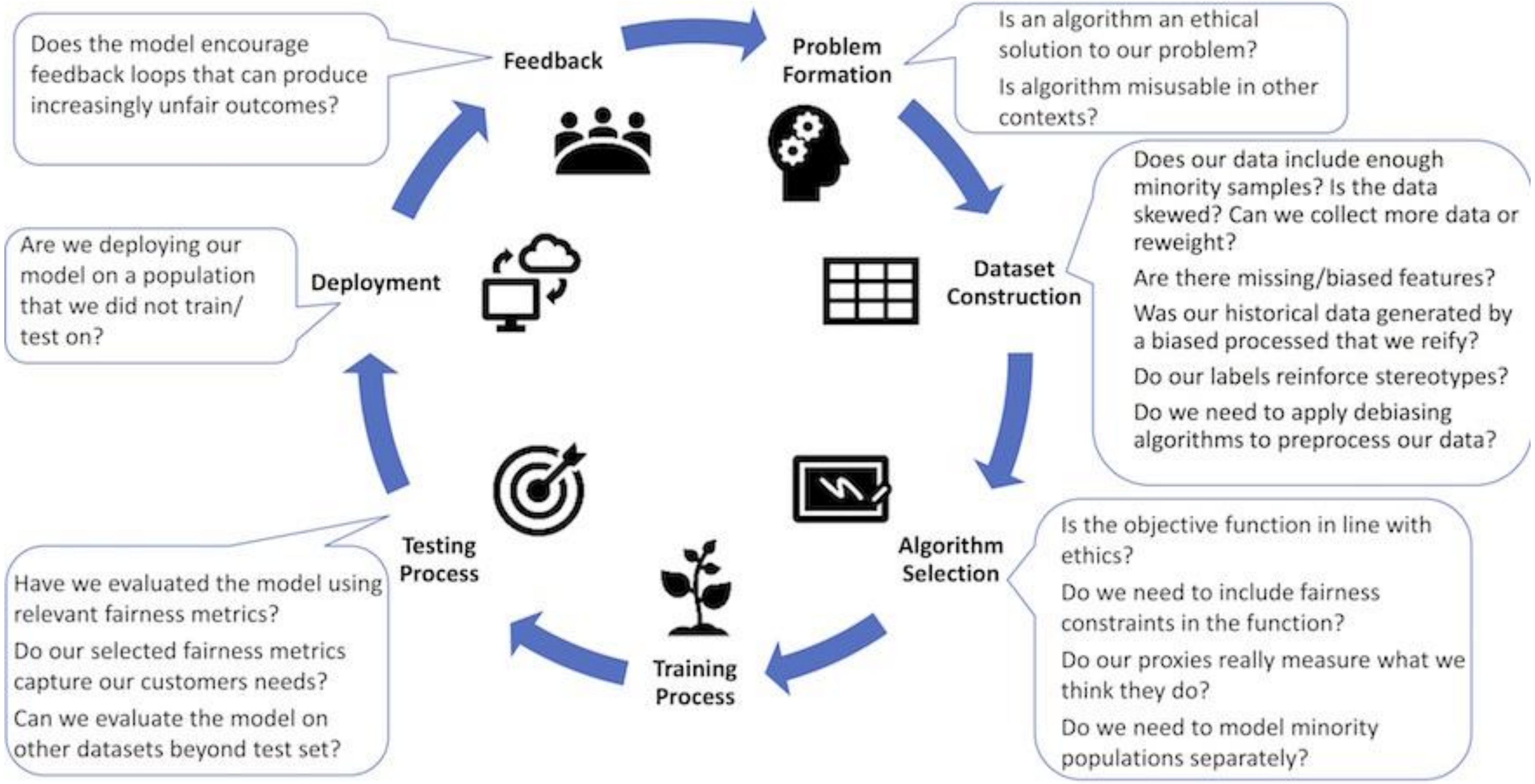
True positives (TPs): 3	False positives (FPs): 16
False negatives (FNs): 7	True negatives (TNs): 474

Female Patient Results

True positives (TPs): 13	False positives (FPs): 5
False negatives (FNs): 2	True negatives (TNs): 480

Achieving fairness

Fairness throughout the lifecycle



Fairness-aware Machine Learning, Bennett et al., WSDM Tutorial (2019).

Challenges for achieving fair systems

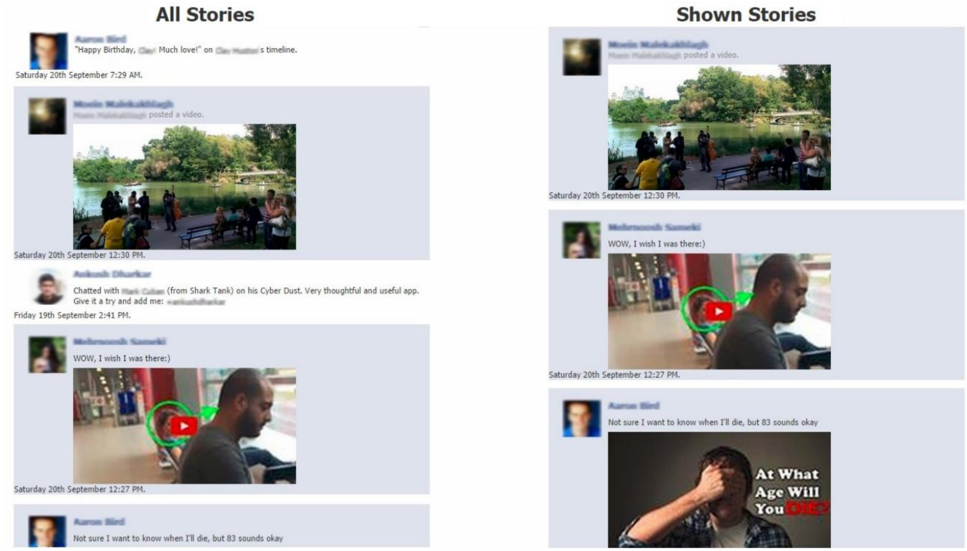
- Fairness is a system-level property
  - Consider goals, user interaction design, data collection, monitoring, model interaction, …
- Fairness-aware data collection and fairness testing for training data
- Identifying blind spots
  - Proactive vs. reactive
  - Team bias and checklists
- Fairness auditing processes and tools
- Diagnosis and debugging
  - Outliers? Systemic problems? Causes?
- Guiding interventions
  - Adjust goals? More data? Better data? Side effects? Redesign?
- Assessing bias of humans in the loop

All of this costs money, which means we need (strong!) incentives for fairness!

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "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. 2019.

Transparency

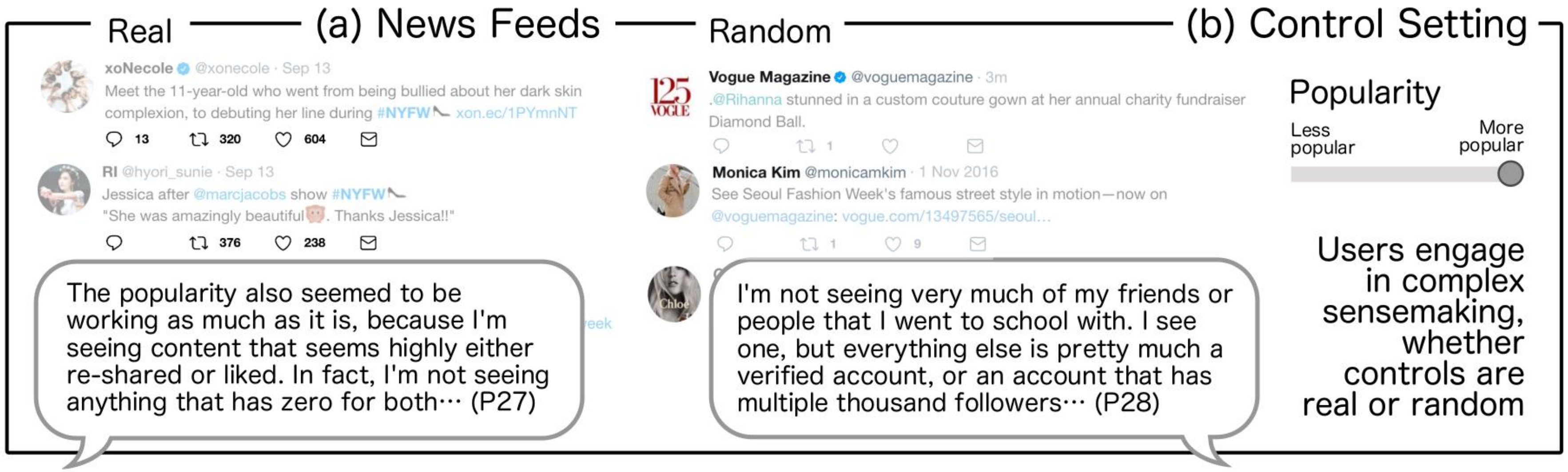
Example: Facebook‘s feed curation



- 62% of people interviewed were not aware that there is a curation algorithm
- Many were surprised and angry when learning about this
  - “Participants were most upset when close friends and family were not shown in their feeds […] participants often attributed mussing stories to their friends‘ decisions to exclude them rather to Facebook News Feed algorithm”
- Learning about the algorithm did not change the satisfaction level
- More active engagement and more feeling of control desired

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 153-162. ACM, 2015.

Useful transparency is difficult



- User may feel influence and control, even if controls are only placebo controls (no real effect)
- Companies often only give vague or generic explanations to appease regulators

Vaccaro, Kristen, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. "The illusion of control: Placebo effects of control settings." In Proc CHI, 2018.

Level of transparency

- Transparency has side effects
    - Intellectual property
    - Trade secrets
    - Fairness
    - Perceptions
    - Ethics
    - Privacy
    - …

- How to determine the level of transparency? How to design the system? How much control should be given?

Live exercise:
Attacking models through explanations

- Would a detailed explanation of how a results was achieved help to hack the model?
  - Loan applications
  - Unlocking mobile phones with a face scan
  - Automatic grading
  - Cancer diagnosis
  - Spam detection

Weak proxies as problem

- Attackable models often use weak proxy features
- Protections require to make the model hard to observe (e.g., expensive to query)
    - Similar to security by obscurity
- Transparency is the opposite!

Human oversight and appeals

- Unavoidable that ML models will make mistakes

- Informing users about this may not comfort them
  - Imagine a bank telling you that they may falsely reject the loan for your dream house with 5% probability

- Often not possible to appeal

Can humans in the loop help?

- … if ML is used because human decisions are the bottleneck that should be reduced/removed?

- … if ML is used because human decisions are biased and inconsistent?

- … if ML is used because the data is extremely complex and hard to understand for humans?

Human in the loop may cause more problems than it solves!

Designing human oversight

- Consider entire system and consequences of mistakes
- Deliberately design mitigation strategies for handling mistakes
- Consider keeping humans in the loop while balancing harms and costs
    - Determine possible pathways for appeals and complaints, as well as how to respond
    - Determine possibility to review ML decisions and how they may be overridden by humans
    - Track telemetry data to enable investigation of (common and uncommon) mistakes
    - Consider if auditing models and decision process is a better choice than an appeals process

Call for transparent and audited models

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › perspectives › article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

No black box should be deployed when there exists an interpretable model with the same level of performance

- High stakes decisions with government involvement (e.g., recidivism, policing, city planning)
- High stakes decisions in medicine (e.g., treatments)
- High stakes decisions with discrimination concerns (e.g., hiring, loans, housing)
- Decisions that influence society and discourse (e.g., content curation, targeted advertisement)

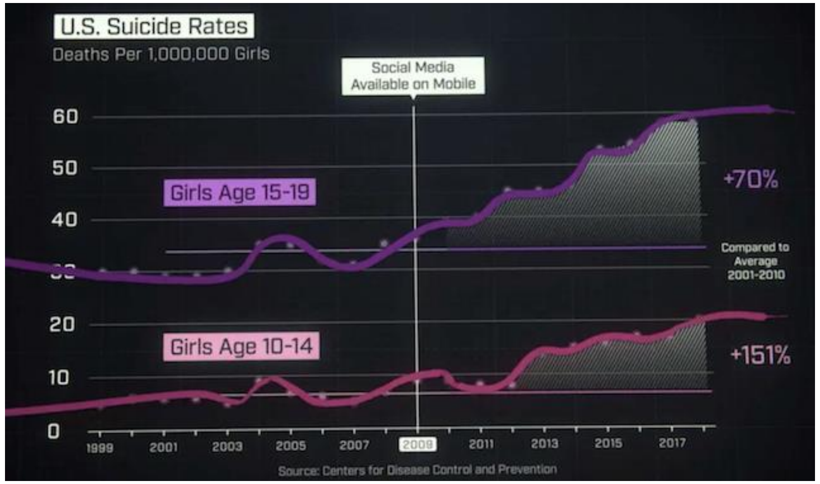
Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.

Accountability

Terminology

- Accountability, responsibility, liability, and culpability all overlap in common use
  - All about assigning blame and responsibility for fixing something or paying for damages
- Liability and culpability have a legal connotation
  - You may be *liable* for damages because you are held *culpable* for a problem
- Accountability and responsibility rather tend to descript ethical aspirations
- Similar to “legal vs. ethical” discussed at the beginning of this chapter

Who is responsible?



https://leftronic.com/blog/social-media-addiction-statistics/

https://www.theverge.com/2021/10/14/22726111/robot-dogs-with-guns-sword-international-ghost-robotics

The Washington Post

Democracy Dies in Darkness

PostEverything • Perspective

How U.S. surveillance technology is propping up authoritarian regimes

(iStock)

By Robert Morgus and Justin Sherman

Jan. 17, 2019 at 6:00 a.m. EST

How software engineers usually handle this …

15. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

GPL 3.0

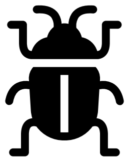
```
7. Disclaimer of Warranty. Unless required by applicable law or
   agreed to in writing, Licensor provides the Work (and each
   Contributor provides its Contributions) on an "AS IS" BASIS,
   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
   implied, including, without limitation, any warranties or conditions
   of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A
   PARTICULAR PURPOSE. You are solely responsible for determining the
   appropriateness of using or redistributing the Work and assume any
   risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory,
   whether in tort (including negligence), contract, or otherwise,
   unless required by applicable law (such as deliberate and grossly
   negligent acts) or agreed to in writing, shall any Contributor be
   liable to You for damages, including any direct, indirect, special,
   incidental, or consequential damages of any character arising as a
   result of this License or out of the use or inability to use the
   Work (including but not limited to damages for loss of goodwill,
   work stoppage, computer failure or malfunction, or any and all
   other commercial damages or losses), even if such Contributor
   has been advised of the possibility of such damages.
```

Apache License 2.0

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

MIT License

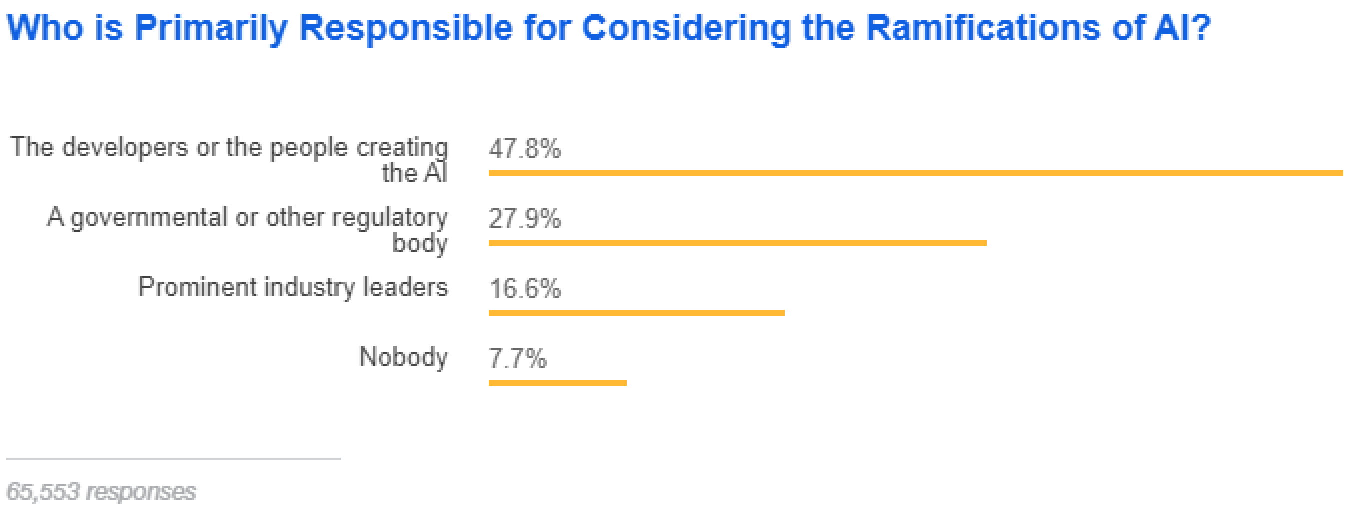
Easy to blame “The Algorithm” / “The Data” / “Software



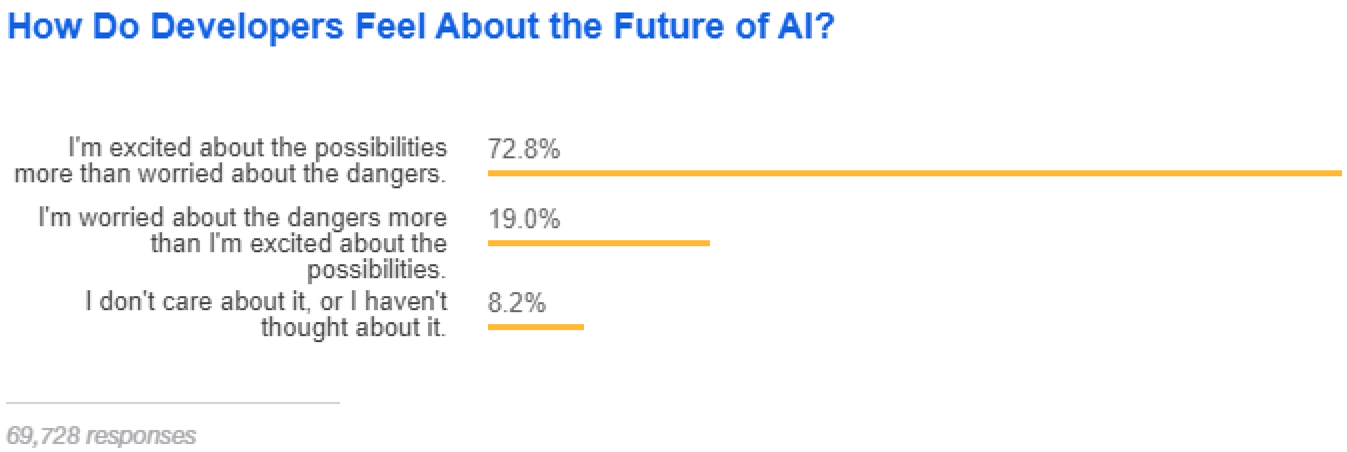
It‘s just a bug! Such things happen and we can do nothing about it.

- But the system was designed by humans
- Humans did not anticipate possible mistakes and did not design mitigations
    - (or they did not care about them…)
- Humans made decisions about what level of quality assurance is sufficient
- Humans designed (or ignored) the process for developing the software
- Humans gave/sold poor quality software to other humans
- Humans used software without understanding it
- …

Developers think that they are responsible …



… but still excited!



https://insights.stackoverflow.com/survey/2018/#technology-and-society

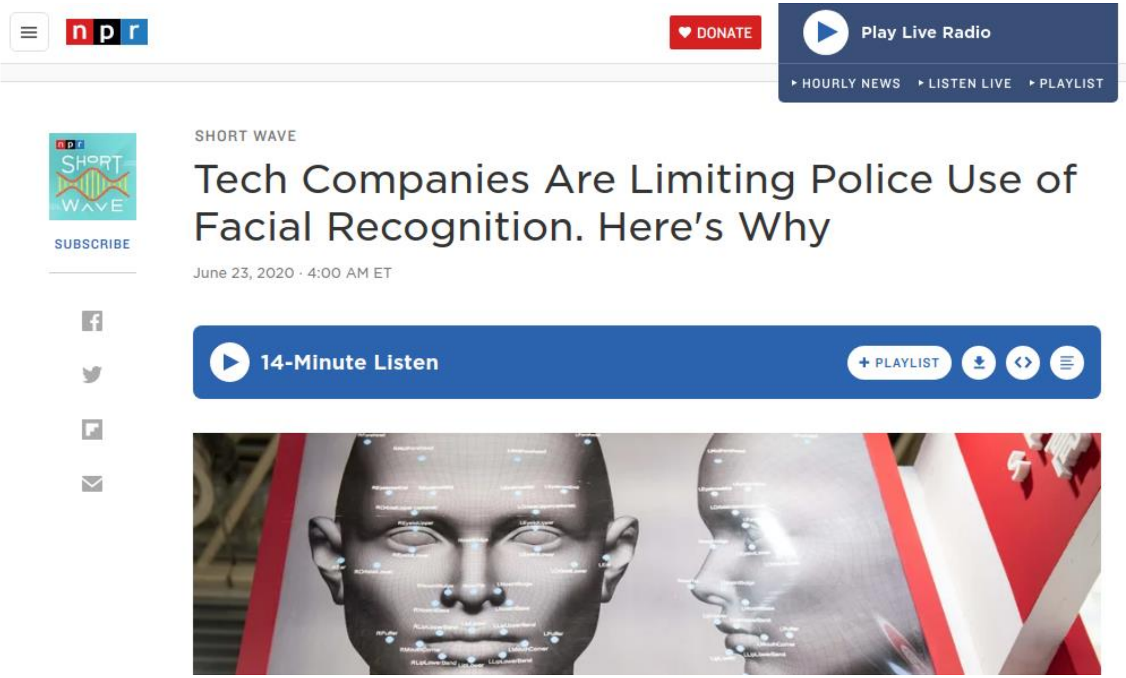
What to do?

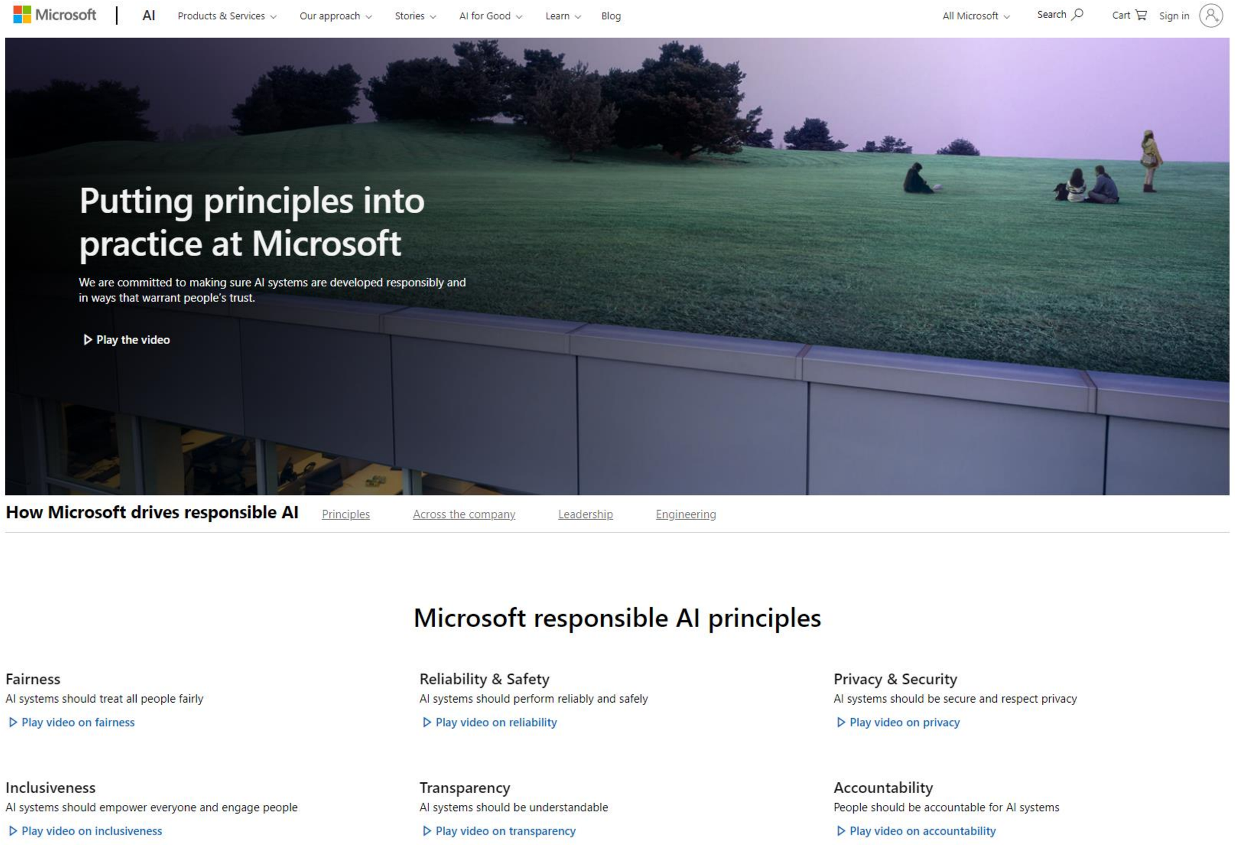
- Responsible organizations embed risk analysis, quality control, and ethical considerations into their process
- Establish and communicate policies defining responsibilities
- Work from aspirations toward culture change
  - Baseline awareness for everyone supported by experts
- Document tradeoffs and decisions
- Consider controlling/restricting how software may be used

- And follow the law …

- Possible starting point:
  - https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/

Self-regulation in practice





https://www.npr.org/2021/02/17/968710172/why-tech-companies-are-limiting-police-use-of-facial-recognition?t=1658309672028

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

Government regulation

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

European Parliament

Regulating facial recognition in the EU

(on the other hand, China requires state control over valuable data…)

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA(2021)698021_EN.pdf

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Questions?