Machine Learning in Production

Transparency and Accountability

More Explainability, Policy, and Politics

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:
System and model goals
User requirements
Environment assumptions
Quality beyond accuracy
Measurement
Risk analysis
Planning for mistakes

Architecture + design:
Modeling tradeoffs
Deployment architecture
Data science pipelines
Telemetry, monitoring
Anticipating evolution
Big data processing
Human-AI design

Quality assurance:
Model testing
Data quality
QA automation
Testing in production
Infrastructure quality
Debugging

Operations:
Continuous deployment
Contin. experimentation
Configuration mgmt.
Monitoring
Versioning
Big data
DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance, versioning, reproducibility

Safety

Security and privacy

Fairness

Interpretability and explainability

Transparency and trust

Ethics, governance, regulation, compliance, organizational culture

Readings

Required reading:

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

Recommendedr hoeading:

- Metcalf, Jacob, and Emanuel Moss. "Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics." *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.

Learning Goals

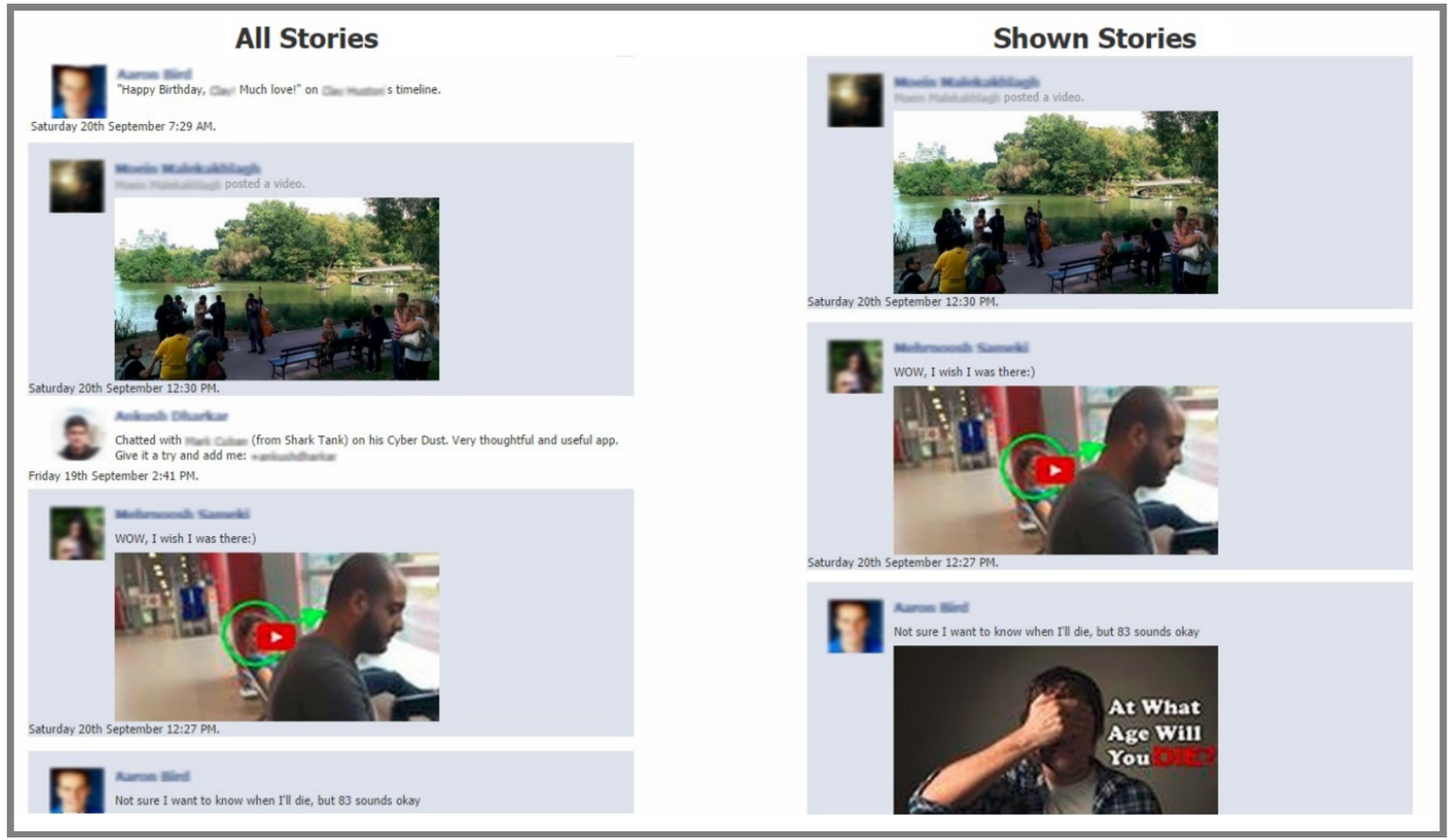
- Explain key concepts of transparency and trust
- Discuss whether and when transparency can be abused to game the system
- Design a system to include human oversight
- Understand common concepts and discussions of accountability/culpability
- Critique regulation and self-regulation approaches in ethical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

Transparency

Transparency: users know that algorithm exists / users know how the algorithm works


Listening to The Cure and thinking about the Bomb
@TheWrongNoel · Follow
A friend of mine has been trying to hire a new employee for her department in a medium-sized org. After advertising several times with few applicants, and a couple of rounds of interviews, the new employee is less than great. Then she discovered there were other applicants ...
5:01 AM · Nov 14, 2019
11.7K
Reply
Copy link
Read 351 replies

Case Study: Facebook's Feed Curation



Eslami, Motahhare, et al. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In Proc. CHI, 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

The Dark Side of Transparency

- Users may feel influence and control, even with placebo controls
- Companies give vague 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.

Appropriate Level of Algorithmic Transparency

IP/Trade Secrets/Fairness/Perceptions/Ethics?

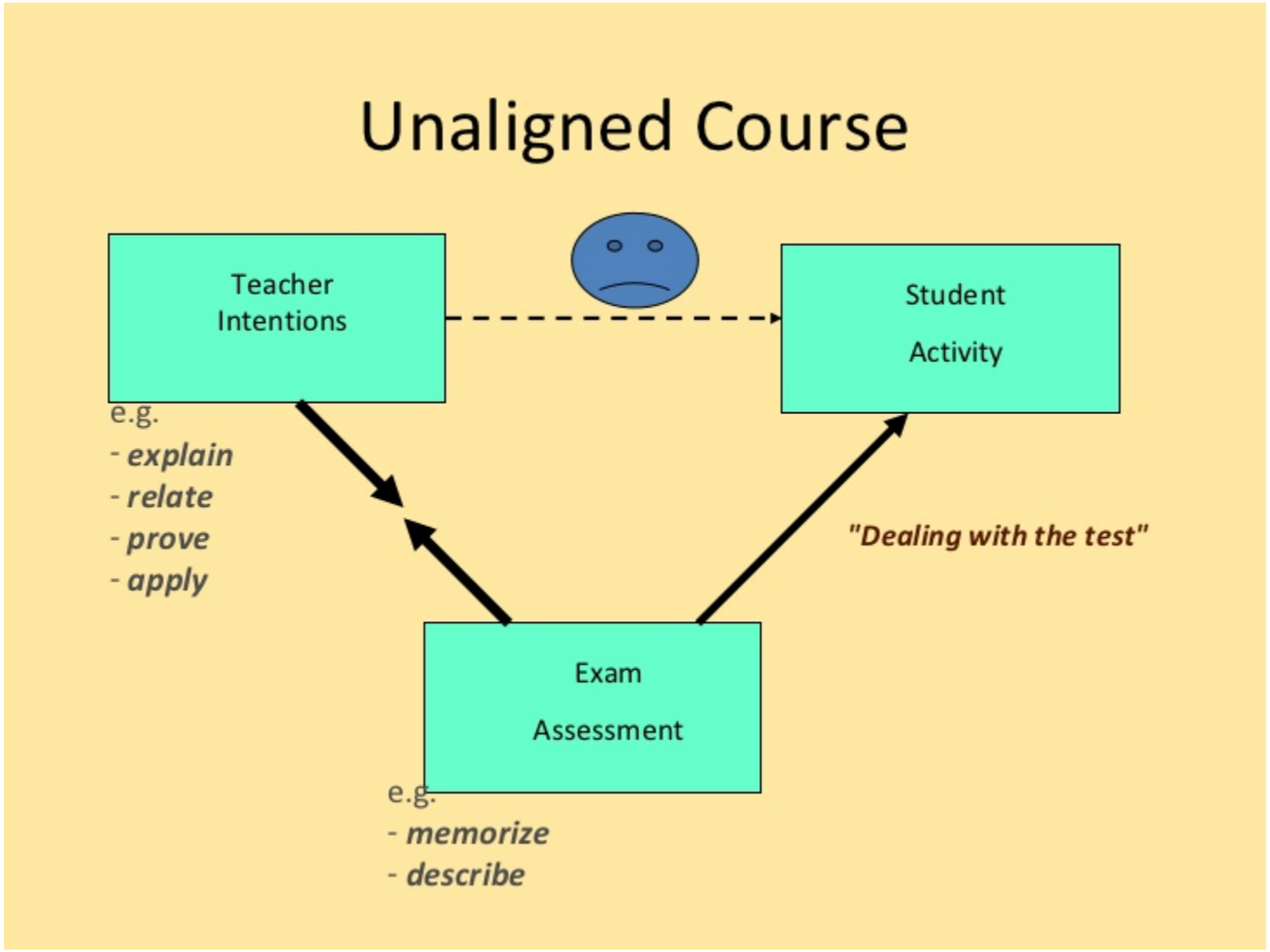
How to design? How much control to give?

Gaming/Attacking the Model with Explanations?

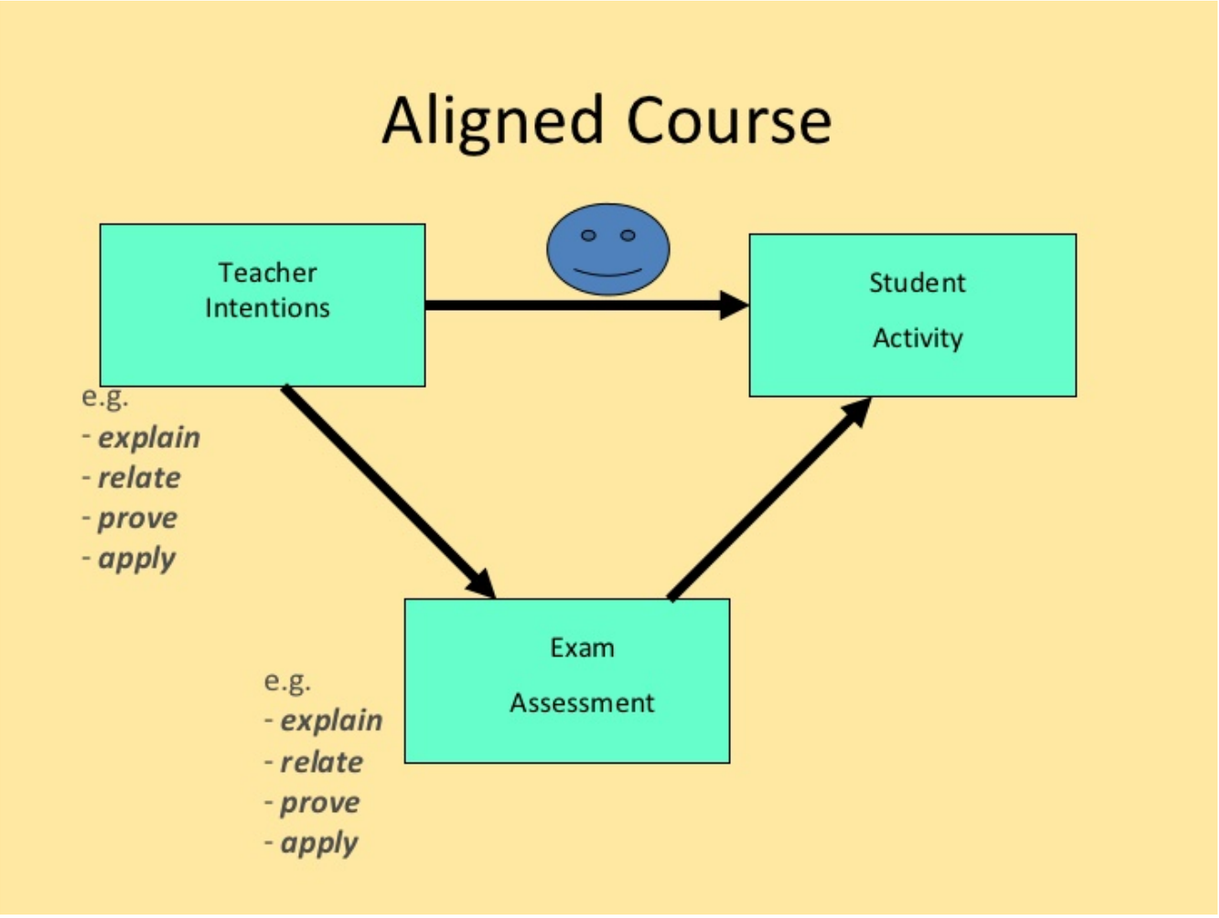
Does providing an explanation allow customers to 'hack' the system?

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

Gaming the Model with Explanations?



Constructive Alignment in Teaching



see also Claus Brabrand. Teaching Teaching & Understanding Understanding. Youtube 2009

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 relate causally to the outcome <- hard to game*

```
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
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
- often about assigning *blame* -- responsible for fixing or liable for paying for damages
- liability, culpability have *legal* connotation
- responsibility tends to describe *ethical* aspirations
- accountability often defined as oversight relationship, where actor is accountable to some "forum" that can impose penalties
- see also legal vs ethical earlier

On Terminology

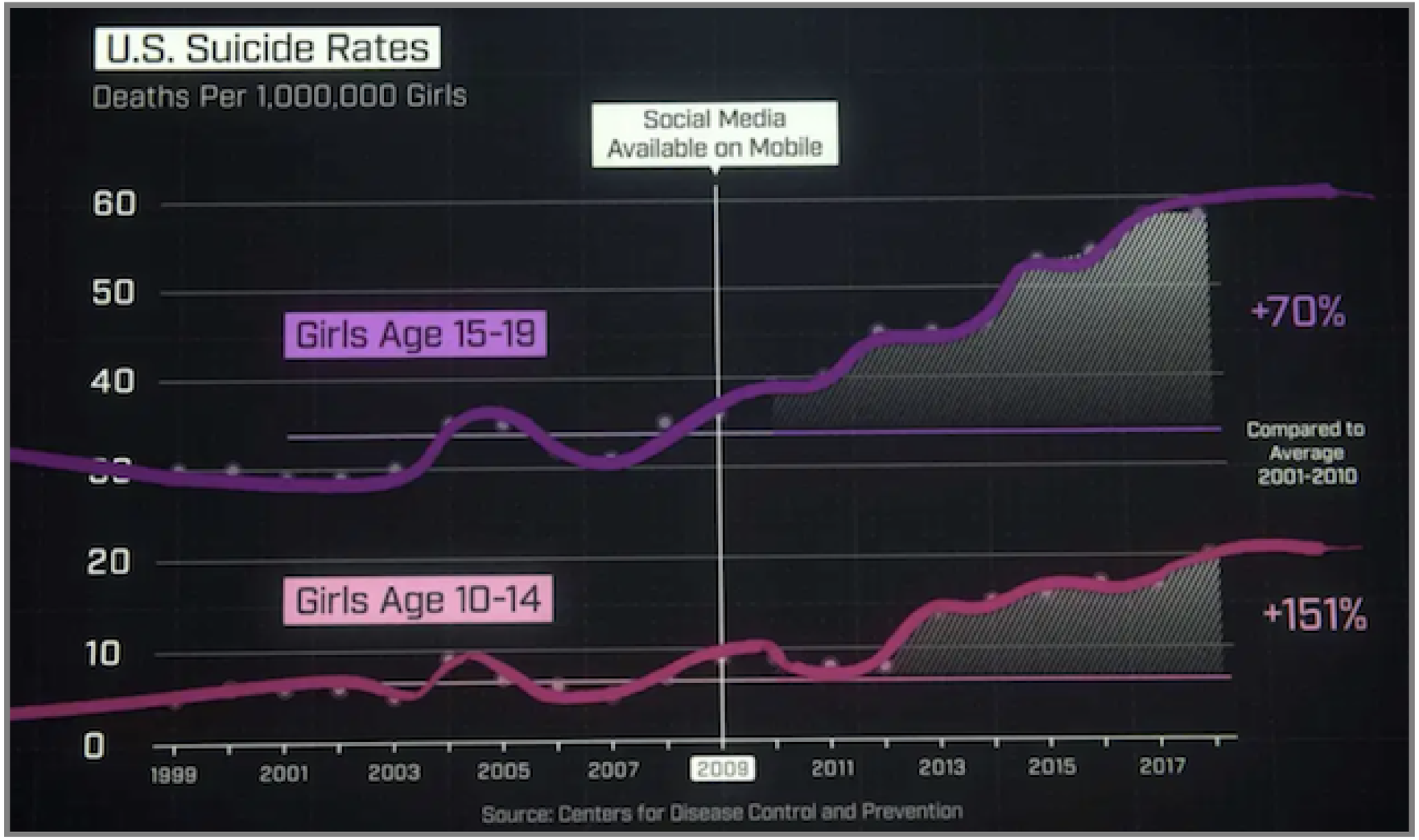
Academic definition of accountability:

> *A relationship between an **actor** and a **forum**, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor **may face consequences**.*

That is accountability implies some oversight with ability to penalize

Wieringa, Maranke. "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 1-18. 2020.

Who is responsible?



Who is responsible?

Who is responsible?

Who is responsible?

Faceswap's README "FaceSwap has ethical uses"

> *[...] as is so often the way with new technology emerging on the internet, it was immediately used to create inappropriate content.*

> *[...] it was the first AI code that anyone could download, run and learn by experimentation without having a Ph.D. in math, computer theory, psychology, and more. Before "deepfakes" these techniques were like black magic, only practiced by those who could understand all of the inner workings as described in esoteric and endlessly complicated books and papers.*

> *[...] the release of this code opened up a fantastic learning opportunity.*

> *Are there some out there doing horrible things with similar software? Yes. And because of this, the developers have been following strict ethical standards. Many of us don't even use it to create videos, we just tinker with the code to see what it does. [...]*

> *FaceSwap is not for creating inappropriate content. FaceSwap is not for changing faces without consent or with the intent of hiding its use. FaceSwap is not for any illicit, unethical, or questionable purposes. [...]*

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

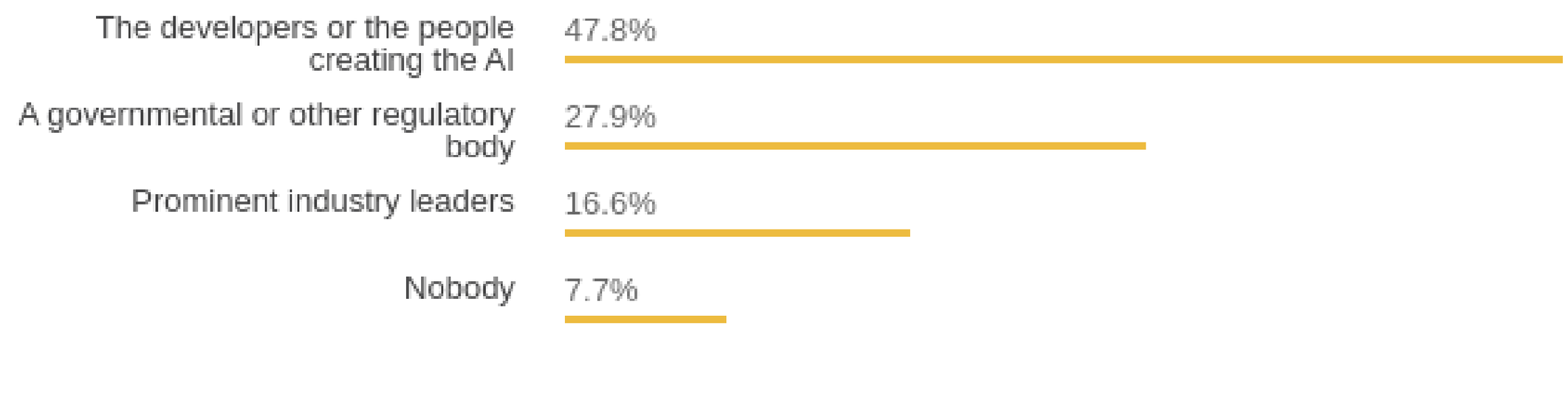
Software engineers got (mostly) away with declaring not to be liable

Easy to Blame "The Algorithm" / "The Data" / "Software"

> *"Just a bug, things happen, nothing we could have done"*

- But system was designed by humans
- But humans did not anticipate possible mistakes, did not design to mitigate mistakes
- But humans made decisions about what quality was good enough
- But humans designed/ignored the development process
- But humans gave/sold poor quality software to other humans
- But humans used the software without understanding it
- ...

Who is Primarily Responsible for Considering the Ramifications of AI?



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, whether it should be built at all
- And... follow the law
- Get started with existing guidelines, e.g., in AI Ethics Guidelines

(Self-)Regulation and Policy

npr

DONATE

Play Live Radio

HOURLY NEWS | LISTEN LIVE | PLAYLIST

SUBSCRIBE

SHORT WAVE

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

June 23, 2020 · 4:00 AM ET

14-Minute Listen

+ PLAYLIST

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?


@emilymbender@dair-community.social on Mastodon
@emilymbender · Follow
Okay, so that AI letter signed by lots of AI researchers calling for a "Pause [on] Giant AI Experiments"? It's just dripping with #Aihype. Here's a quick rundown.
>>
3:36 AM · Mar 29, 2023
1.5K
Reply
Copy link
Read 41 replies

Arvind Narayanan
@random_walker · **Follow**

This open letter — ironically but unsurprisingly — further fuels AI hype and makes it harder to tackle real, already occurring AI harms. I suspect that it will benefit the companies that it is supposed to regulate, and not society. Let's break it down. 🧵

futureoflife.org
Pause Giant AI Experiments: An Open Letter - Future of Life Institute
We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

1:58 PM · Mar 29, 2023

1.2K Reply Copy link

Read 35 replies

"Wishful Worries"

We are distracted with worries about fairness and safety of hypothetical systems

Most systems fail because they didn't work in the first place; don't actually solve a problem or address impossible tasks

Wouldn't help even if they solved the given problem (e.g., predictive policing?)

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The fallacy of AI functionality." In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 959-

Forbes

This Is The Year Of AI Regulations

Kathleen Walch Contributor
COGNITIVE WORLD Contributor Group ⓘ
AI

The world of artificial intelligence is constantly evolving,

“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."

Speaker notes

- 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

Call for Transparent and Audited Models

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

For high-stakes decisions

- ... with government involvement (recidivism, policing, city planning, ...)
- ... in medicine
- ... with discrimination concerns (hiring, loans, housing, ...)
- ... that influence society and discourse? (algorithmic content amplifications, targeted advertisement, ...)

Regulate possible conflict: Intellectual property vs public 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

Further Readings

- Jacovi, Alon, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635. 2021.
- 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.
- Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices." Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW1 (2021): 1–23.
- Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. "Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning." In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019).
- Metcalf, Jacob, and Emanuel Moss. "Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics." *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.
- Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The fallacy of AI functionality." In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 959-972. 2022.