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

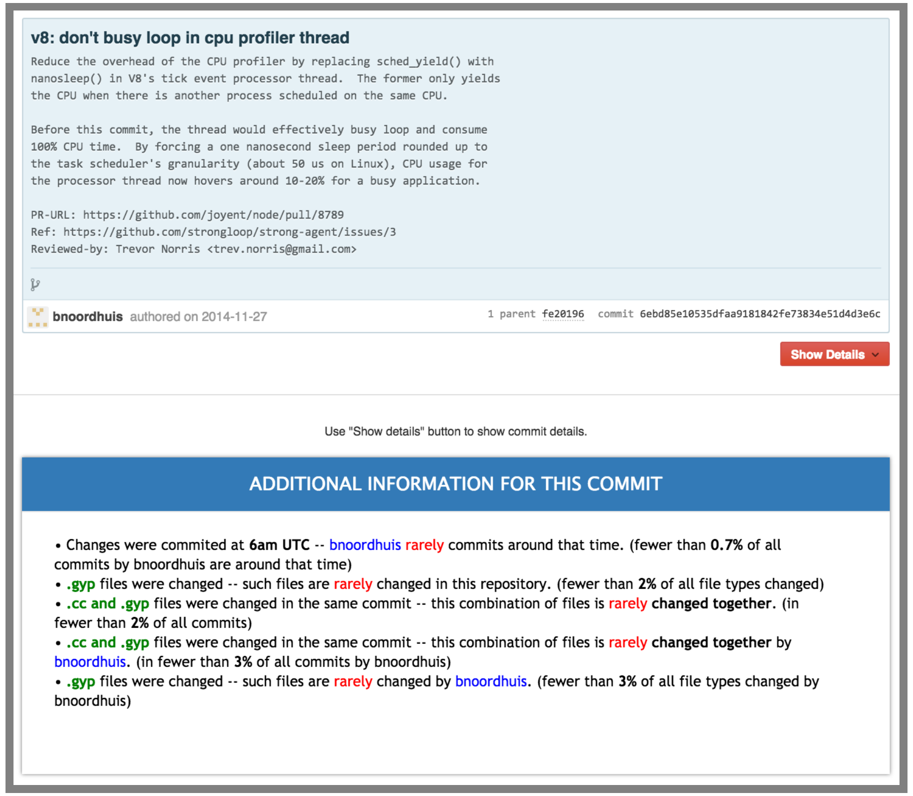
- Explain key concepts of transparency and trust
- Discuss whether and when transparency can be abused to game the system
- Design a system to include human oversight
- Understand common concepts and discussions of accountability/culpability
- Critique regulation and self-regulation approaches in ethical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

Explainability vs Transparency

- Explainability & Interpretability: Tools to understand the model, mostly debugging, mostly developer-focused
- Transparency: Users know that algorithm exists and how the algorithm works

*Transparency is focused on **users** -- Human-AI interaction, oversight, appeals, audits*

Recall: Explaining for Human-AI Interaction, Trust



Goyal, Raman, Gabriel Ferreira, Christian Kästner, and James Herbsleb. "Identifying unusual commits on GitHub." Journal of Software: Evolution and Process 30, no. 1 (2018): e1893.

Recall: Explainability for Auditing

- Understand safety implications
- Ensure predictions use objective criteria and reasonable rules
- Inspect fairness properties
- Reason about biases and feedback loops
- Validate "learned specifications/requirements" with stakeholders

```
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
```

Transparency of the Model's Existance

eradicate billionaires
@TheWrongNoel · Follow

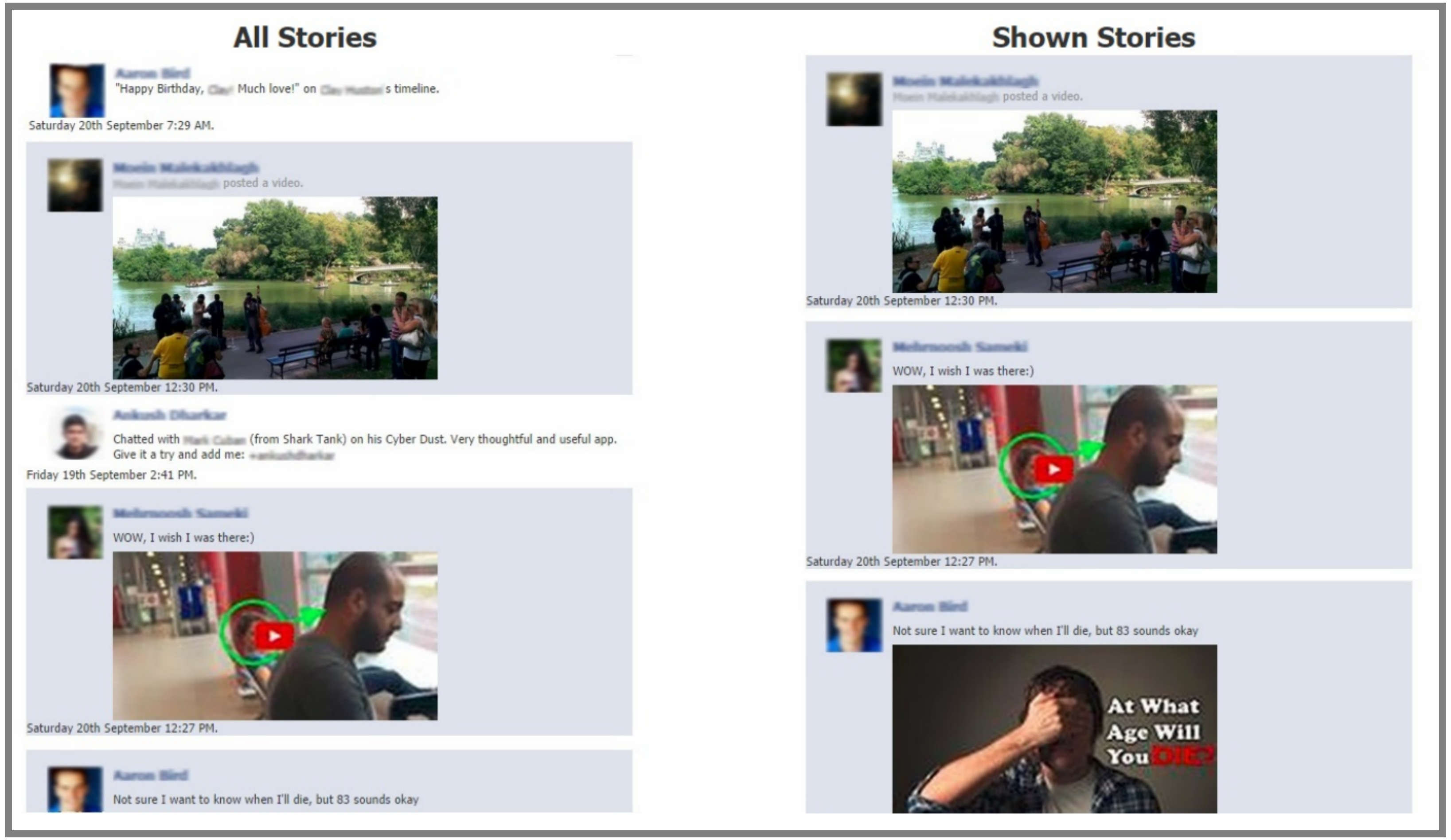
A friend of mine has been trying to hire a new employee for her department in a medium-sized org. After advertising several times with few applicants, and a couple of rounds of interviews, the new employee is less than great. Then she discovered there were other applicants ...

5:01 AM · Nov 14, 2019

10.8K Reply Copy link

Read 324 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."*

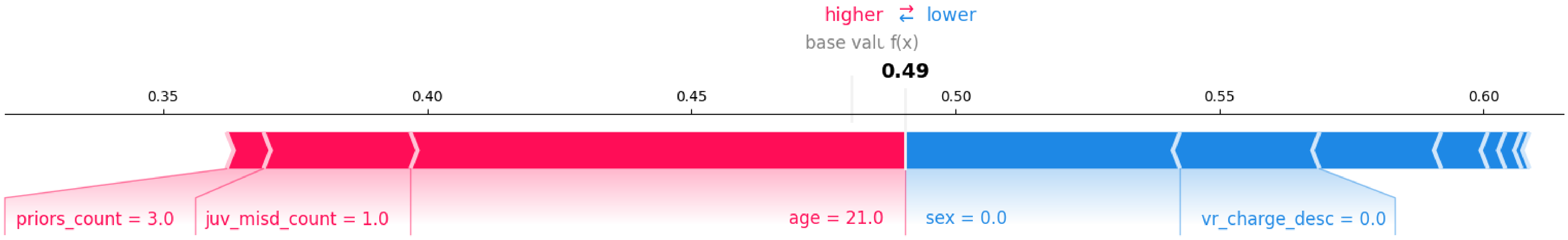
- Learning about algorithm did not change satisfaction level
- More active engagement, more feeling of control

Transparency of How the Model Works

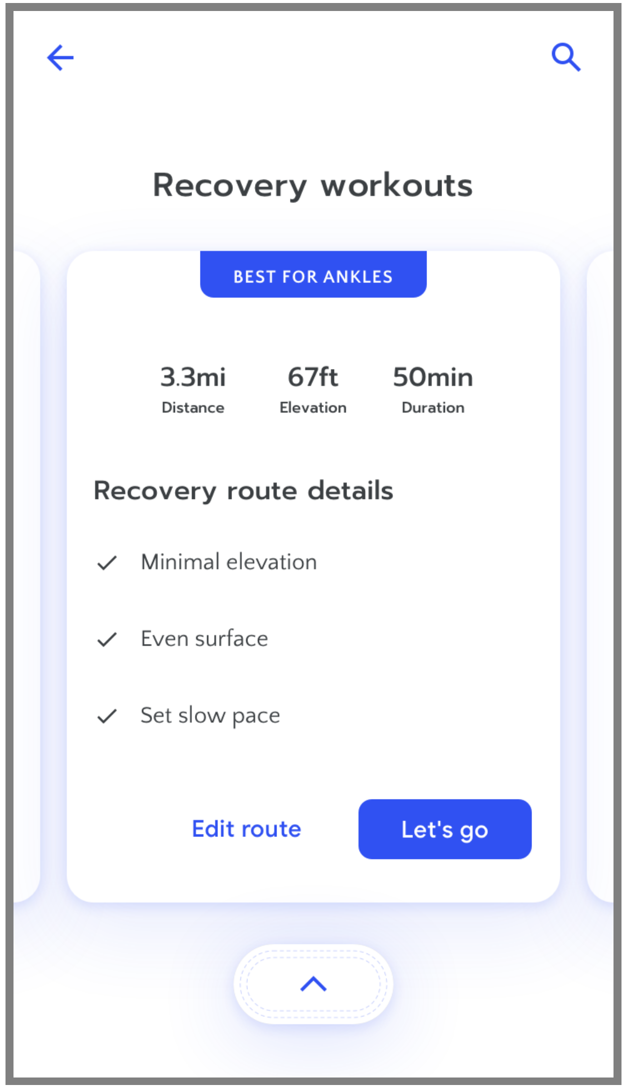
Enabling Oversight and Appeals

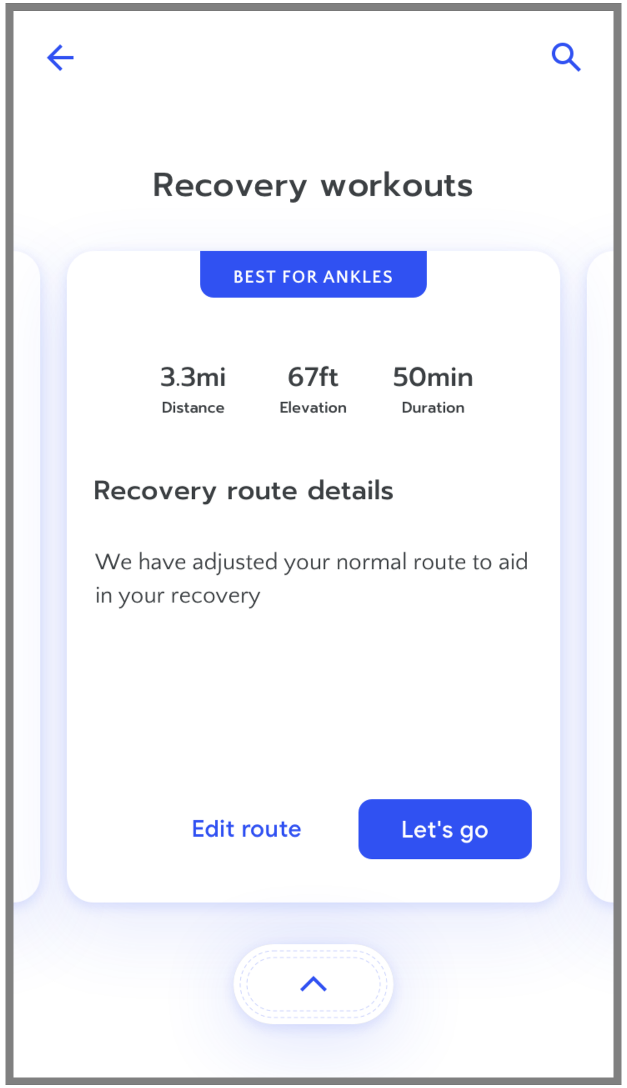
- What is this figure showing?
- Who want to get what information from this plot?
- Who can read this plot? What kinds of expertise? Training?

Human is the key!

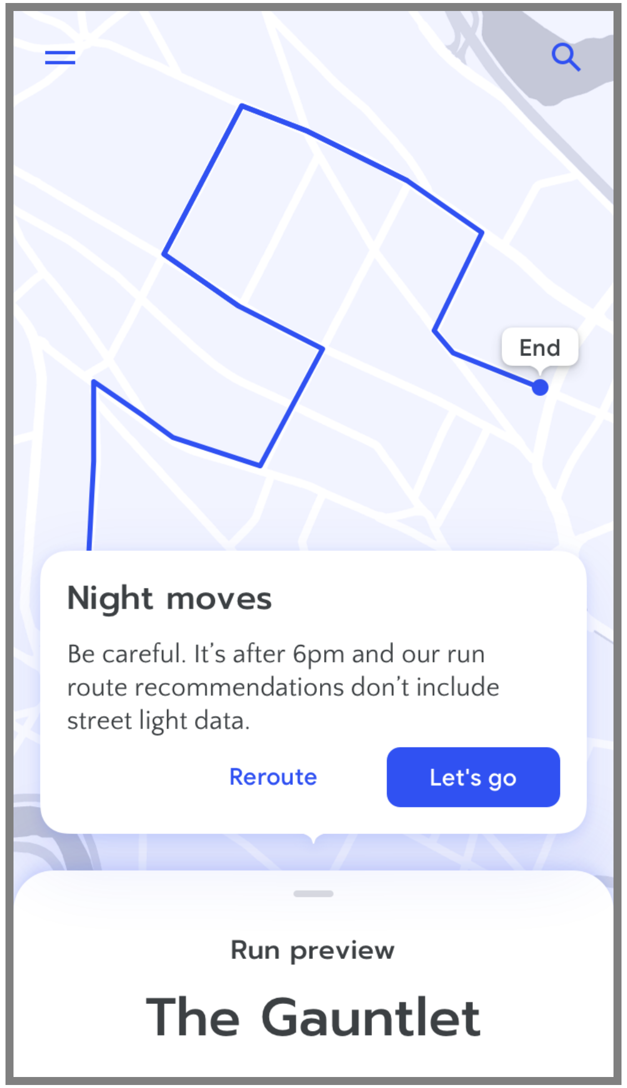


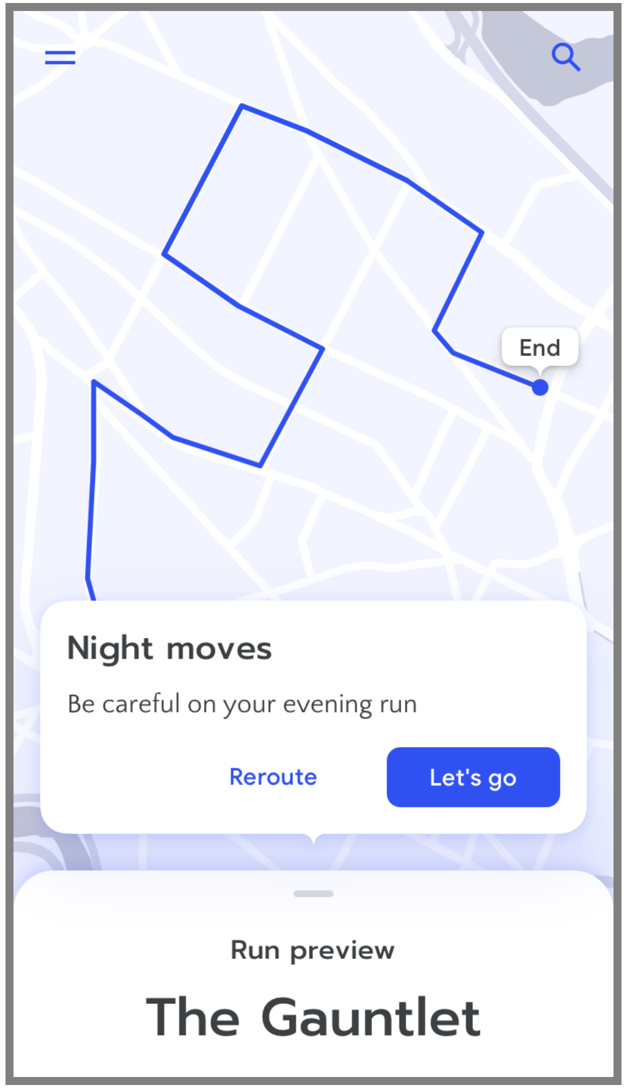
Expl. for Human-AI Interaction





Give the user details about why a prediction was made in a high stakes scenario. Here, the user is exercising after an injury and needs confidence in the app's recommendation.





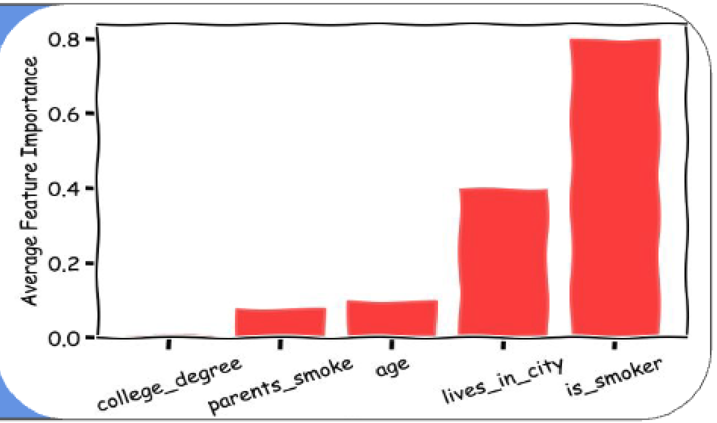
Tell the user when a lack of data might mean they'll need to use their own judgment. Don't be afraid to admit when a lack of data could affect the quality of the AI recommendations.

Source: People + AI Guidebook, Google

Express Explanation Intuitively

For women predicted high risk for lung cancer that are older than 65, why did the model decide to predict them as high risk?

Good question! It looks like the model predicted these individuals as high risk mostly because they were smokers but also because they live in large cities. I'm highly confident these are the reasons because the explanations have high fidelity. Here's the average feature importance for these people (higher means more important).

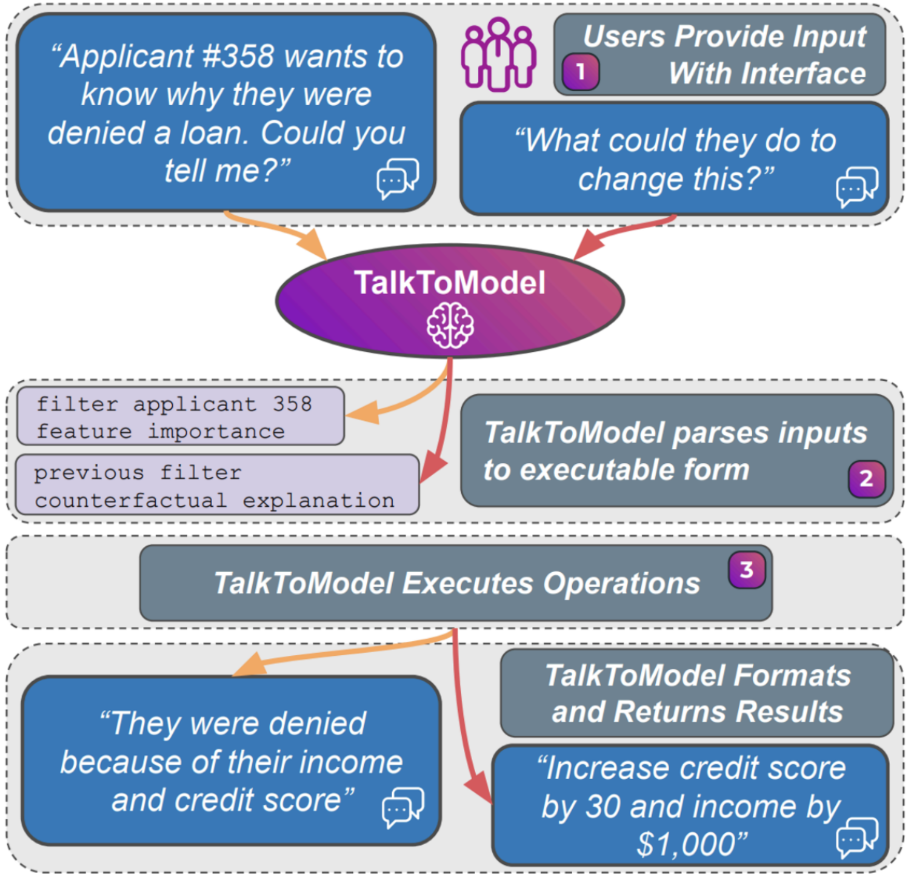


Wow, it's surprising that whether the person lives is in a city is so important.

Yes, lives_in_city has a significant effect on the predictions for these individuals. Perturbing this feature can flip the prediction for 4 of 15 of the instances in this group.

Lakkaraju et al., Rethinking Explainability as a Dialogue: A Practitioner's Perspective, 2022

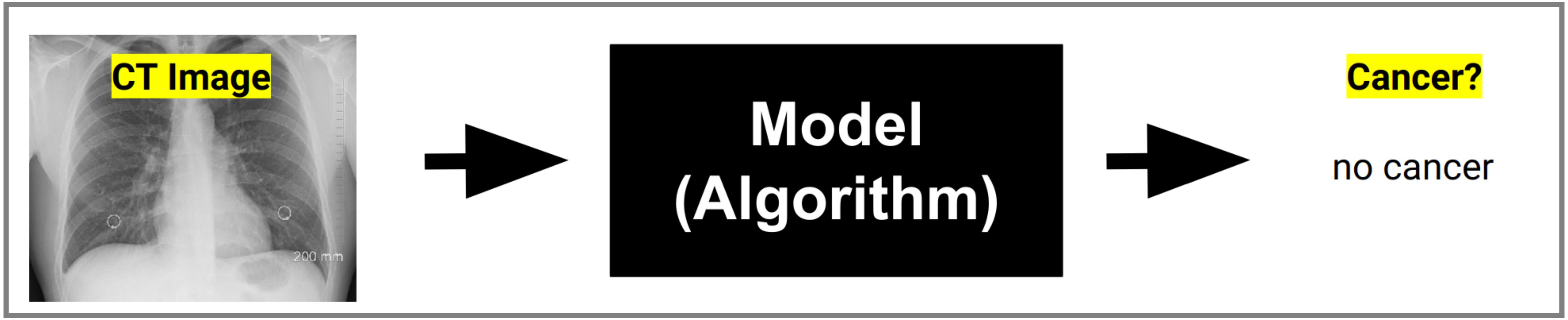
Express Explanation Intuitively



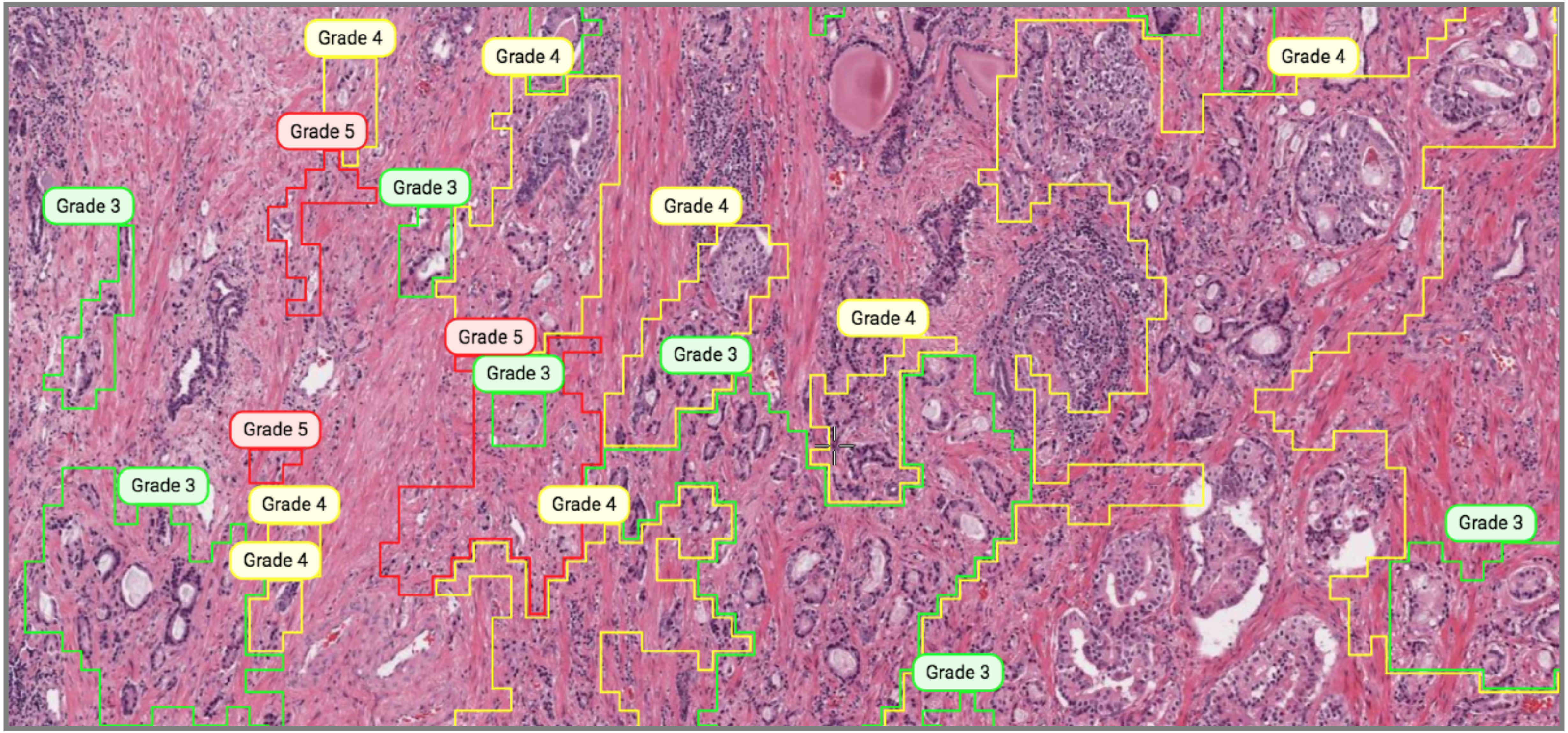
Slack, Dylan, et al. "TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations." (2022).

	operation, arguments, and description
Data	filter(dataset, feature, value, comparison): filters dataset by using value and comparison operator
	change(dataset, feature, value, variation): Changes dataset by increasing, decreasing, or setting feature by value
	show(list): Shows items in list in the conversation
	statistic(dataset, metric, feature): Computes summary statistic for feature
	count(list): Length of list
	and(op1, op2): Logical "and" of two operations
	or(op1, op2): Logical "or" of two operations
Explainability	explain(dataset, method, class=predicted): Feature importances on dataset
	cfe(dataset, number, class=opposite): Gets number counterfactual explanations
	topk(dataset, k): Top k most important features
	important(dataset, feature): Importance ranking of feature
	interaction(dataset): Interaction effects between features
	mistakes(dataset): Patterns in the model's errors on dataset
ML	predict(dataset): Model predictions on dataset
	likelihood(dataset): Prediction probabilities on dataset
	incorrect(dataset): Incorrect predictions
	score(dataset, metric): Scores the model with metric
Conv.	prev_filter(conversation): Gets last filters
	prev_operation(conversation): Gets last non-filtering operations
	followup(conversation): Respond to system followups
Description	function(): Overview of the system's capabilities
	data(dataset): Summary of dataset
	model(): Description of model
	define(term): Defines term

Setting Cancer Imaging -- What explanations do radiologists want?



- *Past attempts often not successful at bringing tools into production. Radiologists do not trust them. Why?*
- Wizard of oz study to elicit requirements


Grade 4
Grade 4
Grade 4
Grade 5
Grade 3
Grade 3
Grade 4
Grade 4
Grade 5
Grade 3
Grade 3
Grade 5
Grade 3
Grade 4
Grade 4
Grade 4
Grade 3
Grade 3

Radiologists' questions

- How does it perform compared to human experts?
- "What is difficult for the AI to know? Where is it too sensitive? What criteria is it good at recognizing or not good at recognizing?"
- What data (volume, types, diversity) was the model trained on?
- "Does the AI have access to information that I don’t have? Does it have access to ancillary studies?" Is all used data shown in the UI?
- What kind of things is the AI looking for? What is it capable of learning? ("Maybe light and dark? Maybe colors? Maybe shapes, lines?", "Does it take into consideration the relationship between gland and stroma? Nuclear relationship?")
- "Does it have a bias a certain way?" (compared to colleagues)

Radiologists' questions

- Capabilities and limitations: performance, strength, limitations; e.g. how does it handle well-known edge cases
- Functionality: What data used for predictions, how much context, how data is used
- Medical point-of-view: calibration, how liberal/conservative when grading cancer severity
- Design objectives: Designed for few false positives or false negatives? Tuned to compensate for human error?
- Other considerations: legal liability, impact on workflow, cost of use

Radiologists Study Insights

- AI literacy important for trust
- Be transparent about data used
- Describe training data and capabilities
- Give mental model, examples, human-relatable test cases
- Communicate the AI’s point-of-view and design goal

Cai, Carrie J., Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. ""Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making." Proceedings of the ACM on Human-computer Interaction 3, no. CSCW (2019): 1-24.

Designing Transparency

- Be explicit about the goal
- Tailor explanation to specific user needs and user's AI literacy
- Partial explanations or justifications often sufficient
- Test effectiveness of transparency mechanisms

The Dark Side of Transparency

Many explanations are wrong

Approximations of black-box models, often unstable

Explanations necessarily partial, social

Often multiple explanations possible (Rashomon effect)

Possible to use inherently interpretable models instead?

When explanation desired/required: What quality is needed/acceptable?

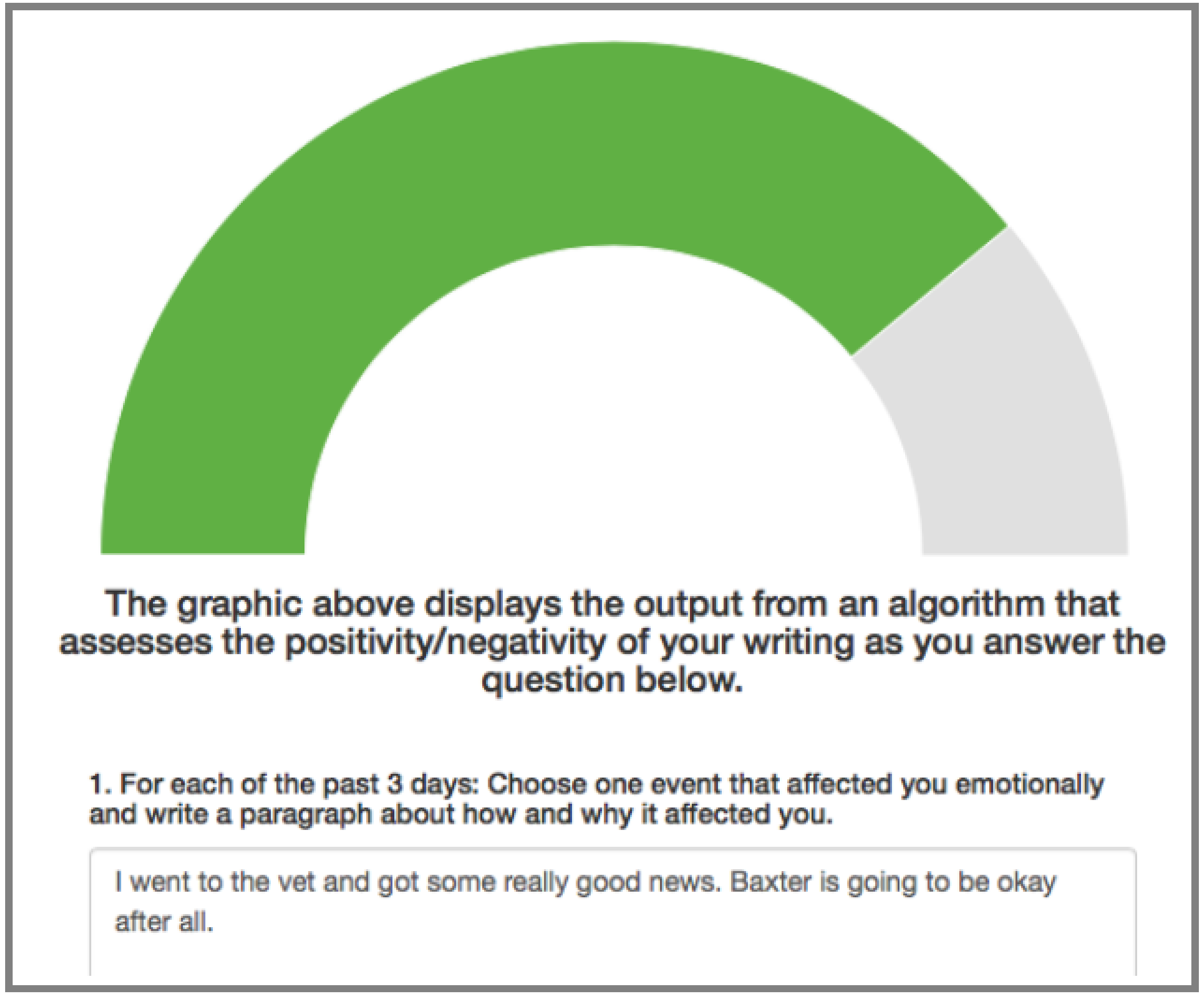
Explanations foster Trust

Users are less likely to question the model when explanations provided

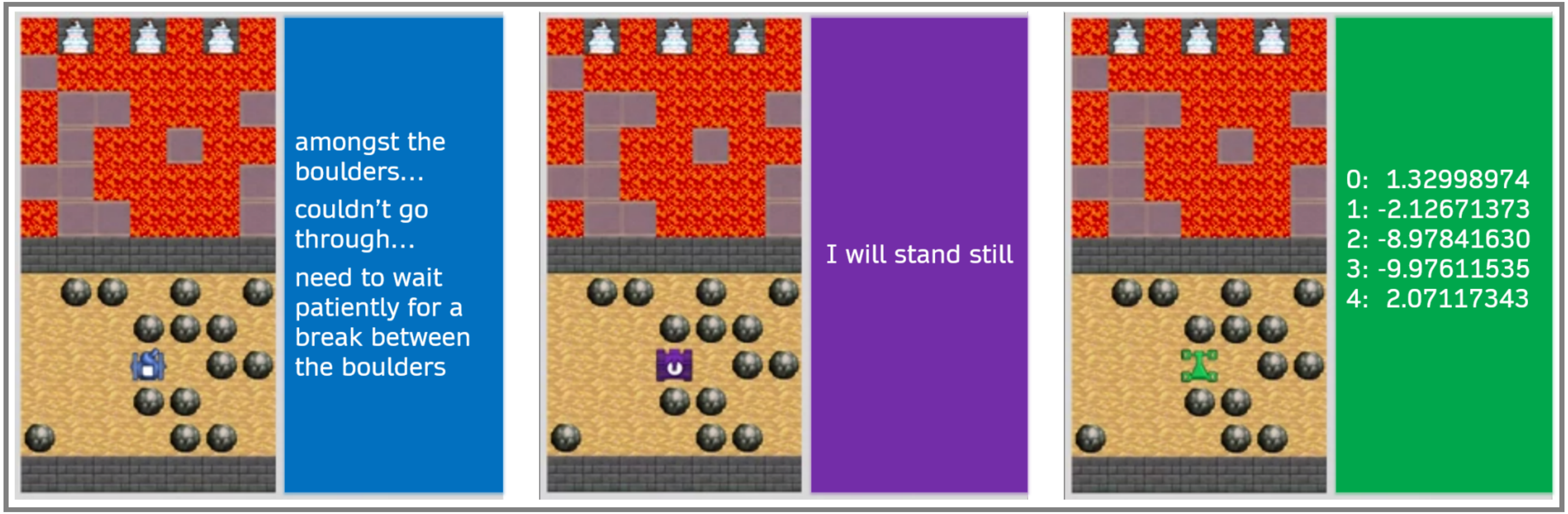
- Even if explanations are unreliable
- Even if explanations are nonsensical/incomprehensible

Danger of overtrust and intentional manipulation

Stumpf, Simone, Adrian Bussone, and Dympna O’sullivan. "Explanations considered harmful? user interactions with machine learning systems." In Proceedings of the ACM SIGCHI Conference on



Springer, Aaron, Victoria Hollis, and Steve Whittaker. "Dice in the black box: User experiences with an inscrutable algorithm." In 2017 AAAI Spring Symposium Series. 2017.



(a) Rationale, (b) Stating the prediction, (c) Numerical internal values

Observation: Both experts and non-experts overtrust numerical explanations, even when inscrutable.

Illusion of Control

Users may feel influence and control, even with placebo controls

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.

Regulatory Compliance

Companies give vague generic explanations to appease regulators

Checkbox compliance: Provide some mechanism without ensuring effectiveness

Example: FairCredit act requires explanation for declined credit applications -- Explanations generic and ineffective for fighting discrimination, at most ensure that input data was correct

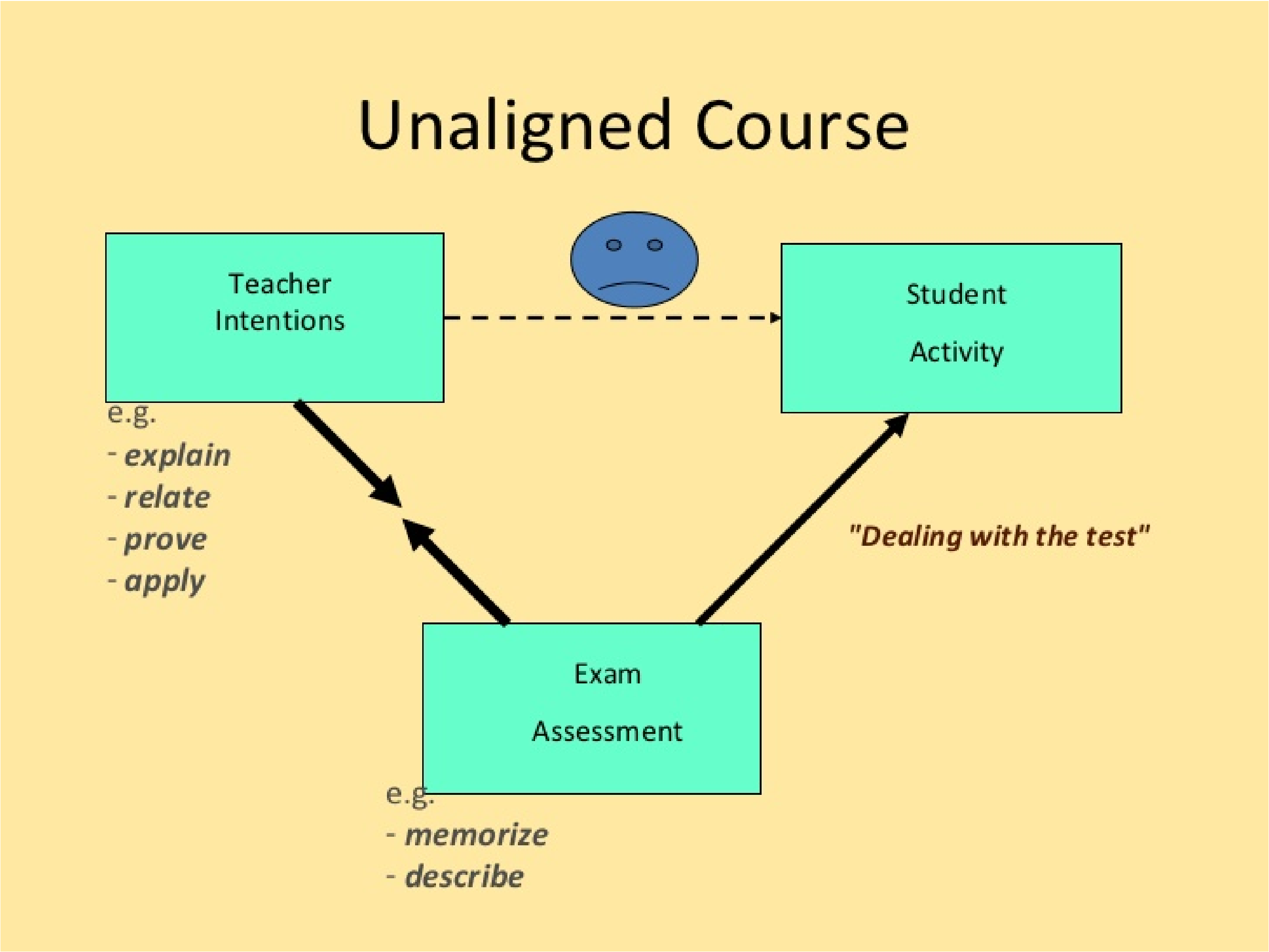
Selbst, Andrew D., and Solon Barocas. "The intuitive appeal of explainable machines." Fordham L. Rev. 87 (2018): 1085.

Gaming/Attacking the Model with Explanations?

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

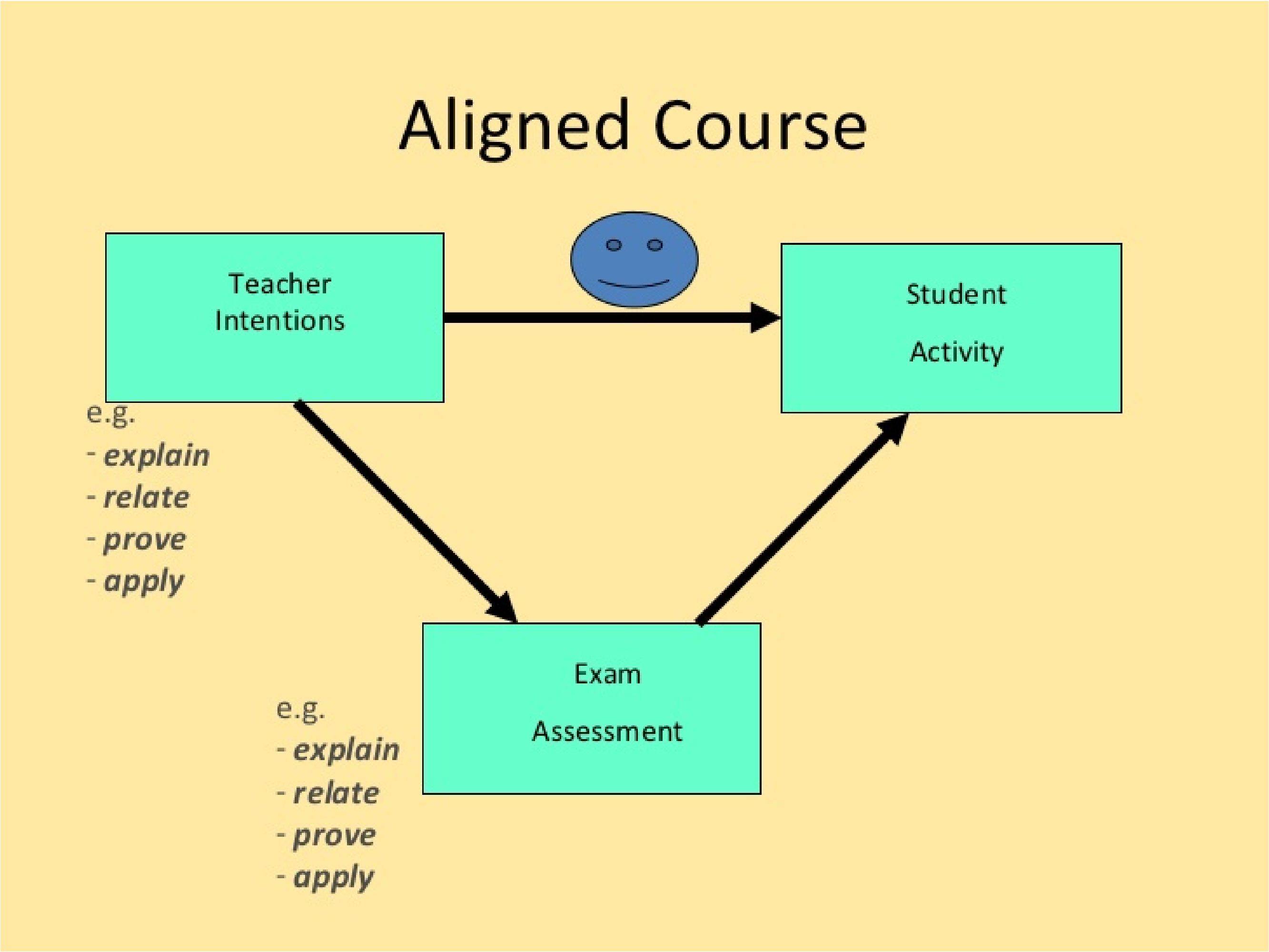
- Loan applications?
- Apple FaceID?
- Recidivism?
- Auto grading?
- Cancer diagnosis?
- Spam detection?

Gaming the Model with Explanations?



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

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?

Breakout: Transparency in Admissions

For a automated Master's admission support system, consider what you would make transparent and to whom.

In groups, tagging group members, respond in `#lecture`:

- What information (global, local) would you provide to applicants? What's the purpose?
- What information (global, local) would you provide to the admissions committee? What's the purpose?

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

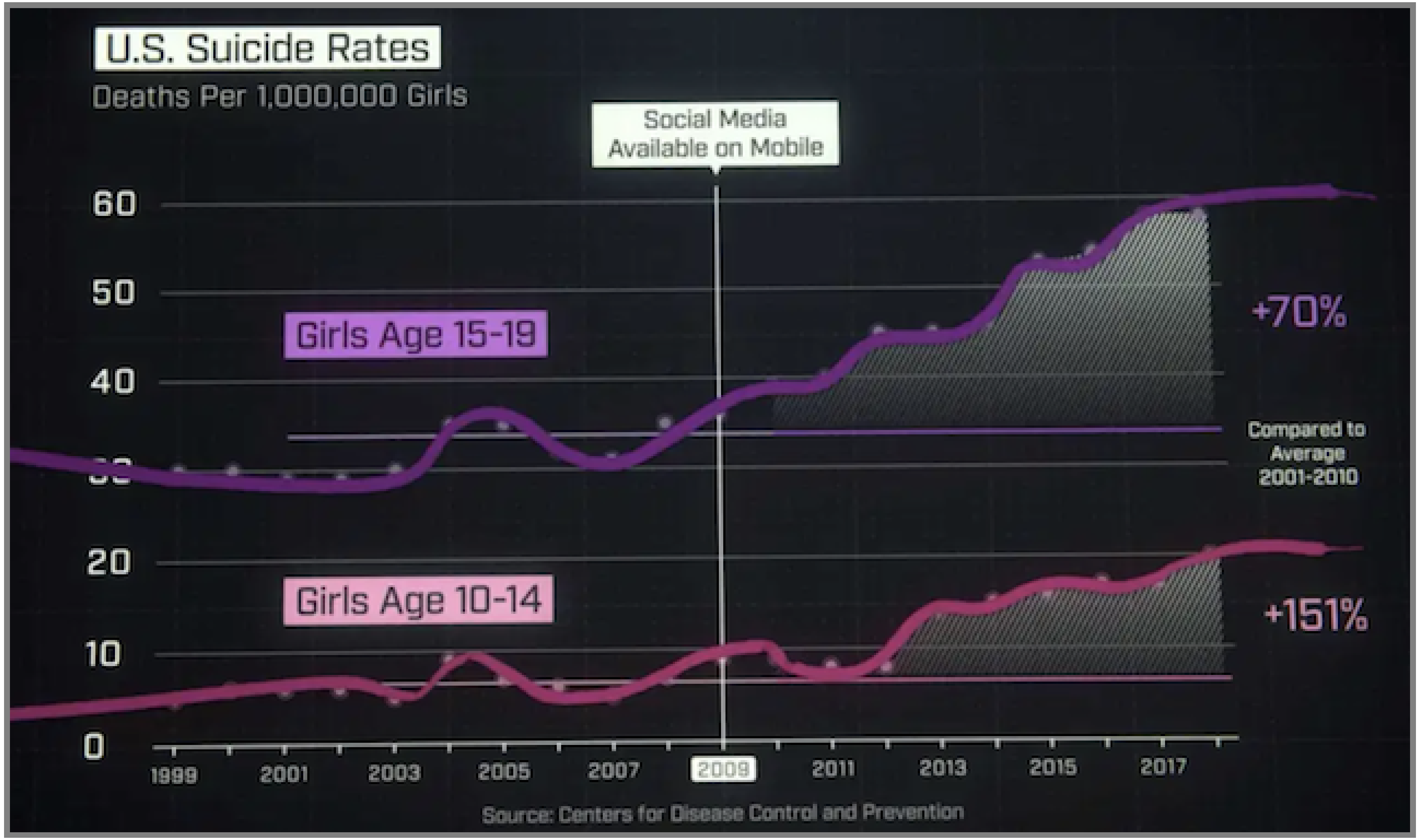
Academic definition of accountability:

> *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?

Who is responsible for Faceswap / Deepfake?

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. [...]*

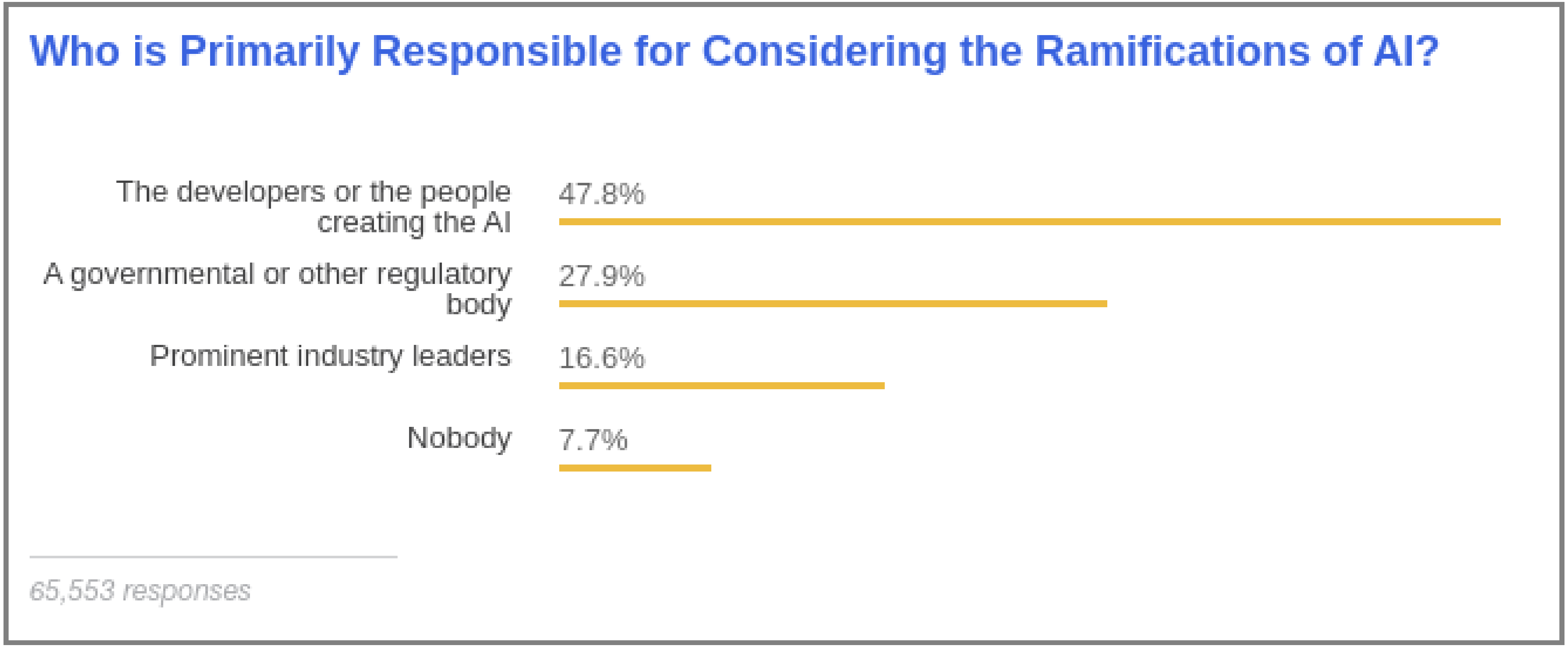
Speaker notes

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
- ...



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

SHORT WAVE

SUBSCRIBE

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.bsky.social
@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

1.4K Reply Copy link

Read 34 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 34 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

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

This Is The Year Of AI Regulations

Kathleen Walch Contributor
COGNITIVE WORLD Contributor Group
AI

f

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."

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.*"

EU AI Act

Broad regulatory framework, passed March 13, 2024

Risk-based framework:

2023 WH Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Instructs agencies to study risks and develop standards

Broad scope, touches on quality assurance standards (incl. red teaming) and marking AI-generated content

Domain-specific guidelines for chemical, biological, radiological, nuclear, and cybersecurity risks to be explored

Committees, reports, guidance, research instead of enforceable rules

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.