SAFETY

Christian Kaestner

With slides from Eunsuk Kang

Required Reading: Salay, Rick, Rodrigo Queiroz, and Krzysztof Czarnecki. "An analysis of ISO 26262: Using machine learning safely in automotive software." arXiv preprint arXiv:1709.02435 (2017).

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

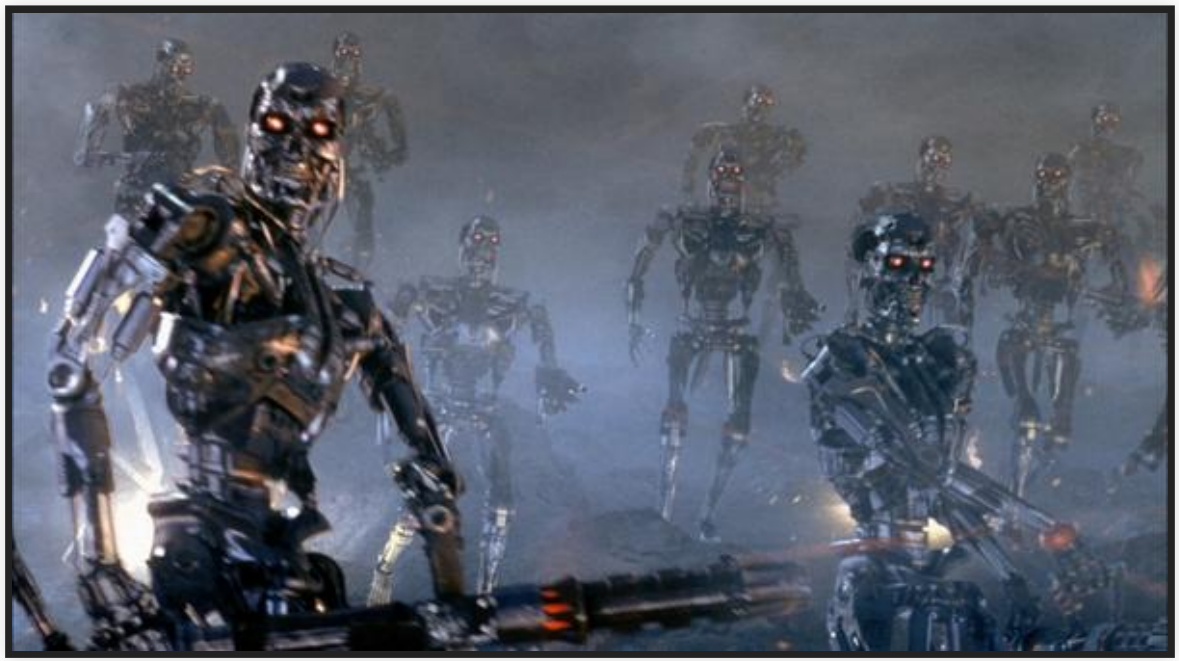
SAFETY

DEFINING SAFETY

- 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 != 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 is a system concept

EXAMPLES OF HARM FROM AI-ENABLED SYSTEMS?

SAFETY



SAFETY

Tweet

SAFETY

Tweet

SAFETY CHALLENGE WIDELY RECOGNIZED

Being able to apply ML in safety-critical applications will be important to my organization in the future	a)	
V&V of features that rely on ML is recognized as a particularly challenging area in my organization	b)	
My organization is well-prepared for a future in which V&V of safety-critical ML is commonplace	c)	

(survey among automotive engineers)

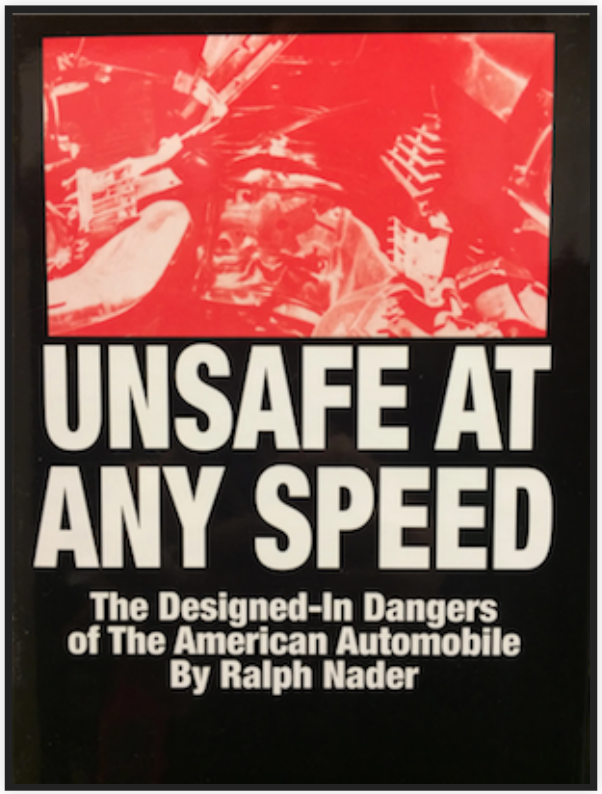
Borg, Markus, et al. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." arXiv preprint arXiv:1812.05389 (2018).

SAFETY IS A BROAD CONCEPT

- Includes harm to mental health
- Includes polluting the environment, including noise pollution
- Includes harm to society, e.g. poverty, polarization

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?

Challenges?

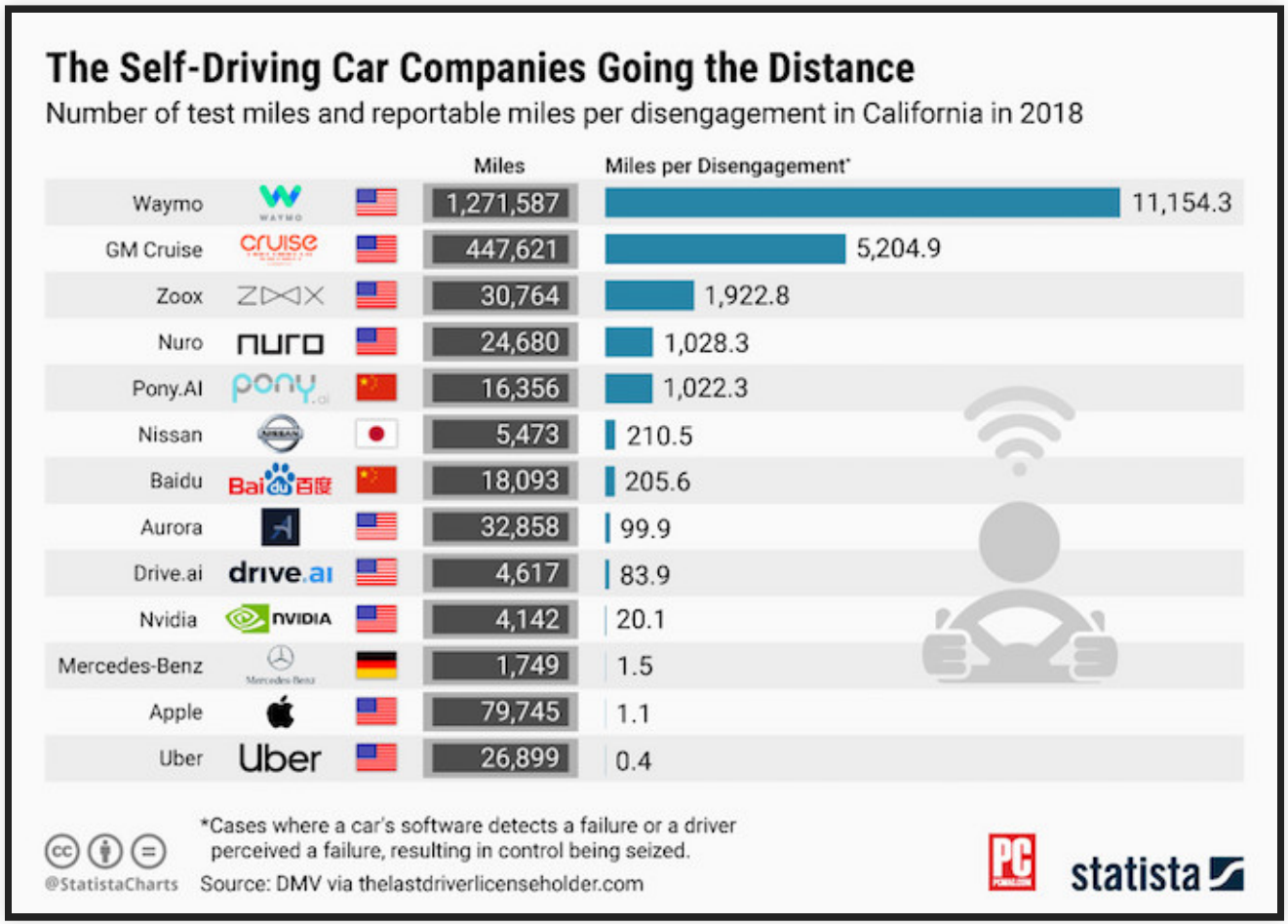
AUTONOMOUS VEHICLES: WHAT'S DIFFERENT?

- 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 to cover all unknown cases
- **Why is this a unique problem for AI**? What about humans?

DEMONSTRATING SAFETY



More miles tested => safer?

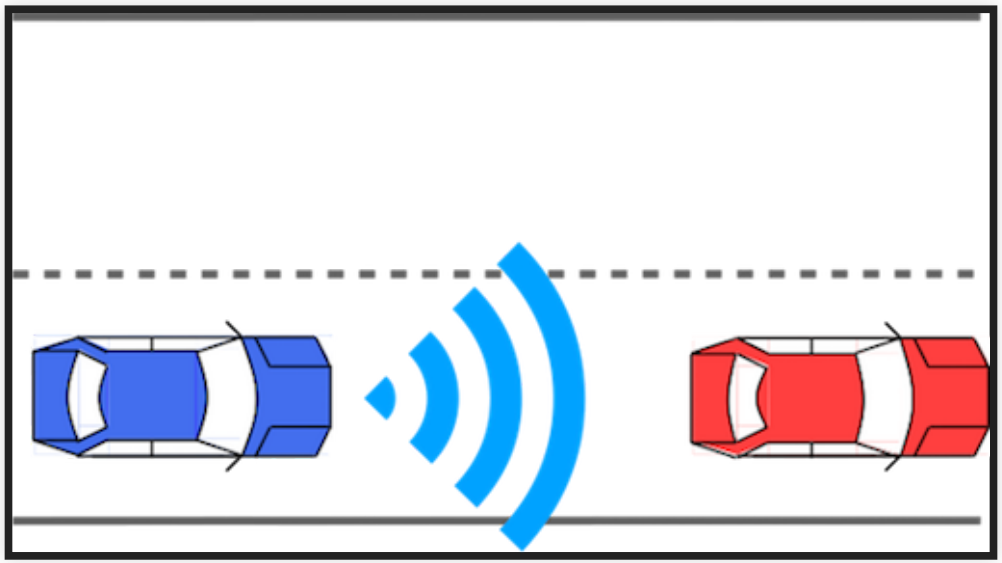
APPROACH FOR DEMONSTRATING SAFETY

- 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

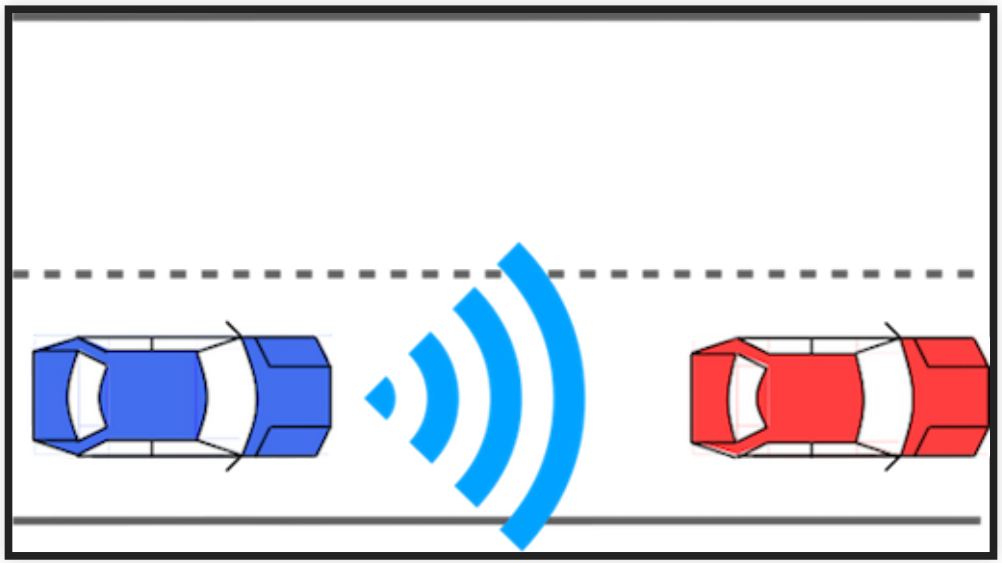
(system level!)

WHAT IS HAZARD ANALYSIS?



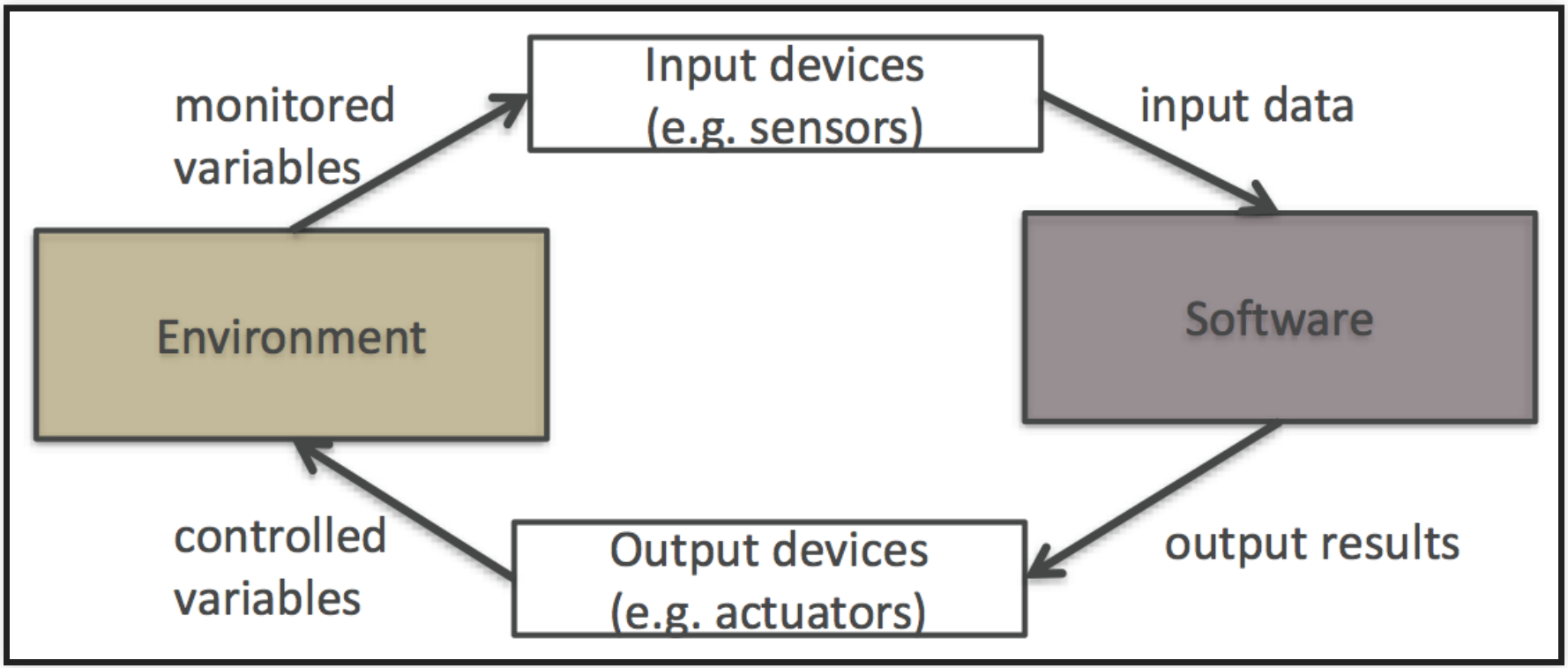
- **Hazard**: A condition or event that may result in undesirable outcome
  - e.g., "Ego vehicle is in risk of a collision with another vehicle."
- **Safety requirement**: Intended to eliminate or reduce one or more hazards
  - "Ego vehicle must always maintain some minimum safe distance to the leading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

RECALL: 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.

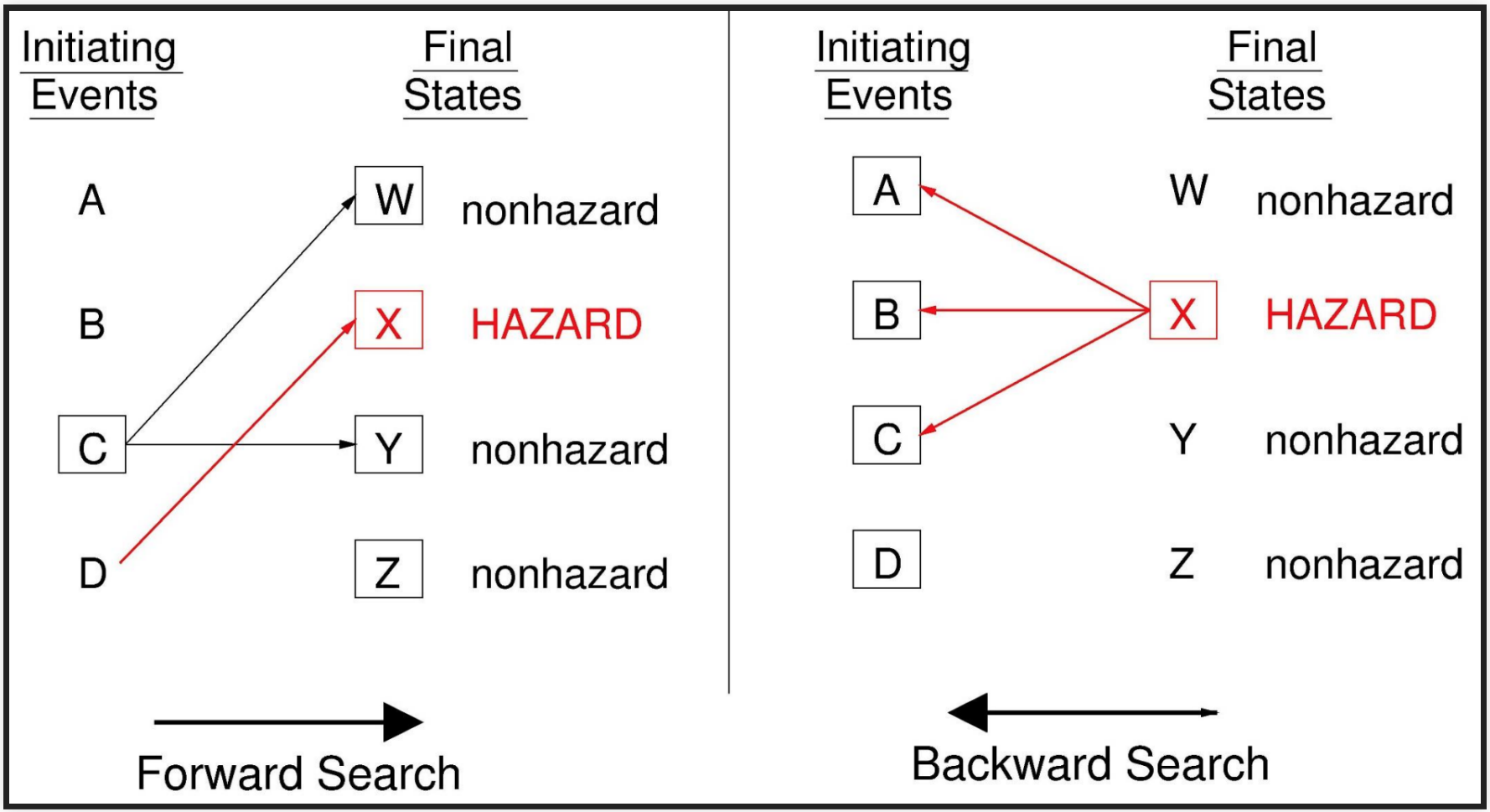
RECALL: WORLD VS MACHINE



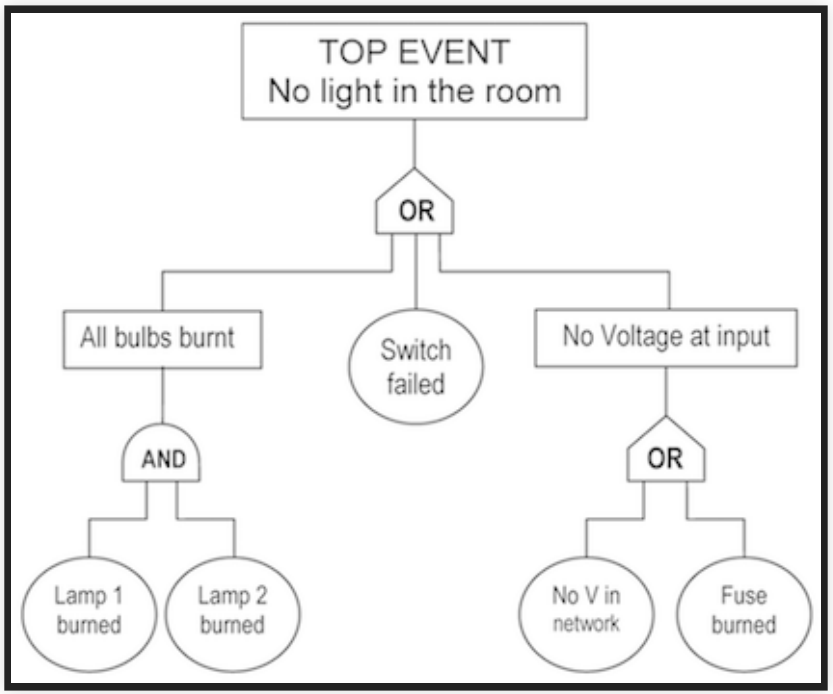
Software is not unsafe; the control signals it generates may be

Root of unsafety usually in wrong requirements

FORWARD VS BACKWARD SEARCH



RECALL: FAULT TREE ANALYSIS (FTA)



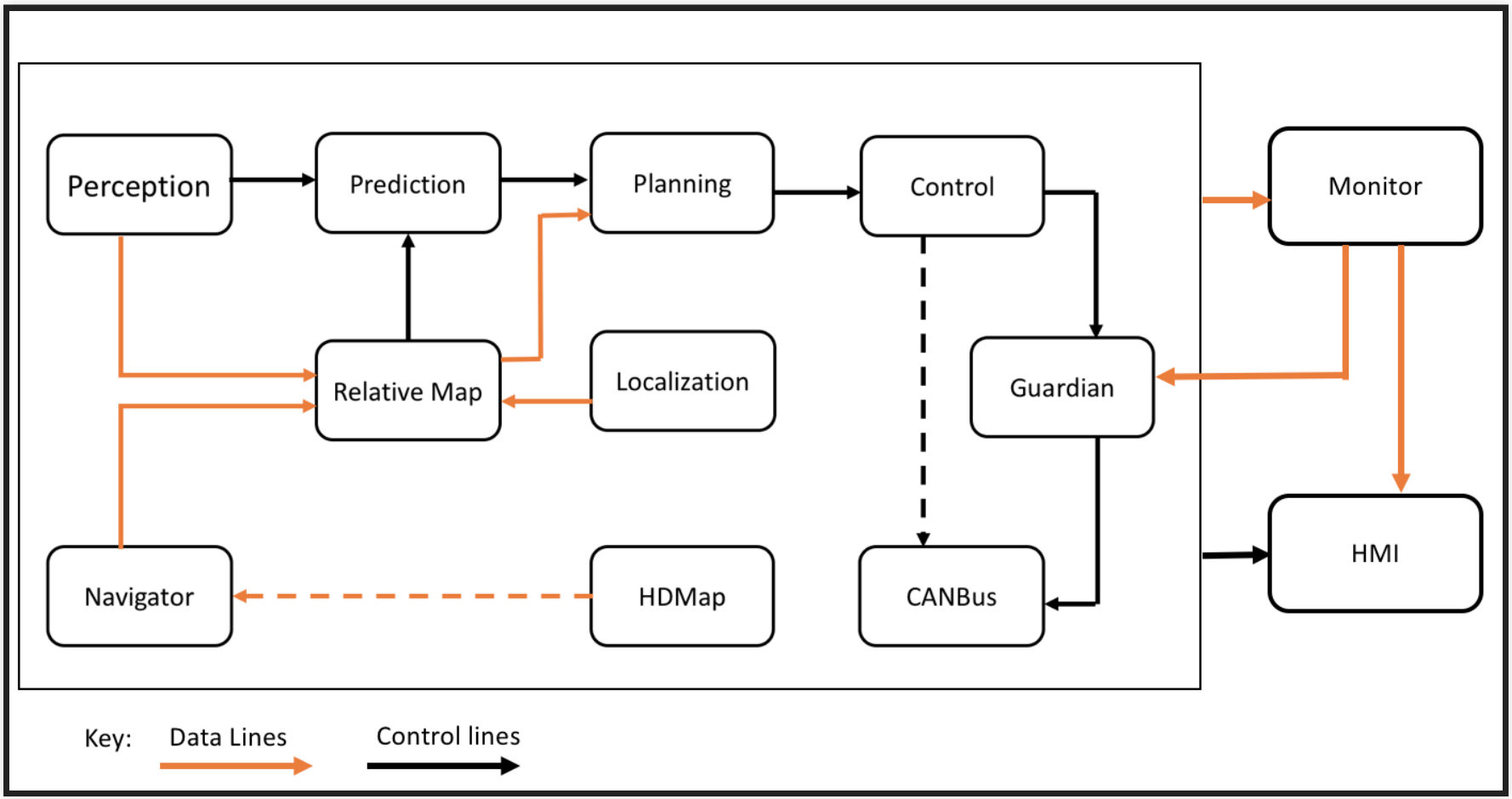
- Top-down, **backward** search method for root cause analysis
  - Start with a given hazard (top event), derive a set of component faults (basic events)
  - Compute minimum cutsets as potential root causes

RECALL: FAILURE MODE AND EFFECTS ANALYSIS

	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
- 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

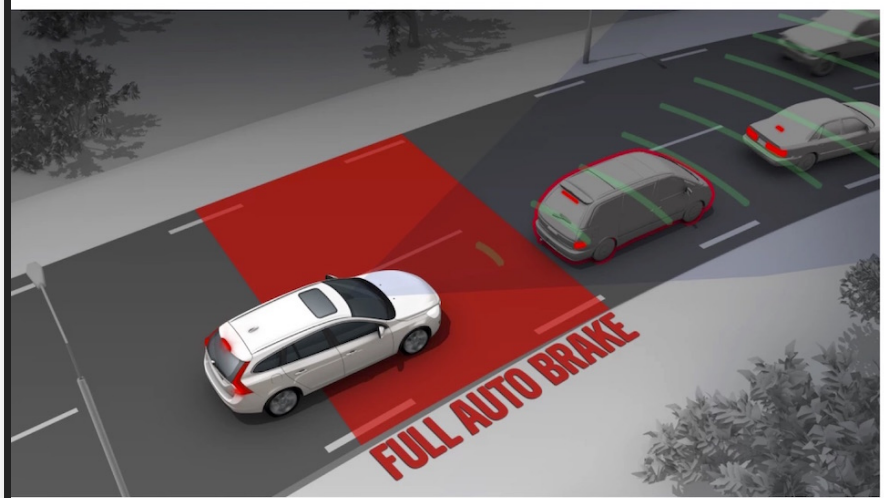
Component	Failure Mode	Failure Effects	Detection	Mitigation
Perception	Failure to detect an object	Risk of collision	Human operator (if present)	Deploy secondary classifier
Perception	Detected but misclassified	"	"	"
Lidar Sensor	Mechanical failure	Inability to detect objects	Monitor	Switch to manual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

RECALL: HAZARD AND OPERABILITY STUDY

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?

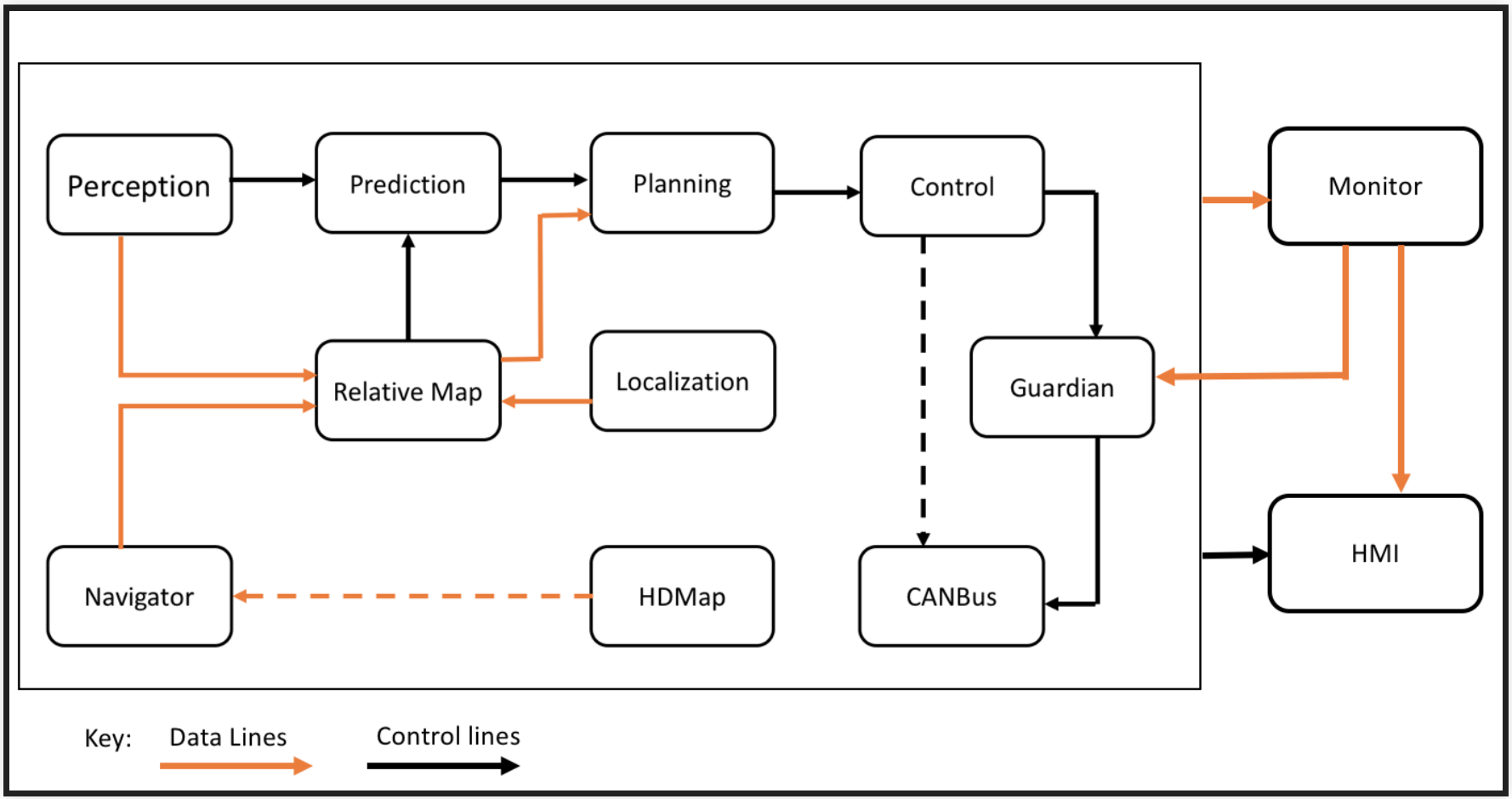
HAZOP EXAMPLE: EMERGENCY BRAKING (EB)



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

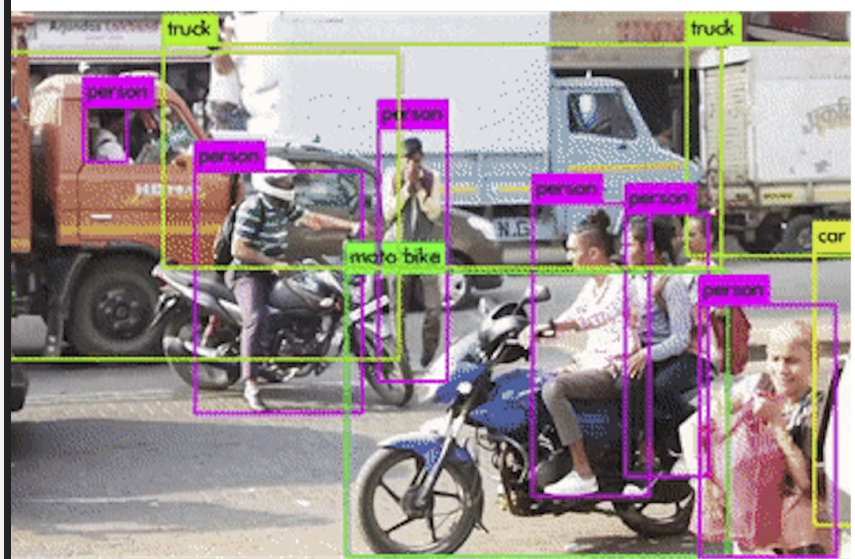
- Specification: EB must apply a maximum braking command to the engine.
  - **NONE**: EB does not generate any braking command.
  - **LESS**: EB applies less than max. braking.
  - **LATE**: EB applies max. braking but after a delay of 2 seconds.
  - **REVERSE**: EB generates an acceleration command 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 method 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

RECALL: 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 IN A SAFETY SETTING

- Does the model reliably detect stop signs?
- Also in poor lighting? In fog? With a tilted camera?
- With stickers taped to the sign?

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

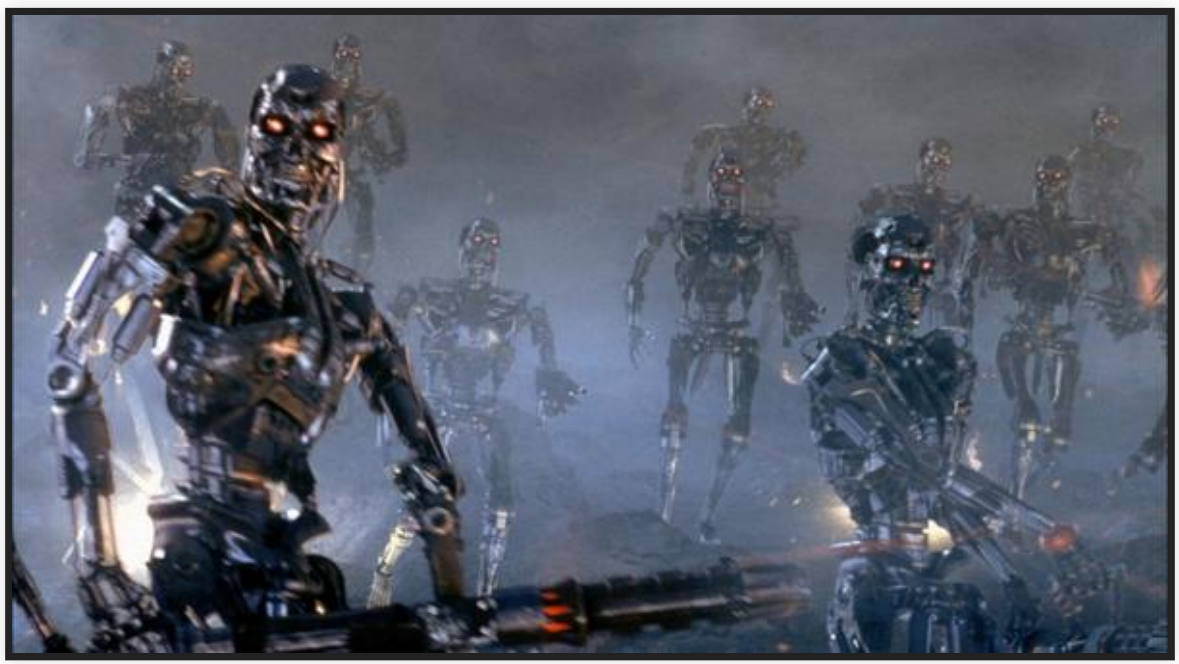
ROBUSTNESS VERIFICATION FOR SAFETY

- Rely only on predictions that are robust
    - online verification, smoothing
- Detect outliers in inputs
- Learn more robust models
    - data augmentation, simulation
    - and many other strategies (see security lecture)

TESTING FOR SAFETY

- Curate data sets for critical scenarios (see model quality lecture)
- Create test data for difficult settings (e.g. fog)
- Simulation feasible? Shadow deployment feasible?

OTHER AI SAFETY CONCERNS



NEGATIVE SIDE EFFECTS

https://www.decisionproblem.com/paperclips/index2.html

```
.
.
.
. Welcome to Universal Paperclips
> AutoClippers available for purchase
```

Paperclips: 148

Make Paperclip

Business

Available Funds: $ 9.50
Unsold Inventory: 89
lower raise Price per Clip: $ .25
Public Demand: 32%

Marketing Level: 1
Cost: $ 100.00

Manufacturing

Clips per Second: 1

Wire 852 inches
Cost: $ 26

AutoClippers 1
Cost: $ 6.10

NEGATIVE SIDE EFFECTS

- Challenge: Define good goal/cost function
- Design in system context, beyond the model
- "Perform X" --> "perform X *subject to common-sense constraints on the environment*" or "perform X *but avoid side effects to the extent possible*"

Other examples?

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

Speaker notes

An self-driving car may break laws in order to reach a destination faster

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

Self-driving car figures out that it can avoid getting penalized for driving too close to other cars by exploiting certain sensor vulnerabilities so that it can't "see" how close it is getting

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 intend*
- Many reasons, incl. partially observed goals, abstract rewards, proxies, feedback loops
- Challenging to specify goal and reward function properly

Other examples?

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

- Scalable Oversight
  - Cannot provide human oversight over every action (or label all possible training data)
  - Use indirect proxies in telemetry to assess success/satisfaction
  - Training labels may not align well with goals
  - -> Semi-supervised learning? Distant supervision?
- 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

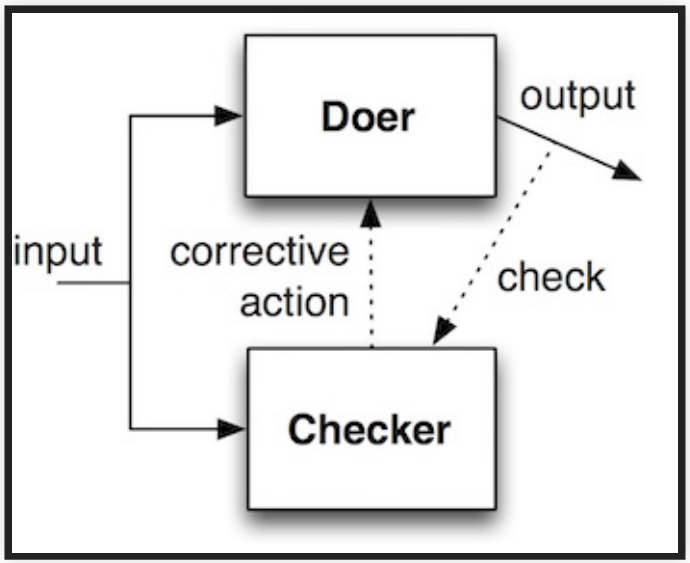
Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).

DESIGNING FOR SAFETY

ELEMENTS OF SAFE DESIGN

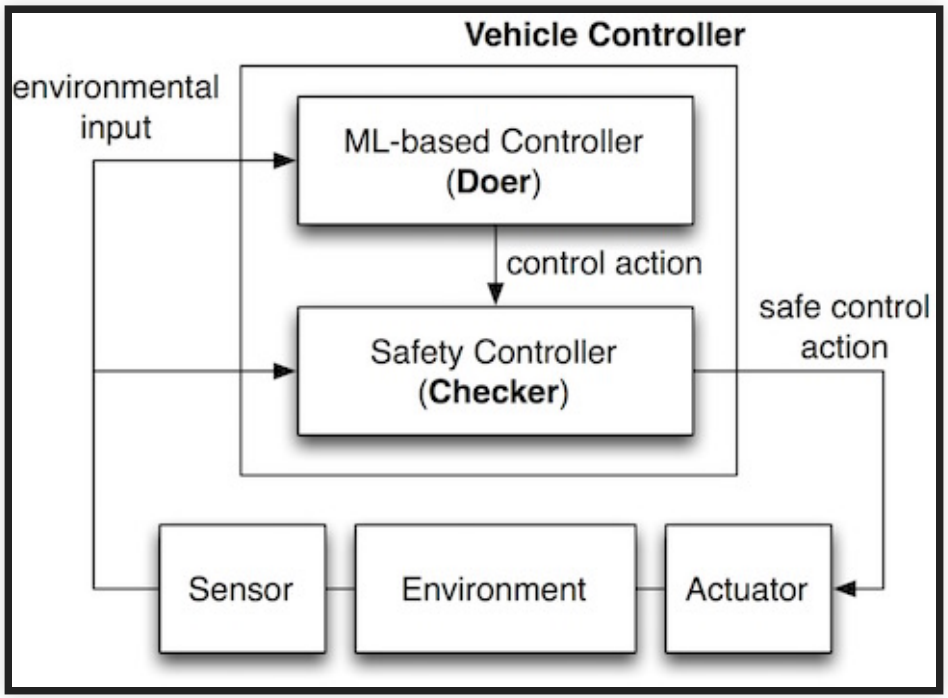
- **Assume**: Components will fail at some point
- **Goal**: Minimize the impact of failures on safety
- **Detection**
    - Monitoring
- **Control**
    - Graceful degradation (fail-safe)
    - Redundancy (fail over)
- **Prevention**
    - Decoupling & isolation

DETECTION: MONITORING



- **Goal**: Detect when a component failure occurs
- **Heartbeat** pattern
  - Periodically sends diagnostic message to monitor
- **Doer-Checker** pattern
  - Doer: Perform primary function; untrusted and potentially faulty
  - Checker: If doer output faulty, perform corrective action (e.g., default safe output, shutdown); trusted and verifiable

DOER-CHECKER EXAMPLE: AUTONOMOUS VEHICLE

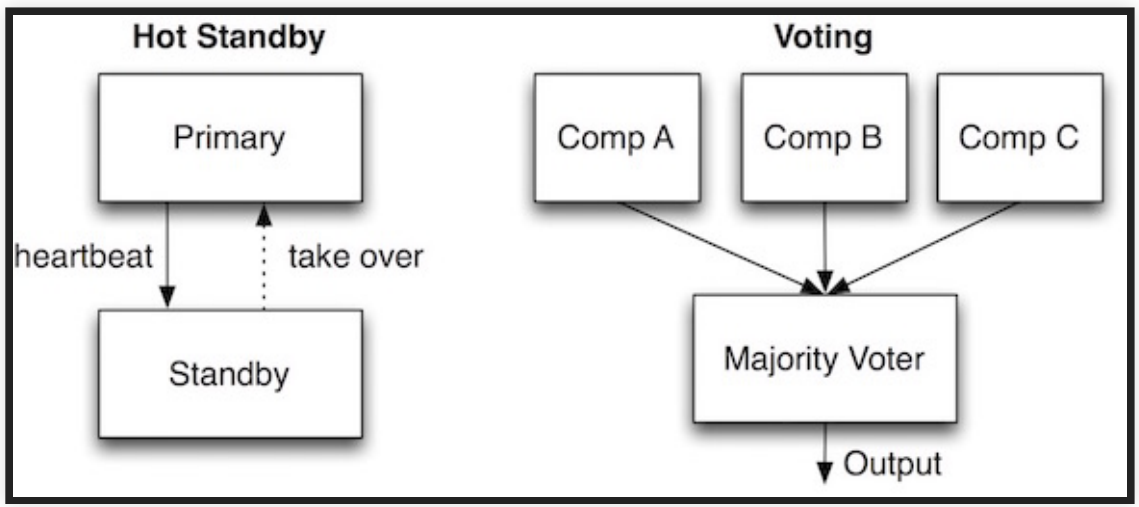


- ML-based controller (**doer**): Generate commands to maneuver vehicle
  - Complex DNN; makes performance-optimal control decisions
- Safety controller (**checker**): Checks commands from ML controller; overrides it with a safe default command if maneuver deemed risky
  - Simpler, based on verifiable, transparent logic; conservative control

RESPONSE: GRACEFUL DEGRADATION (FAIL-SAFE)

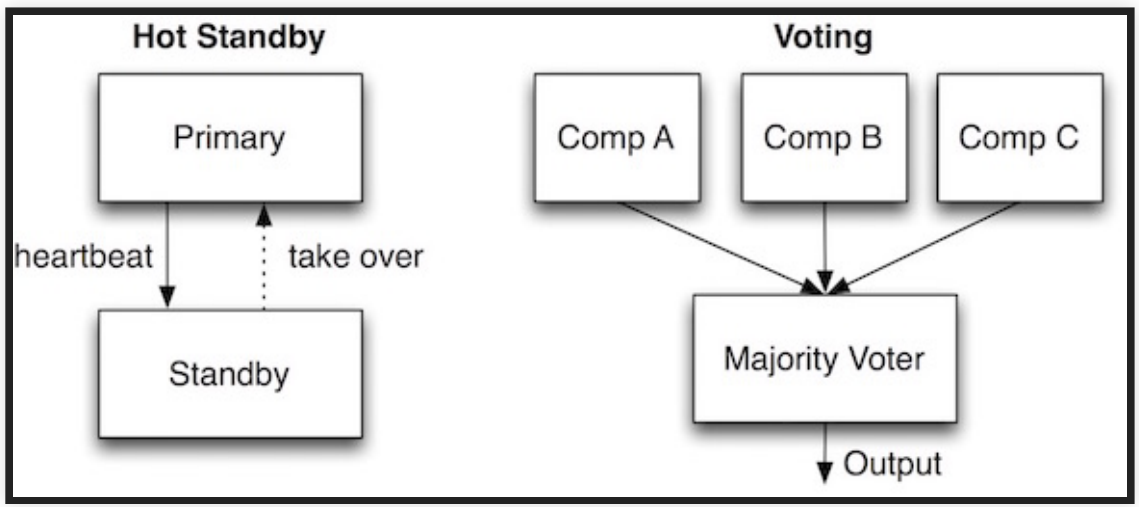
- **Goal**: When a component failure occurs, continue to provide safety (possibly at reduced functionality and performance)
- Relies on a monitor to detect component failures
- Example: Perception in autonomous vehicles
    - If Lidar fails, switch to a lower-quality detector; be more conservative
    - **But what about other types of ML failures? (e.g., misclassification)**

RESPONSE: REDUNDANCY (FAILOVER)



- **Goal**: When a component fails, continue to provide the same functionality
- **Hot Standby**: Standby watches & takes over when primary fails
- **Voting**: Select the majority decision
- Caution: Do components fail independently?
    - Reasonable assumption for hardware/mechanical failures
    - **Q. What about software?**

RESPONSE: REDUNDANCY (FAILOVER)



- **Goal**: When a component fails, continue to provide the same functionality
- **Hot Standby**: Standby watches & takes over when primary fails
- **Voting**: Select the majority decision
- Caution: Do components fail independently?
    - Reasonable assumption for hardware/mechanical failures
    - Software: Difficult to achieve independence even when built by different teams (e.g., N-version programming)
    - **Q. ML components?**

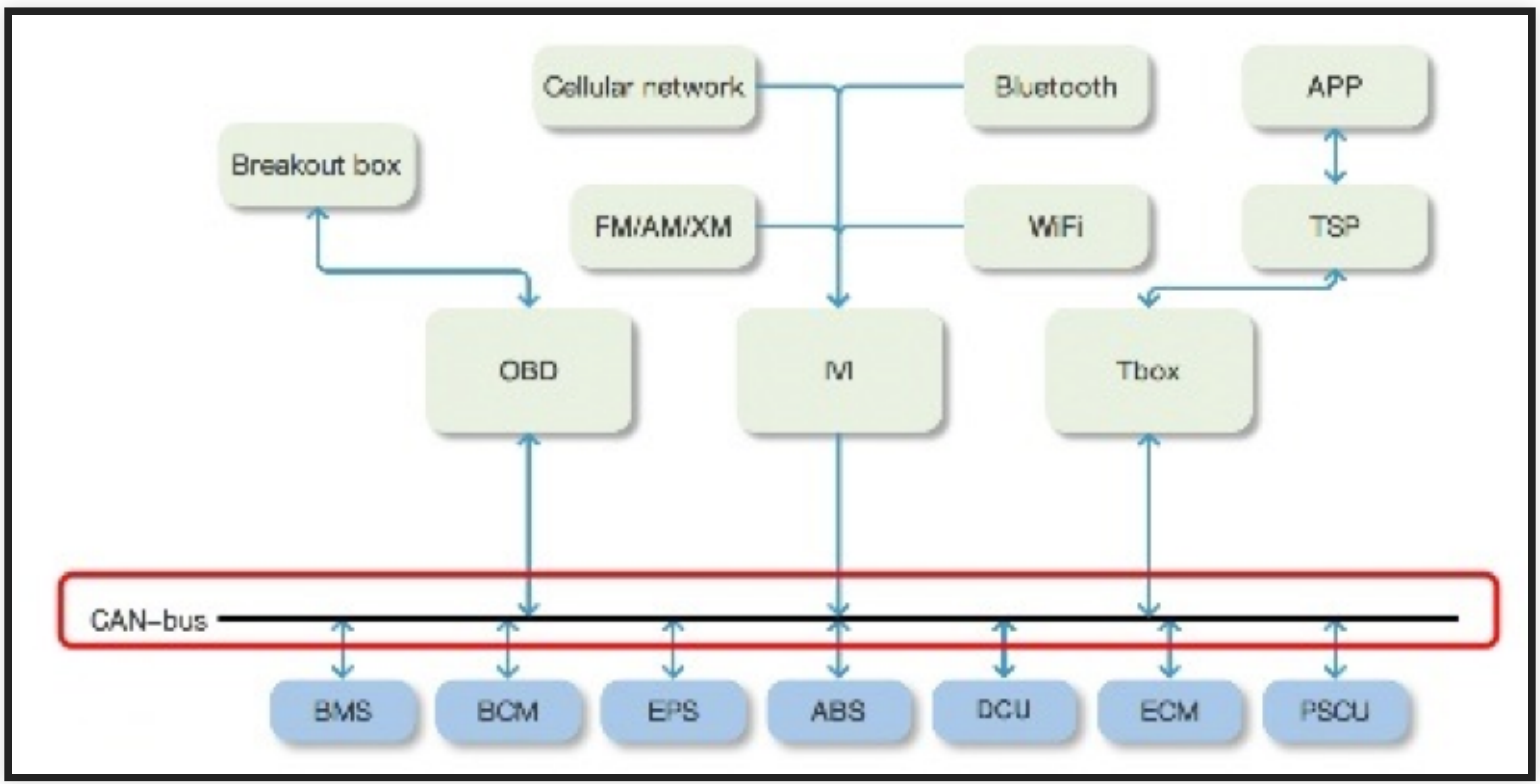
PREVENTION: DECOUPLING & ISOLATION

- **Goal**: Faults in a low-critical (LC) components should not impact high-critical (HC) components

POOR DECOUPLING: USS YORKTOWN (1997)

- Invalid data entered into DB; divide-by-zero crashes entire network
- Required rebooting the whole system; ship dead in water for 3 hours
- **Lesson**: Handle expected component faults; prevent propagation

POOR DECOUPLING: AUTOMOTIVE SECURITY

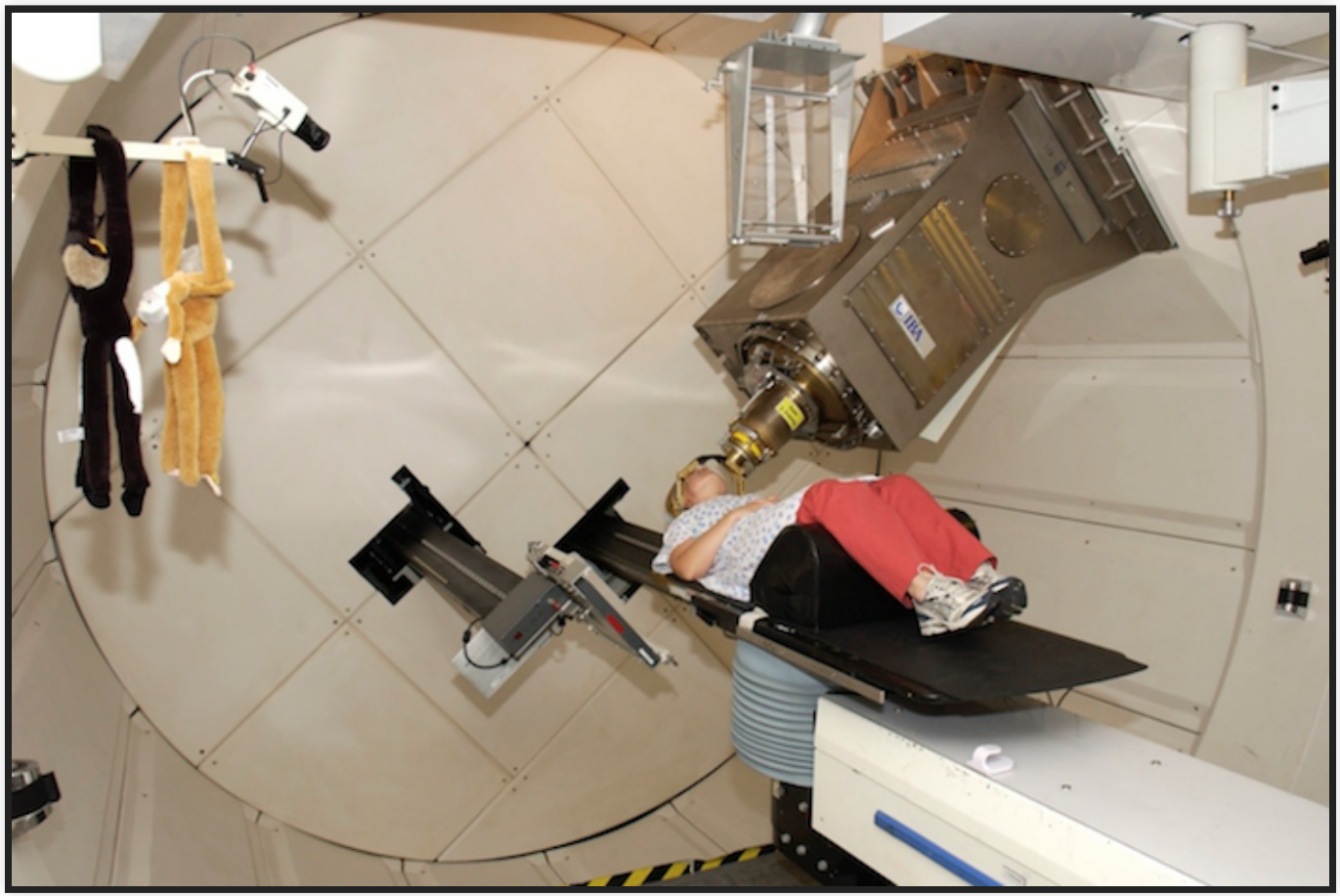


- Main components connected through a common CAN bus
  - Broadcast; no access control (anyone can read/write)
- Can control brake/engine by playing a malicious MP3 (Stefan Savage, UCSD)

PREVENTION: DECOUPLING & ISOLATION

- Goal: Faults in a low-critical (LC) components should not impact high-critical (HC) components
- Apply the principle of least privilege
    - LC components should be allowed to access min. necessary data
- Limit interactions across criticality boundaries
    - Deploy LC & HC components on different networks
    - Add monitors/checks at interfaces
- Identify and eliminate implicit interactions
    - Memory: Shared memory, global variables
    - CPU resources: LC tasks running at high-priority, starving HC tasks
- Is AI in my system performing an LC or HC task?
    - If HC, can we "demote" it into LC?

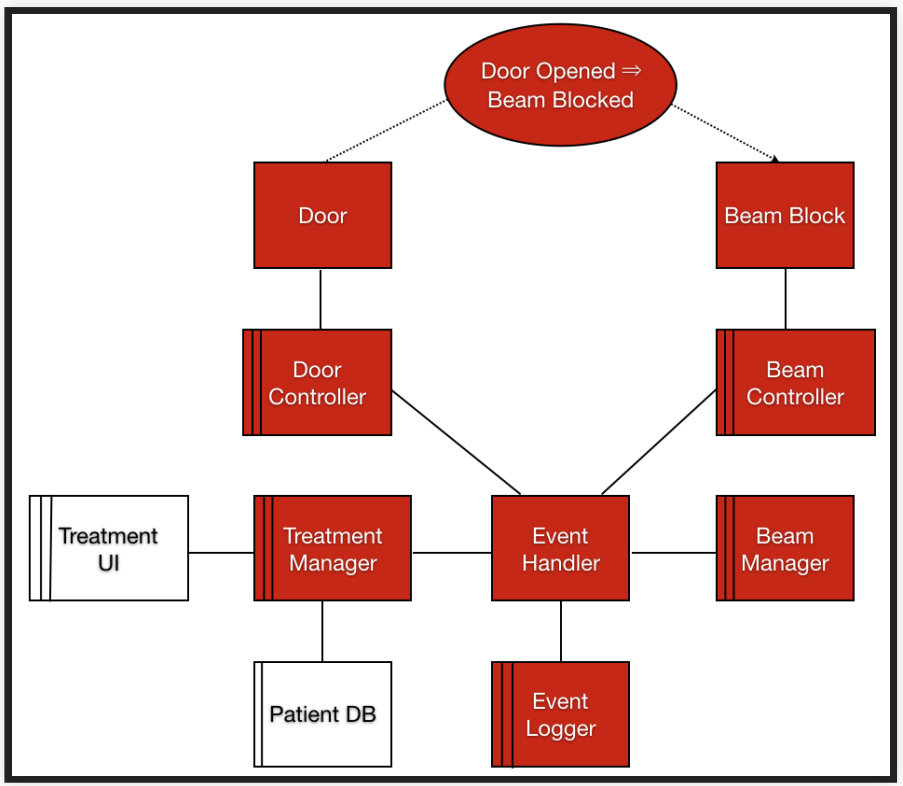
EXAMPLE: RADIATION THERAPY



- **Safety requirement**: If door opens during treatment, insert beam block.

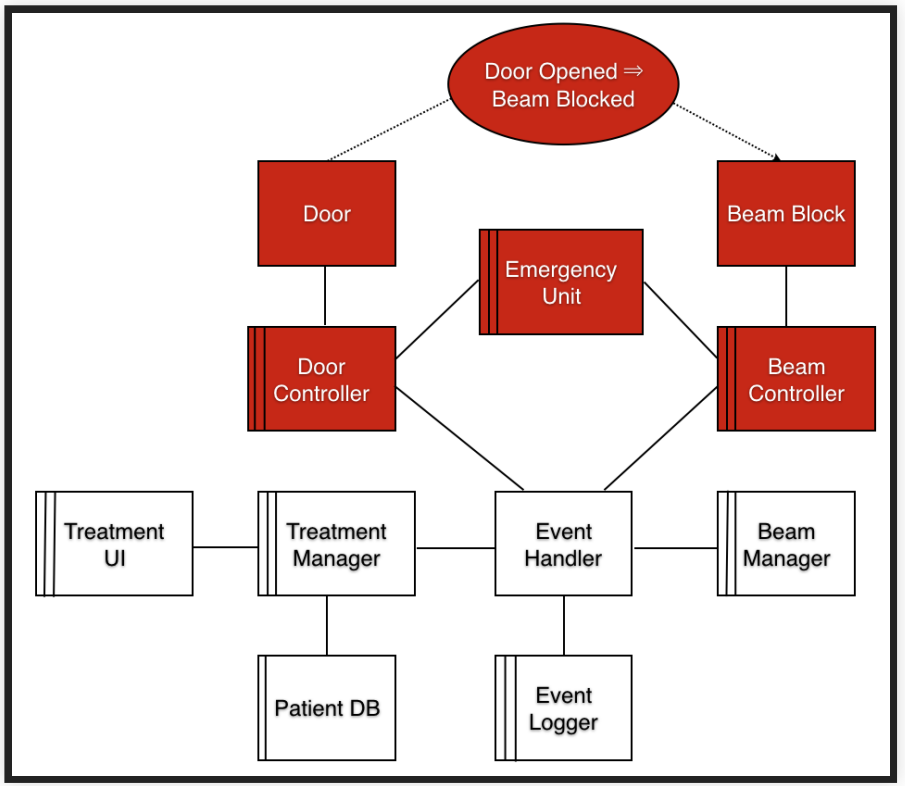
EXISTING DESIGN

- Which components are responsible for establishing this safety requirement (e.g., high critical)?
- Existing design includes:
  - Pub/sub event handler: 3rd-party library; missing source code; company went bankrupt
  - Event logging: May throw an error if disk full
  - Event handler/logging used by all tasks, including LC ones
- Is it possible to achieve high confidence that these HC components don't fail?



ALTERNATIVE DESIGN

- Build in an emergency unit
  - Bypass event handler for HC tasks
- Still needs to rely on door & beam controllers
  - Can't eliminate the risk of failure, but significantly reduce it
  - Emergency unit simpler, can be verified & tested



ML AS UNRELIABLE COMPONENTS

- Symbolic AI can provide guarantees
- ML models may make mistakes, no specifications
    - see also ML as requirements engineering?
- Mistakes are hard to predict or understand
    - Does interpretability help?
- Mistakes are not independent or uniformly distributed
    - Classic redundancy mechanisms may not work?

SELF-DRIVING CARS


168

Speaker notes

Driving in controlled environments vs public roads

ISO 26262

- Current standards not prepared for machine learning
- Assume specifications and corresponding testing

- Salay, Rick, Rodrigo Queiroz, and Krzysztof Czarnecki. "An analysis of ISO 26262: Using machine learning safely in automotive software." arXiv preprint arXiv:1709.02435 (2017).
- 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).

ML-SPECIFIC FAULT TOLERANCE PATTERNS

- Ensemble learning methods
  - e.g. multiple classifiers for pedestrian detection
- Safety envelope (hard-coded constraints on safe solutions)
  - e.g. combine ML-based pedestrian detector with programmed object detector for obstacle avoidance
- Simplex architecture (conservative approach on low-confidence predictions)
  - e.g. slow down if obstacle is detected, but kind/trajectory of obstacle unclear
- Runtime verification + Fail Safety (partial specs)
  - e.g. detect whether detected pedestrian detector behavior violates partial specification at runtime (plausibility checks)
- Data harvesting (keep low confidence data for labeling and training)
  - e.g. pedestrian detector's safe low confidence predictions saved for offline analysis

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).

THE UBER CRASH

Speaker notes

investigators instead highlighted the many human errors that culminated in the death of 49-year-old Elaine Herzberg. Driver was reportedly streaming an episode of The Voice on her phone, which is in violation of Uber's policy banning phone use. In fact, investigators determined that she had been glancing down at her phone and away from the road for over a third of the total time she had been in the car up until the moment of the crash.

woefully inadequate safety culture

federal government also bore its share of responsibility for failing to better regulate autonomous car operations

The company also lacked a safety division and did not have a dedicated safety manager responsible for risk assessment and mitigation. In the weeks before the crash, Uber made the fateful decision to reduce the number of safety drivers in each vehicle from two to one. That decision removed important redundancy that could have helped prevent Herzberg's death.

(from https://www.theverge.com/2019/11/20/20973971/uber-self-driving-car-crash-investigation-human-error-results)

SAE SELF-DRIVING LEVELS

- Level 0: No automation
- Level 1: Driver assistance
  - Speed xor steering in certain conditions; e.g. adaptive cruise control
  - Driver fully active and responsible
- Level 2: Partial automation
  - Steer, accelerate and break in certain circumstances, e.g. Tesla Autopilot
  - Driver scans for hazards and initiates actions (lane changes)
- Level 3: Conditional automation
  - Full automation in some conditions, Audi Traffic Jam Pilot
  - Driver takes over when conditions not met
- Level 4: High automation
  - Full automation in some areas/conditions, e.g. highways in good weather
  - No driver involvement in restricted areas
- Level 5: Full automation
  - Full automation on any road and any condition where human could drive

SAE Standard J3016

SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	**You <u>are</u> driving** whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			**You <u>are not</u> driving** when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			**These are automated driving features**		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering **OR** brake/ acceleration support to the driver	These features provide steering **AND** brake/ acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering **OR** • adaptive cruise control	• lane centering **AND** • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/ steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

For a more complete description, please download a free copy of SAE J3016: https://www.sae.org/standards/content/j3016_201806/

ROBUSTNESS DEFENSE

Use map with known signs as safety mechanism for hard to recognize signs

BUGS IN SELF-DRIVING CARS

- Study of 499 bugs of autonomous driving systems during development
- Many traditional development bugs, including configuration bugs (27%), build errors (16%), and documentation bugs
- All major components affected (planning 27%, perception 16%, localization 11%)
- Bugs in algorithm implementations (28%), often nontrivial, many symptoms
- Few safety-relevant bugs

Garcia, Joshua, Yang Feng, Junjie Shen, Sumaya Almanee, Yuan Xia, and Qi Alfred Chen. "A Comprehensive Study of Autonomous Vehicle Bugs." ICSE 2020

SAFETY CHALLENGES WIDELY RECOGNIZED

Being able to apply ML in safety-critical applications will be important to my organization in the future	a)	
V&V of features that rely on ML is recognized as a particularly challenging area in my organization	b)	
My organization is well-prepared for a future in which V&V of safety-critical ML is commonplace	c)	

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).

CHALLENGES DISCUSSED FOR SELF-DRIVING CARS

- No agreement on how to best develop safety-critical DNN
- Research focus on showcasing attacks or robustness improvements rather than (system-level) engineering practices and processes
- Pioneering spirit of AI clashes with conservatism of safety engineering
- Practitioners prefer simulation and tests over formal/probabilistic methods
- No consensus on certification and regulation, gap in safety standards

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).

SAFETY CAGES

- Encapsulate ML component
- Observe, monitor with supervisor
- Anomaly/novelty/out-of-distribution detection
- Safe-track backup solution with traditional safety engineering without ML

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).

AUTOMATION COMPLACENCY

IF TRADITIONAL VERIFICATION DOESN'T WORK, NOW WHAT?

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

FOLLOW RESEARCH

- Understand safety problems and safety properties
- Understand verification techniques (testing, formal, and probabilistic)
- Understand adversarial attack and defense mechanisms
- Anomaly detection, out of distribution detection, drift detection
- Advances in interpretability and explainability
- Human-ML interaction, humans in the loop designs and problems

Starting point: 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): 100270.

DON'T FORGET THE BASICS

- Hazard analysis
- Configuration management
- Requirements and design specifications
- Testing

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


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?

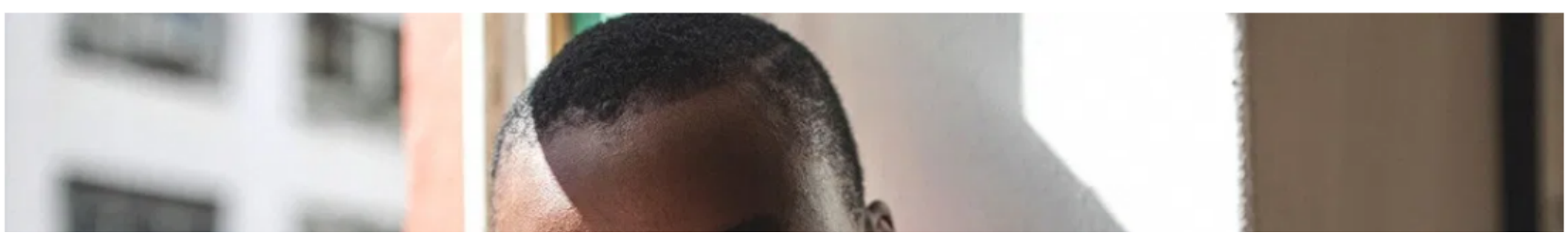
MENTAL HEALTH

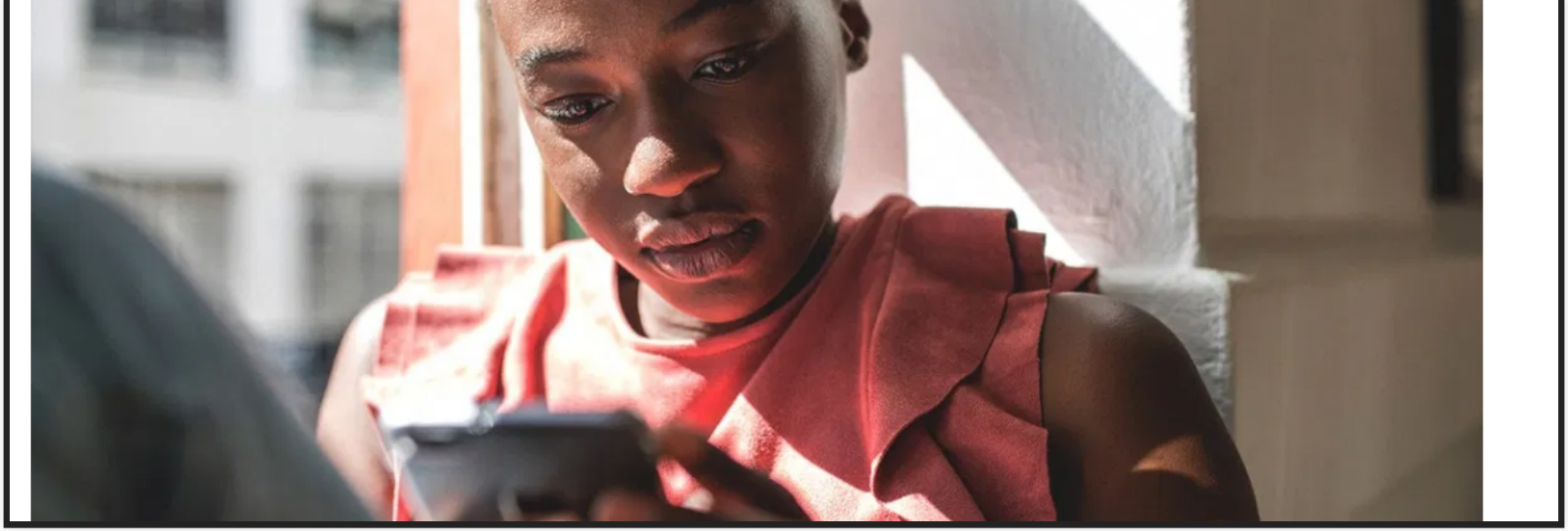
HEALTH NEWS ✓ Fact Checked

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

Written by Gigen Mammoser on December 10, 2018

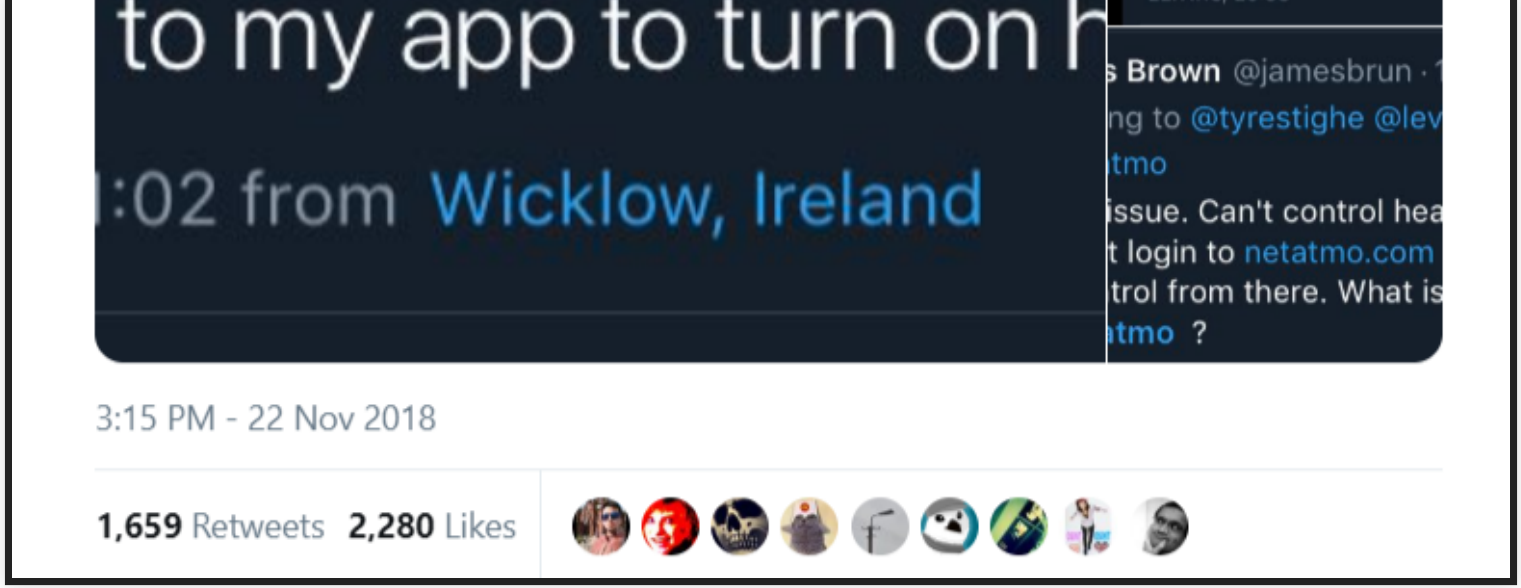
New research reveals how social media platforms like Facebook can greatly affect your mental health.



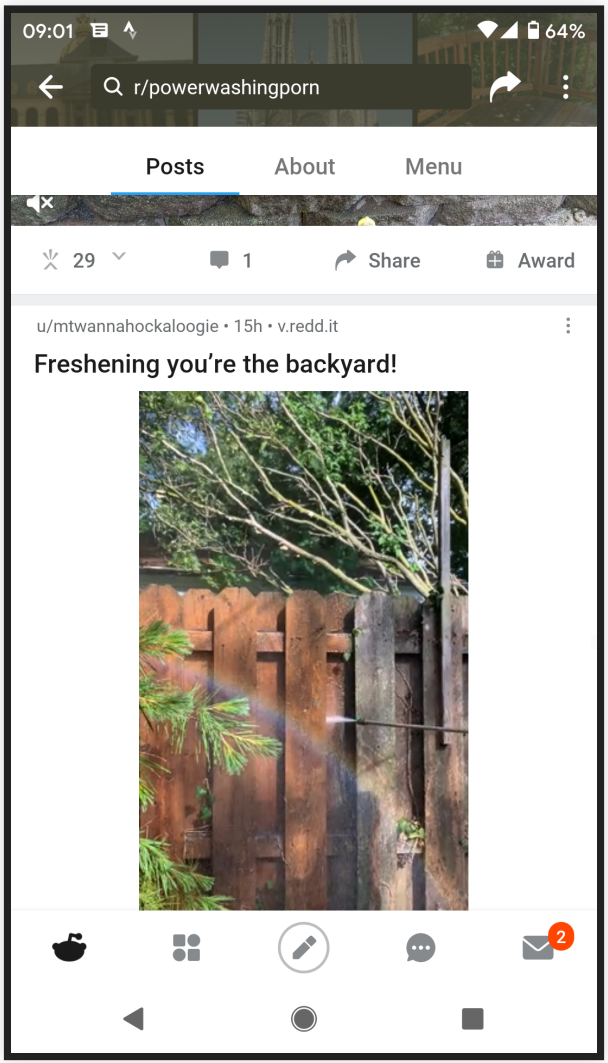


IOT


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
eran
DivemasterK
no Are your servers do
Kiran vadgama
@kiran_vadgama
netatmo hi my manual ovrride
ermostat is not working and wh
ng the app it comes up with ar
h servers down. Can i overide
d?
1.18, 20:58
Andy Mc
@ITakeSugar
Replying to @levisleedaniel and
Is there a way to contro
the servers are down, it
moment
22.11.18, 20:38


to my app to turn on h
:02 from Wicklow, Ireland
s Brown @jamesbrun · 1
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.

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

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
- AI goals are difficult to specify precisely, reward hacking
- Hazard analysis to identify safety risks and requirements; classic safety design at the system level
- Model robustness can help with some problems
- Self-driving cars are challenging and evolving