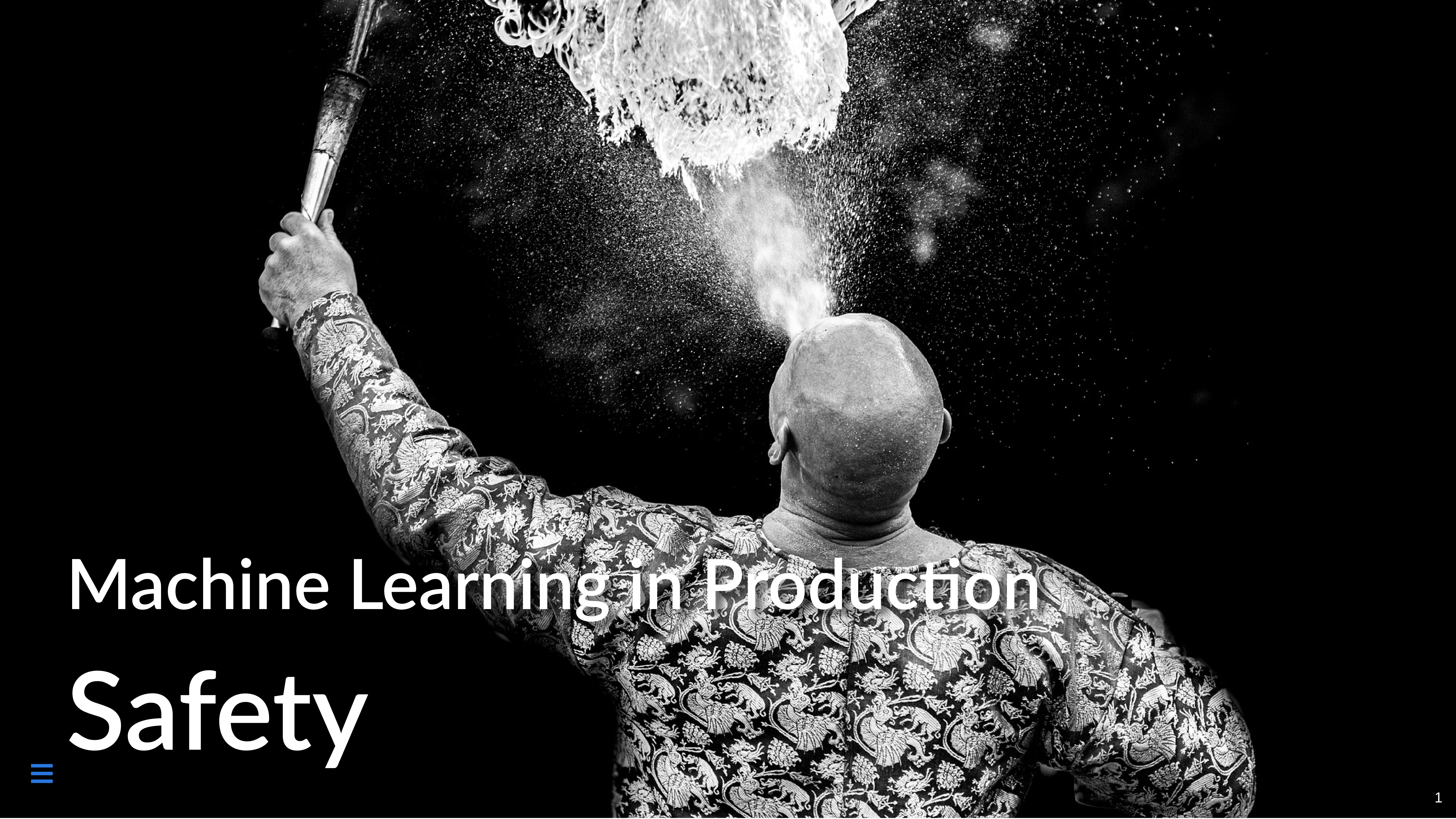


Machine Learning in Production

Safety

Mitigating more mistakes...

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

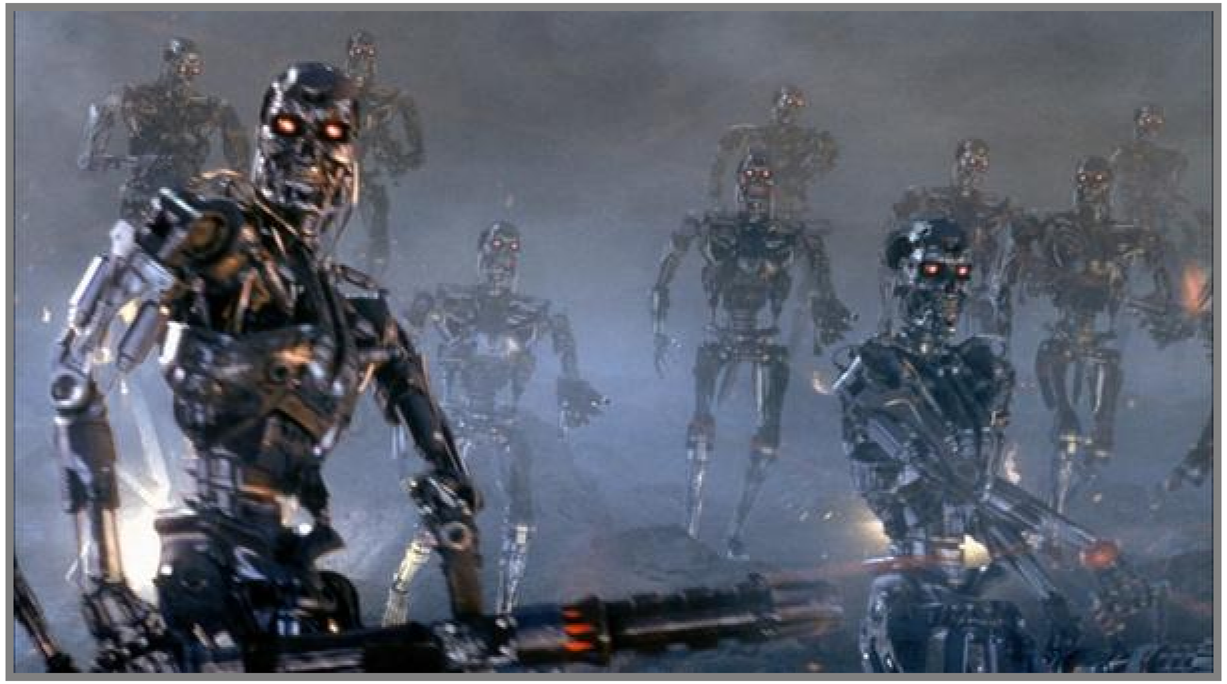
Reading

S. Mohseni et al., Practical Solutions for Machine Learning Safety in Autonomous Vehicles. SafeAI Workshop@AAAI (2020).

Learning Goals

- Understand safety concerns in traditional and AI-enabled systems
- Apply hazard analysis to identify risks and requirements and understand their limitations
- Discuss ways to design systems to be safe against potential failures
- Suggest safety assurance strategies for a specific project
- Describe the typical processes for safety evaluations and their limitations

AI Safety



Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).

Your Favorite AI Dystopia?

The AI Alignment Problem

AI is optimized for a specific objective/cost function

- Inadvertently cause undesirable effects on the environment
- e.g., Transport robot: Move a box to a specific destination
- Side effects: Scratch furniture, bump into humans, etc.,

Side effects may cause ethical/safety issues (e.g., social media optimizing for clicks, causing teen depression)

Difficult to define sensible fitness functions:

- Perform X *subject to common-sense constr. on the environment*
- Perform X *but avoid side effects to the extent possible*

Reward Hacking

PlayFun algorithm pauses the game of Tetris indefinitely to avoid losing

When about to lose a hockey game, the PlayFun algorithm exploits a bug to make one of the players on the opposing team disappear from the map, thus forcing a draw.

Self-driving car rewarded for speed learns to spin in circles

Example: Coast Runner

Reward Hacking

- AI can be good at finding loopholes to achieve a goal in unintended ways
- Technically correct, but does not follow *designer's informal intent*
- Many possible causes, incl. partially observed goals, abstract rewards, feedback loops
- In general, a very challenging problem!
  - Difficult to specify goal & reward function to avoid all possible hacks
  - Requires careful engineering and iterative reward design

Reward Hacking -- Many Examples

Victoria Krakovna
@vkrakovna · Follow

New resource: a master list of examples of AI systems gaming their objective specification: docs.google.com/spreadsheets/d…
Accompanying blog post: vkrakovna.wordpress.com/2018/04/02/spe…
Thanks @gwern and @catherineols for the inspiration and feedback on putting this together!

vkrakovna.wordpress.com
Specification gaming examples in AI
Update: for a more detailed introduction to specification ga…

12:37 PM · Apr 2, 2018

Read the full conversation on Twitter

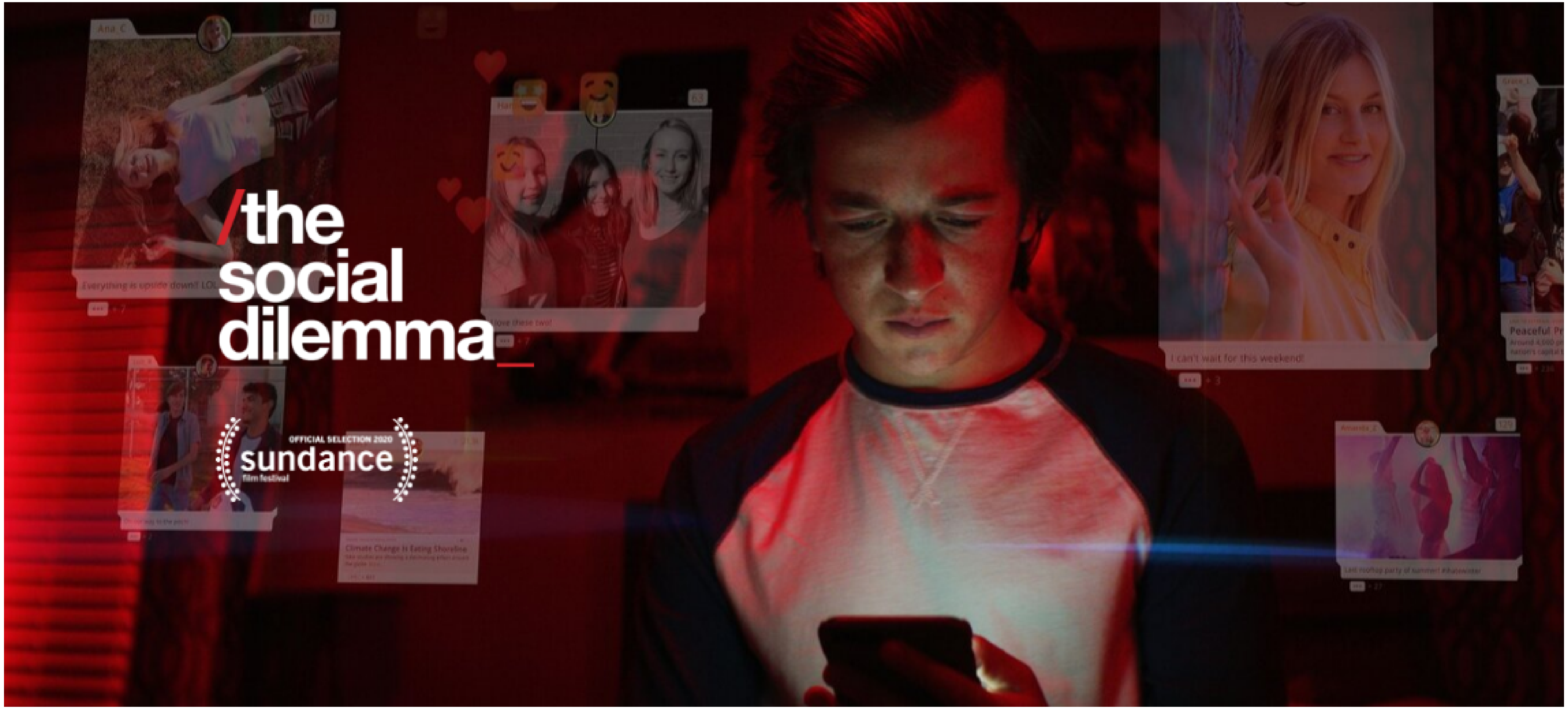
Exploiting Human Weakness

The Morality Of A/B Testing

Josh Constine @joshconstine / 11:50 PM EDT • June 29, 2014

We don't use the "real" Facebook. Or Twitter. Or Google, Yahoo,

Exploiting Human Weakness



See also Center for Humane Technology

AI Alignment Problem = Requirements Problem

Recall: "World vs. machine"

Identify stakeholders in the environment & possible effects on them

Anticipate side effects, feedback loops

Constrain scope of the system

Perfect contracts usually infeasible, undesirable

But more requirements engineering unlikely to be only solution

Other Challenges

- Safe Exploration
  - Exploratory actions "in production" may have consequences
  - e.g., trap robots, crash drones
- Robustness to Drift
  - Drift may lead to poor performance that may not even be recognized
- Scalable Oversight
  - Cannot provide human oversight over every action (or label all possible training data)
  - Use indirect proxies in telemetry to assess success/satisfaction

Existential AI Risk

Existential risk and AI alignment common in research

Funding through *longtermism* branch of effective altruism
(Longtermism is the view that positively influencing the longterm future is a key moral priority of our time.)

Ord estimates 10% existential risk from unaligned AI in 100 years

Our view: AI alignment not a real concern for the kind of ML-enabled products we consider here

Ord, Toby. The precipice: Existential risk and the future of humanity. Hachette Books, 2020.

Speaker notes

Relevant for reinforcement learning and AGI

Practical Alignment Problems

Does the model goal align with the system goal? Does the system goal align with the user's goals?

- Profits (max. accuracy) vs fairness
- Engagement (ad sales) vs enjoyment, mental health
- Accuracy vs operating costs

Test model *and* system quality *in production*

(see requirements engineering and architecture lectures)

Model Robustness

Defining Robustness:

- A prediction for input x is robust if the outcome is stable under minor perturbations to the input:
  - $\forall x'.\, d(x, x') < \epsilon \Rightarrow f(x) = f(x')$
  - distance function d and permissible distance ϵ depends on the problem domain!
- A model is said to be robust if most predictions are robust
- An important concept in safety and security settings
  - In safety, perturbations tend to be random or predictable (e.g., sensor noise due to weather conditions)
  - In security, perturbations are intentionally crafted (e.g., adversarial attacks)

Robustness and Distance for Images

- Slight rotation, stretching, or other transformations
- Change many pixels minimally (below human perception)
- Change only few pixels
- Change most pixels mostly uniformly, e.g., brightness

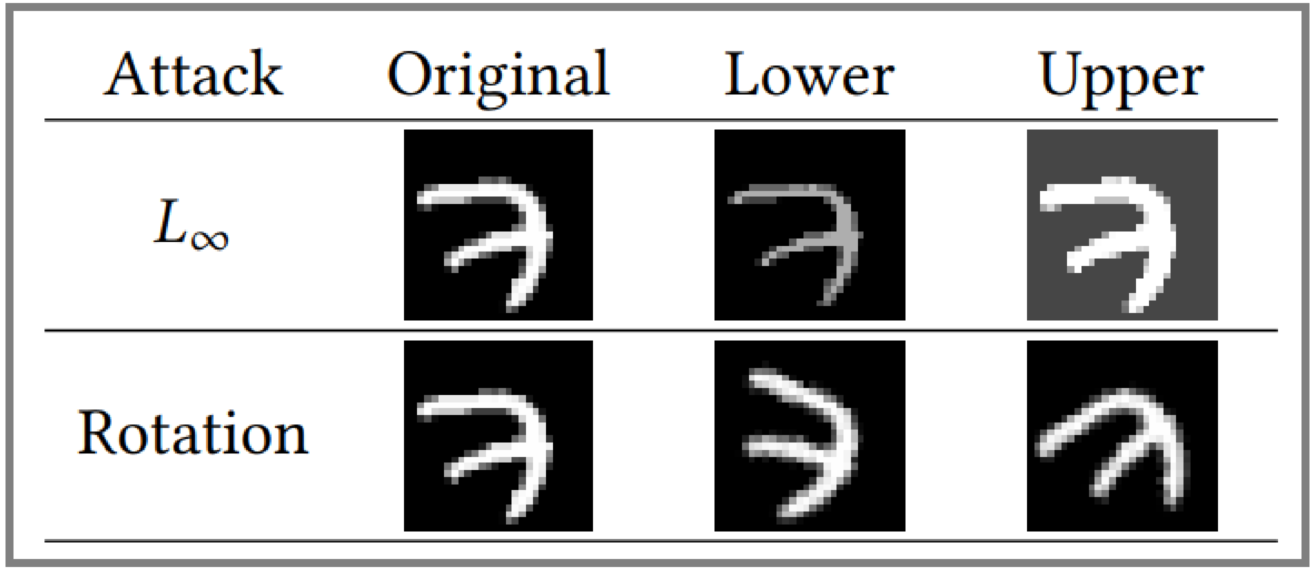


Image: *An abstract domain for certifying neural networks*. Gagandeep et al., POPL (2019).

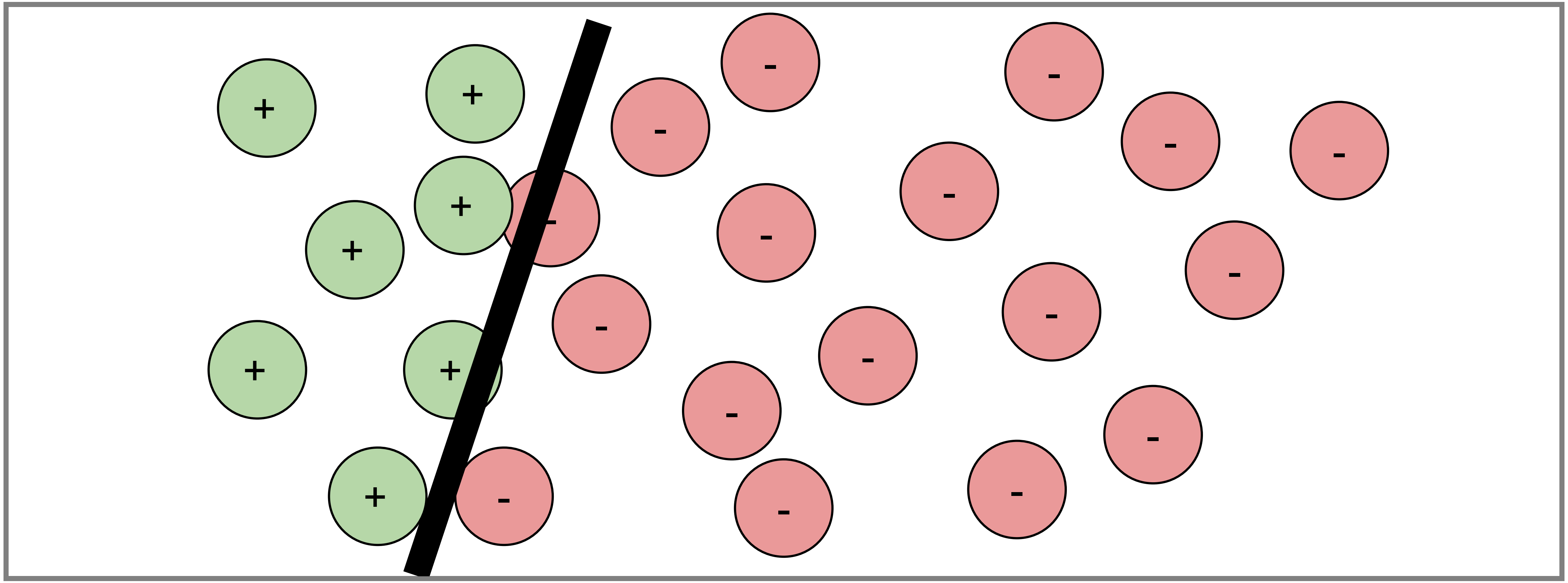
Robustness in a Safety Setting

- Does the model reliably detect stop signs?
- Also in poor lighting? In fog? With a tilted camera? Sensor noise?
- With stickers taped to the sign? (adversarial attacks)

Image: David Silver. Adversarial Traffic Signs. Blog post, 2017

No Model is Fully Robust

- Every useful model has at least one decision boundary
- Predictions near that boundary are not (and should not) be robust



Robustness of Interpretable Models

Is this model robust?

Is the prediction for a 20 year old male with 2 priors robust?

```
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 arres
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Evaluating Robustness

- Lots of on-going research (especially for DNNs)
- Formal verification
  - Constraint solving or abstract interpretation over computations in neuron activations
  - Conservative abstraction, may label robust inputs as not robust
  - Currently not very scalable
  - Example: *An abstract domain for certifying neural networks*. Gagandeep et al., POPL (2019).
- Sampling
  - Sample within distance, compare prediction to majority prediction
  - Probabilistic guarantees possible (with many queries, e.g., 100k)
  - Example: *Certified adversarial robustness via randomized smoothing*. Cohen, Rosenfeld, and Kolter, ICML (2019).
- Lots of tools that provide a robustness number

Improving Robustness for Safety

Robustness checking at inference time

- Handle inputs with non-robust predictions differently (e.g. discard or output low confidence score)
- Downside: Significantly raises cost of prediction; may not be suitable for time-sensitive applications (e.g., self-driving cars)

Design mechanisms

- Deploy redundant components for critical tasks (e.g., vision + map)
- Ensemble learning: Combine models with different biases
- Multiple, independent sensors (e.g., LiDAR + radar + cameras)

Improving Robustness for Safety

(a) Non Foggy Condition (b) Foggy Condition

Learning more robust models

- Test/think about domain-specific scenarios that might result in perturbations to model input (e.g., fogs, snow, sensor noise)
- Curate data for those abnormal scenarios or augment training data with transformed inputs

Image: *Automated driving recognition technologies for adverse weather conditions.* Yoneda et al., IATSS Research (2019).

Breakout: Robustness

Scenario: Medical use of transcription service, dictate diagnoses and prescriptions

As a group, tagging members, post to `#lecture`:

> 1. *What safety concerns can you anticipate?*
> 2. *What notion of robustness are you concerned about (i.e., what distance function)?*
> 3. *How could you use robustness to improve the product (i.e., when/how to check robustness)?*

Reality-Based Safety

Defining Safety

Prevention of a system failure or malfunction that results in:

- Death or serious injury to people
- Loss or severe damage to equipment/property
- Harm to the environment or society

Safety is a system concept

- Can't talk about software/ML being "safe"/"unsafe" on its own
- Safety is defined in terms of its effect on the **environment**

Safety != Reliability

Reliability = absence of defects, mean time between failure

Safety = prevents accidents, harms

Can build safe systems from unreliable components (e.g. redundancy, safeguards)

System may be unsafe despite reliable components (e.g. stronger gas tank causes more severe damage in incident)

Accuracy and robustness are about reliability!

Safety of AI-Enabled Systems

Dr. Emily Slackerman Ackerman
@EmilyEAckerman · **Follow**

i (in a wheelchair) was just trapped *on* forbes ave by one of these robots, only days after their independent roll out. i can tell that as long as they continue to operate, they are going to be a major accessibility and safety issue. [thread]

pittnews.com
Everything we know about the Starship food delivery robots
The white, 2-foot tall battery-powered delivery robots will be…

7:27 PM · Oct 21, 2019

Read the full conversation on Twitter

4K Reply Copy link

Safety of AI-Enabled Systems

skoops 🤨😐😵
@skoops · **Follow**

The @netatmo servers are down and twitter is already full of freezing people not able to control their heating :D (via [protected]) / cc @internetofshit

8:15 PM · Nov 22, 2018

2K Reply Copy link

Read 67 replies

Safety is a broad concept

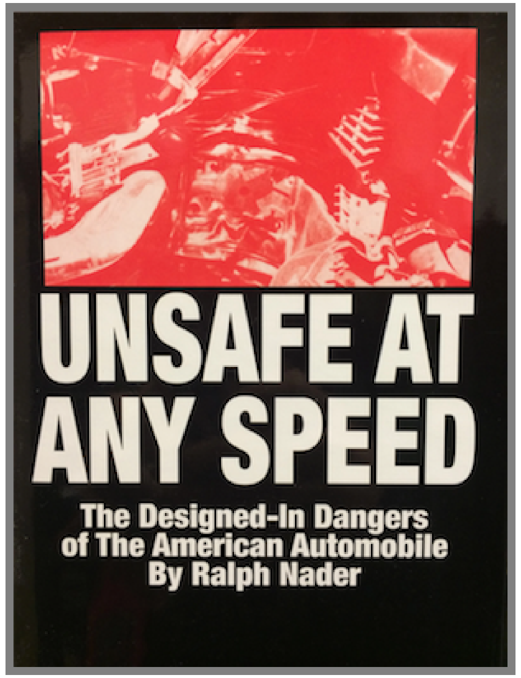
Not just physical harms/injuries to people

Includes harm to mental health

Includes polluting the environment, including noise pollution

Includes harm to society, e.g. poverty, polarization

How did traditional vehicles become safer?



National Traffic & Motor Safety Act (1966):

- Mandatory design changes (head rests, shatter-resistant windshields, safety belts)
- Road improvements (center lines, reflectors, guardrails)

Improving Safety of ML-Enabled Systems

Anticipate problems (hazard analysis, FTA, FMEA, HAZOP, ...)

Anticipate the existence of unanticipated problems

Plan for mistakes, design mitigations

- Human in the loop
- Undoable actions, failsoft
- Guardrails
- Mistaked detection
- Redundancy, ...

Improve reliability (accuracy, robustness)

Challenge: Edge/Unknown Cases

- Gaps in training data; ML unlikely to cover all unknown cases
- Why is this a unique problem for AI? What about humans?

Safety Engineering

Safety Engineering: An engineering discipline which assures that engineered systems provide acceptable levels of safety.

Typical safety engineering process:

- Identify relevant hazards & safety requirements
- Identify potential root causes for hazards
- For each hazard, develop a mitigation strategy
- Provide evidence that mitigations are properly implemented

Demonstrating and Documenting Safety

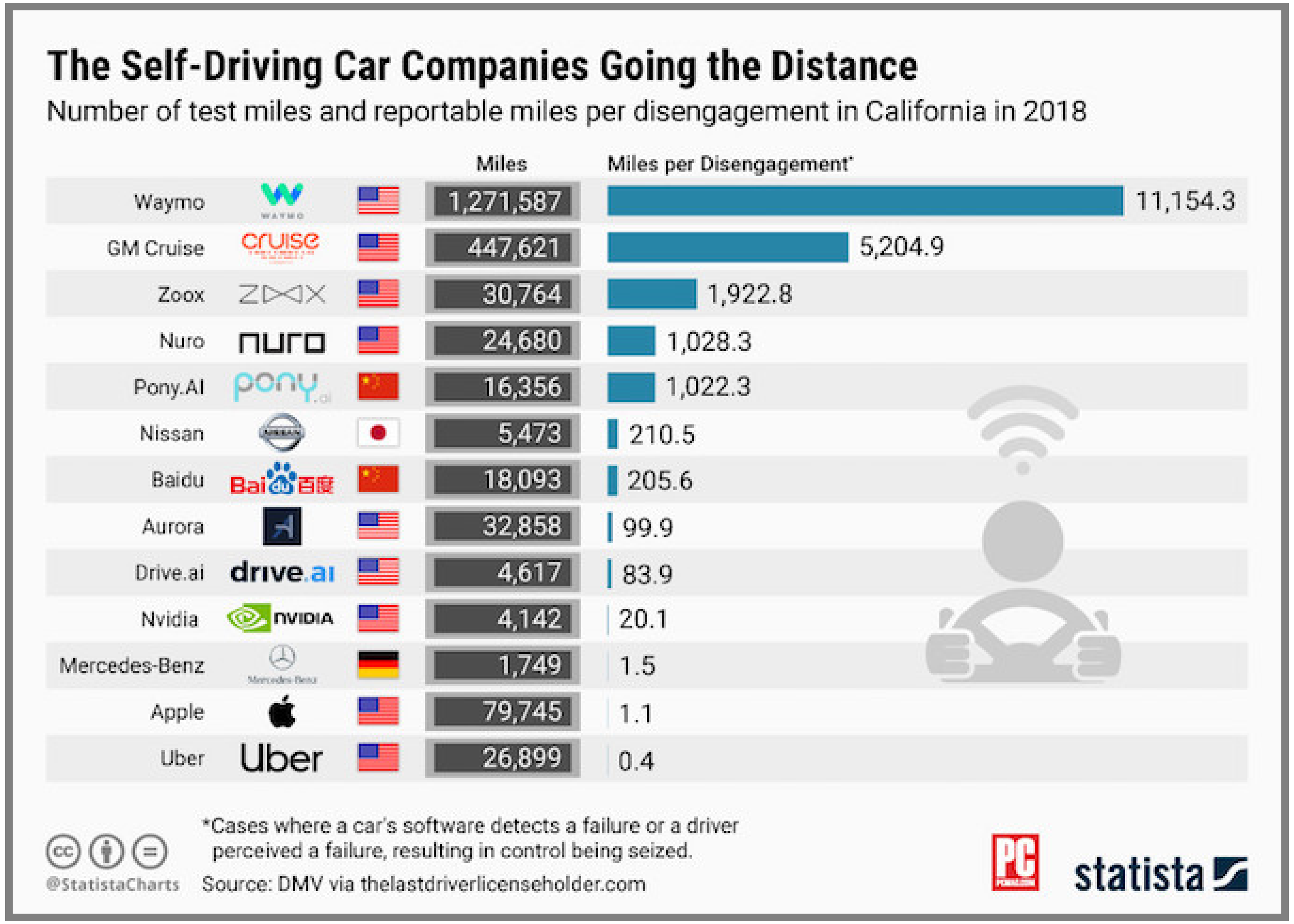
Demonstrating Safety

Two main strategies:

1. **Evidence of safe behavior in the field**
   - Extensive field trials
   - Usually expensive
2. **Evidence of responsible (safety) engineering process**
   - Process with hazard analysis, testing mitigations, etc
   - Not sufficient to assure safety

Most standards require both

Demonstrating Safety

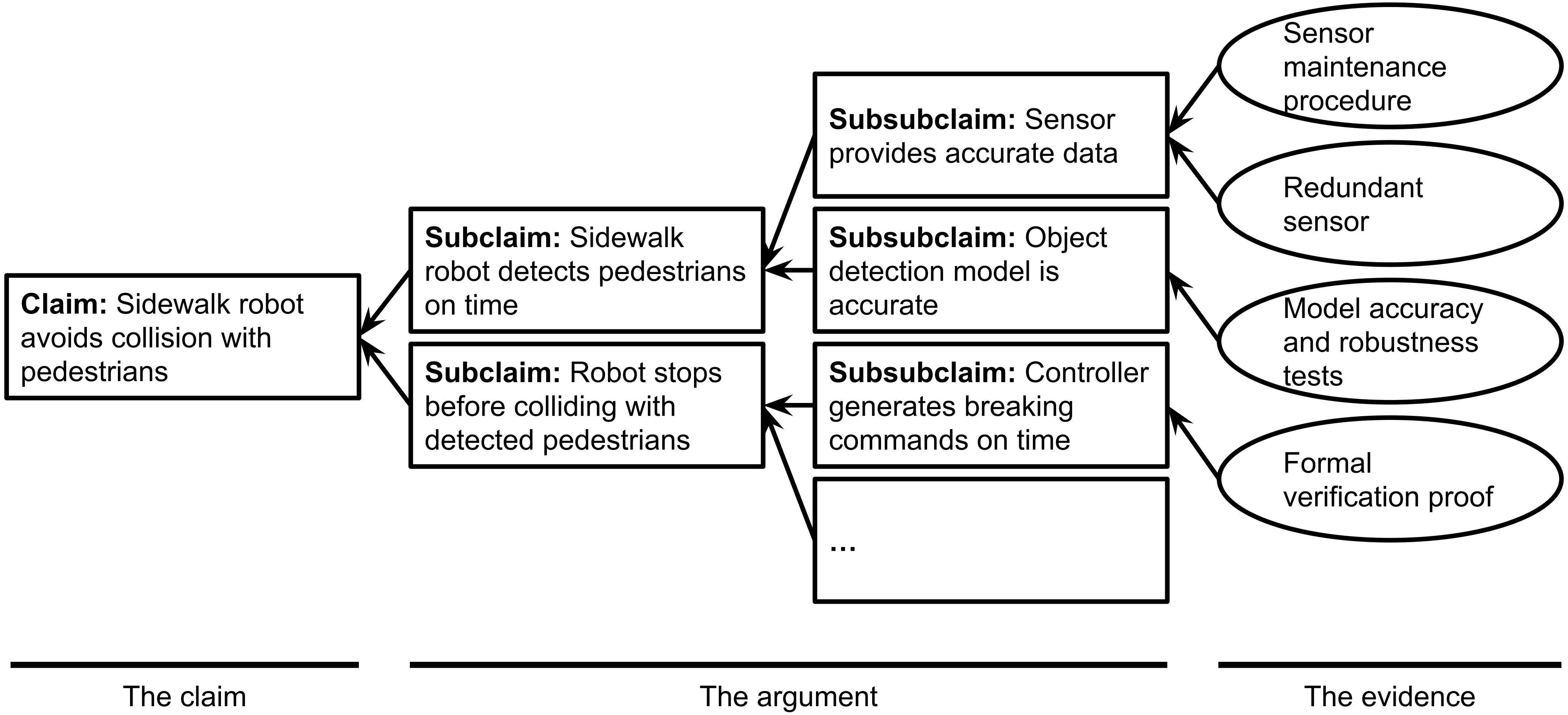


How do we demonstrate to a third-party that our system is safe?

Safety & Certification Standards

- Guidelines & recommendations for achieving an acceptable level of safety
- Examples: DO-178C (airborne systems), ISO 26262 (automotive), IEC 62304 (medical software), Common Criteria (security)
- Typically, **prescriptive & process-oriented**
  - Recommends use of certain development processes
  - Requirements specification, design, hazard analysis, testing, verification, configuration management, etc.,
- Limitations
  - Most not designed to handle ML systems (exception: UL 4600)
  - Costly to satisfy & certify, but effectiveness unclear (e.g., many FDA-certified products recalled due to safety incidents)
- Good processes are important, but not sufficient; provides only indirect evidence for system safety

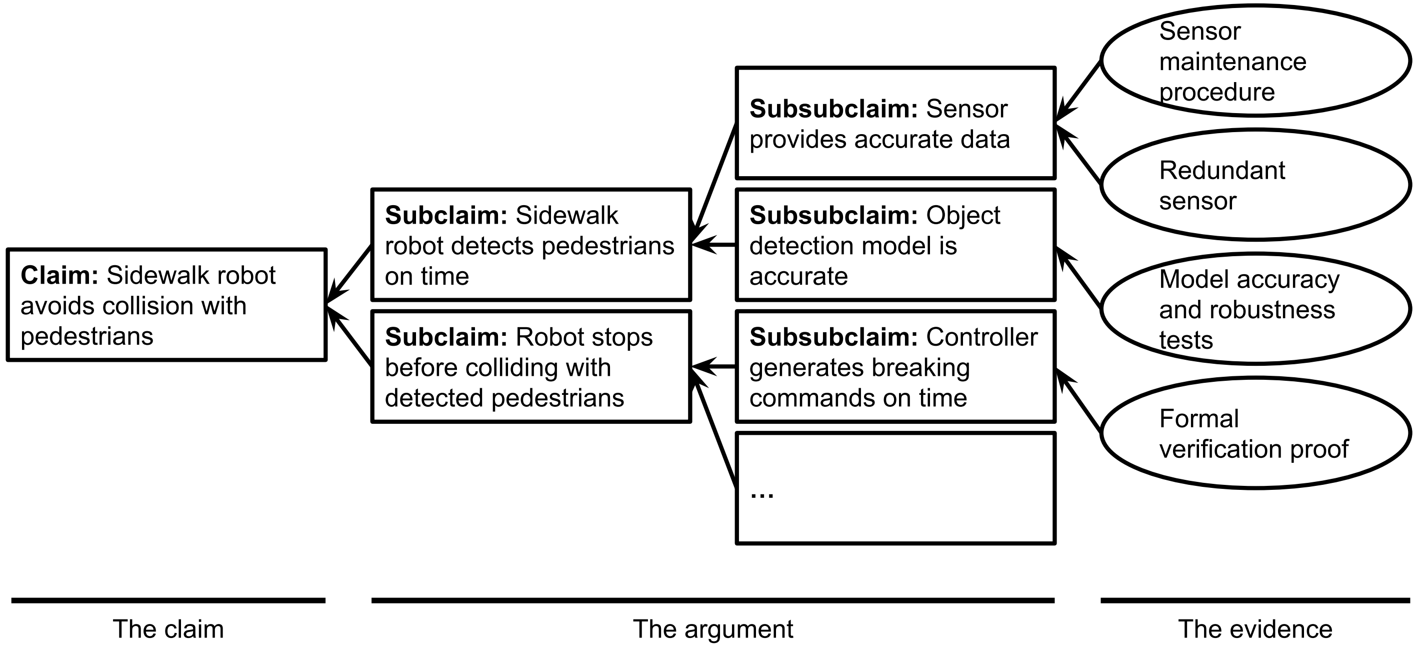
Documenting Safety with Assurance (Safety) Cases



Assurance (Safety) Cases

- An explicit argument that a system achieves a desired safety requirement, along with supporting evidence
- Structure:
  - Argument: A top-level claim decomposed into multiple sub-claims
  - Evidence: Testing, software analysis, formal verification, inspection, expert opinions, design mechanisms...

Assurance Cases: Example



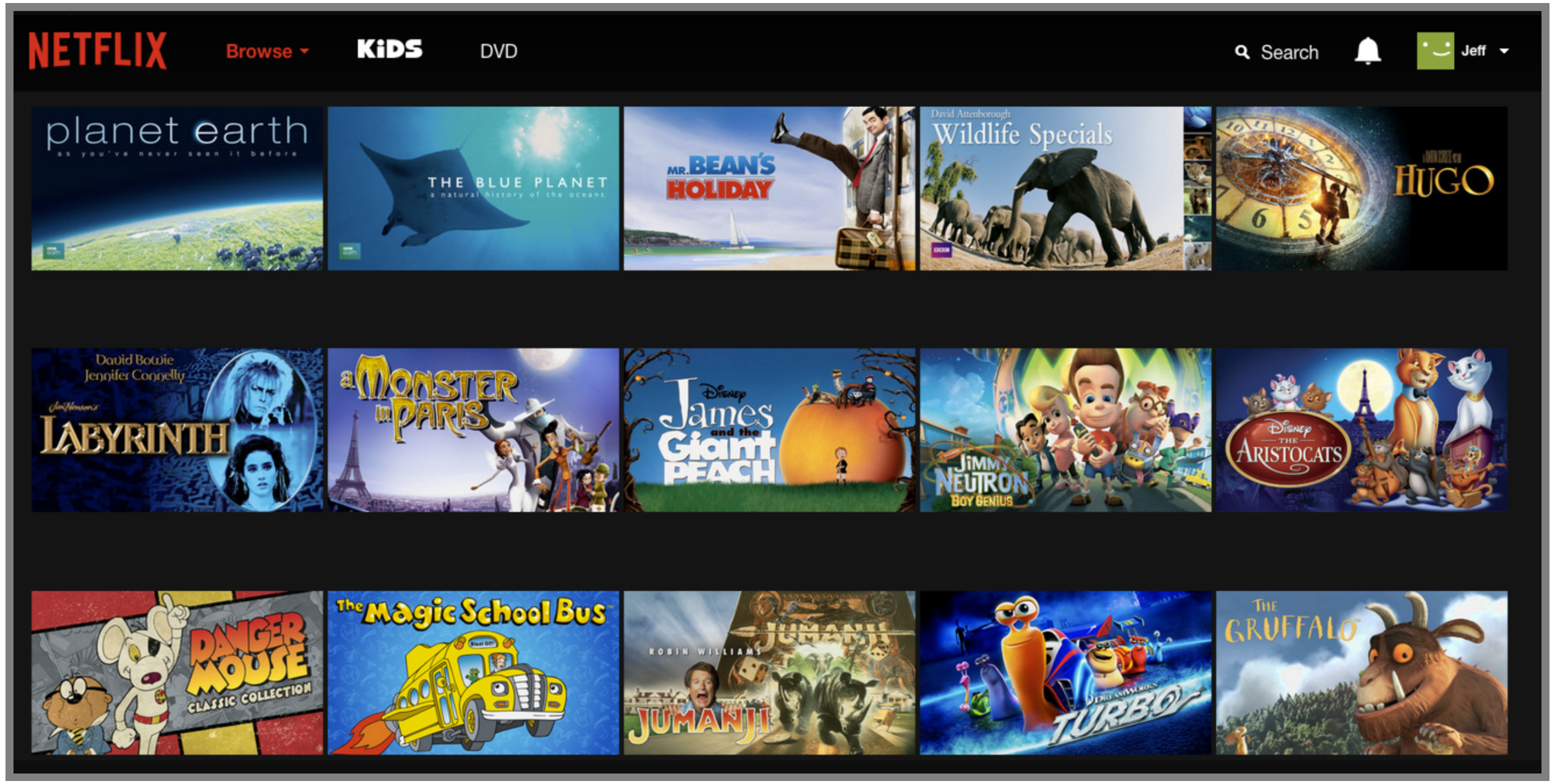
Questions to think about:

- Do sub-claims imply the parent claim?
- Am I missing any sub-claims?
- Is the evidence strong enough to discharge a leaf claim?

Assurance Cases: Example

Aurora Safety Case

Exercise: Assurance Case for Recommender



Build a safety case to argue that your movie recommendation system provides at least 95% availability. Include evidence to support your

Assurance Cases: Benefits & Limitations

- Provides an explicit structure to the safety argument
  - Easier to navigate, inspect, and refute for third-party auditors
  - Provides traceability between system-level claims & low-level evidence
  - Can also be used for other types of system quality (security, reliability, etc.,)
- Challenges and pitfalls
  - Informal links between claims & evidence, e.g., Does the sub-claims actually imply the top-level claim?
  - Effort in constructing the case & evidence: How much evidence is enough?
  - System evolution: If system changes, must reproduce the case & evidence
- Tools for building & analyzing safety cases available
  - e.g., ASCE/GSN from Adelard
  - But ultimately, can't replace domain knowledge & critical thinking

Beyond Traditional Safety Critical Systems

Beyond Traditional Safety Critical Systems

- Recall: Legal vs ethical
- Safety analysis not only for regulated domains (nuclear power plants, medical devices, planes, cars, ...)
- Many end-user applications have a safety component

Examples?

Mental Health

healthline

SUBSCRIBE

HEALTH NEWS ✓ Fact Checked

The FOMO Is Real: How Social Media Increases Depression and Loneliness

Written by Gigen Mammoser on December 10, 2018

New research reveals how social media platforms like Facebook can greatly affect your mental health.

IoT


skoops
@skoops
Follow
The @netatmo servers are down and twitter is already full of freezing people not able to control their heating :D (via [protected]) / cc @internetofshit
Kiran vadgama
@kiran_vadgama
Andy Mc
@ITakeSugar
Are your servers do

Addiction

NO MERCY NO MALICE

Robinhood Has Gamified Online Trading Into an Addiction

Tech's obsession with addiction will hurt us all

Scott Galloway Follow

Jun 23 · 7 min read ★

Warning: This post contains a discussion of suicide.

ddiction is the inability to stop consuming a chemical or pursuing an

Society: Unemployment Engineering / Deskilling

Speaker notes

The dangers and risks of automating jobs.

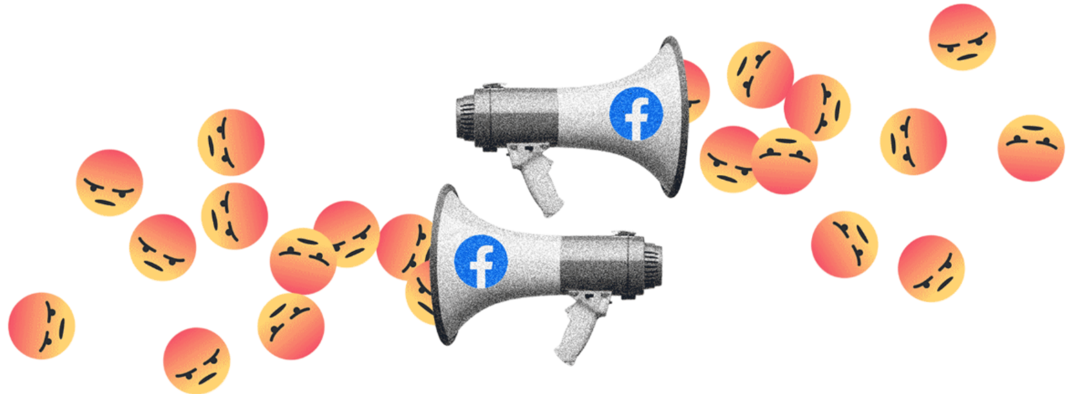
Discuss issues around automated truck driving and the role of jobs.

See for example: Andrew Yang. The War on Normal People. 2019

Society: Polarization

THE WALL STREET JOURNAL.

SUBSCRIBE SIGN IN



TECH

Facebook Executives Shut Down Efforts to Make the Site Less Divisive

The social-media giant internally studied how it polarizes users, then largely shelved the research

By Jeff Horwitz and Deepa Seetharaman

May 26, 2020 11:38 am ET

Speaker notes

Recommendations for further readings: https://www.nytimes.com/column/kara-swisher, https://podcasts.apple.com/us/podcast/recode-decode/id1011668648

Also isolation, Cambridge Analytica, collaboration with ICE, ...

Environmental: Energy Consumption

Exercise

Look at apps on your phone. Which apps have a safety risk and use machine learning?

Consider safety broadly: including stress, mental health, discrimination, and environment pollution

Takeaway

- Many systems have safety concerns
- ... not just nuclear power plants, planes, cars, and medical devices
- Do the right thing, even without regulation
- Consider safety broadly: including stress, mental health, discrimination, and environment pollution
- Start with requirements and hazard analysis

Designing for Safety

See Lecture **Planning for Mistakes**

Safety Assurance with ML Components

- Consider ML components as unreliable, at most probabilistic guarantees
- Testing, testing, testing (+ simulation)
  - Focus on data quality & robustness
- *Adopt a system-level perspective!*
- Consider safe system design with unreliable components
  - Traditional systems and safety engineering
  - Assurance cases
- Understand the problem and the hazards
  - System level, goals, hazard analysis, world vs machine
  - Specify *end-to-end system behavior* if feasible

Summary

- Defining safety: absence of harm to people, property, and environment -- consider broadly; safety != reliability
- *Adopt a safety mindset!*
- Assume all components will eventually fail in one way or another, especially ML components
- Hazard analysis to identify safety risks and requirements; classic safety design at the system level
- Model robustness can help with some problems
- AI alignment: AI goals are difficult to specify precisely; susceptible to negative side effect & reward hacking

Further Readings

- Borg, Markus, Cristofer Englund, Krzysztof Wnuk, Boris Duran, Christoffer Levandowski, Shenjian Gao, Yanwen Tan, Henrik Kaijser, Henrik Lönn, and Jonas Törnqvist. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." Journal of Automotive Software Engineering. 2019
- Leveson, Nancy G. Engineering a safer world: Systems thinking applied to safety. The MIT Press, 2016.
- Salay, Rick, and Krzysztof Czarnecki. "Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262." arXiv preprint arXiv:1808.01614 (2018).
- Mohseni, Sina, Mandar Pitale, Vasu Singh, and Zhangyang Wang. "Practical Solutions for Machine Learning Safety in Autonomous Vehicles." SafeAI workshop at AAAI'20, (2020).
- Huang, Xiaowei, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability." Computer Science Review 37 (2020).
- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).