SAFETY

Eunsuk Kang

Required Reading: Practical Solutions for Machine Learning Safety in Autonomous Vehicles. S. Mohseni et al., 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

SAFETY

DEFINING 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

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 being "safe"/"unsafe" on its own
  - Safety is defined in terms of its effect on the **environment**

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 being "safe"/"unsafe" on its own
    - Safety is defined in terms of its effect on the **environment**
- Safety != Reliability
    - Can build safe systems from unreliable components (e.g. redundancies)
    - Reliable components may be unsafe (e.g. stronger gas tank causes more severe damage in incident)

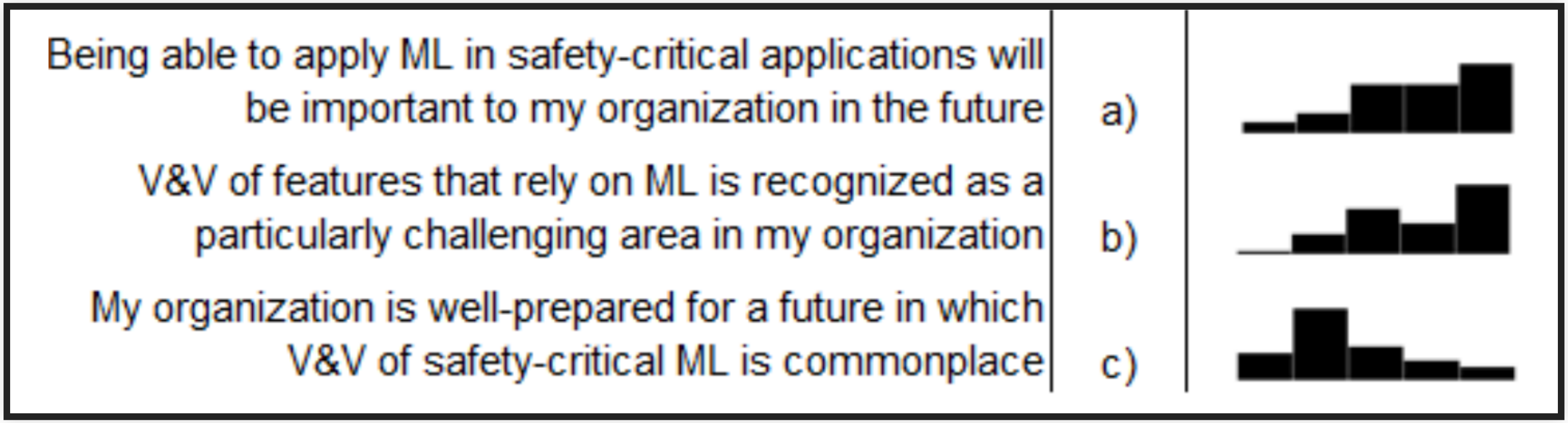
SAFETY OF AI-ENABLED SYSTEMS

SAFETY OF AI-ENABLED SYSTEMS

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

SAFETY CHALLENGE WIDELY RECOGNIZED

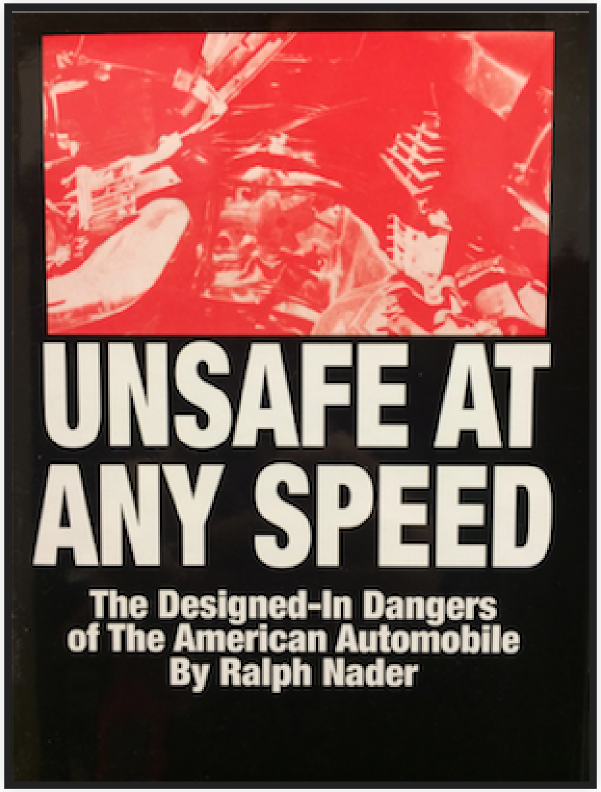


(survey among automotive engineers)

Borg, Markus, et al. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." arXiv preprint arXiv:1812.05389 (2018).

CASE STUDY: SELF-DRIVING CAR

HOW DID TRADITIONAL VEHICLES BECOME SAFE?



- National Traffic & Motor Safety Act (1966): Mandatory design changes (head rests, shatter-resistant windshields, safety belts); road improvements (center lines, reflectors, guardrails)

AUTONOMOUS VEHICLES: WHAT'S DIFFERENT?

Ford Taps the Brakes on the Arrival of Self-Driving Cars

HYPE CYCLE —

The hype around driverless cars came crashing down in 2018

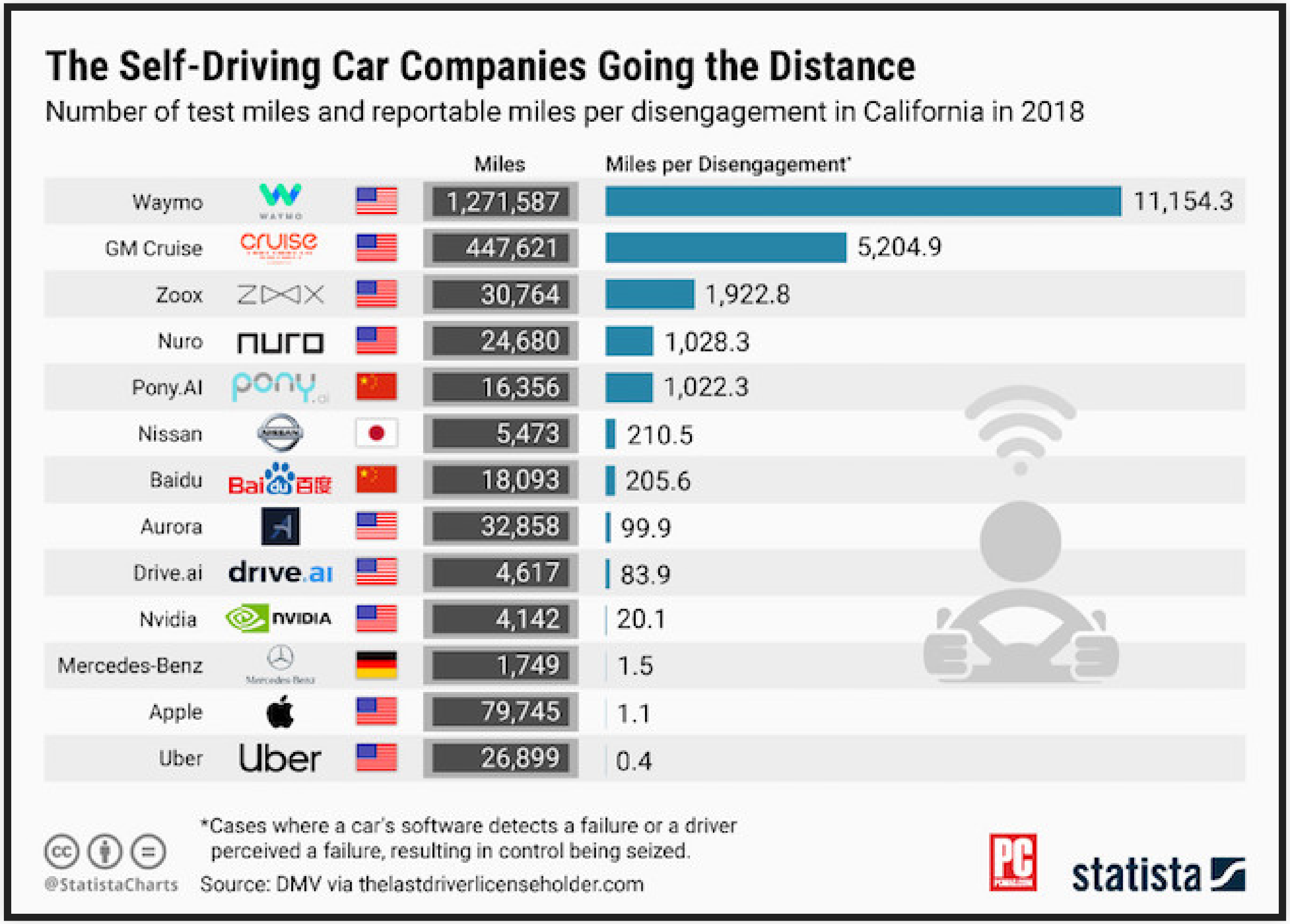
Top Toyota expert throws cold water on the driverless car hype

- In traditional vehicles, humans ultimately responsible for safety
  - Some safety features (lane keeping, emergency braking) designed to help & reduce risks
  - i.e., safety = human control + safety mechanisms
- Use of AI in autonomous vehicles: Perception, control, routing, etc.,
  - Inductive training: No explicit requirements or design insights
  - **Can ML achieve safe design solely through lots of data?**

CHALLENGE: EDGE/UNKNOWN CASES

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

DEMONSTRATING SAFETY



More miles tested => safer?

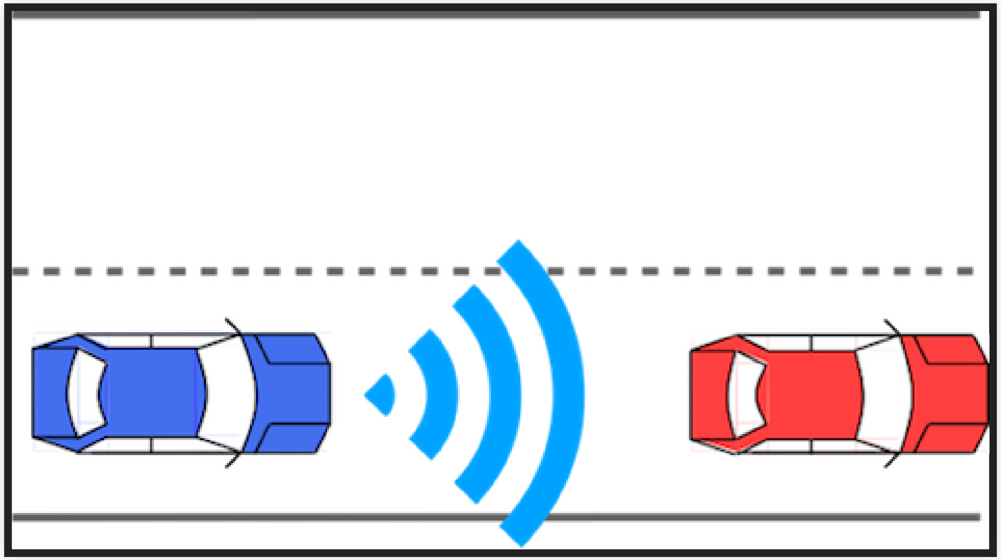
APPROACH FOR DEMONSTRATING SAFETY

- 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

HAZARD ANALYSIS

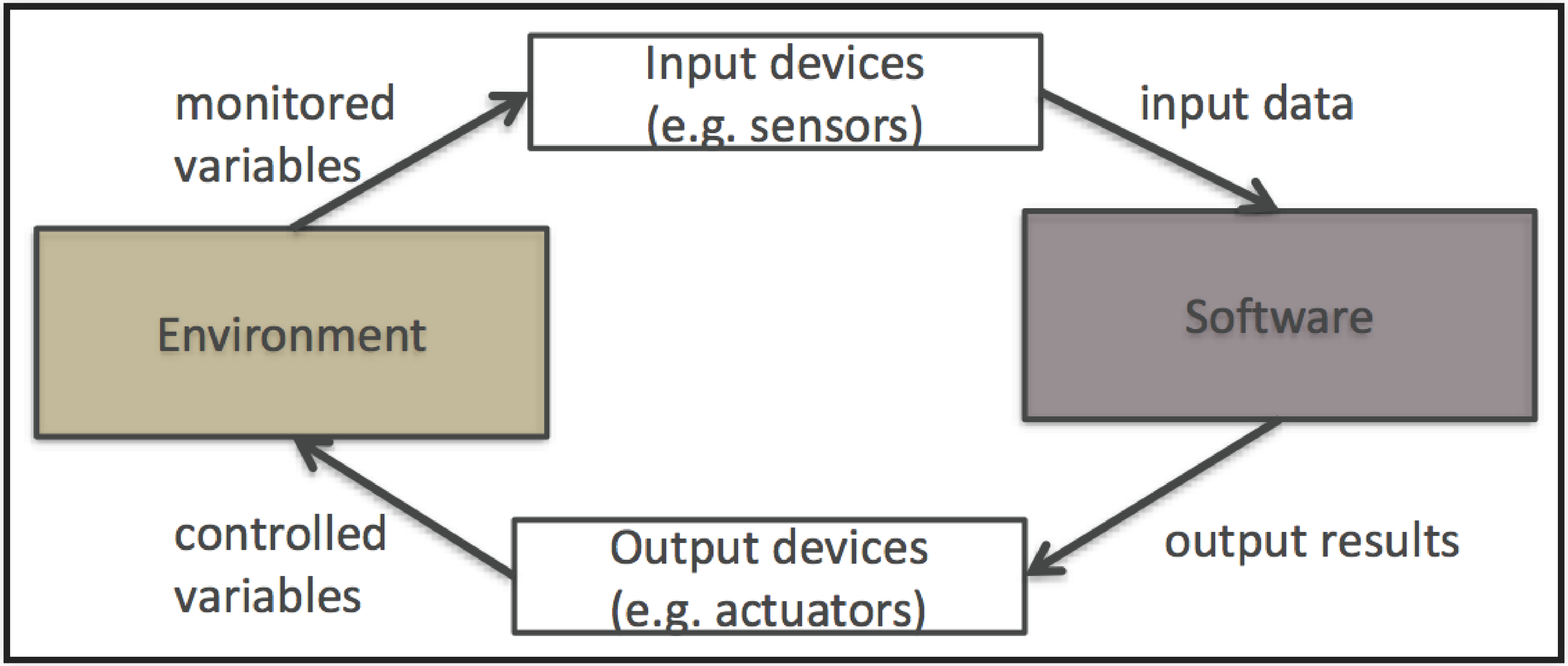
(system level!)

WHAT IS HAZARD ANALYSIS?



- **Hazard**: A condition or event that may result in undesirable outcome
  - e.g., "Ego vehicle is in risk of a collision with another vehicle."
- **Safety requirement**: Intended to eliminate or reduce one or more hazards
  - "Ego vehicle must always maintain some minimum safe distance to the leading vehicle."
- **Hazard analysis**: Methods for identifying hazards & potential root causes

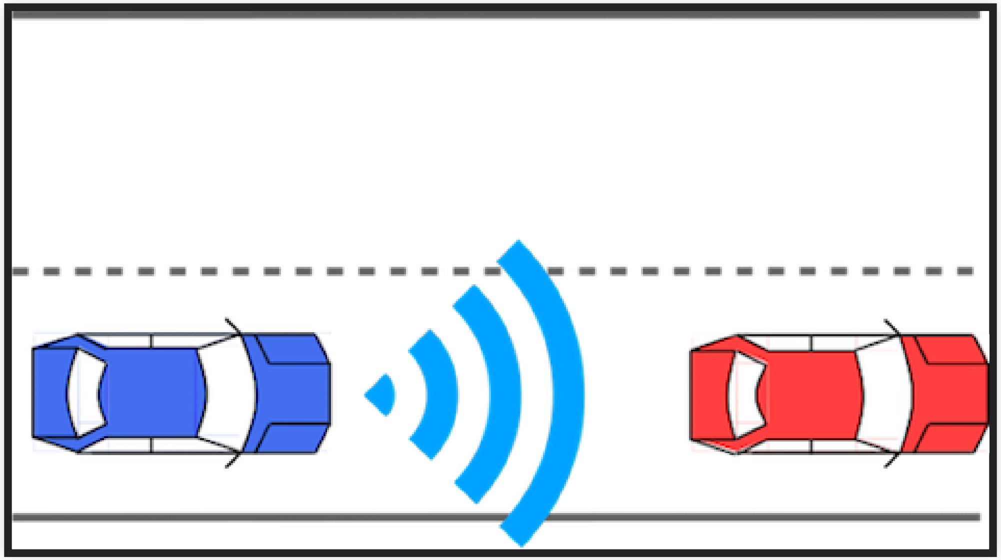
RECALL: WORLD VS MACHINE



Software is not unsafe on its own; the control signals it generates may be

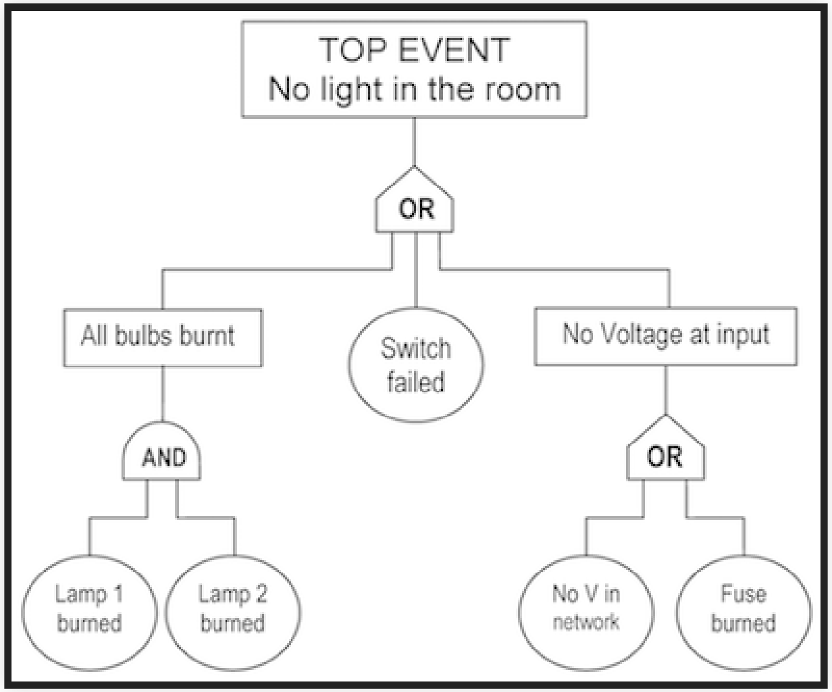
Root of unsafety usually in wrong requirements & environmental assumptions

RECALL: REQUIREMENT VS SPECIFICATION



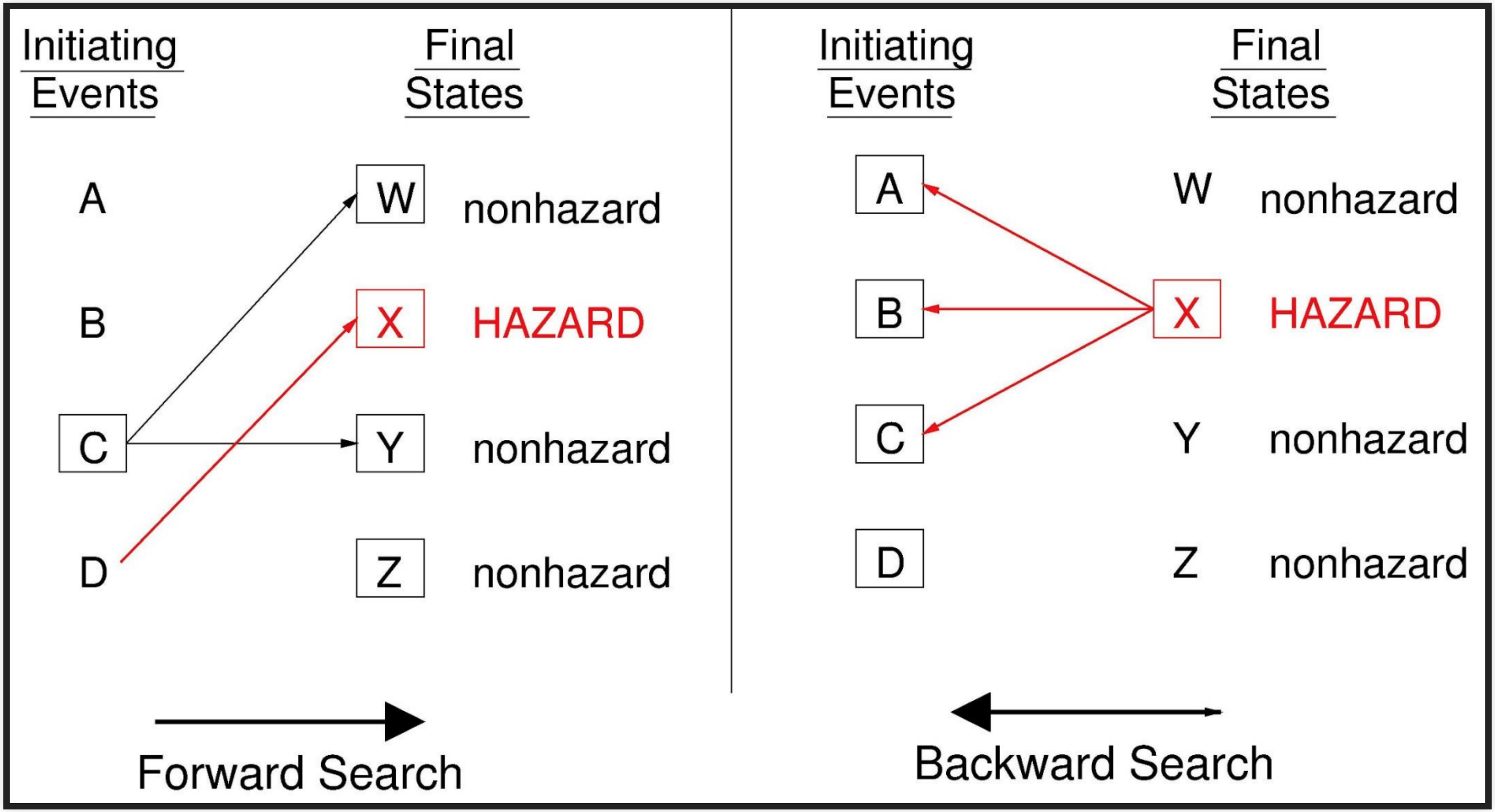
- **REQ:** Ego vehicle must always maintain some minimum safe distance to the leading vehicle.
- **ENV**: Engine is working as intended; sensors are providing accurate information about the leading car (current speed, distance...)
- **SPEC**: Depending on the sensor readings, the controller must issue an actuator command to accelerate/decelerate the vehicle as needed.

REVIEW: FAULT TREE ANALYSIS (FTA)



- Top-down, **backward** search method for root cause analysis
    - Start with a given hazard (top event), derive a set of components faults (basic events)
    - Compute minimum cutsets as potential root causes
    - **Q. But how do we identify relevant hazards in the first place?**

FORWARD VS BACKWARD SEARCH

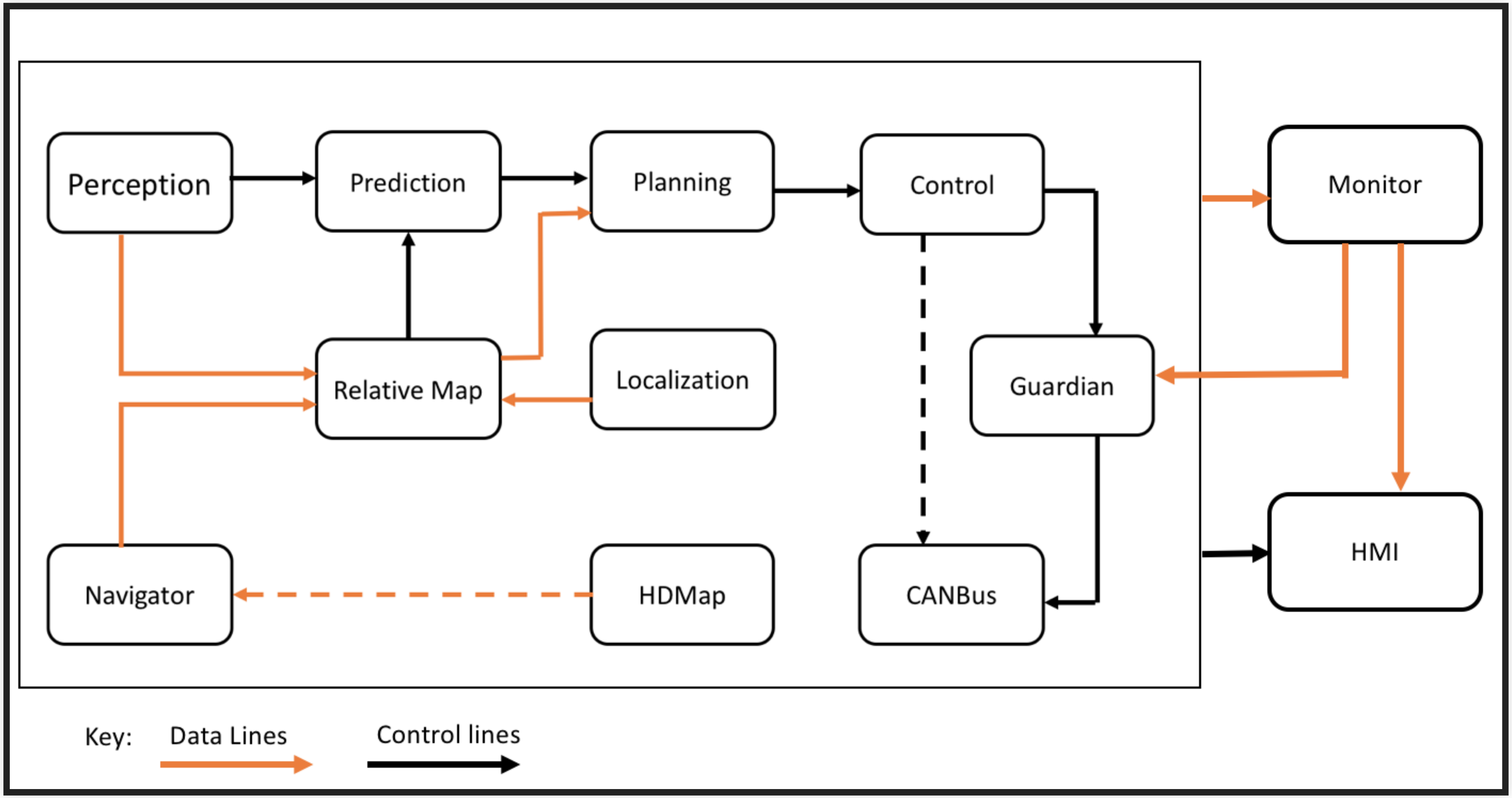


FAILURE MODE AND EFFECTS ANALYSIS (FMEA)

	Function	Potential Failure Mode	Potential Effect(s) of Failure	SEV i	Potential Cause(s) of Failure	OCC i	Current Design Controls (Prevention)	Current Design Controls (Detection)	DET i	RPN i	Recommended Action(s)
1	Provide required levels of radiation	Radiation level too high for the required intervention	Over radiation of the patients.		Technician did not set the radiation at the right level.			Current algorithm resets to normal levels after imaging each patient.			Modify software to alert technician to unusually high radiation levels before activating.
2		Radiation at lower level than required	Patient fails to receive enough radiation.		Software does not respond to hardware mechanical setting.			Failure detection included in software			Include visual / audio alarm in the code when lack of response.
3											Improve recovery protocol.
4	Protect patients from unexpected high radiation	Higher radiation than required	Radiation burns		sneak paths in software			Shut the system if radiation level does not match the inputs.			Perform traceability matrix.

- A **forward search** technique to identify potential hazards
- For each function, (1) enumerate possible *failure modes* (2) possible safety impact (*effects*) and (3) mitigation strategies.
- Widely used in aeronautics, automotive, healthcare, food services, semiconductor processing, and (to some extent) software

FMEA EXAMPLE: AUTONOMOUS VEHICLES



- Architecture of the Apollo autonomous driving platform

FMEA EXAMPLE: AUTONOMOUS VEHICLES

Component	Failure Mode	Failure Effects	Detection	Mitigation
Perception	?	?	?	?
Perception	?	?	?	?
Lidar Sensor	Mechanical failure	Inability to detect objects	Monitor	Switch to manual control mode
...	...	...	...	...

FMEA EXAMPLE: AUTONOMOUS VEHICLES

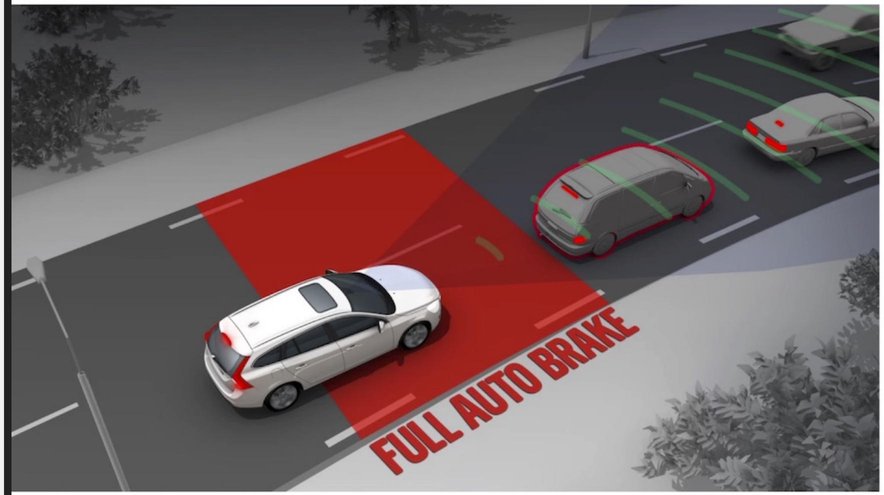
Component	Failure Mode	Failure Effects	Detection	Mitigation
Perception	Failure to detect an object	Risk of collision	Human operator (if present)	Deploy secondary classifier
Perception	Detected but misclassified	"	"	"
Lidar Sensor	Mechanical failure	Inability to detect objects	Monitor	Switch to manual control mode
...	...	...	...	...

HAZARD AND OPERABILITY STUDY (HAZOP)

Guide Word	Meaning
NO OR NOT	Complete negation of the design intent
MORE	Quantitative increase
LESS	Quantitative decrease
AS WELL AS	Qualitative modification/increase
PART OF	Qualitative modification/decrease
REVERSE	Logical opposite of the design intent
OTHER THAN / INSTEAD	Complete substitution
EARLY	Relative to the clock time
LATE	Relative to the clock time
BEFORE	Relating to order or sequence
AFTER	Relating to order or sequence

- A **forward search** method to identify potential hazards
- For each component, use a set of **guide words** to generate possible deviations from expected behavior
- Consider the impact of each generated deviation: Can it result in a system-level hazard?

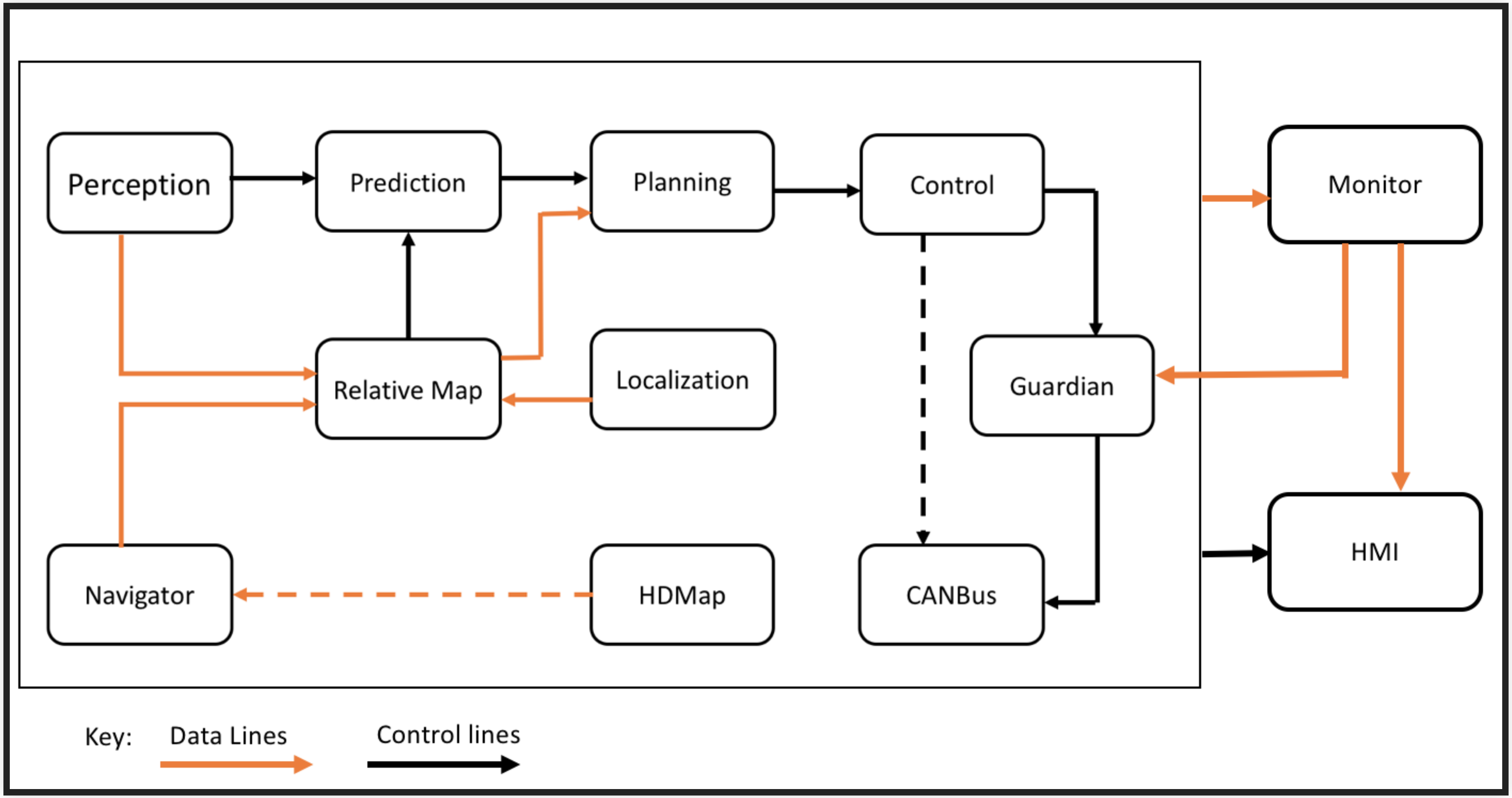
HAZOP EXAMPLE: EMERGENCY BRAKING (EB)



Guide Word	Meaning
NO OR NOT	Complete negation of the design intent
MORE	Quantitative increase
LESS	Quantitative decrease
AS WELL AS	Qualitative modification/increase
PART OF	Qualitative modification/decrease
REVERSE	Logical opposite of the design intent
OTHER THAN / INSTEAD	Complete substitution
EARLY	Relative to the clock time
LATE	Relative to the clock time
BEFORE	Relating to order or sequence
AFTER	Relating to order or sequence

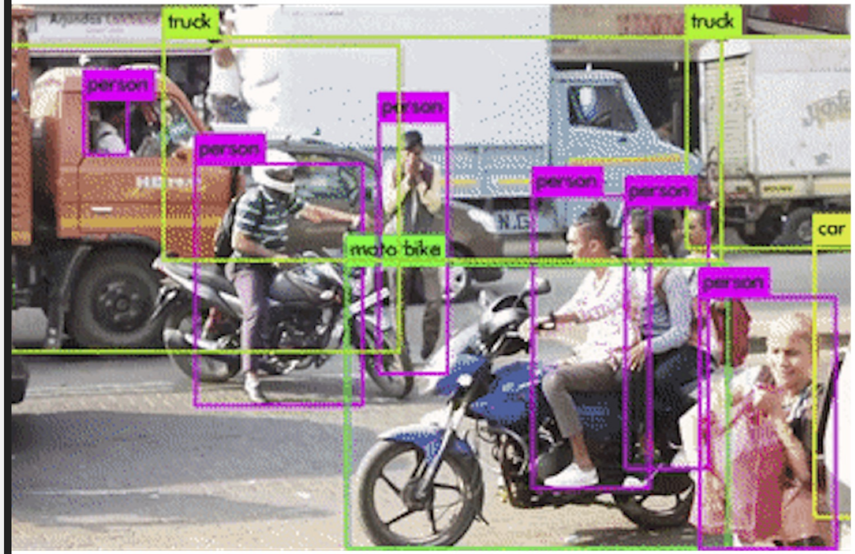
- Specification: EB must apply a maximum braking command to the engine.
  - **NO OR NOT**: EB does not generate any braking command.
  - **LESS**: EB applies less than max. braking.
  - **LATE**: EB applies max. braking but after a delay of 2 seconds.
  - **REVERSE**: EB generates an acceleration command instead of braking.
  - **BEFORE**: EB applies max. braking before a possible crash is detected.

HAZOP EXERCISE: AUTONOMOUS VEHICLES



- Architecture of the Apollo autonomous driving platform

HAZOP EXERCISE: PERCEPTION



Guide Word	Meaning
NO OR NOT	Complete negation of the design intent
MORE	Quantitative increase
LESS	Quantitative decrease
AS WELL AS	Qualitative modification/increase
PART OF	Qualitative modification/decrease
REVERSE	Logical opposite of the design intent
OTHER THAN / INSTEAD	Complete substitution
EARLY	Relative to the clock time
LATE	Relative to the clock time
BEFORE	Relating to order or sequence
AFTER	Relating to order or sequence

- What is the specification of the perception component?
- Use HAZOP to answer:
    - What are possible deviations from the specification?
    - What are potential hazards resulting from these deviations?

HAZOP: BENEFITS & LIMITATIONS

Guide Word	Meaning
NO OR NOT	Complete negation of the design intent
MORE	Quantitative increase
LESS	Quantitative decrease
AS WELL AS	Qualitative modification/increase
PART OF	Qualitative modification/decrease
REVERSE	Logical opposite of the design intent
OTHER THAN / INSTEAD	Complete substitution
EARLY	Relative to the clock time
LATE	Relative to the clock time
BEFORE	Relating to order or sequence
AFTER	Relating to order or sequence

- Easy to use; encourages systematic reasoning about component faults
- Can be combined with FTA/FMEA to generate faults (i.e., basic events in FTA)
- Potentially labor-intensive; relies on engineer's judgement
- Does not guarantee to find all hazards (but also true for other techniques)

REMARKS: HAZARD ANALYSIS

- None of these methods guarantee completeness
  - You may still be missing important hazards, failure modes
- Intended as structured approaches to thinking about failures
  - But cannot replace human expertise and experience
- When available, leverage prior domain knowledge
  - **Safety standards**: A set of design and process guidelines for establishing safety
  - ISO 26262, ISO 21448, IEEE P700x, etc.,
  - Most do not consider AI; new standards being developed (e.g., UL 4600)

MODEL ROBUSTNESS

DEFINING ROBUSTNESS:

- A prediction for x is robust if the outcome is stable under minor perturbations of the input
  - $\forall x'. d(x, x') < \epsilon \Rightarrow f(x) = f(x')$
  - distance function d and permissible distance ϵ depends on problem
- A model is robust if most predictions are robust

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

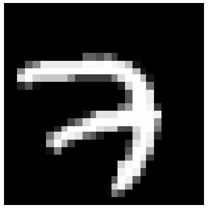
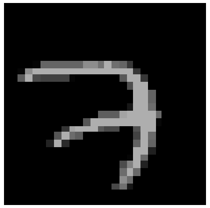
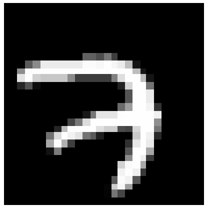
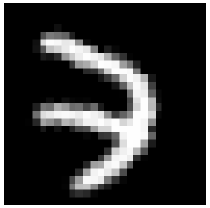
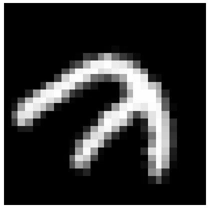
Attack	Original	Lower	Upper
L_∞			
Rotation			

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

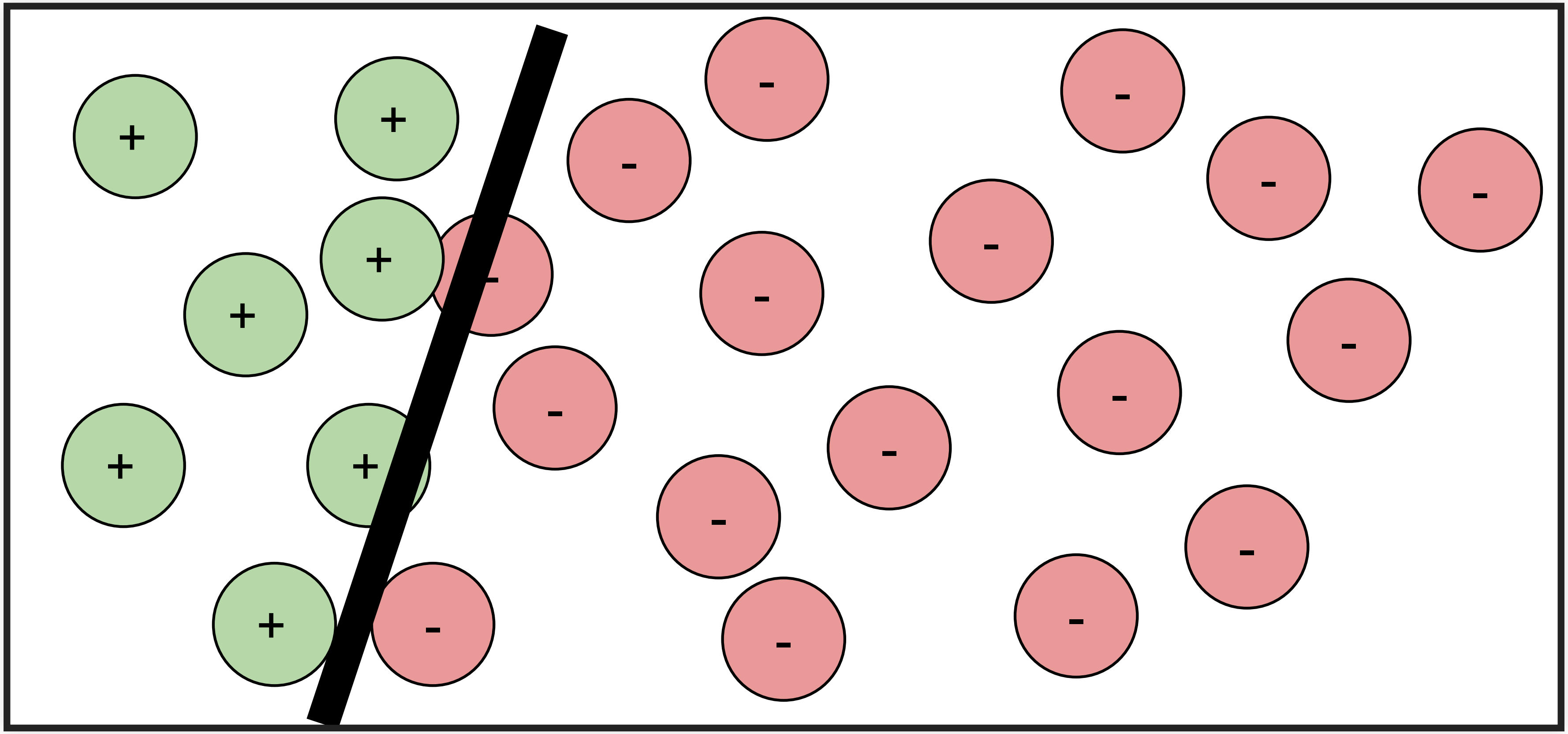
ROBUSTNESS IN A SAFETY SETTING

- 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)

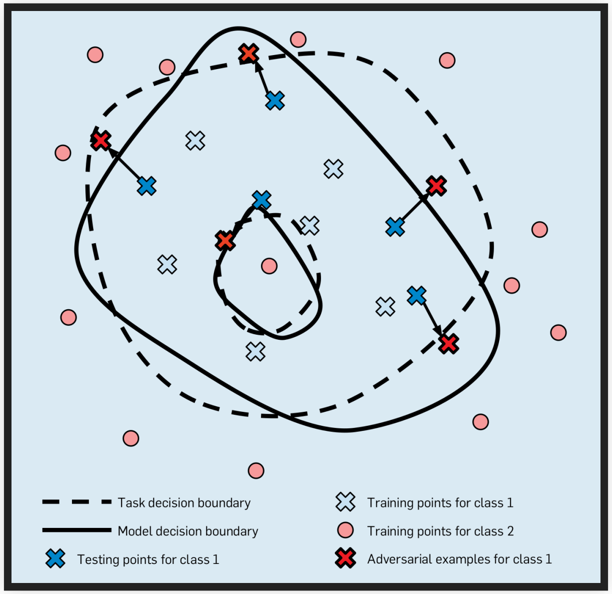
Image: David Silver. Adversarial Traffic Signs. Blog post, 2017

NO MODEL IS FULLY ROBUST

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



TASK DECISION BOUNDARY VS MODEL BOUNDARY



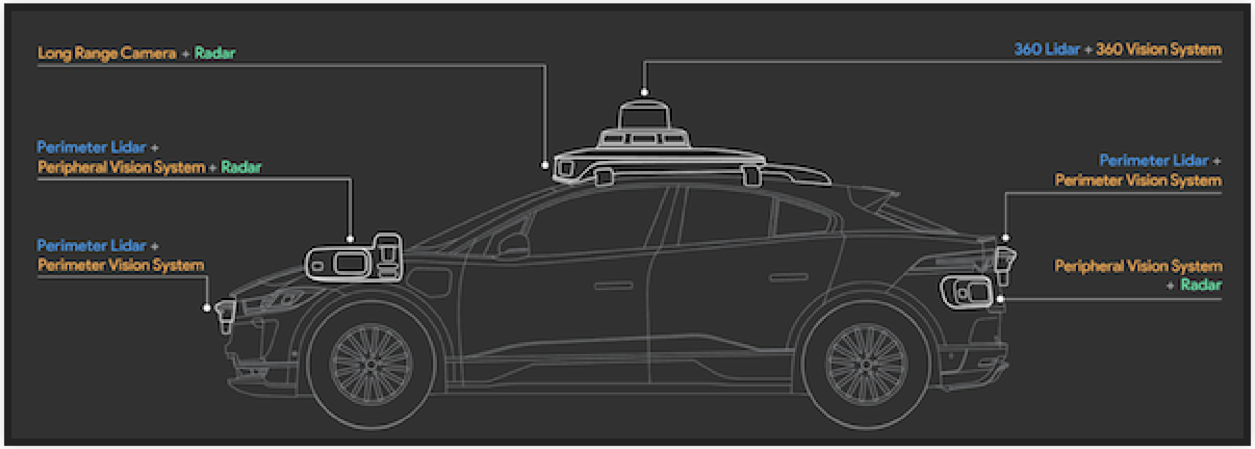
- Decision boundary: Ground truth; often unknown and not specifiable
- Model boundary: What the model learns; an approximation of decision boundary
- Often, learned & actual decision boundaries do not match!

From Goodfellow et al (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, *61*(7), 56-66.

EVALUATING ROBUSTNESS

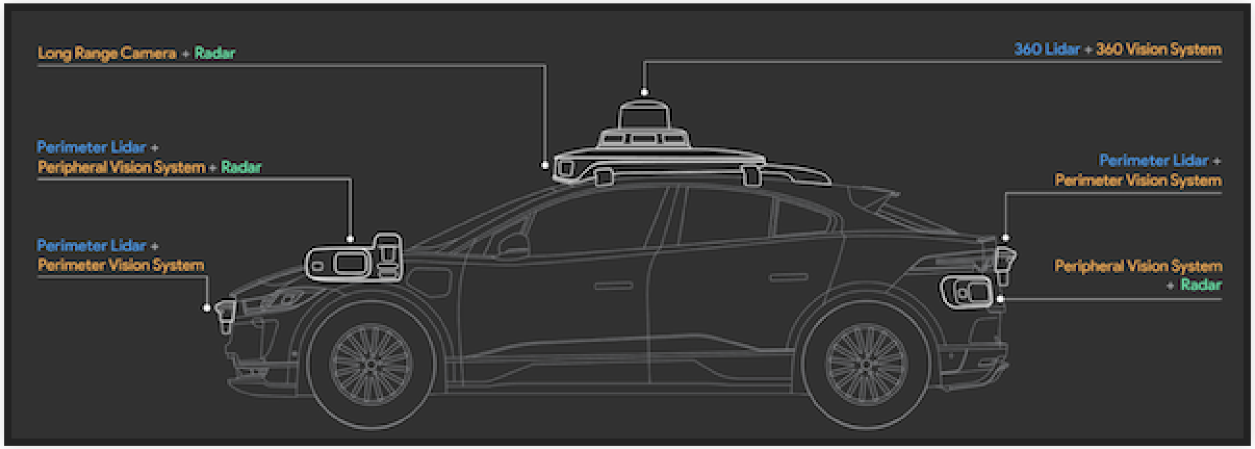
- Lots of on-going research (especially for DNNs)
- Formal verification
    - Constraint solving or abstract interpretation over computations in neuron activations
    - Conservative abstraction, may label robust inputs as not robust
    - Currently not very scalable
    - Example: *An abstract domain 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).

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)

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
  - Ensemble learning: Combine models with different biases
  - Multiple, independent sensors (e.g., lidar + radar + cameras)

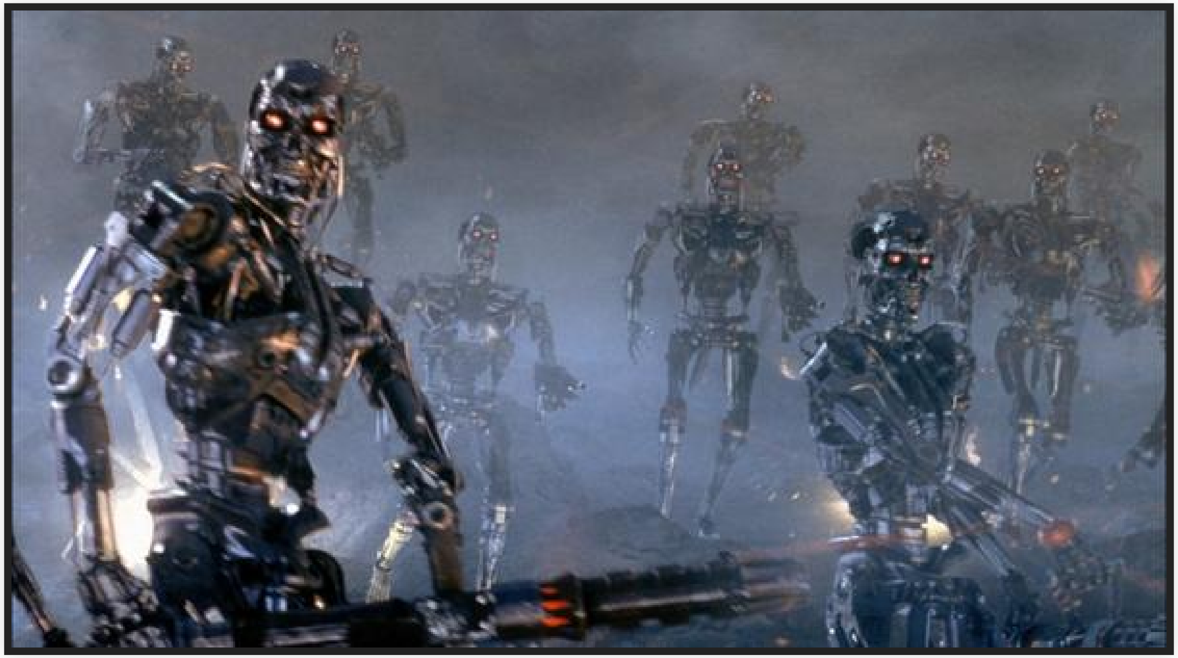
IMPROVING ROBUSTNESS FOR SAFETY


(a) Non Foggy Condition (b) Foggy Condition

- Learning more robust models
  - Curate data for abnormal scenarios (e.g., fogs, snow, sensor noise)
  - Augment training data with transformed versions (but same label)
- Testing and debugging
  - Identify training data near model's decision boundary (i.e., is the model robust around all training data?)
  - Check robustness on test data

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

OTHER AI SAFETY CONCERNS



NEGATIVE SIDE EFFECTS

- 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 example from the Ethics lecture)
- Again, **requirements** problem!
    - Recall: "World vs. machine"
    - Identify stakeholders in the environment & possible effects on them
- Modify the AI goal from "Perform Task X" to:
    - Perform X *subject to common-sense constraints on the environment*
    - Perform X *but avoid side effects to the extent possible*

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

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

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

REWARD HACKING -- MANY EXAMPLES

Tweet

OTHER CHALLENGES

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

OTHER CHALLENGES

- Safe Exploration
  - Exploratory actions "in production" may have consequences
  - e.g., trap robots, crash drones
  - -> Safety envelopes and other strategies to explore only in safe bounds (see also chaos engineering)

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

OTHER CHALLENGES

- Safe Exploration
    - Exploratory actions "in production" may have consequences
    - e.g., trap robots, crash drones
    - -> Safety envelopes and other strategies to explore only in safe bounds (see also chaos engineering)
- Robustness to Drift
    - Drift may lead to poor performance that may not even be recognized
    - -> Check training vs production distribution (see data quality lecture), change detection, anomaly detection

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

OTHER CHALLENGES

- Safe Exploration
  - Exploratory actions "in production" may have consequences
  - e.g., trap robots, crash drones
  - -> Safety envelopes and other strategies to explore only in safe bounds (see also chaos engineering)
- Robustness to Drift
  - Drift may lead to poor performance that may not even be recognized
  - -> Check training vs production distribution (see data quality lecture), change detection, anomaly detection
- Scalable Oversight
  - Cannot provide human oversight over every action (or label all possible training data)
  - Use indirect proxies in telemetry to assess success/satisfaction
  - -> Semi-supervised learning? Distant supervision?

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

DESIGNING FOR SAFETY

REVIEW: ELEMENTS OF SAFE DESIGN

(See Mitigation Strategies from the Lecture on Risks)

- **Assume**: Components will fail at some point
- **Goal**: Minimize the impact of failures
- **Detection**
  - Monitoring
- **Response**
  - Graceful degradation (fail-safe)
  - Redundancy (fail over)
- **Containment**
  - Decoupling & isolation

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
- Recent research on adversarial learning and safety in reinforcement learning

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?

TWITTER

Home | Your profile | Invite | Public timeline | Badges | Settings | Help | Sign out
What are you doing?
Characters available: 140
Update
Archive
Recent
What You And Your Friends Are Doing
RonLandreth building an xml page out of a MySQL database half a minute ago from web
Fitz Just got off the phone with Lopez. He's gona go easter egg hunting on sunday. half a minute ago from web
Sofia legend half a minute ago from im
nzkoz thinks gardening is house-owning-1.0. Gotta be some kinda social tag cloud house keeping half a minute ago from twitterrific
GeekLady Leo Laporte is nuts. Aye tutis, they'll confuse an acronym with a
Welcome back
Currently: Reading: "Tech Blot » Blog Archive » Why It's So Easy To Impersonate On Twitter" (http://tinyurl...)
0 Direct Messages
0 Favorites
2669 Friends
715 Followers
7 Updates
Send Notifications To:
web-only
Activate Phone!
Activate your IM!

Speaker notes

What consequences should Twitter have foreseen? How should they intervene now that negative consequences of interaction patterns are becoming apparent?

MENTAL HEALTH

healthline

SUBSCRIBE

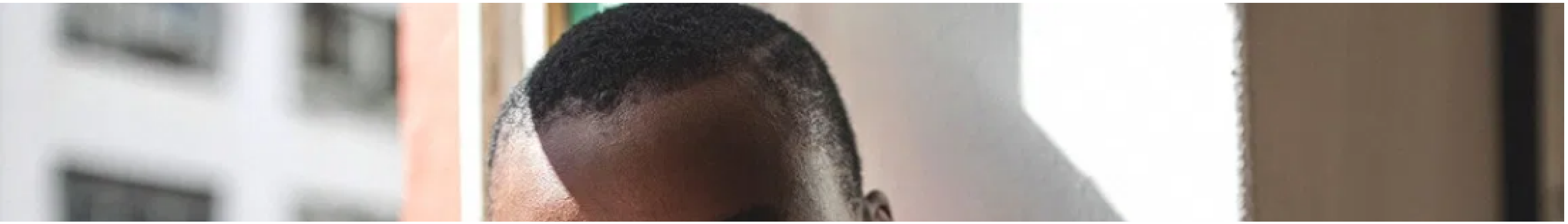
HEALTH NEWS

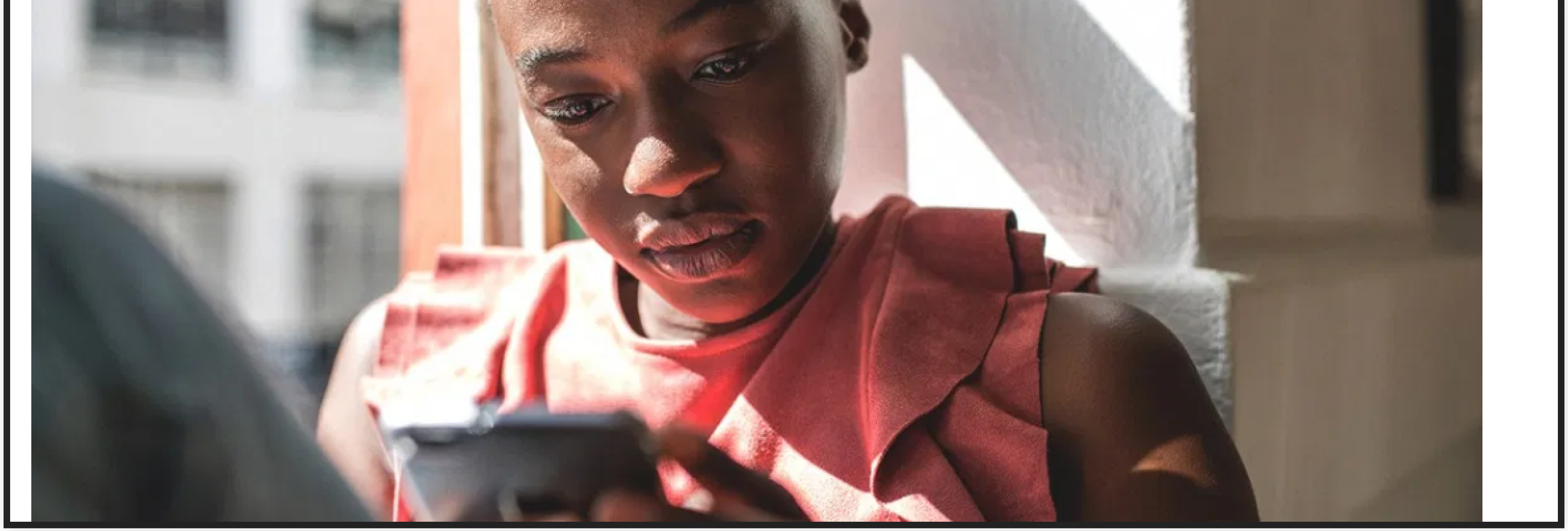
✓ Fact Checked

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.



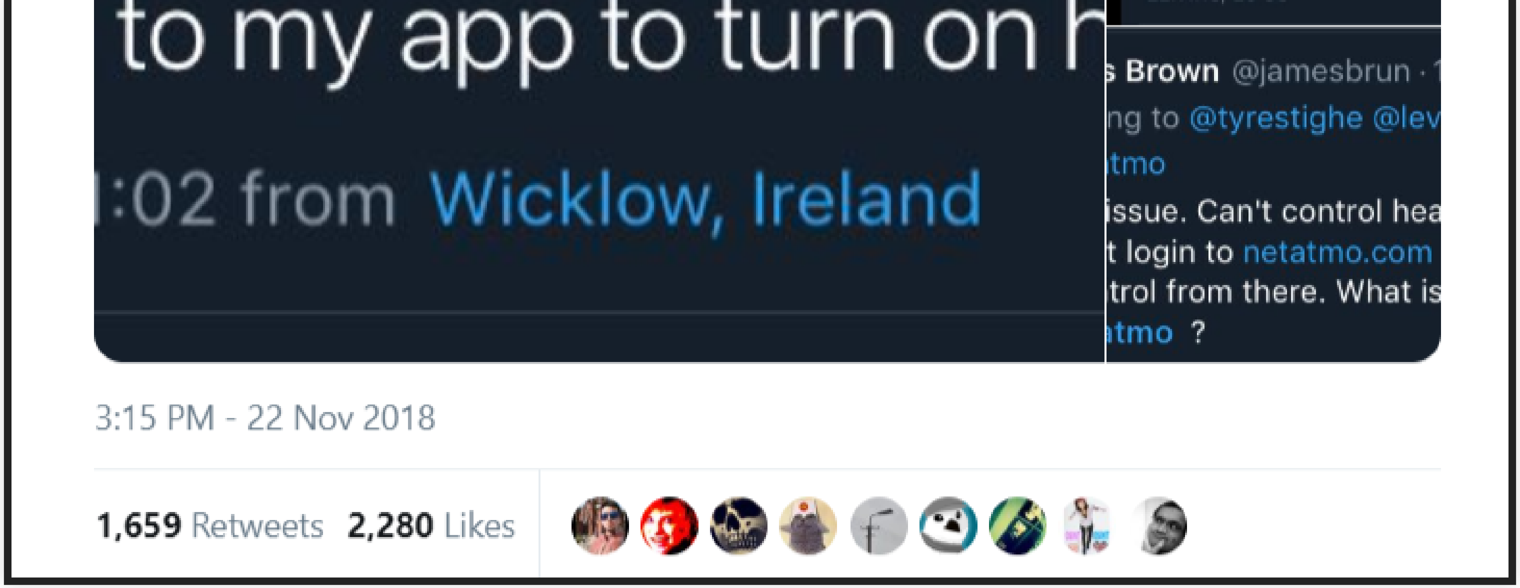


IOT

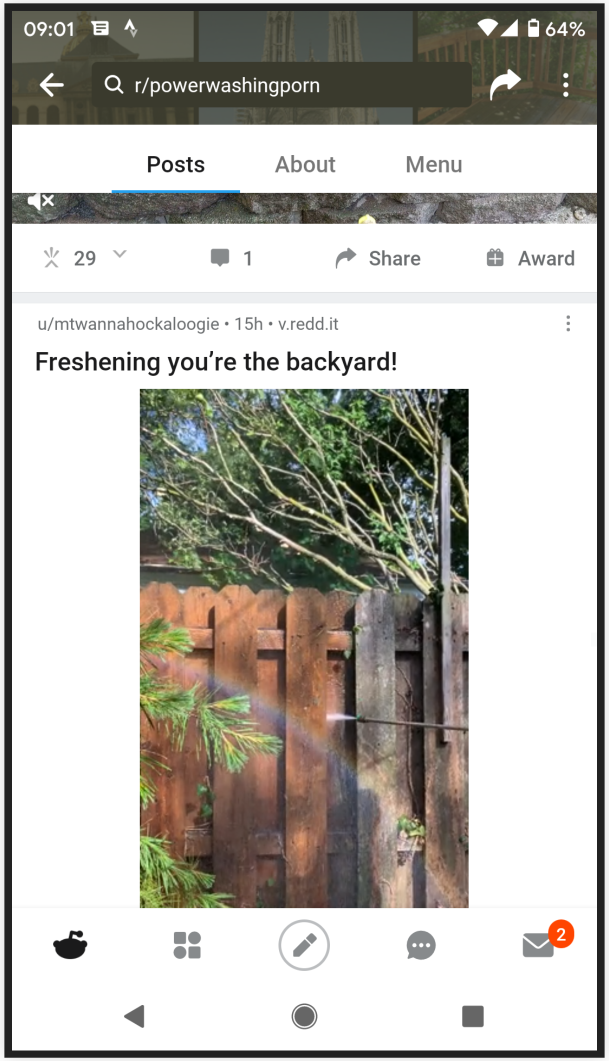
skoops 🧸 🏴‍☠️
@skoops

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


to my app to turn on
:02 from Wicklow, Ireland
Brown @jamesbrun
ng to @tyrestighe @lev
tmo
issue. Can't control hea
t login to netatmo.com
trol from there. What is
tmo ?
3:15 PM - 22 Nov 2018
1,659 Retweets 2,280 Likes

ADDICTION


09:01
64%
r/powerwashingporn
Posts
About
Menu
29
1
Share
Award
u/mtwannahockaloogie • 15h • v.redd.it
Freshening you're the backyard!
2

Speaker notes

Infinite scroll in applications removes the natural breaking point at pagination where one might reflect and stop use.

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.

Addiction is the inability to stop consuming a chemical or pursuing an activity although it's causing harm.

I engage with almost every substance or behavior associated with addiction: alcohol, drugs, coffee, porn, sex, gambling, work, spending,

SOCIETY: UNEMPLOYMENT ENGINEERING / DESKILLING

Speaker notes

The dangers and risks of automating jobs.

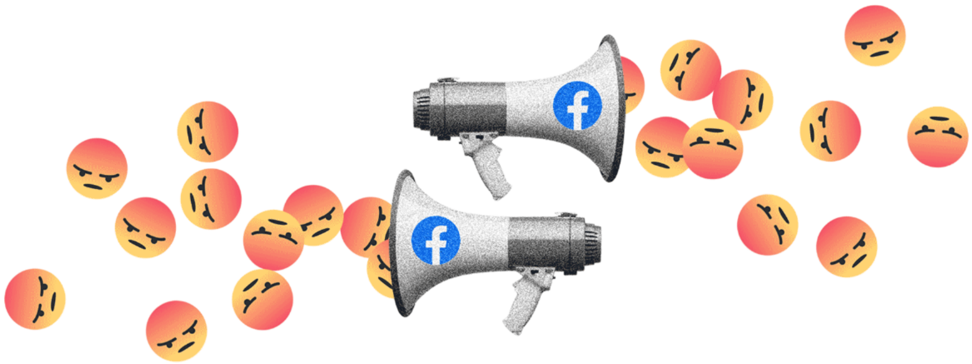
Discuss issues around automated truck driving and the role of jobs.

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

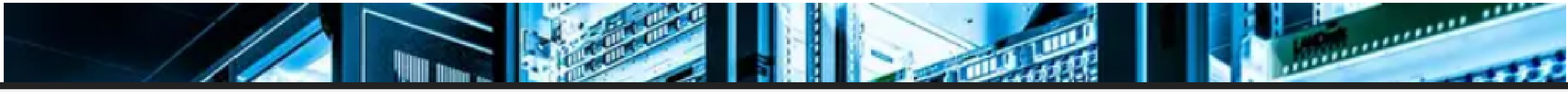
NewScientist

SUBSCRIBE AND SAVE 69%

Creating an AI can be five times worse for the planet than a car

TECHNOLOGY 6 June 2019

By Donna Lu



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

SUMMARY

- *Adopt a safety mindset!*
- Defining safety: absence of harm to people, property, and environment
    - Beyond traditional safety critical systems, affects many apps and web services
- 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
- AI goals are difficult to specify precisely; susceptible to negative side effect & reward hacking
- Model robustness can help with some problems