TRANSPARENCY AND ACCOUNTABILITY

Christian Kaestner

Required reading: Google PAIR. People + AI Guidebook. Chapter: Explainability and Trust. 2019.

Recommended supplementary reading: Christoph Molnar. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2019

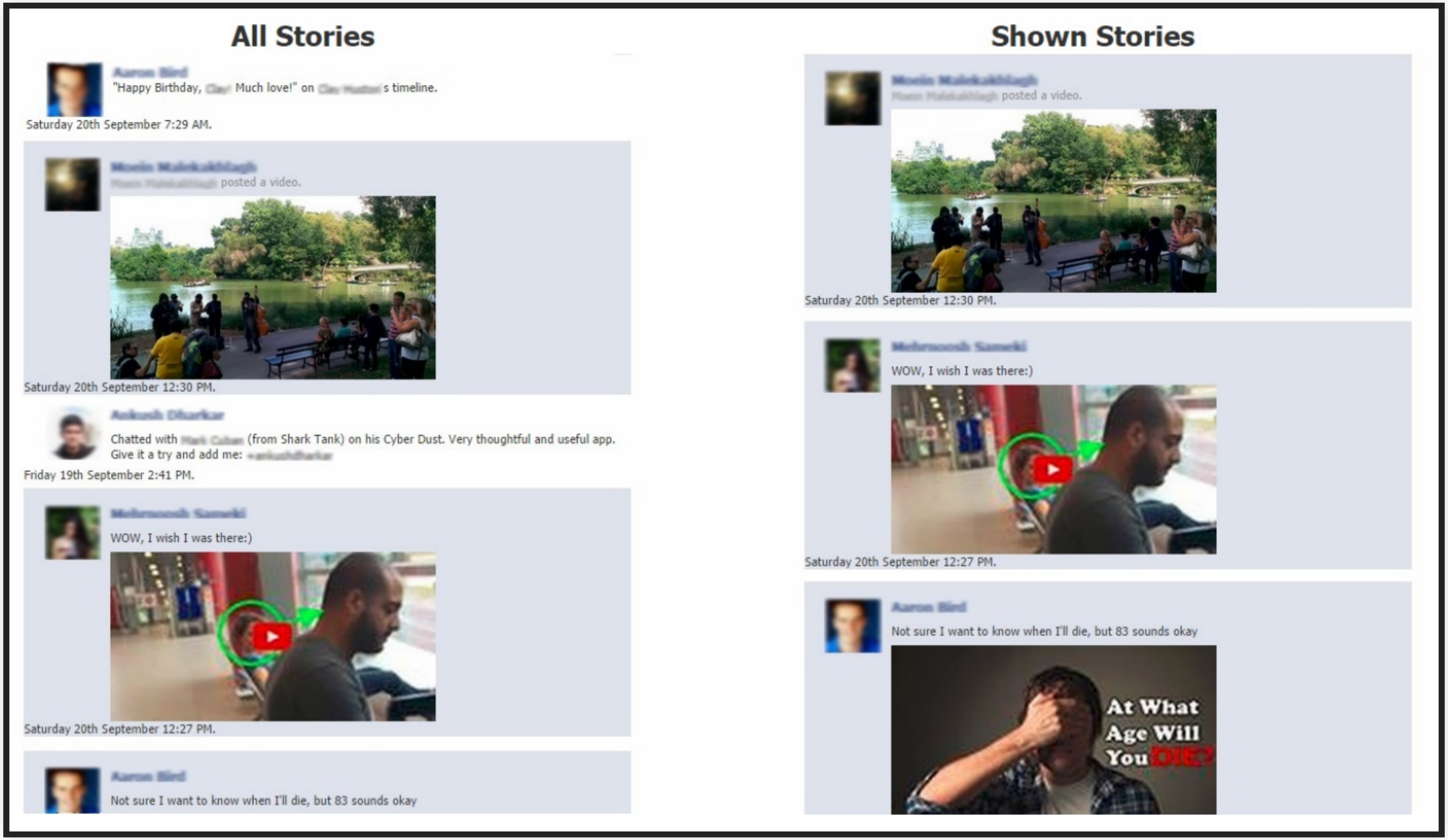
LEARNING GOALS

- Explain key concepts of transparency and trust
- Discuss whether and when transparency can be abused to game the system
- Design a system to include human oversight
- Understand common concepts and discussions of accountability/culpability
- Critique regulation and self-regulation approaches in ethical machine learning

TRANSPARENCY

(users know that algorithm exists / users know how the algorithm works)

CASE STUDY: FACEBOOK'S FEED CURATION



All Stories

[redacted]
"Happy Birthday, [redacted] Much love!" on [redacted]s timeline.
Saturday 20th September 7:29 AM.

[redacted]
[redacted] posted a video.
Saturday 20th September 12:30 PM.

[redacted]
Chatted with [redacted] (from Shark Tank) on his Cyber Dust. Very thoughtful and useful app.
Give it a try and add me: [redacted]
Friday 19th September 2:41 PM.

[redacted]
WOW, I wish I was there:)
Saturday 20th September 12:27 PM.

[redacted]
Not sure I want to know when I'll die, but 83 sounds okay

Shown Stories

[redacted]
[redacted] posted a video.
Saturday 20th September 12:30 PM.

[redacted]
WOW, I wish I was there:)
Saturday 20th September 12:27 PM.

[redacted]
Not sure I want to know when I'll die, but 83 sounds okay

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.

CASE STUDY: FACEBOOK'S FEED CURATION

- 62% of interviewees were not aware of curation algorithm
- Surprise and anger when learning about curation

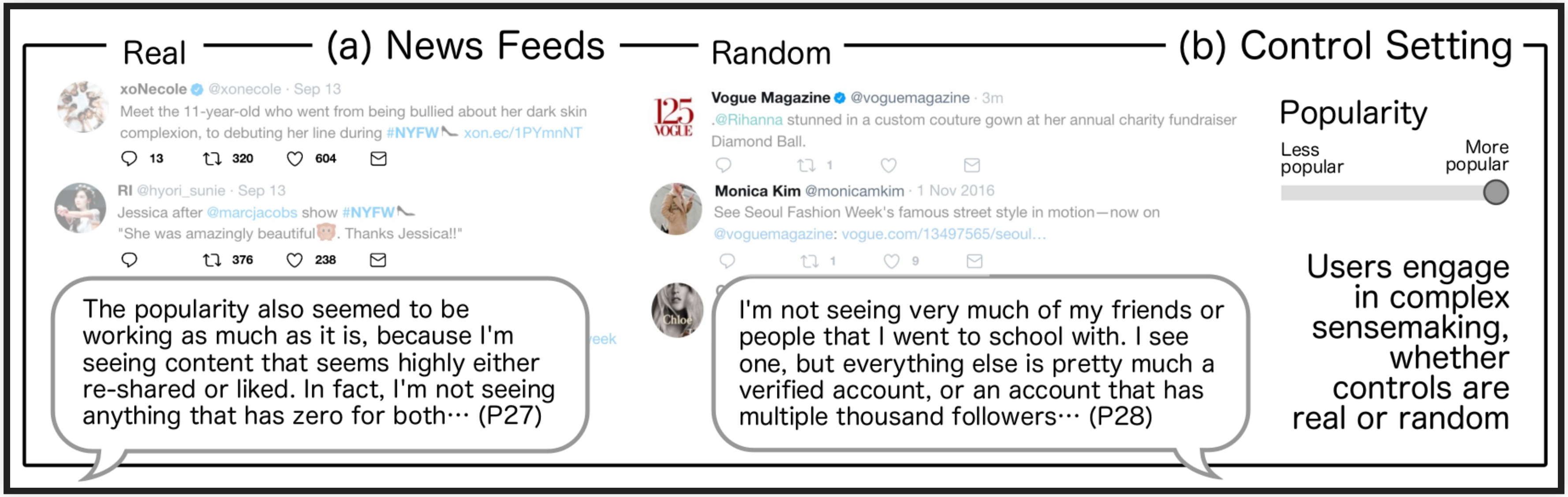
> *"Participants were most upset when close friends and family were not shown in their feeds [...] participants often attributed missing stories to their friends' decisions to exclude them rather than to Facebook News Feed algorithm."*

- Learning about algorithm did not change satisfaction level
- More active engagement, more feeling of control

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.

THE DARK SIDE OF TRANSPARENCY

- Users may feel influence and control, even with placebo controls
- Companies give vague generic explanations to appease regulators



xoNecole @xonecole · Sep 13
Meet the 11-year-old who went from being bullied about her dark skin complexion, to debuting her line during #NYFW xon.ec/1PYmnNT
13 320 604

RI @hyori_sunie · Sep 13
Jessica after @marcjacobs show #NYFW
"She was amazingly beautiful. Thanks Jessica!!"
376 238

Vogue Magazine @voguemagazine · 3m
.@Rihanna stunned in a custom couture gown at her annual charity fundraiser Diamond Ball.
1

Monica Kim @monicamkim · 1 Nov 2016
See Seoul Fashion Week's famous street style in motion—now on @voguemagazine: vogue.com/13497565/seoul...
1 9

The popularity also seemed to be working as much as it is, because I'm seeing content that seems highly either re-shared or liked. In fact, I'm not seeing anything that has zero for both··· (P27)

I'm not seeing very much of my friends or people that I went to school with. I see one, but everything else is pretty much a verified account, or an account that has multiple thousand followers··· (P28)

Users engage in complex sensemaking, whether controls are real or random

- 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.

APPROPRIATE LEVEL OF ALGORITHMIC TRANSPARENCY

IP/Trade Secrets/Fairness/Perceptions/Ethics?

How to design? How much control to give?

GAMING/ATTACKING THE MODEL WITH EXPLANATIONS?

Does providing an explanation allow customers to 'hack' the system?

- Loan applications?
- Apple FaceID?
- Recidivism?
- Auto grading?
- Cancer diagnosis?
- Spam detection?

GAMING THE MODEL WITH EXPLANATIONS?

GAMING THE MODEL WITH EXPLANATIONS?

- A model prone to gaming uses weak proxy features
- Protections requires to make the model hard to observe (e.g., expensive to query predictions)
- Protecting models akin to "security by obscurity"

- Good models rely on hard facts that are hard to game and relate causally to the outcome

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE
IF age between 21–23 and 2–3 prior offenses THEN predict arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

HUMAN OVERSIGHT AND APPEALS

HUMAN OVERSIGHT AND APPEALS

- Unavoidable that ML models will make mistakes
- Users knowing about the model may not be comforting
- Inability to appeal a decision can be deeply frustrating

CAPACITY TO KEEP HUMANS IN THE LOOP?

- ML used because human decisions as a bottleneck
- ML used because human decisions biased and inconsistent

- Do we have the capacity to handle complaints/appeals?
- Wouldn't reintroducing humans bring back biases and inconsistencies

DESIGNING HUMAN OVERSIGHT

- Consider the entire system and consequences of mistakes
- Deliberately design mitigation strategies for handling mistakes
- Consider keeping humans in the loop, balancing harms and costs
    - Provide pathways to appeal/complain? Respond to complains?
    - Review mechanisms? Can humans override tool decision?
    - Tracking telemetry, investigating common mistakes?
    - Audit model and decision process rather than appeal individual outcomes?

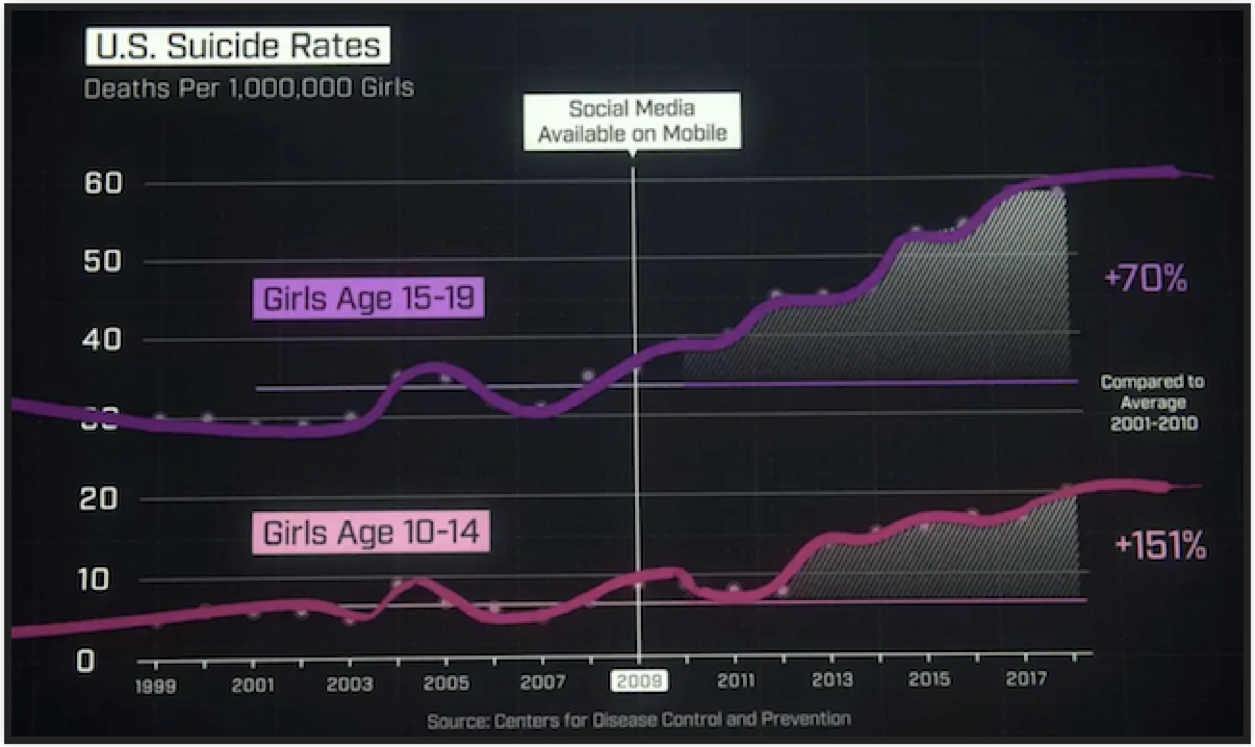
ACCOUNTABILITY AND CULPABILITY

Who is held accountable if things go wrong?

ON TERMINOLOGY

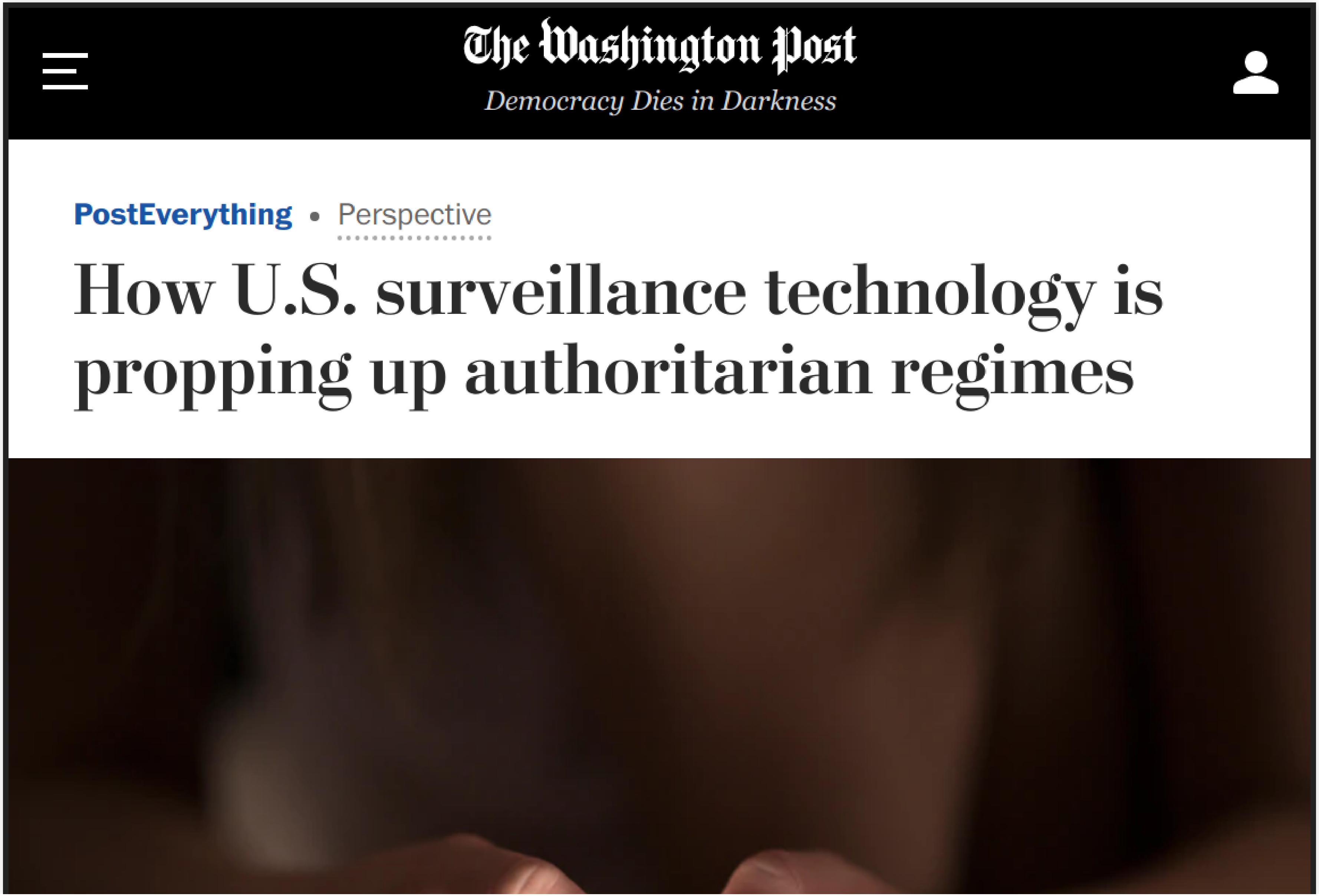
- accountability, responsibility, liability, and culpability all overlap in common use
- all about assigning *blame* -- responsible for fixing or liable for paying for damages
- liability, culpability have *legal* connotation
- accountability, responsibility tend to describe *ethical* aspirations
- see legal vs ethical earlier

WHO IS RESPONSIBLE?



Source: Centers for Disease Control and Prevention

WHO IS RESPONSIBLE?

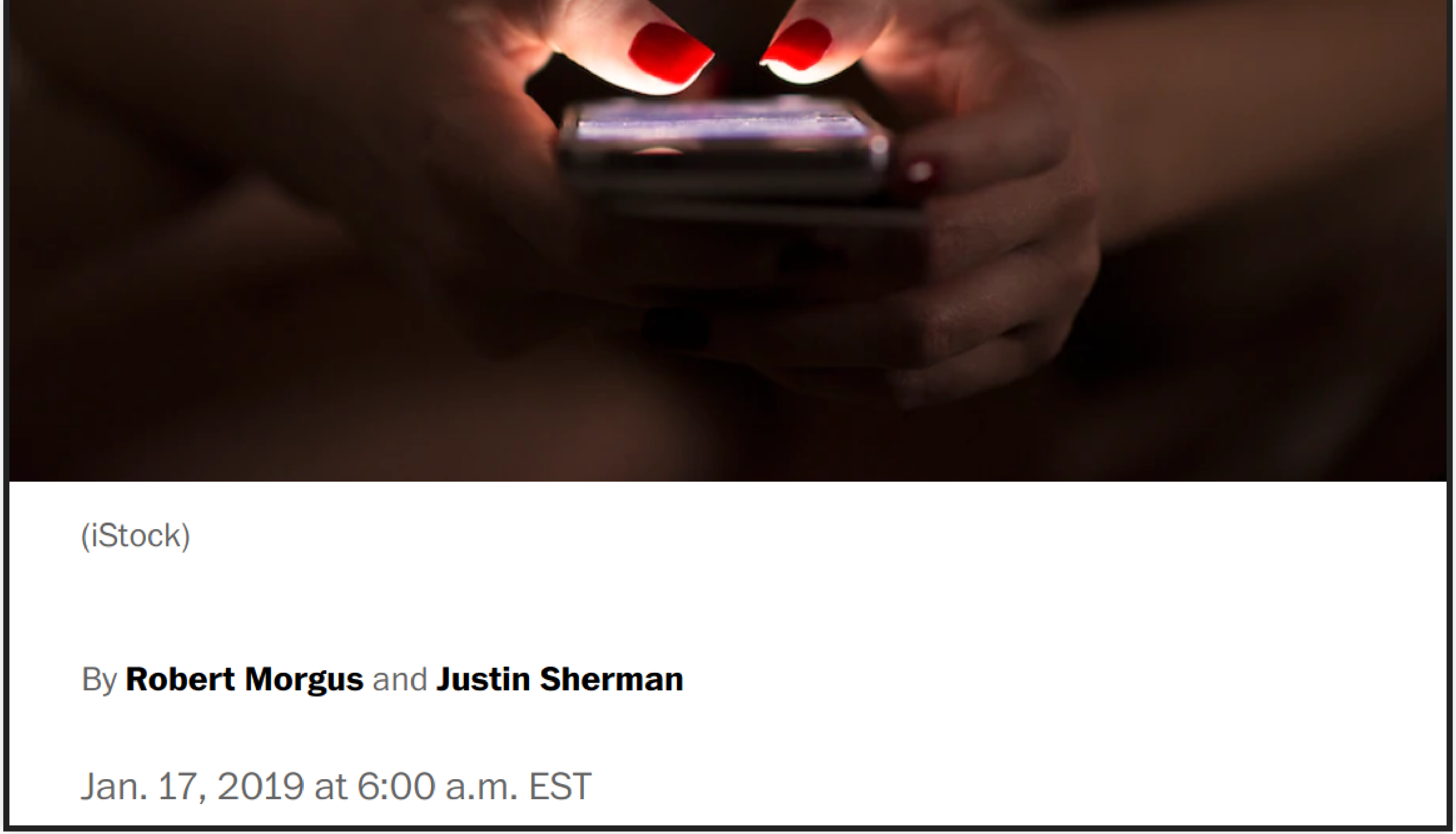


The Washington Post

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

WHO IS RESPONSIBLE?

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.

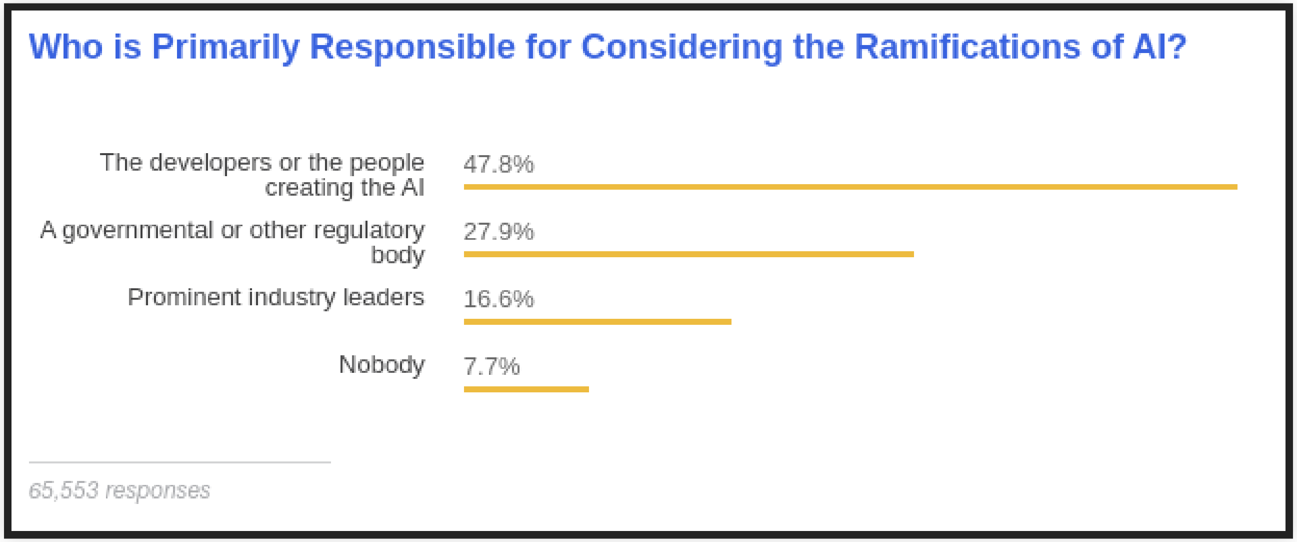
Speaker notes

Software engineers got (mostly) away with declaring not to be liable

EASY TO BLAME "THE ALGORITHM" / "THE DATA" / "SOFTWARE"

> *Just a bug, things happen, nothing we can do about it*

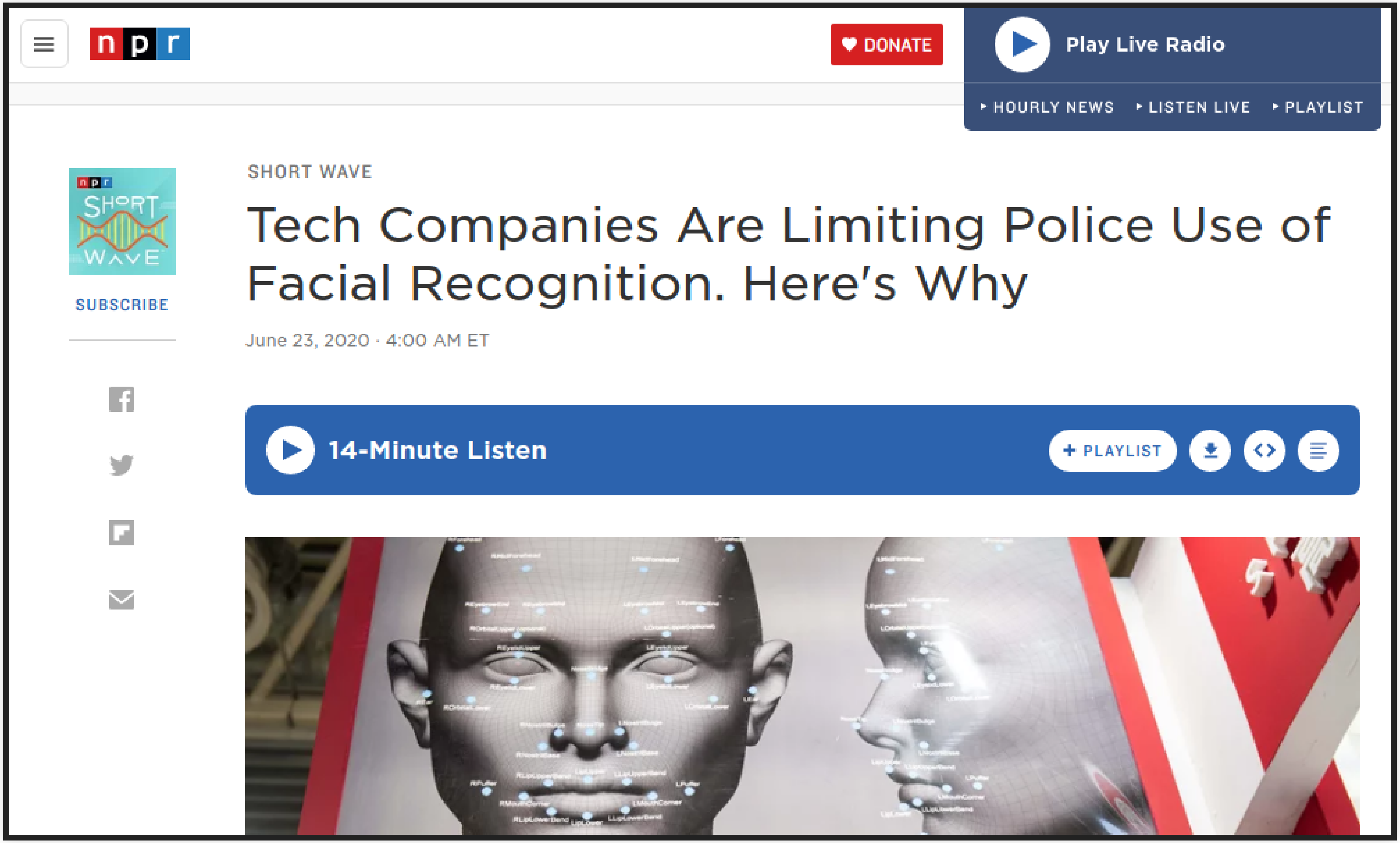
- But system was designed by humans
- Humans did not anticipate possible mistakes, did not design to mitigate mistakes
- Humans made decisions about what quality assurance would be sufficient
- Humans designed (or ignored) the process for developing the software
- Humans gave/sold poor quality software to other humans
- Humans used the software without understanding it
- ...



Results from the 2018 StackOverflow Survey

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 + experts
- Document tradeoffs and decisions (e.g., datasheets, model cards)
- Continuous learning
- Consider controlling/restricting how software may be used
- And... follow the law
- Get started with existing guidelines, e.g., in AI Ethics Guidelines Global Inventory

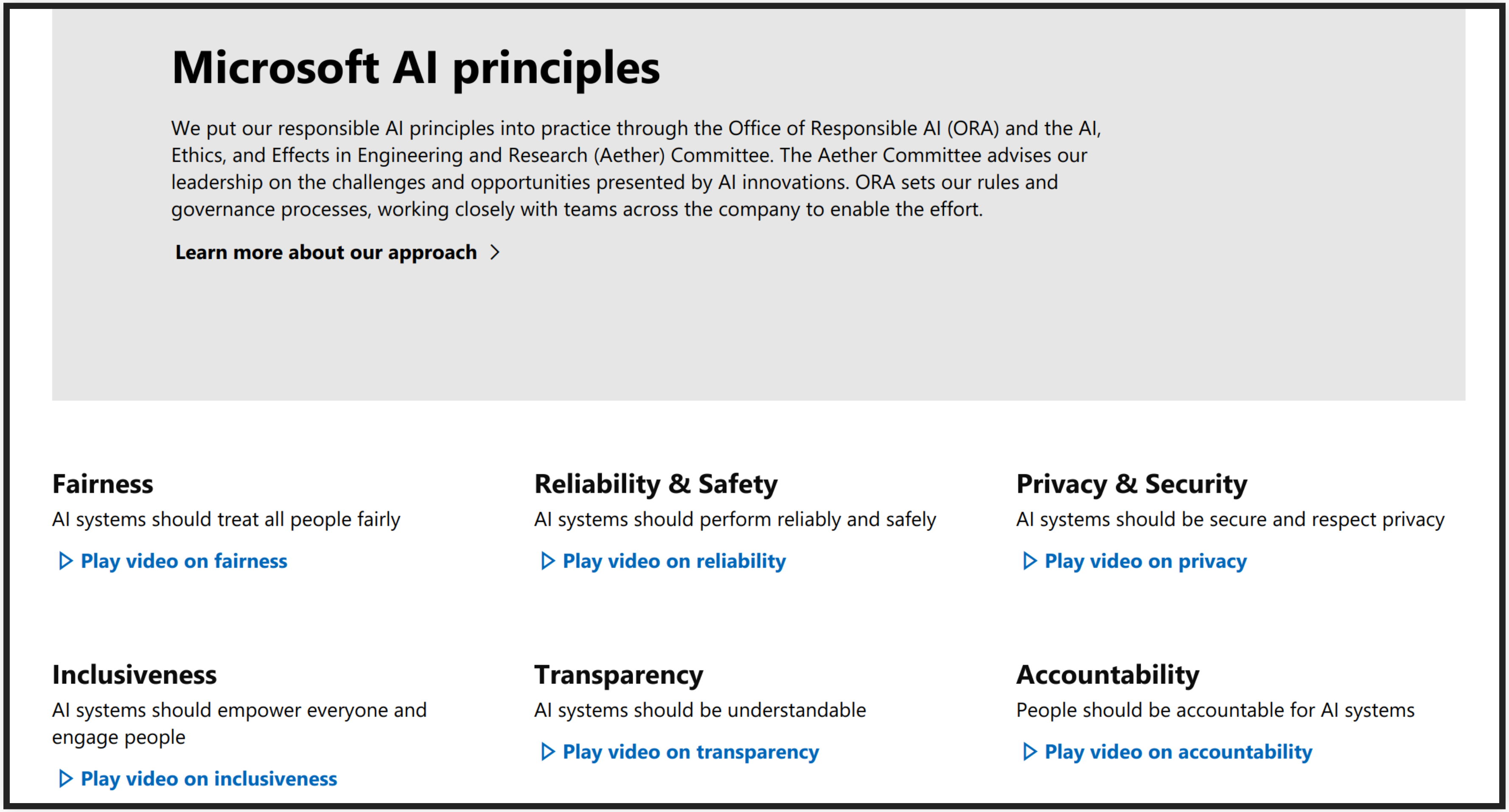


SHORT WAVE

Tech Companies Are Limiting Police Use of Facial Recognition. Here's Why

June 23, 2020 · 4:00 AM ET

(SELF-)REGULATION AND POLICY



Microsoft AI principles

We put our responsible AI principles into practice through the Office of Responsible AI (ORA) and the AI, Ethics, and Effects in Engineering and Research (Aether) Committee. The Aether Committee advises our leadership on the challenges and opportunities presented by AI innovations. ORA sets our rules and governance processes, working closely with teams across the company to enable the effort.

Learn more about our approach ›

Fairness

AI systems should treat all people fairly

▷ **Play video on fairness**

Reliability & Safety

AI systems should perform reliably and safely

▷ **Play video on reliability**

Privacy & Security

AI systems should be secure and respect privacy

▷ **Play video on privacy**

Inclusiveness

AI systems should empower everyone and engage people

▷ **Play video on inclusiveness**

Transparency

AI systems should be understandable

▷ **Play video on transparency**

Accountability

People should be accountable for AI systems

▷ **Play video on accountability**

POLICY DISCUSSION AND FRAMEING

- Corporate pitch: "Responsible AI" (Microsoft, Google, Accenture)
- Counterpoint: Ochigame "The Invention of 'Ethical AI': How Big Tech Manipulates Academia to Avoid Regulation", The Intercept 2019
    - "*The discourse of "ethical AI" was aligned strategically with a Silicon Valley effort seeking to avoid legally enforceable restrictions of controversial technologies.*"
- Self-regulation vs government regulation? Assuring safety vs fostering innovation?

Forbes

4,576 views | Mar 1, 2020, 01:00am EST

This Is The Year Of AI Regulations

Kathleen Walch Contributor

COGNITIVE WORLD Contributor Group

AI

The world of artificial intelligence is constantly evolving, and certainly so is the legal and regulatory environment

“ACCELERATING AMERICA’S LEADERSHIP IN ARTIFICIAL INTELLIGENCE”

> *“the policy of the United States Government [is] to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI.” -- White House Executive Order Feb. 2019*

Tone: "When in doubt, the government should not regulate AI."

- 3. Setting AI Governance Standards: "*foster public trust in AI systems by establishing guidance for AI development. [...] help Federal regulatory agencies develop and maintain approaches for the safe and trustworthy creation and adoption of new AI technologies. [...] NIST to lead the development of appropriate technical standards for reliable, robust, trustworthy, secure, portable, and interoperable AI systems.*"

JAN 13 2020 DRAFT RULES FOR PRIVATE SECTOR AI

- *Public Trust in AI*: Overarching theme: reliable, robust, trustworthy AI
- *Public participation:* public oversight in AI regulation
- *Scientific Integrity and Information Quality:* science-backed regulation
- *Risk Assessment and Management:* risk-based regulation
- *Benefits and Costs:* regulation costs may not outweigh benefits
- *Flexibility:* accommodate rapid growth and change
- *Disclosure and Transparency:* context-based transparency regulation
- *Safety and Security:* private sector resilience

Draft: Guidance for Regulation of Artificial Intelligence Applications

OTHER REGULATIONS

- *China:* policy ensures state control of Chinese companies and over valuable data, including storage of data on Chinese users within the country and mandatory national standards for AI
- *EU:* Ethics Guidelines for Trustworthy Artificial Intelligence; Policy and investment recommendations for trustworthy Artificial Intelligence; draft regulatory framework for high-risk AI applications, including procedures for testing, record-keeping, certification, ...
- *UK:* Guidance on responsible design and implementation of AI systems and data ethics

Source: https://en.wikipedia.org/wiki/Regulation_of_artificial_intelligence

CALL FOR TRANSPARENT AND AUDITED MODELS

> *"no black box should be deployed when there exists an interpretable model with the same level of performance"*

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

Regulate possible conflict: Intellectual property vs public health/welfare

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. (Preprint)

CRITICISM: ETHICS WASHING, ETHICS BASHING, REGULATORY CAPTURE

SUMMARY

- Transparency goes beyond explaining predictions
- Plan for mistakes and human oversight
- Accountability and culpability are hard to capture, little regulation
- Be a responsible engineer, adopt a culture of responsibility
- Regulations may be coming