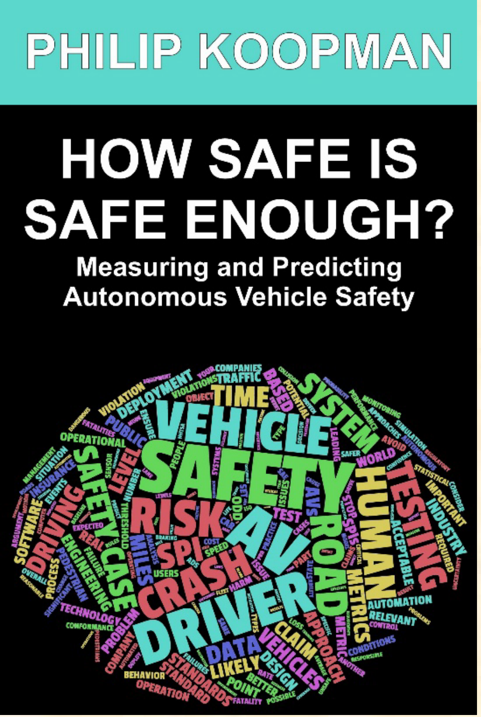


PHILIP KOOPMAN

HOW SAFE IS SAFE ENOUGH?

Measuring and Predicting Autonomous Vehicle Safety

How Safe Is Safe Enough for Autonomous Vehicles?

Prof. Philip Koopman

November 2022

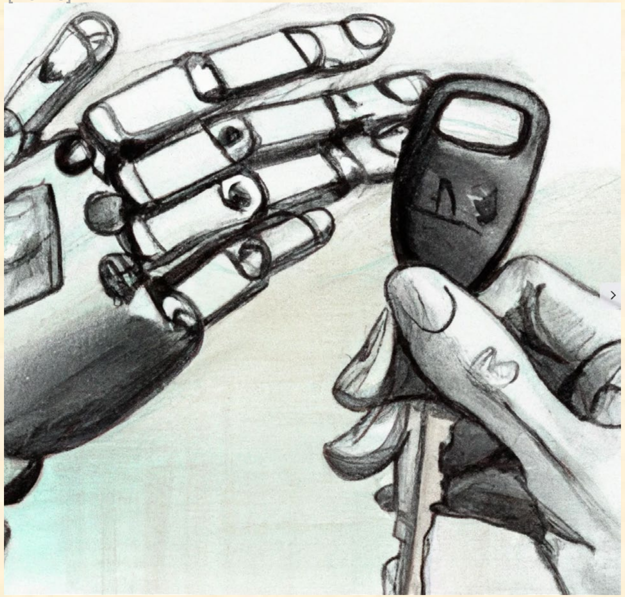
www.Koopman.us

Carnegie Mellon University

Overview

- Risk management frameworks
  - Which human is a baseline driver?
  - Risk mitigation is not safety
- Uncertainty as a limiting factor
  - Predicting safety before deployment
  - Field feedback to manage uncertainty
- A broader view of Safe Enough
  - Ethical considerations
  - Hierarchical model of safety needs
- Deployment criteria

[Dall-e]



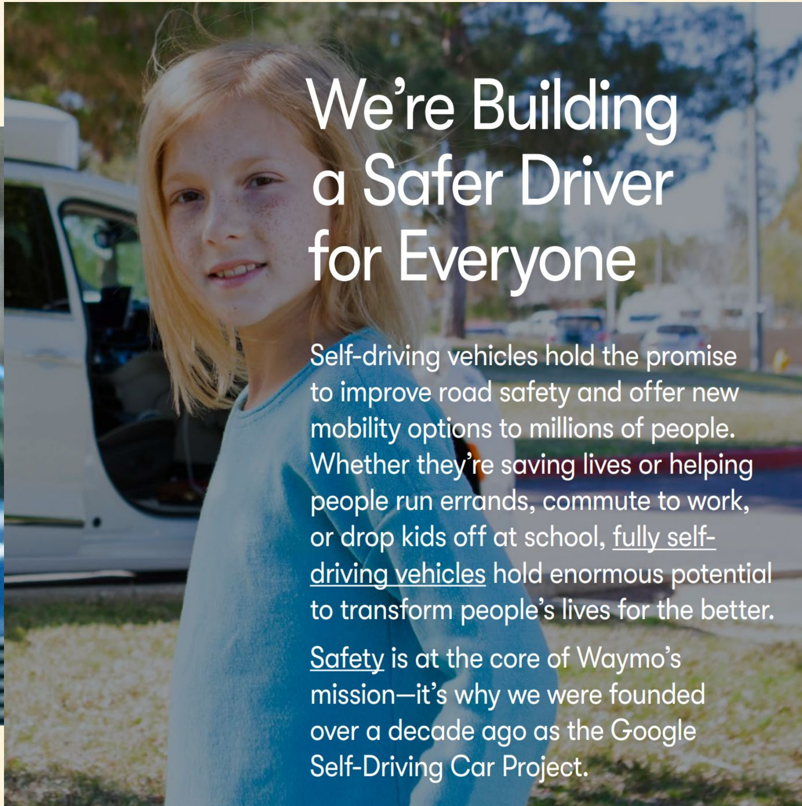
ADS = Automated Driving System

(Car drives; people can sleep)

ADS Technology: Sold Based on Safety

Ford VSSA   https://bit.ly/3njionT

Waymo VSSA https://bit.ly/2QuYhai



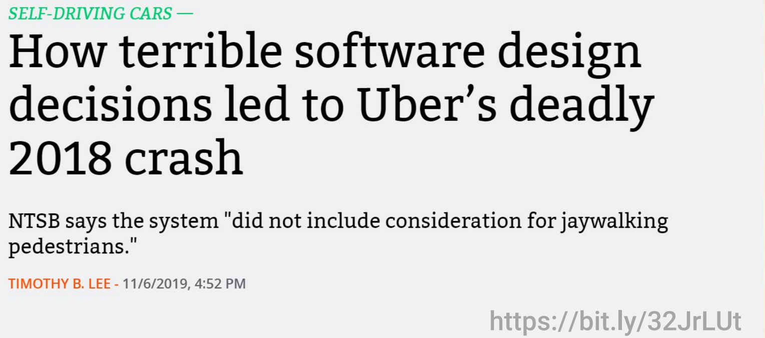
We're Building a Safer Driver for Everyone

Self-driving vehicles hold the promise to improve road safety and offer new mobility options to millions of people. Whether they're saving lives or helping people run errands, commute to work, or drop kids off at school, fully self-driving vehicles hold enormous potential to transform people's lives for the better.

Safety is at the core of Waymo's mission—it's why we were founded over a decade ago as the Google Self-Driving Car Project.

Safe Enough Based On News Cycle?

- Newsworthy crashes might not predict safety
  - Crewed testing is not autonomous
  - Crash reports need a denominator
- Need a framework for evaluating safety beyond the news cycle



SELF-DRIVING CARS —

How terrible software design decisions led to Uber's deadly 2018 crash

NTSB says the system "did not include consideration for jaywalking pedestrians."

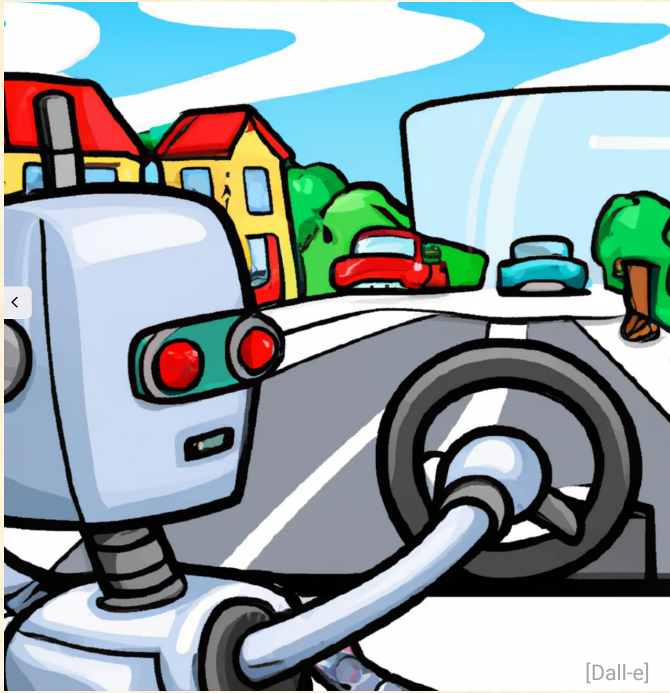
TIMOTHY B. LEE - 11/6/2019, 4:52 PM

https://bit.ly/32JrLUt

https://bit.ly/3AupcWb

Ethics: The Blame Game

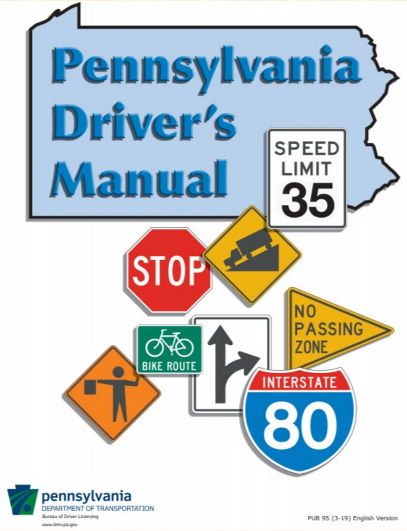
- Companies blame human drivers for bad news
  - Humans are terrible at supervising automation
  - Maybe driver monitoring helps(?)
- The Moral Crumple Zone:
  - Blame the most convenient human for failing to mitigate technical malfunctions
- Regulatory strategy: computer is driver
  - Not a legal person, so … crashes are nobody's fault (???)



[Dall-e]

How About A Robot Driver Test

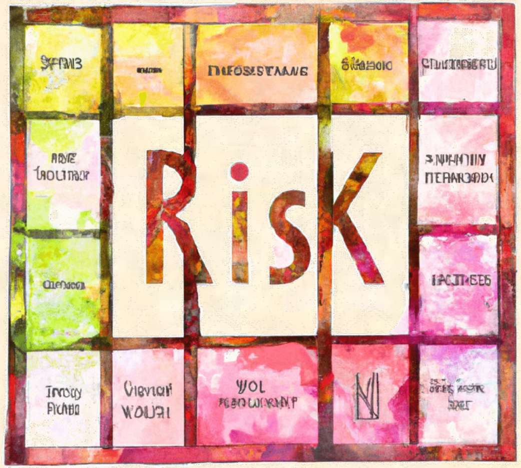
- Written test
  - Does ADS know traffic laws & behaviors?
- Road test
  - Can ADS obey traffic laws?
  - Can ADS negotiate effectively with human drivers?
  - Can ADS resolve potentially ambiguous situations?
- Being a 16 year old human
  - How do we measure ADS judgment maturity?
  - Autonomous systems struggle with novelty, unknowns


Pennsylvania Driver's Manual
SPEED LIMIT 35
STOP
NO PASSING ZONE
BIKE ROUTE
INTERSTATE 80
pennsylvania DEPARTMENT OF TRANSPORTATION
PUB 95 (3-19) English Version

➔ Need safety engineering, not just a driver test

Setting The Risk Goal

- MEM – Minimum Endogenous Mortality
  - System risk has minimal effect on overall risk
- ALARP – As Low As Reasonably Practicable
  - Reduce identified risks unless cost is extreme
- NMAU – “Nicht Mehr Als Unvermeidbar”
  - Reduce identified risks within reasonable cost
- SIL – Safety Integrity Level approaches
  - Engineering rigor applied to mitigate risks
- GAMAB – “Globalement Au Moins Aussi Bon”
  - At least as good as an existing system (e.g., a human driver)



[Dall-e]

Positive Risk Balance (PRB)

- Utilitarian GAMAB approach
  - 36,096 fatalities (1.10/100M miles)
  - 2,740,000 injuries
  - 6,756,000 police-reported crashes
  - Data includes drunk drivers, speeders, no seat belts

U.S. Department of Transportation
National Highway Traffic Safety Administration

TRAFFIC SAFETY FACTS
Research Note

NHTSA

DOT HS 813 060 December 2020

[DOT HS 813 060 & DOT HS 813 021]

2019 Data

➔ Expect zero deaths in a 10M mile testing campaign

- The averages do not necessarily apply
  - Which driver?
  - Under what conditions?
  - Driving which vehicle?

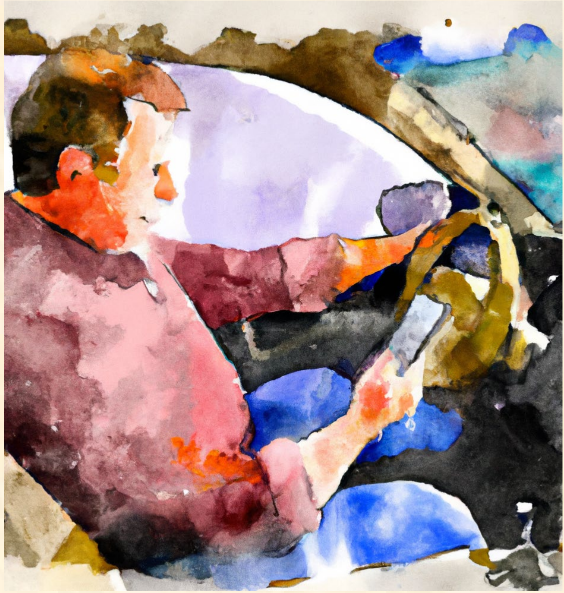
Which Driver Are We Better Than?

Carnegie Mellon University

- ~100M miles/fatal mishap for human drivers
  - 28% Alcohol impaired/Driving Under Influence
  - 26% Speed-related
  - 9% distracted driving
  - 2% drowsy ... [DOT HS 813 060 & DOT HS 813 021]

(total > 100% due to multiple factors in some mishaps)

- Fully functional drivers are much safer
- New AV has better safety than 10+ year old "average" car



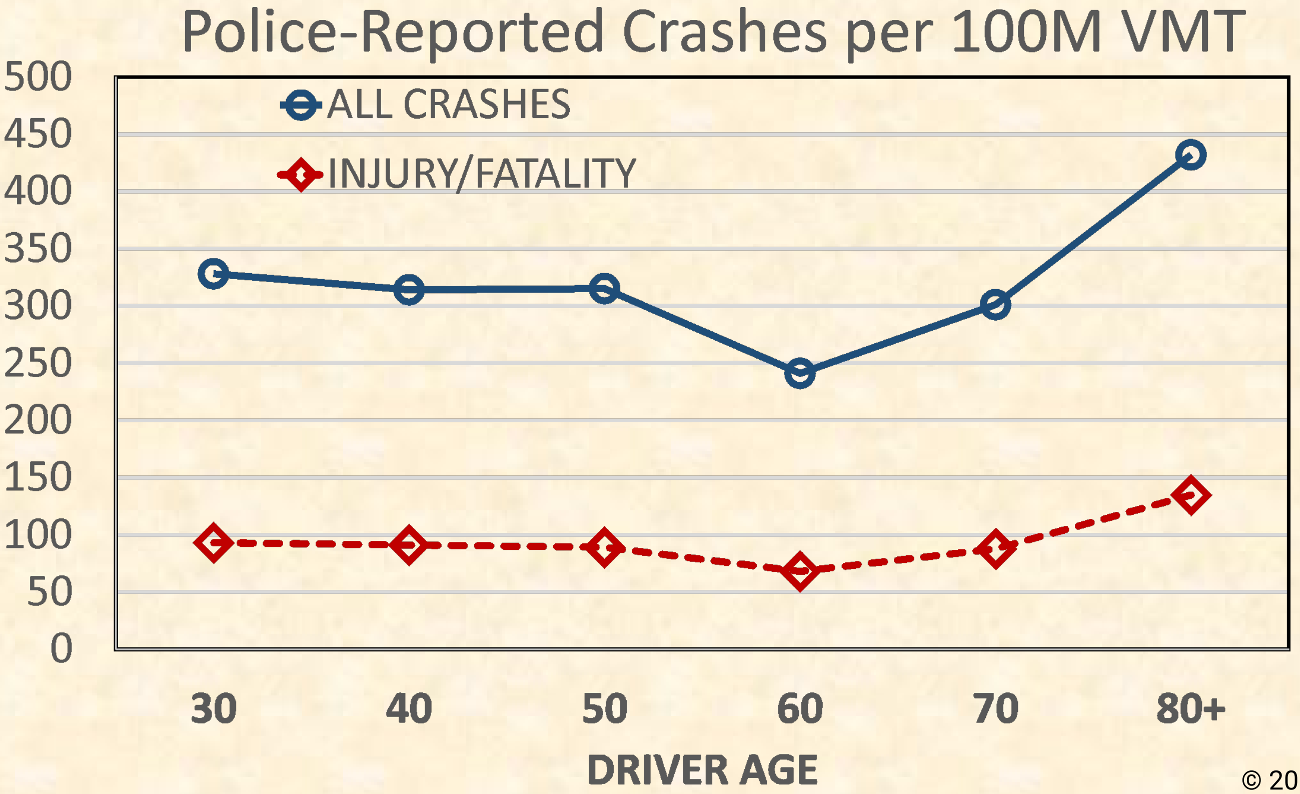
[Dall-e]

➔ Better than an unimpaired, undistracted driver in new car

Driver Age Affects Crash Rates

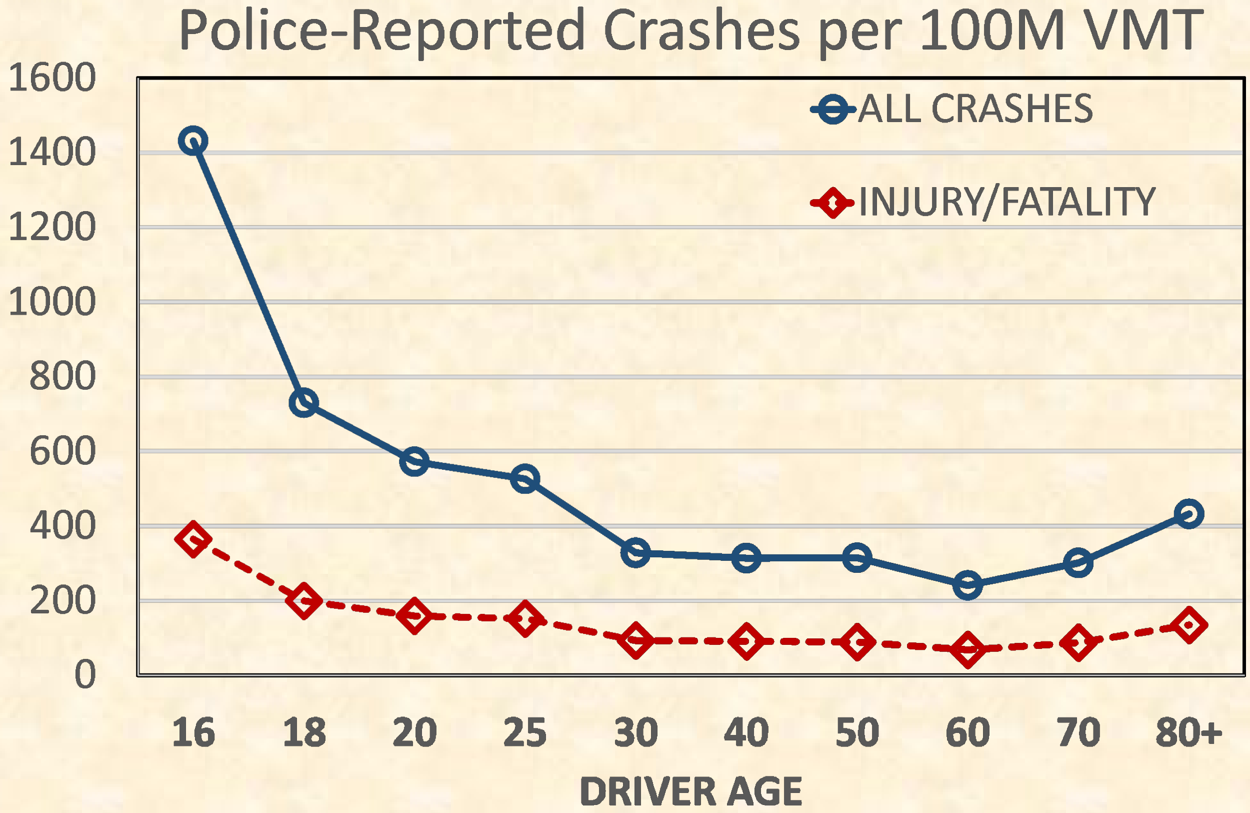
Carnegie Mellon University

- Are older drivers worse? (caution – not the whole story!)



Driver Age Affects Crash Rates

➔ Better than a middle-aged driver



Region Affects “Safe Enough” Value

- Fatality averages for 2019 (IIHS)

Location	Deaths/100K people		Deaths/100M miles		
DC	3.3		MA	0.51	
US	11.0	7.7x	US	1.11	3.4x
WY	25.4		SC	1.73	

- Fatal crash type

[IIHS Fatality Fact Sheets State by State; DOT HS 813 060]

  - DC: highest pedestrian rate (39%)
  - NY, FL, DE: highest bicycle rate (5%)
  - Fatalities per 100M miles: Urban 0.86 vs. Rural 1.65
  - What about day/night, weather, etc.?

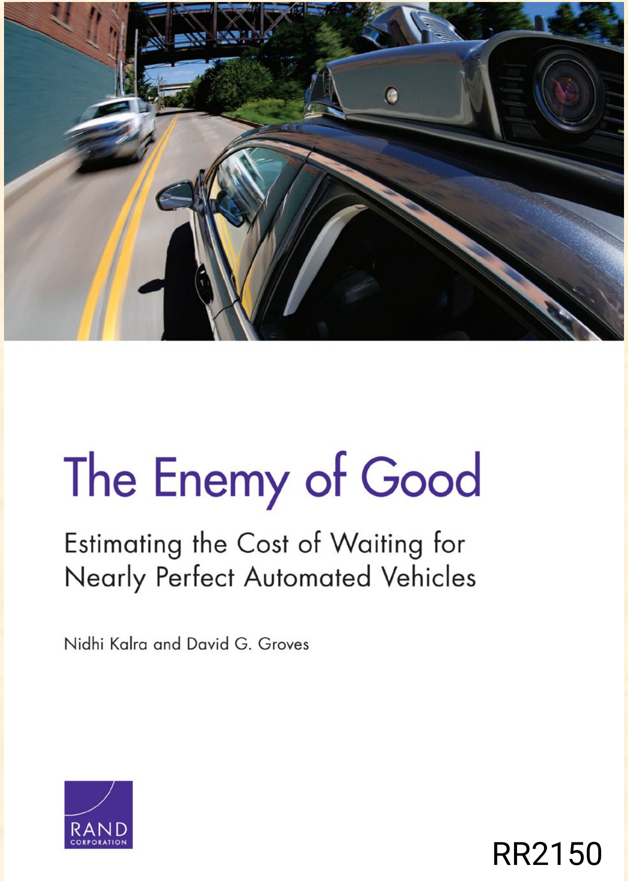
➔ Better in same conditions as AV operations

https://bit.ly/3CJm7nP

When Do We Deploy?

- Assume we determined a human driver baseline for comparison
  - Competent, unimpaired middle-age driver
  - Same operational conditions as AV (location, time of day, weather, …)
- RAND report says only 10% better than human driver is a safety win
  - But, this assumes accurate estimate of safety is available before deployment
  - What if estimate is 5x too optimistic?

➔ Need to address uncertainty



The Enemy of Good

Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles

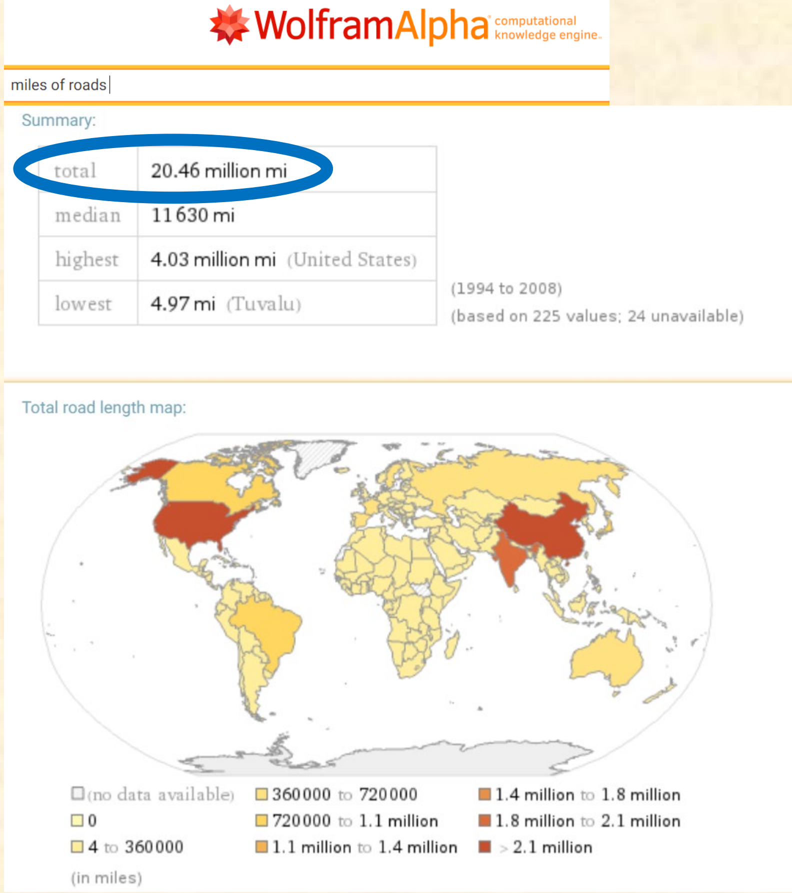
Nidhi Kalra and David G. Groves

RAND CORPORATION

RR2150

Validation Via Brute Force Road Testing?

- If 200M miles/critical mishap…
  - Test 3x–10x longer than mishap rate
    ➔ Need 2 Billion miles of testing
- That's ~50 round trips
  on every road in the world
  - With fewer than 10 critical mishaps
  - Even more testing if you find a defect and redo some testing
- Road testing leaves uncertainty



miles of roads

Summary:

total	20.46 million mi
median	11630 mi
highest	4.03 million mi (United States)
lowest	4.97 mi (Tuvalu)

(1994 to 2008)
(based on 225 values; 24 unavailable)

Total road length map:

□ (no data available)	🟨 360000 to 720000	🟧 1.4 million to 1.8 million
□ 0	🟨 720000 to 1.1 million	🟧 1.8 million to 2.1 million
□ 4 to 360000	🟧 1.1 million to 1.4 million	🟥 > 2.1 million

(in miles)

Do Lots of Simulation

- Highly scalable; fidelity vs. cost tradeoff
  - Need to build highly detailed models (modeling errors?)
  - Challenge of matching real world data into simulation models
  - Only tests things you have thought of ➔ residual uncertainty

Mairie
de St FORGET

[ANSYS]

How Much Do You Trust Validation?

- Would you put a child in front of an AV validated with:
  - 10,000M mile sims
    … perhaps with a simulator error?
  - Based on 100M miles road data collected
    … perhaps with scenario analysis errors?
  - Validated by 10M miles of road testing
    … that missed the above errors?
  - And 10K repetitions of closed course testing
    … with standard dummies instead of people
  - Built with biased perception training data?
  - Using software binaries & tools
    … with no safety qualification?

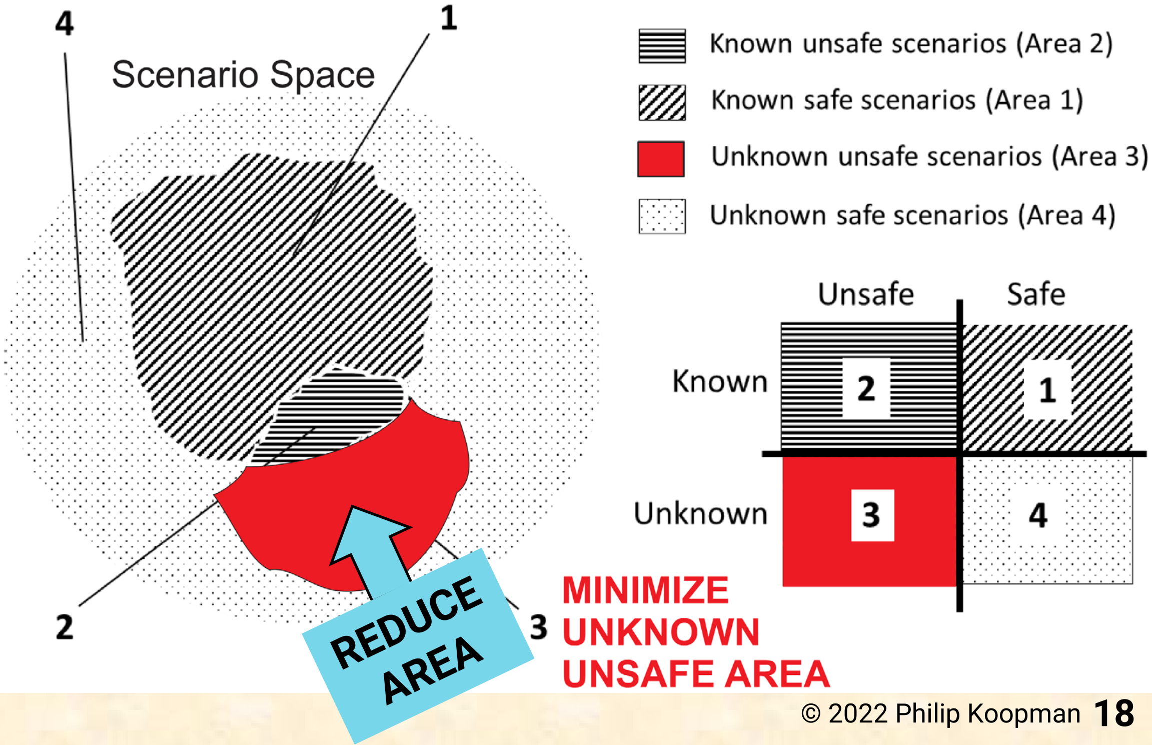
Engineering Rigor

- Testing alone is insufficient for life-critical systems
  - So we use also use engineering rigor
- Can you trust the system itself?
  - Is it engineered for safety?
  - Were standards and best practices used?
  - Is there a safety case documenting all this?
- Can you trust your validation process?
  - Did you engineer the simulations properly?
  - Did you design the validation campaign properly?

Identifying & Mitigating Hazards

- ISO 26262: Hazard and Risk Analysis (HARA)
  - Identify and mitigate risks per ASIL requirements

- ISO 21448: Identify and mitigate unsafe scenarios
  - Safety of the Intended Function (SOTIF)
  - Reduce "unknown unsafe" area
  - Deploy at acceptable residual risk



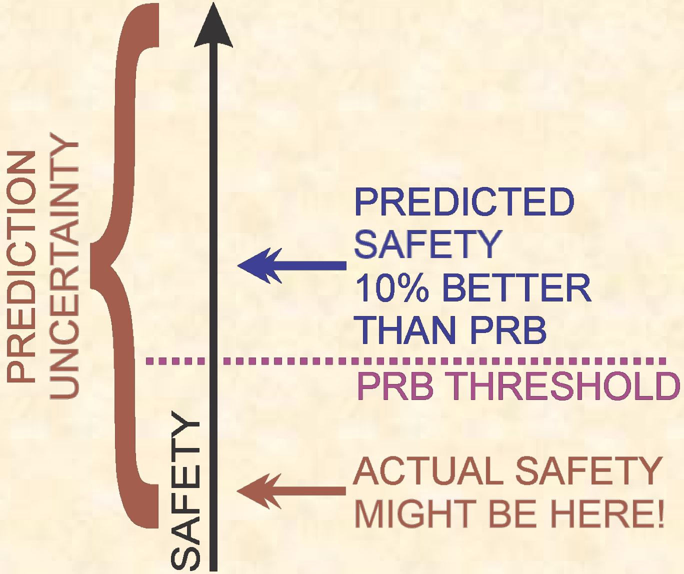
- ▤ Known unsafe scenarios (Area 2)
- ▨ Known safe scenarios (Area 1)
- 🟥 Unknown unsafe scenarios (Area 3)
- ⬚ Unknown safe scenarios (Area 4)

	Unsafe	Safe
Known	2	1
Unknown	3	4

MINIMIZE UNKNOWN UNSAFE AREA

Field Engineering Feedback

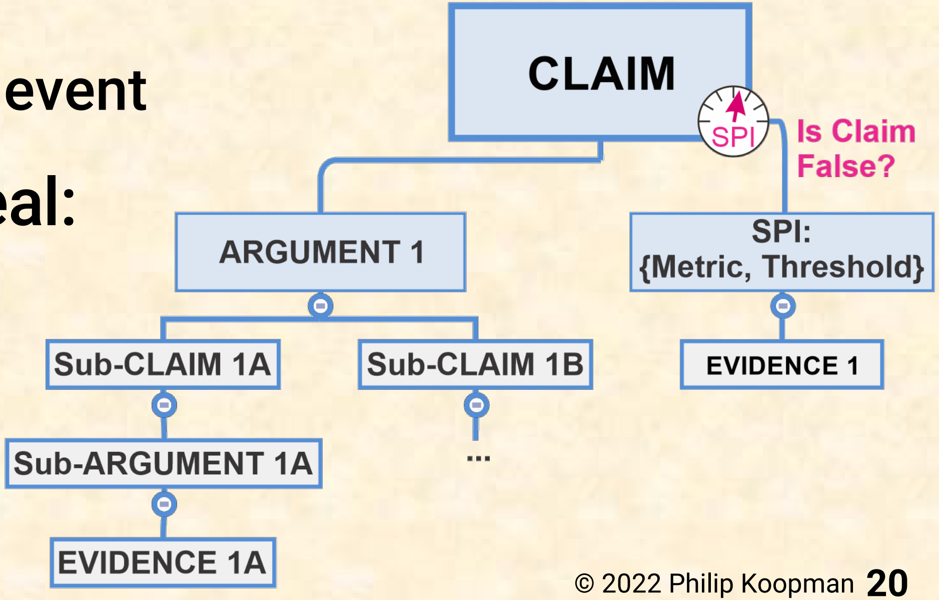
- Expected risk has a mean + uncertainty
  - Deploy only when mean is acceptable
  - But there will be uncertainty
    - Missed edge cases during road testing
    - Unknown gaps in validation plan
    - Unknown unknowns in general


PREDICTION UNCERTAINTY
SAFETY
PREDICTED SAFETY 10% BETTER THAN PRB
PRB THRESHOLD
ACTUAL SAFETY MIGHT BE HERE!

- Solution: manage uncertainty
  - Safety Performance Indicators (SPIs)
    - SPI violation means safety argument has a defect (surprise!)
  - “Surprise” arrival rates could help estimate safety case uncertainty
    - Start during validation; continue after deployment

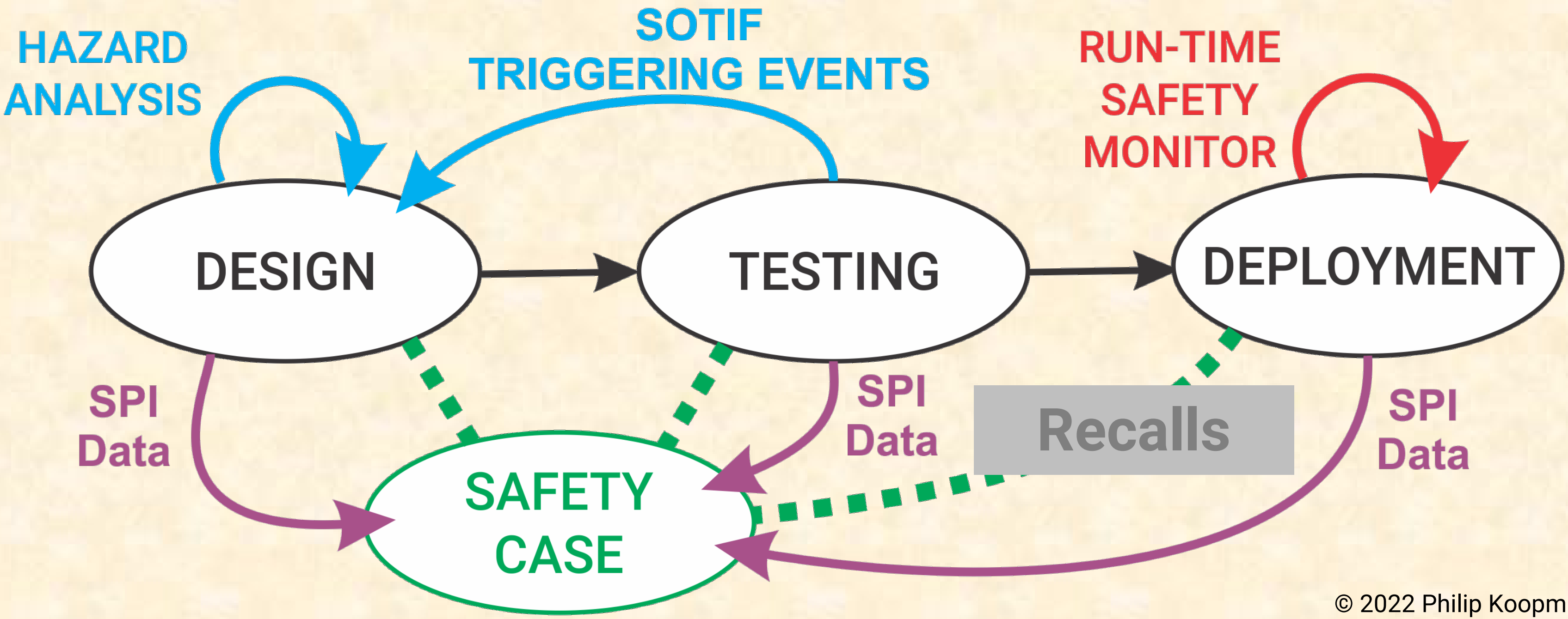
ANSI/UL 4600 SPIs and Lifecycle Feedback

- SPI: direct measurement of safety case claim failure
  - Independent of reasoning (“claim is X … yet here is ~X”)
- A falsified safety case claim:
  - Safety case has some defect
  - Not (necessarily) imminent loss event
- Root cause analysis might reveal:
  - Product or process defect
  - Invalid safety argument
  - Issue with supporting evidence
  - Assumption error



Field Engineering Feedback

- Architectural support for lifecycle field feedback
  - Safety Performance Indicators (SPI) data linked to safety case
    - Transition from recall model to continuous improvement



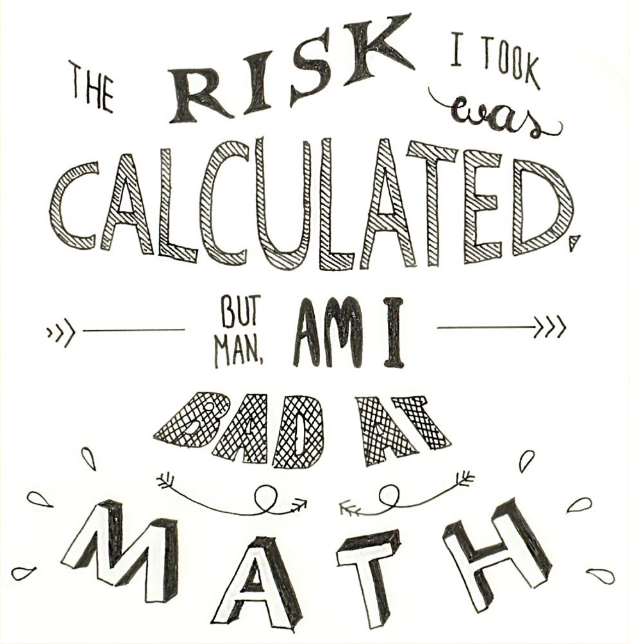
Ethics: Risk vs. Safety

- Cost of excessive risk drives improvement
  - Reducing risk tends to improve safety, but…

- Affordable risk might exceed acceptable safety
  - Life insurance for combat military personnel
  - Commercial space launch insurance
  - Cost of fatality settlement compared to $2M-$5M/day burn rate
- Risk management is not enough for acceptable safety
  - Risk transfer (occupants vs. pedestrians)
  - Existential pressure for company to deploy with unproven safety

Ethics: Deployment Governance

- #1 ethical issue in AVs is deployment governance
  - Who decides when to deploy based on what?
- Pressure for aggressive deployments
  - Missing independent technical oversight
- Ethical deployment should address:
  - Publicly disclosed safety prediction
  - Inclusion of stakeholder concerns
  - Transparency of data & processes
  - Accountability for any losses
  - Non-discrimination in operational concept



https://bit.ly/3rJeaJ4

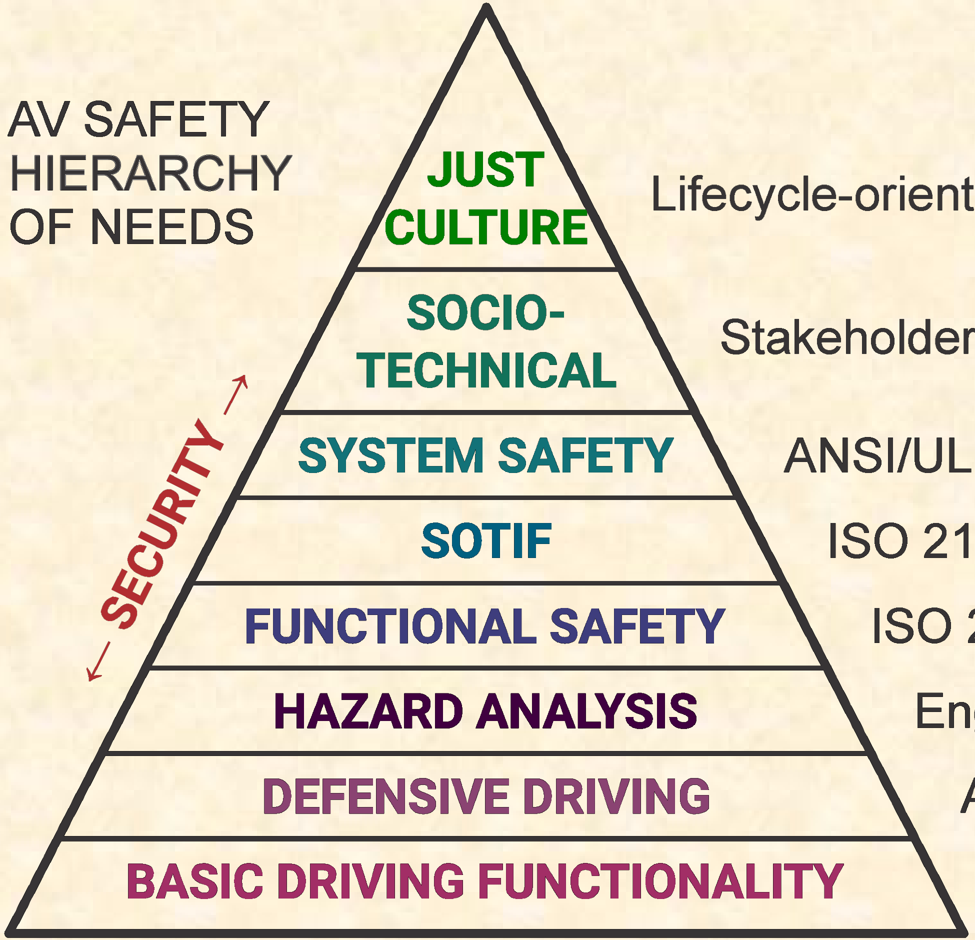
What People Mean By "Safe"

- Human drivers are bad, so computers will be safe
  - Industry rhetorical talking points are ubiquitous
- "Safety is our #1 priority"
- Safe driving behavior
  - Follows traffic laws; good roadmanship
- Tested/simulated for millions of miles
- Risk is managed via insurance
- Conforms to safety standards
- Positive Risk Balance
- Safety cases supported by evidence

[Dall-e]

Hierarchy of Concurrent Safety Needs

AV SAFETY
HIERARCHY
OF NEEDS

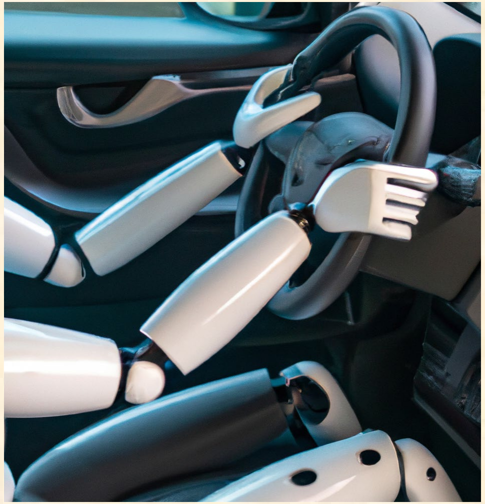


JUST CULTURE	Lifecycle-oriented safety culture
SOCIO-TECHNICAL	Stakeholder expectations
SYSTEM SAFETY	ANSI/UL 4600 – safety case
SOTIF	ISO 21448 – insufficiencies
FUNCTIONAL SAFETY	ISO 26262 – internal faults
HAZARD ANALYSIS	Engineering risk mitigation
DEFENSIVE DRIVING	AV avoids driving risk
BASIC DRIVING FUNCTIONALITY	Can the AV drive?

Summary: Safe Enough AV Deployment

Carnegie Mellon University

- Don’t forget safety while public road testing – SAE J3018
- Acceptable safety is more than just a risk number
  - Good human PRB + safety factor for unknowns
  - Safety & security industry engineering standards
  - Ethical & stakeholder concerns addressed
- Safety case
  - Transparent argument based on evidence
  - Lifecycle uncertainty management via feedback
- Deployment Governance – #1 ethical issue
  - Stakeholders involved in safety criteria & decision
  - Safety culture assures fair dealing on decision



[Dall-e]