Prof. Philip Koopman

Autonomous Vehicles and Software Safety Engineering

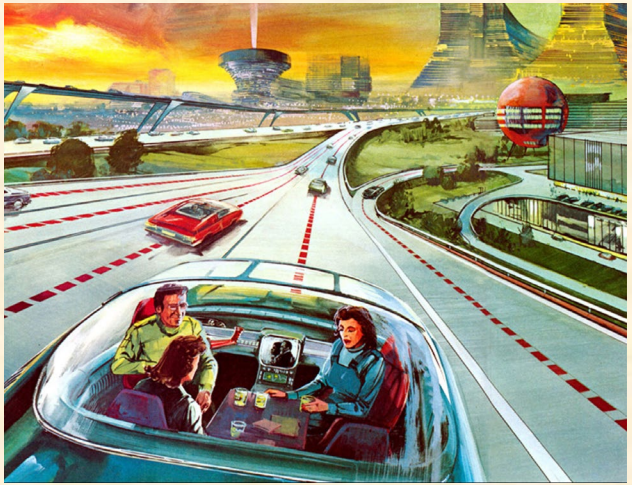
ICSE Keynote, May 2022

Carnegie Mellon University

Overview

Carnegie Mellon University

- Autonomous Vehicles almost “solved”
  - But … “almost” is misleading
- Huge challenge: safety
  - AVs present additional challenges
  - Perception edge cases are a limiting factor
  - Testing alone won’t get us to safety
- Safety requires a standards + safety case approach
  - Life cycle argument supporting deployment safety
  - ANSI/UL 4600 standard for #DidYouThinkofThat ?



[General Motors]

AV Problem 98% Solved For 25+ Years

Carnegie Mellon University

- D.C. to San Diego
  - CMU Navlab 5
  - Dean Pomerleau & Todd Jochem
    https://www.cs.cmu.edu/~tjochem/nhaa/nhaa_home_page.html
  - AHS San Diego demo Aug 1997
- Remaining challenges:
  - That last 2% … and the safety driver

CMU NREC: 35+ Years Of Cool Robots

Carnegie Mellon University

Machinery Safety

Software Safety

DARPA SC-ALV

NASA Lunar Rover

Auto Excavator

Auto Forklift

DARPA PerceptOR

DARPA LAGR

DARPA UPI

Army FCS

Laser Paint Removal

NATIONAL ROBOTICS NREC ENGINEERING CENTER

Carnegie Mellon University

1985 1990 1995 2000 2005 2010

ARPA Demo II

NASA Dante II

Auto Harvesting

DARPA Grand Challenge

Mars Rovers

Urban Challenge

Auto Haulage

Auto Spraying

Software Safety Engineering

- Safety is a system property
  - Correctness is not enough for safety
- Safety engineering emphasis on hazard mitigation
  - Identify hazards: if X goes wrong, could result in loss event
    - Includes hardware failures, tool defects, environmental surprises
  - Predict risk = probability * consequence
    - The tricky part is: “Probably Never * Catastrophic”
  - Mitigate risk via:
    - Engineering rigor: process quality, analysis, test, redundancy patterns
    - Functional safety: detect and shut down malfunctioning equipment
    - Safety of Intended Function (SOTIF): resilience to requirements gaps, inconsistent sensor data, unexpected environments

Why Is AV Safety Complicated?

Carnegie Mellon University

- Public expectations
  - Expect super-human machine performance
  - Trust too easily given, backlash when broken
- Technical challenges
  - Machine Learning safety is work in progress
  - Statistical approach vs. high severity rare events
- Historical industry culture clash
  - Autonomy researchers: it's all about the cool small-scale demo
  - Silicon Valley: move fast + break things
  - Automotive: blame driver for not mitigating equipment failures
  - Regulators: test-centric; weak digital safety expertise

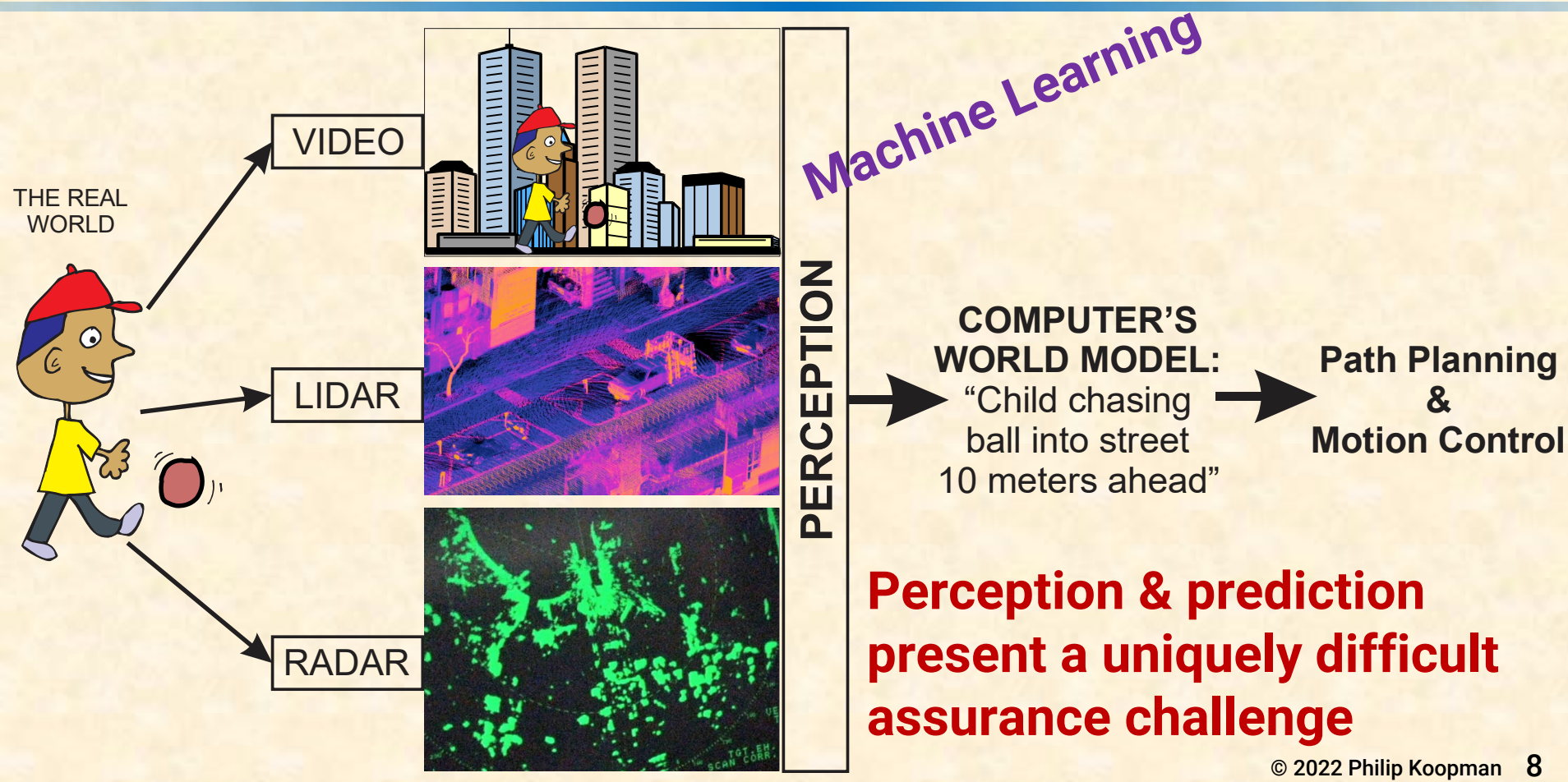
Should You Trust an AV?

Carnegie Mellon University

- Heaviest technical lift is perception/prediction safety

Perception Builds the World Model

Carnegie Mellon University



Perception & prediction present a uniquely difficult assurance challenge

Edge Cases As A Limiting Factor

Carnegie Mellon University

- Machine learning is best at what it has already seen
  - But the world is full of novelty
  - Perception/prediction poor at recognizing it is just guessing
- Is this a Person or Chicken?
- Edge Case are surprises
  - You won't see these in testing

PREDICTED CONCEPT	PROBABILITY
bird	0.997
no person	0.990
one	0.975
feather	0.970
nature	0.963
poultry	0.954
outdoors	0.936
color	0.910
animal	0.908

https://www.clarifai.com/demo

➔ Edge cases are the stuff you didn't think of!

The Challenge Is Covering Everything

Carnegie Mellon University

- Have you covered the possible unknowns?

https://goo.gl/J3SSyu

Brute Force AV Validation: Public Road Testing

Carnegie Mellon University

- Good for identifying “easy” cases
  - Expensive and potentially dangerous

Autonomy Testing Risks

Carnegie Mellon University

- Uber ATG fatality, Tempe AZ/US: March 2018
  - Uber ATG closed: January 2021
- Local Motors injury, Whitby CA: Dec. 2021
  - Company closed: Jan. 2022
- Pony.AI crash: CA/US: Oct. 2021
  - Uncrewed test permit revoked
- WeRide sleeping test driver: Oct. 2021
  - Company deflects issue / no apparent regulator action
- Easymile shuttle phantom braking injuries: (2019, 2020)
- SAE J3018 standard for testing safety (2015; 2020 update)
  - Only Argo.AI publicly pledges conformance

https://bit.ly/3AupcWb

Brute Force Road Testing

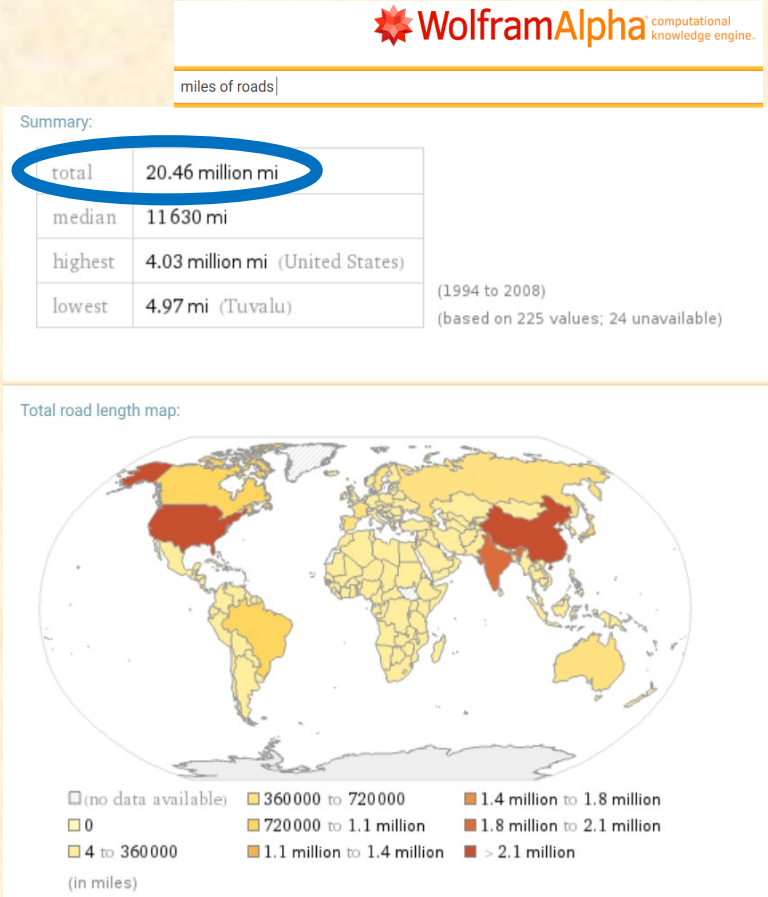
Carnegie Mellon University

- If 100M miles/critical mishap…
  - Test 3x–10x longer than mishap rate
    ➔ Need 1 Billion miles of testing
- That's ~25 round trips on every road in the world
  - With fewer than 10 critical mishaps

…

  - Start over for each software update

➔ Brute force testing impracticable



Closed Course Testing

Carnegie Mellon University

- Safer, but expensive
  - Not scalable
  - Only tests things you have thought of!

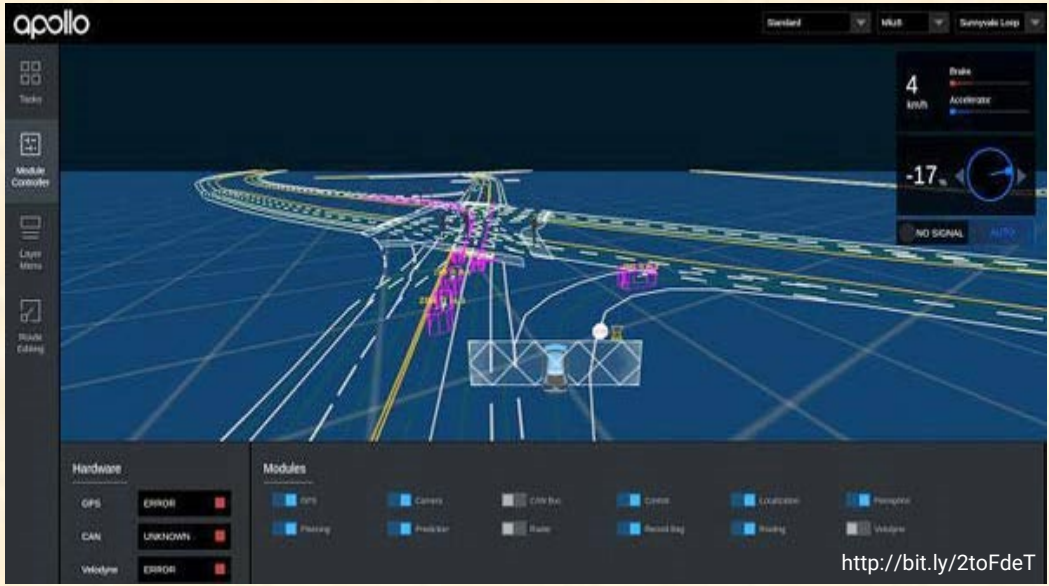
Volvo / Motor Trend

Simulation

Carnegie Mellon University

- Highly scalable; less expensive than road testing
  - Simulation validation ("tool qualification")
  - Only tests things you have thought of!

Udacity



Apollo

How Much Do You Trust Simulation?

Carnegie Mellon University

- Would you put <u>your</u> child in front of this self driving car:
  - 10,000M simulation miles
    … perhaps with a simulator error?
  - 100M miles data collected
    … perhaps missing some relevant scenarios?
  - 10M of road testing
    … that missed high risk situations?
  - Designed with research-quality tooling
    … with no safety qualification?
  - With 5% labeling errors in training data?
- Need simulation and other tool qualification

Industry Safety Standards Can Help

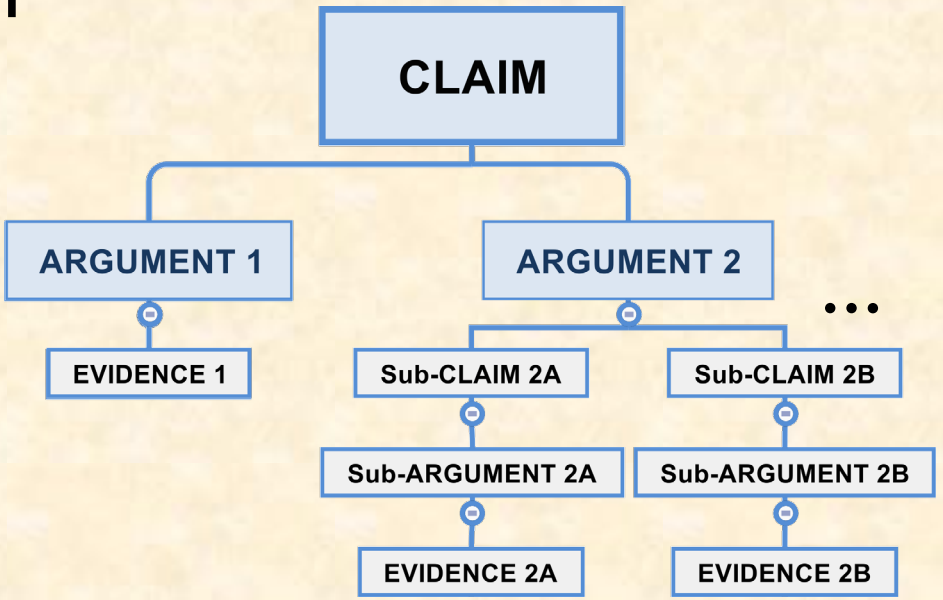
Carnegie Mellon University

- ISO 26262 – Functional Safety
  - Covers run-time faults & design defects
  - Assumes complete requirements known
- ISO 21448 – SOTIF
  - SOTIF: “Safety Of The Intended Function”
  - Iteratively mitigate discovered “unknowns”
- Also need: #DidYouThinkofThat? lists
  - A technically substantive safety argument
  - Evidence of coverage initially + feedback from surprises
  - Continuously improve based on lessons learned
  - A way to organize everything to ensure safety

Safety Cases To Organize Safety Argument

Carnegie Mellon University

- Claim – a property of the system
  - "System avoids pedestrians"
- Argument – why this is true
  - "Detect & maneuver to avoid"
- Evidence – supports argument
  - Tests, analysis, simulations, …
- Sub-claims/arguments address complexity
  - "Detects pedestrians" // evidence
  - "Maneuvers around detected pedestrians" // evidence
  - "Stops if can't maneuver" // evidence



Lifecycle, Maintenance & Supply Chain

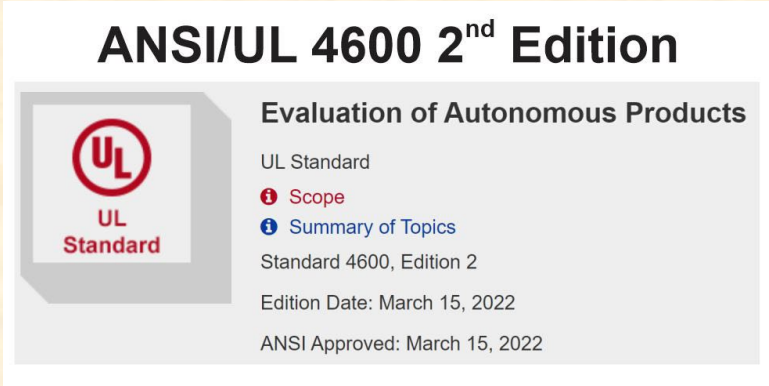
Carnegie Mellon University

- Safety related maintenance
  - What maintenance is required for safety?
  - How do you know it is done effectively?
- Safety related aspects of lifecycle
  - Requirements/design/ML training
  - Handoff to manufacturing; deployment
  - Supply chain
  - Field modifications & updates
  - Operation, retirement & disposal
- Safety case kept updated during system lifecycle

UL 4600 – An Autonomy Safety Standard

Carnegie Mellon University

- Evaluation of a Safety Case
  - Independently assess safety case
  - Mix & match supporting standards
  - Discourages questionable practices
  - Extensive #DidYouThinkofThat? lists
- "Unknowns" are first class citizens
  - Balance between analysis & field experience
  - Field monitoring used for continual safety case improvement
  - Assessment findings & field data used to update practices
- ANSI/UL 4600 2nd Edition issued March 2022
  - 3rd edition to address heavy trucks in progress



The Path To Achieving AV Safety

Carnegie Mellon University

- Cultural reconciliation within industry
  - Safety for on-road testing (driver & vehicle)
  - Mature beyond a rushed demo mentality
- Stakeholder trust for acceptable safety
  - System-level safety for machine learning
  - Independent safety assessments
- Use industry safety standards
  - Reform “standards optional” regulations
  - Traditional software safety … PLUS …
    - Account for unknown unknowns at deployment
  - UL 4600 Autonomous Vehicle Safety Standard

http://bit.ly/2MTbT8F (sign modified)

BoF Discussion Starters

Carnegie Mellon University

Autonomous Vehicles and Software Safety Engineering

- Should software developers share blame for a fatality?
  - Ethics of when to deploy “beta” software on public roads
- Machine learning – how do we:
  - Ensure training data coverage of operational domain
  - Account for high risk heavy tail events (see SEAMS talk)
- Commercial/research software for life critical systems
  - Simulator software & simulation object models
  - Machine Learning development toolchains
  - DevOps, cloud infrastructure, and SaaS toolchains
- Gaps between ICSE research results and AV system level safety

Trolley Problem is irrelevant for practical AVs https://youtu.be/30YiMc1k2Xw