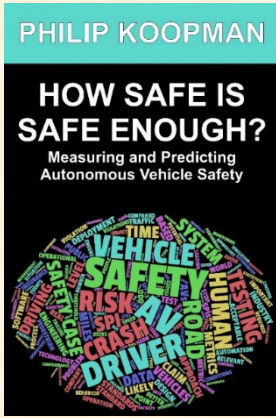
Prof. Philip Koopman

Understanding Self-Driving Vehicle Safety

October 2024



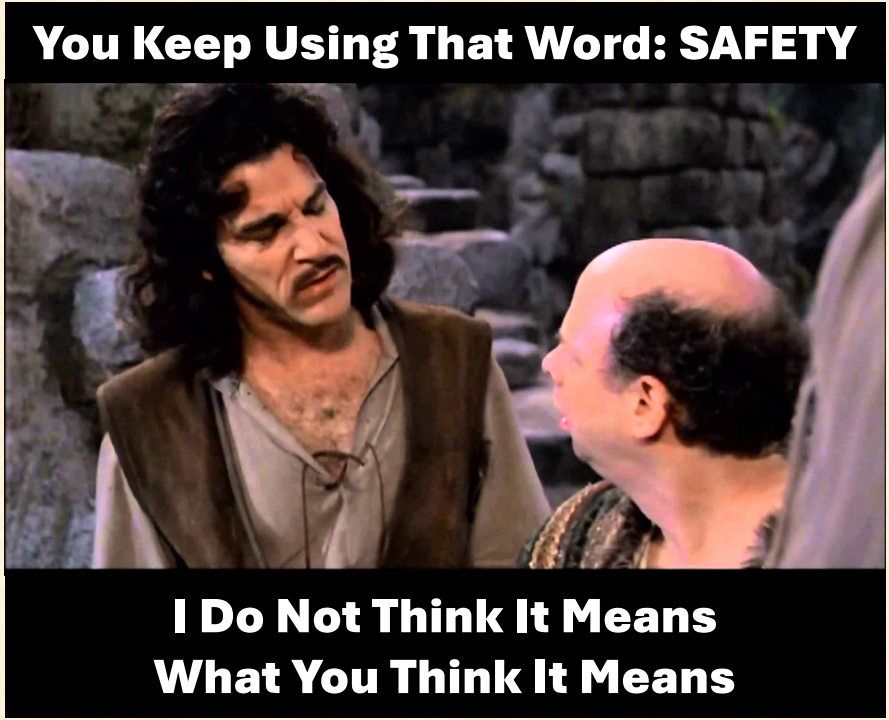
PhilKoopman.Substack.com

Carnegie Mellon University

Overview

Carnegie Mellon University

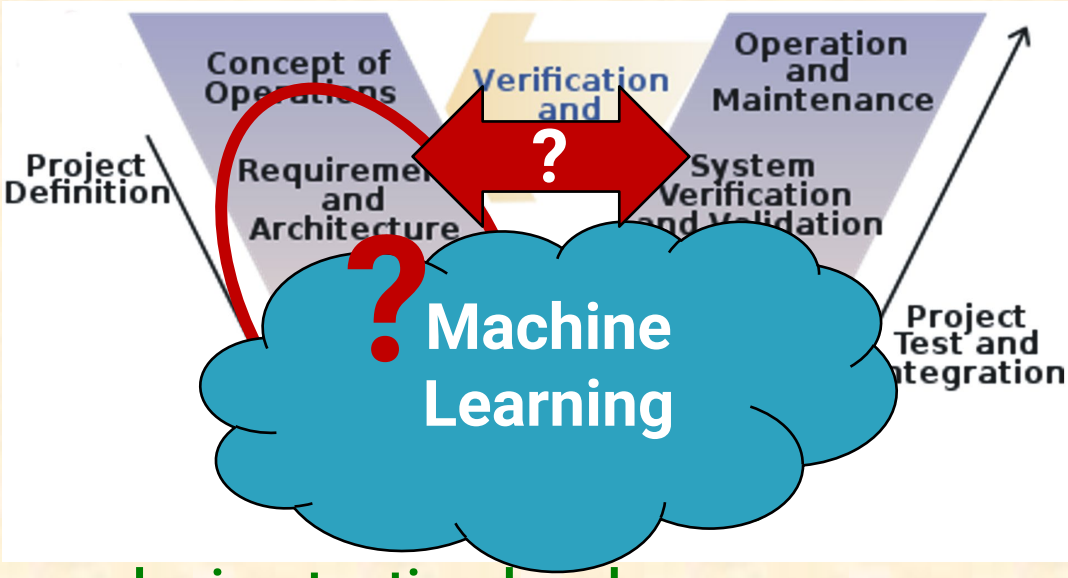
- Understanding “safe enough” for Autonomous Vehicles
  - More than Positive Risk Balance
  - Equity, negligence, recalls, …
- Changing the scope of safety
  - Beyond net risk minimization
    - ➔ Multi-constraint optimization
  - Terminology affects how we think about safety
  - Applies to anything autonomous



Background: Machine Learning & Vee Model

Carnegie Mellon University

- Traditional safety engineering Vee model
  - Trace requirements to implementation
  - Testing validates the engineering process
  - Engineering rigor sets a prior expectation of safety, reducing testing burden



- Machine Learning (inductive training) breaks the Vee
  - More testing required due to degraded prior belief in safe design
  - Using simulation just shifts the validation burden for edge cases

What Should We Mean By Safe?

Carnegie Mellon University

- Are loss events all that matter?
- Is it statistical parity with (sometimes drunk) human drivers?
- In reality, it takes a lot more
  - #1: Positive Risk Balance (PRB)
  - #2: Avoiding risk transfer
  - #3: Avoiding negligent driving
  - #4: Safety standards conformance
  - #5: Specific risk mitigation / recalls
  - #6: Ethical & equity concerns
  - #7: Sustainable trust

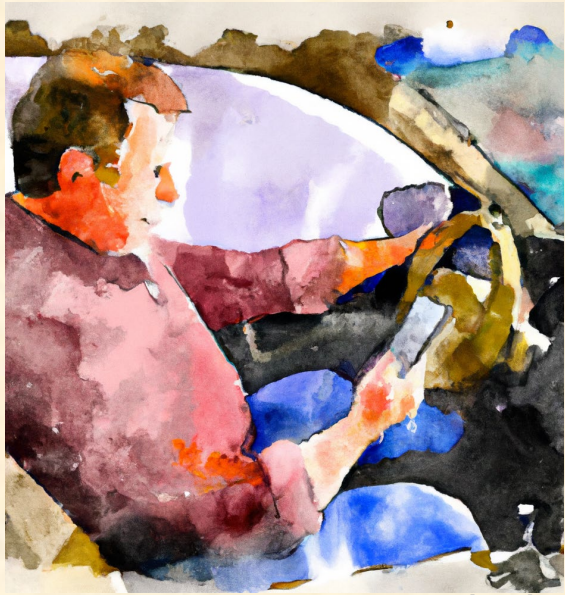
August 2023
Nobody was hurt.
Does that make this safe?

Reference: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634179

#1: PRB – Which, Where, Who?

Carnegie Mellon University

- Positive Risk Balance: safer than a human driver
- But which human driver?
  - 28% Alcohol/driving under influence fatalities
  - 26% speed-related, 9% distracted, 2% drowsy
  - 60 year old driver is ~3.5x better than 16 y.o.
- Where/Who?
  - 3.4x fatality per VMT variation by US state
  - Victim demographic (e.g., pedestrians)
- Which vehicle?
  - New cars have active safety – BUT <u>average</u> car age >12 years



[Dall-e]

[DOT HS 813 060 & DOT HS 813 021] [AAA] [IIHS Fatality Fact Sheets State by State] [DOT HS 813 060]

#2: Avoid Risk Transfer

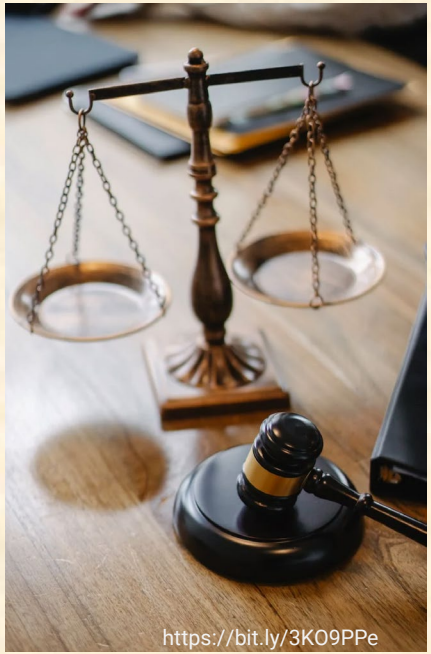
- What if children at greater risk?
  - Or disabled pedestrians?
  - Or bicyclists? Etc.
- Caution – this particular article's conclusion is controversial
  - Regardless, this is illustrates an important safety constraint
- Avoid increasing any group's risk
  - Spend extra effort decreasing risk to vulnerable groups

https://bit.ly/46Gehav

#3: Avoid Negligent Driving

Carnegie Mellon University

- "Negligent" robotaxi driving involves:
  - Establishing a duty of care to other road users
  - Breach of that duty of care causes a loss event
    - Ask: would a human driver have been negligent?
- Statistical safety arguments are irrelevant here
  - "Safe" drivers don't a free pass to run red lights
  - "We're saving lives" does not excuse negligence
- 2023 Cruise pedestrian dragging mishap:
  - Accelerated toward pedestrian in crosswalk
  - Moved after collision with pedestrian under vehicle



#4: Standards Set Expectation of Safety

Carnegie Mellon University

SYSTEM SAFETY	ANSI/UL 4600		**Safety Beyond Dynamic Driving**
DYNAMIC DRIVING FUNCTION	ISO 21448	SaFAD/ISO TR 4804	**Environment & Edge Cases**
FUNCTIONAL SAFETY	ISO 26262		**Equipment Faults**
CYBER-SECURITY	SAE J3061	SAE 21434	**Computer Security**
VEHICLE SAFETY	FMVSS	NCAP	**Basic Vehicle Functions**

REQUIRED

HIGHLY AUTOMATED VEHICLE SAFETY CASE

ANSI/UL 4600

Coming Soon:

- **ISO 8800**
- **ISO 5083**
- **etc.**

#5: Fine-Grain Risk & Regulators

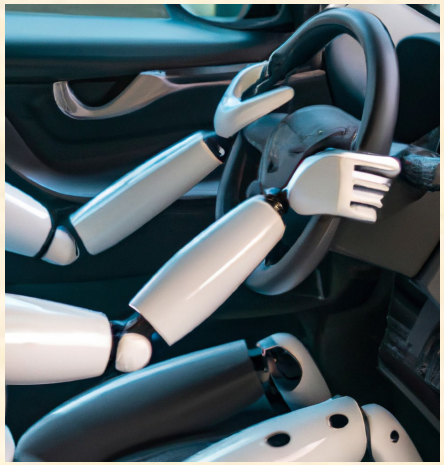
Carnegie Mellon University

- Want to avoid regulatory recalls
  - "Undue Risk" in the small – specific issues
  - Informed by vehicle test-centric standards
- Recalls are for specific defects, not net risk
  - Rolling through stop signs
  - Phantom braking
  - Malfunctioning display console
  - Software quality & net risk are typically beyond regulatory scope
- Regulators struggle with software safety
  - 2020 Proposal to require industry safety standards is inactive

#6: Ethical & Equity Concerns

Carnegie Mellon University

- Ride Hail made promises … with disappointing results
  - Why will robotaxis turn out any differently?
- Equity concerns:
  - Labor issues (e.g., displaced ride-hail/taxi drivers)
  - Will disabled community access really happen?
  - Cheap taxis undermine *safer* public transit



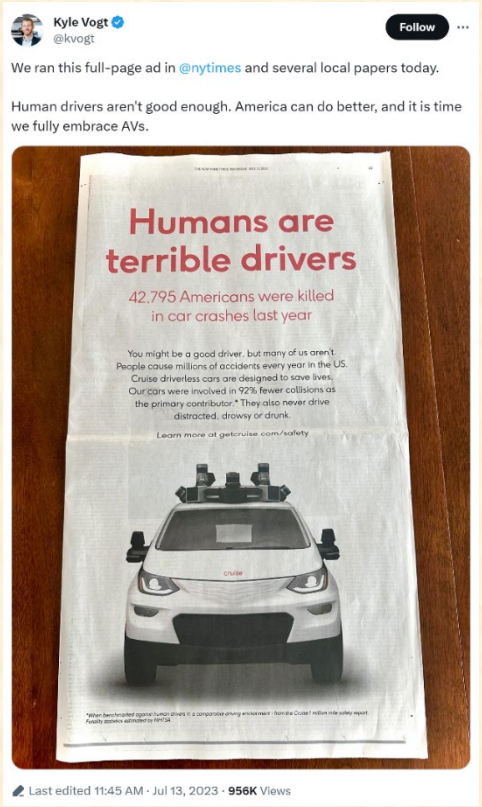
[Dall-e]

- Ethical & related concerns
  - Long-term aspirational safety at the cost of real short-term harm
  - No required independent safety technical oversight
    - Companies themselves make decision when to pull the safety driver

#7: Sustainable Trust

Carnegie Mellon University

- Trust-degrading rhetoric:
  - "Robotaxis won't make stupid driving mistakes"
  - Relentless blame of human drivers
- Trust-degrading actions:
  - Lobbying for municipal preemption
  - Redacting & withholding information
- Toward increasing trust:
  - More transparency on incidents & follow-up
  - Accepting proportional responsibility for losses
  - Stating & tracking release criteria

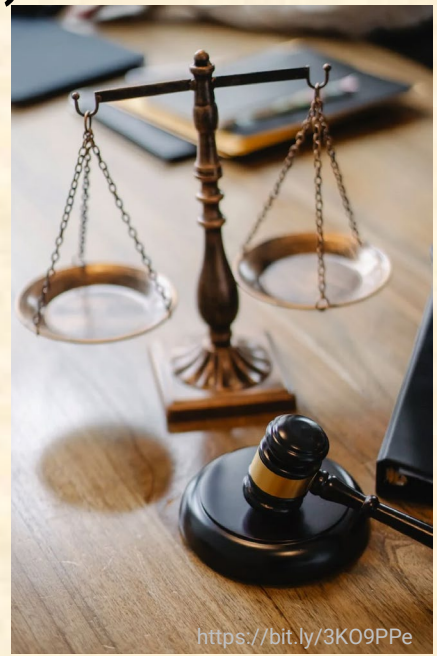


https://twitter.com/kvogt/status/1679517290847694848

Scope of "Acceptably Safe" Claim

Carnegie Mellon University

- Net statistical safety (safer than average driver?)
  - Establishing a baseline is very complex!
- What tolerance for risk transfer?
  - What if pedestrian risk doubles? (etc.)
- What tolerance for negligent behavior?
  - What if breaking a traffic rule results in harm?
- Fine-grain absence of unreasonable risk
  - Recalls tend to be for specific behaviors
- Ethical behavior & equity concerns
  - Consequences of testing & deployment decisions



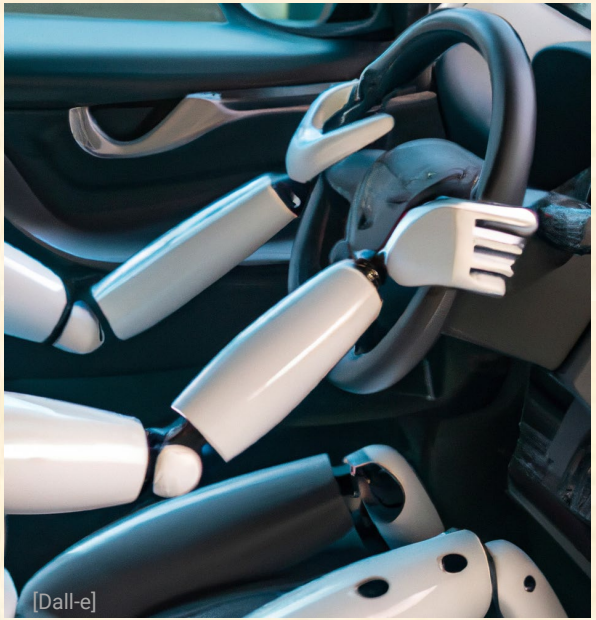
Reference: Redefining Safety for AVs https://arxiv.org/abs/2404.16768

Expanding The Safety Discussion

Carnegie Mellon University

- Time for safety engineering to evolve
  - Autonomous systems show us where to improve
- Definitional build-up:

  Loss
  ➔ Risk
  ➔ Safety Constraint
  ➔ Safety Engineering
  ➔ Safety Case
  ➔ Acceptable safety
- Viewpoint: safety as multi-constraint satisfaction rather than optimization


[Dall-e]

Is "Safety Case" Definition Broken?

Carnegie Mellon University

- DefStan 00-56: "… in a given operating environment"
  - Changing, incompletely defined environments
  - Unexpected obstacles, vehicle types, etc.

Crash into utility pole

Crash into articulated bus

Is "Risk" Definition Broken?

Carnegie Mellon University

- Typically: combination of probability and severity
  - See also *Positive Risk Balance ("safer than human driver")*
  - What about risk redistribution onto vulnerable populations?
- 11 of 74 SF Fire Dept. robotaxi incidents in Tenderloin District
  - Economically distressed
  - High drug use
- Mishaps at edge of Tenderloin:
  - Cruise fire truck crash
  - Cruise pedestrian dragging

Statistic: https://bit.ly/3WtI3NU // annotated Google Map image

Expanding The Scope of "Safety"

Carnegie Mellon University

- Robots can fail even if they do not drive drunk
  - Is negligent driving OK?
  - Is uneven risk distribution OK?
  - Should losses due to "rare" events be OK?
- No human operator to blame
  - Who is responsible for negligent behavior?
  - Who/what monitors "for a given environment"?
  - Social interactions are in-scope for technology
- Let's explore revising safety terminology

Definition of Loss

- ISO 26262 Harm: physical injury to people
  - But what about other incidents?

- Loss: an adverse outcome, including damage to the system itself, negative societal externalities, damage to property, damage to the environment, injury or death to animals, and injury or death to people

Autonomous Waymo car runs over dog in San Francisco

The vehicle was in autonomous mode with a safety driver present in a 25 mph zone.

RON AMADEO - 6/7/2023, 2:24 PM

https://bit.ly/4cLX2s4

Definition of Risk

Carnegie Mellon University

- Classical risk: combination of probability and severity
  - ISO 26262 includes controllability
  - But, we see recalls for *patterns* of losses
- NHTSA EA22002 / Recall 23V838
  - 956 Tesla crashes/ 29 fatalities https://bit.ly/4cChQ4z
  - Avoidable crashes, loss of yaw control
  - Inadvertent AutoSteer override

Federal regulator finds Tesla Autopilot has 'critical safety gap' linked to hundreds of collisions https://bit.ly/3SXklHr

The NHTSA report comes as Tesla signals it is betting its future on autonomous driving.

- Risk: combination of the probability of occurrence of a loss, or pattern of losses, and the importance to stakeholders of the associated consequences

Definition of Safety Constraint

Carnegie Mellon University

- Is safety net minimizing the sum of risks?
  - Near zero probability * catastrophic consequence = ???
- Risk due to negative externalities
  - How does design team assign consequence to blocking a fire truck?
- Rules & regulations help here
  - Reasonable road rule violations??
- Safety constraint: a limitation imposed on risk or other aspects of the system by stakeholder requirements

San Francisco's fire chief is fed up with robotaxis that mess with her firetrucks. And L.A. is next

https://bit.ly/3Wc3bXA

San Francisco Fire Chief Jeanine Nicholson says state regulators are moving too fast on robotaxi expansion, jeopardizing public safety. Meanwhile, Waymo and Motional are planning to begin robotaxi service in Los Angeles. (Lea Suzuki / San Francisco Chronicle via Associated Press)

Definition of Safety Engineering

- Testing alone does not create safe software

But … arguing safety via brute force testing is a pervasive narrative

- Safety engineering: a methodical process of ensuring a system meets all its safety constraints throughout its lifecycle, including at least hazard analysis, risk assessment, risk mitigation, validation, and field engineering feedback

Definition of a Safety Case

Carnegie Mellon University

- Safety case: … "given application in given environment"
  - Who/what enforces operational limits?
  - What if the environment is unknowable in full?
  - Foreseeable Misuse/abuse?

- Safety case: structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and sound argument that safety engineering efforts have ensured a system meets a comprehensive set of safety constraints

Definition of Acceptable Safety

Carnegie Mellon University

- More to safety than positive risk balance
  - Meet ethical constraints (e.g., risk distribution)
  - Non-negligent driving (e.g., justifiable road rule violation)
  - No recallable behaviors (even if net risk is OK)
  - Meet legal restrictions (e.g., passenger drop-off)
- Net acceptability across all stakeholders
  - Auto industry, insurance industry
  - Regulators, legislators
  - Road users, consumer advocates

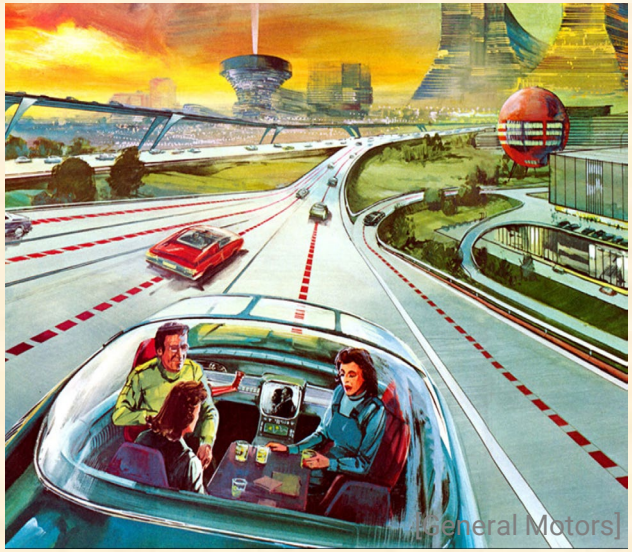
[Dall-e]

Acceptable: meets all safety constraints as shown by a safety case

Summary

Carnegie Mellon University

- New definitions needed – no human driver to handle:
  - Surprises in environment
  - Enforcement of operational limits
  - “Do the right thing” rule interpretation
  - Legal and ethical constraints
- Can’t we just re-interpret existing terms?
  - Minimal-compliance organizations are likely to fall short
  - Terms should say what they mean



- Extended paper compares to specific safety standards

https://arxiv.org/abs/2404.16768

Collected Definitions:

- **Loss:** an adverse outcome, including damage to the system itself, negative societal externalities, damage to property, damage to the environment, injury or death to animals, and injury or death to people
- **Risk:** combination of the probability of occurrence of a loss, or pattern of losses, and the importance to stakeholders of the associated consequences
- **Safety constraint:** a limitation imposed on risk or other aspects of the system by stakeholder requirements
- **Safety engineering:** a methodical process of ensuring a system meets all its safety constraints throughout its lifecycle, including at least hazard analysis, risk assessment, risk mitigation, validation, and field engineering feedback
- **Safety case:** structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and sound argument that safety engineering efforts have ensured a system meets a comprehensive set of safety constraints
- **Acceptable:** meets all safety constraints as shown by a safety case

Resources

Carnegie Mellon University

- Talks & papers on autonomous vehicle safety:
  - Video talks: https://bit.ly/KoopmanTalks
  - Papers: https://bit.ly/KoopmanTalks
- "Safe Enough" book & talk video:
  - https://safeautonomy.blogspot.com/2022/09/book-how-safe-is-safe-enough-measuring.html
- UL 4600 AV safety standard book & talk video:
  - https://safeautonomy.blogspot.com/2022/11/blog-post.html
- Liability-based proposal for state AV regulation & podcast
  - https://safeautonomy.blogspot.com/2023/05/a-liability-approach-for-automated.html
- US Congressional House E&C testimony:
  - https://safeautonomy.blogspot.com/2023/07/av-safety-claims-and-more-on-my.html