Prof. Philip Koopman

Carnegie Mellon University

@PhilKoopman

EDGE CASE RESEARCH

Autonomous Vehicle Safety and Perception Robustness Testing

Australian System Safety Conference
May 23, 2019

Overview

- **Making safe robots**
  - Doer/Checker safety
- **Edge cases matter**
  - Robust perception matters
- **The heavy tail distribution**
  - Fixing stuff you see in testing isn't enough
- **Perception stress testing**
  - Finding the weaknesses in perception
- **UL 4600: autonomy safety standard**

[General Motors]

98% Solved For 20+ Years

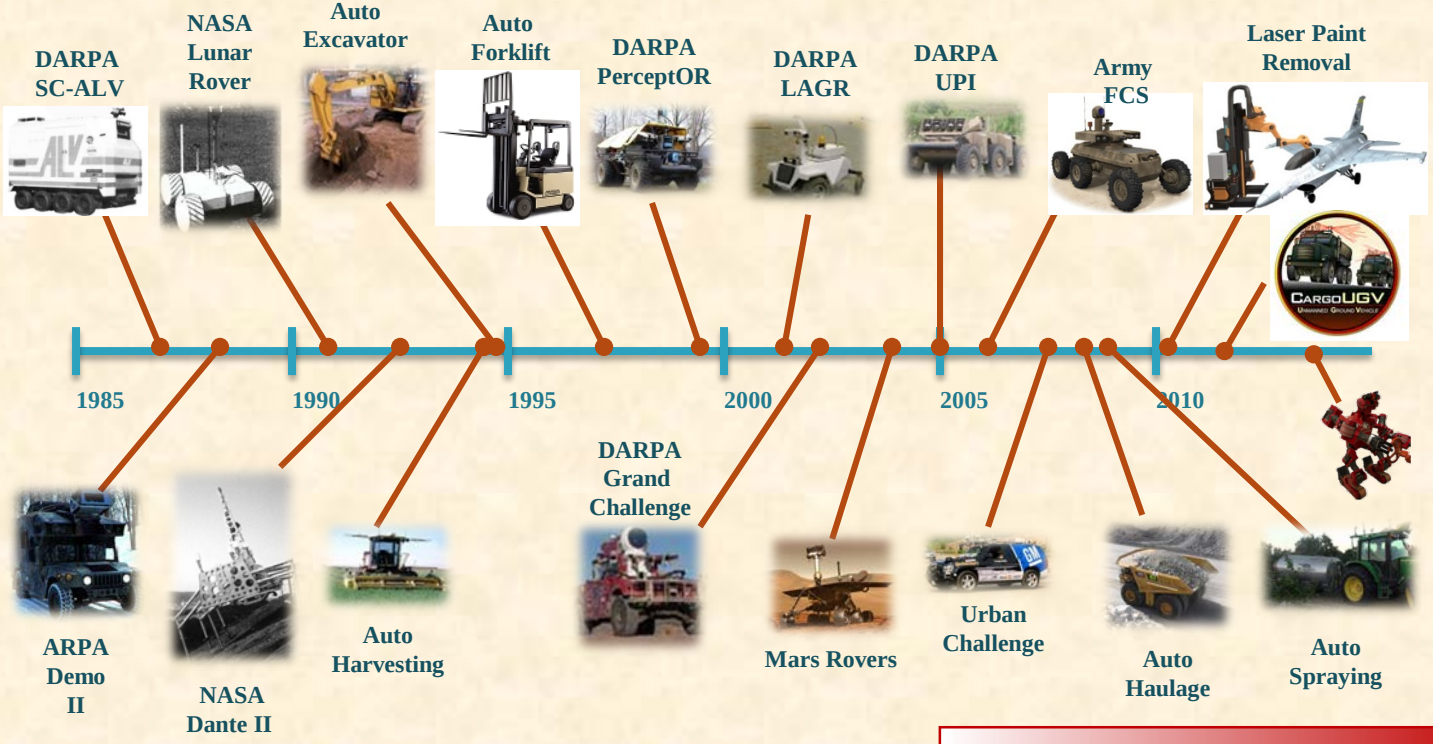
Carnegie Mellon University

- **Washington DC 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**

NREC: 30+ Years Of Cool Robots

NATIONAL ROBOTICS NREC ENGINEERING CENTER



Software Safety

Carnegie Mellon University Faculty, staff, students
Off-campus Robotics Institute facility

Before Autonomy Software Safety

Carnegie Mellon University

- The Big Red Button era

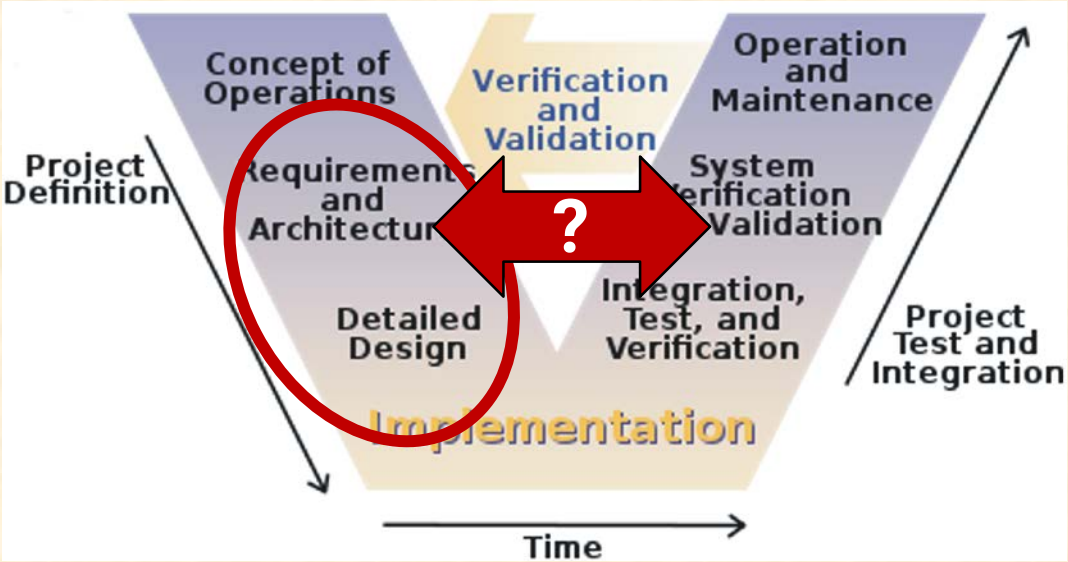
Traditional Validation Vs. Machine Learning

Carnegie Mellon University

- **Use traditional software safety where you can**

..BUT..

- **Machine Learning (inductive training)**
  - **No requirements**
    - Training data is difficult to validate
  - **No design insight**
    - Generally inscrutable; prone to gaming and brittleness



APD (Autonomous Platform Demonstrator)

Carnegie Mellon University

Safety critical speed limit enforcement

NATIONAL ROBOTICS NREC ENGINEERING CENTER

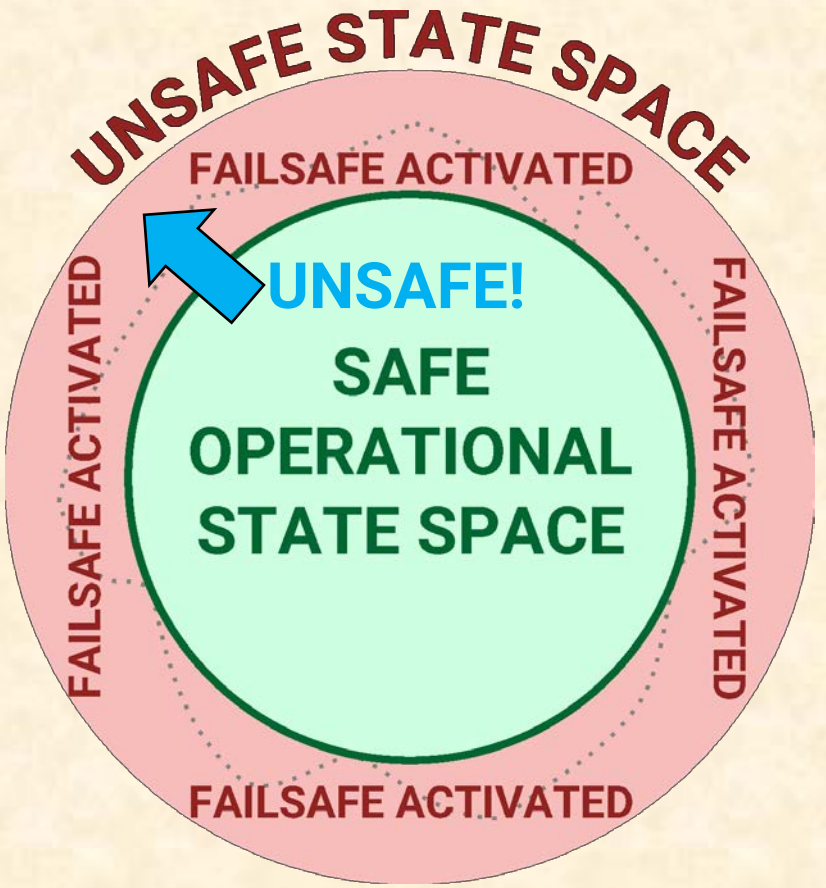
TARGET GVW: 8,500 kg
TARGET SPEED: 80 km/hr

Approved for Public Release. TACOM Case #20247 Date: 07 OCT 2009

Safety Envelope Approach to ML Deployment

Carnegie Mellon University

- **Specify unsafe regions**
- **Specify safe regions**
  - Under-approximate to simplify
- **Trigger system safety response upon transition to unsafe region**

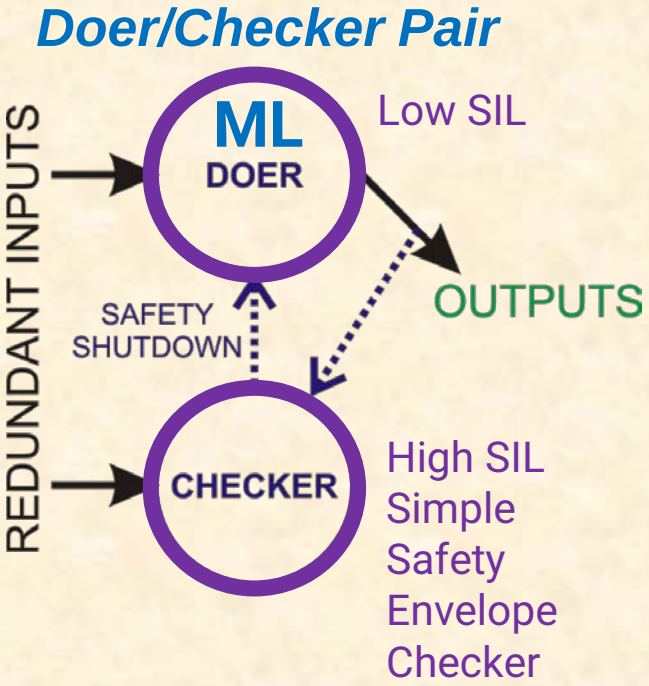


Architecting A Safety Envelope System

Carnegie Mellon University

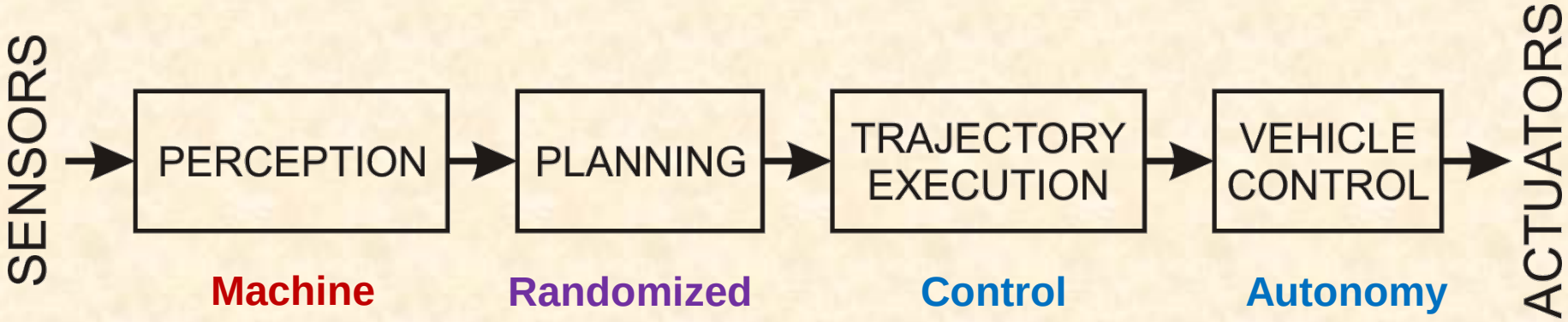
- **"Doer" subsystem**
  - Implements normal, untrusted functionality
- **"Checker" subsystem – Traditional SW**
  - Implements failsafes (safety functions)
- **Checker entirely responsible for safety**
  - Doer can be at low Safety Integrity Level
  - Checker must be at higher SIL

(Also known as a "safety bag" approach;
also monitor/actuator pair)



Validating an Autonomous Vehicle Pipeline

EDGE CASE RESEARCH



Machine Learning Based Approaches

➔ ???

Randomized & Heuristic Algorithms

➔ Run-Time Safety Envelopes

➔ Doer/Checker Architecture

Control Systems

➔ Control Software Validation

➔ Doer/Checker Architecture

Autonomy Interface To Vehicle

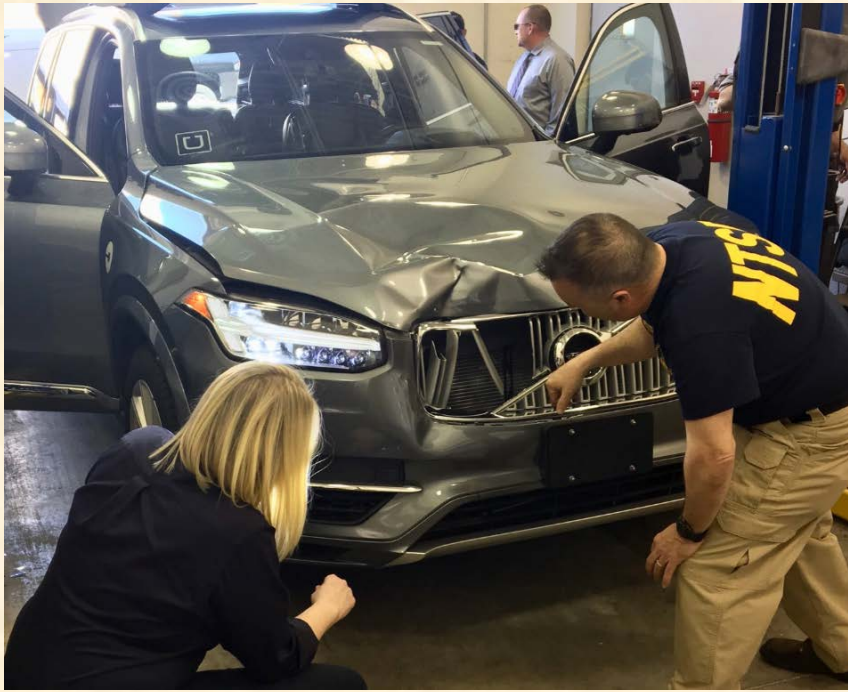
➔ Traditional Software Validation

Perception presents a uniquely difficult assurance challenge

Brute Force AV Validation: Public Road Testing

EDGE CASE RESEARCH

- **Good for identifying "easy" cases**
  - Expensive and potentially ***dangerous***



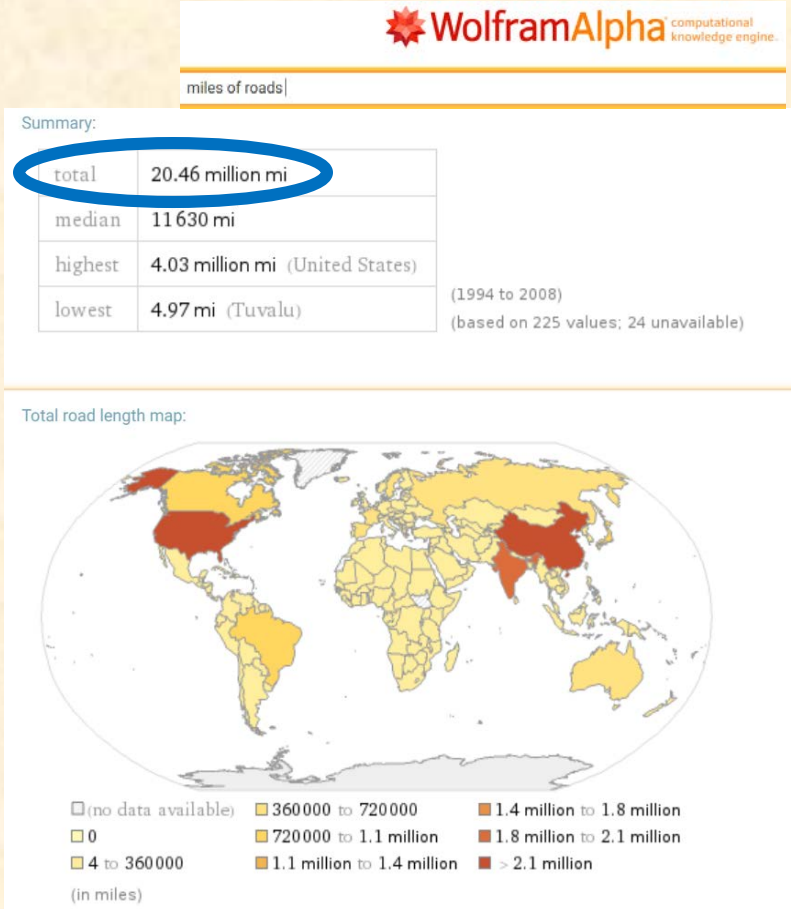
Validation Via Brute Force Road Testing?

EDGE CASE RESEARCH

- **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

…



Closed Course Testing

EDGE CASE RESEARCH

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

Volvo / Motor Trend

Simulation

EDGE CASE RESEARCH

- **Highly scalable; less expensive**
  - Scalable; need to manage fidelity vs. cost
  - Only tests things you have thought of!

Udacity

ANSYS

What About Edge Cases?

EDGE CASE RESEARCH

- **Gaps in training data can lead to perception failure**
  - Safety needs to know: "Is that a person?"
  - Machine learning provides: "Is that thing like the people in my training data?"

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 Case are surprises**
  - You won't see these in training or testing

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

Need An Edge Case "Zoo"

EDGE CASE RESEARCH

- Novel objects (missing from zoo) are triggering events

https://goo.gl/J3SSyu

Why Edge Cases Matter

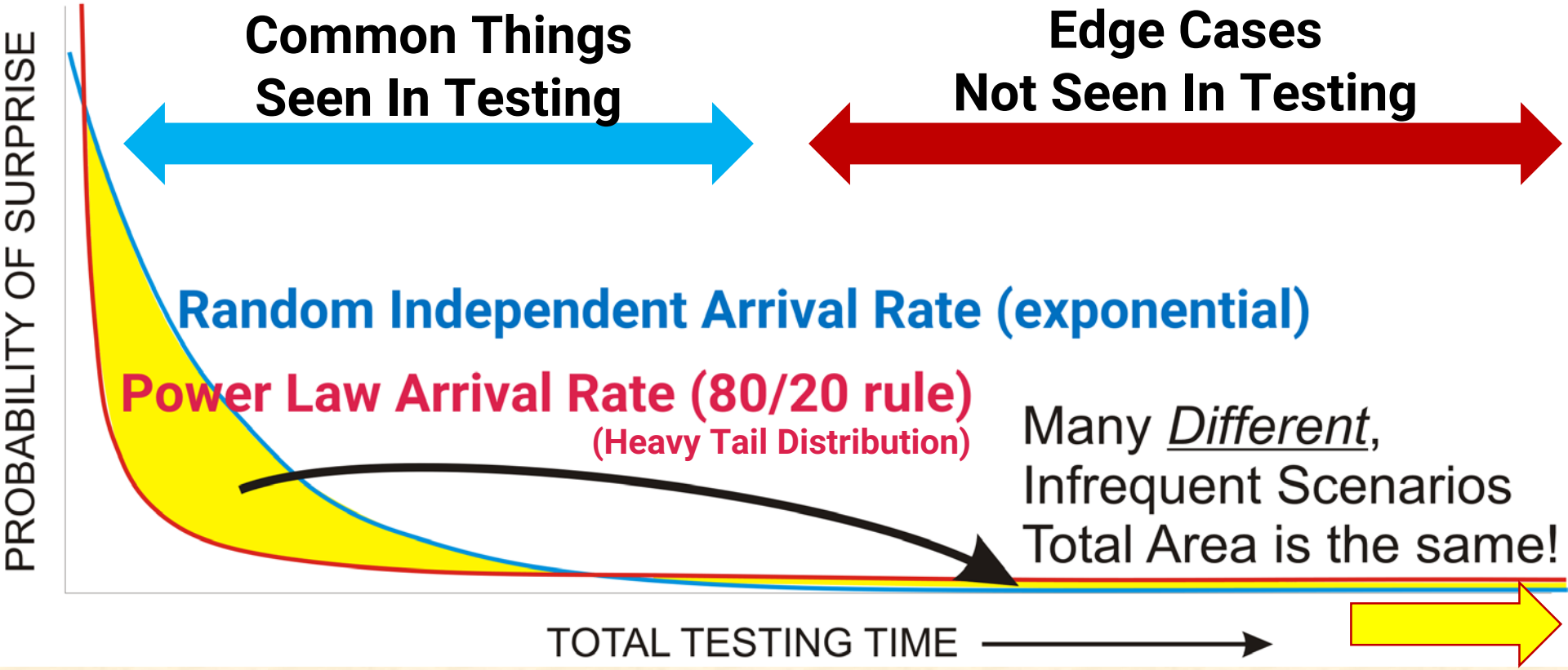
EDGE CASE RESEARCH

- **Where will you be after 1 Billion miles of validation testing?**
- **Assume 1 Million miles between unsafe "surprises"**
  - Example #1:
    100 "surprises" @ 100M miles / surprise
    - All surprises seen about 10 times during testing
    - With luck, all bugs are fixed
  - Example #2:
    100,000 "surprises" @ 100B miles / surprise
    - Only 1% of surprises seen during 1B mile testing
    - Bug fixes give no real improvement (1.01M miles / surprise)

https://goo.gl/3dzguf

Real World: Heavy Tail Distribution(?)

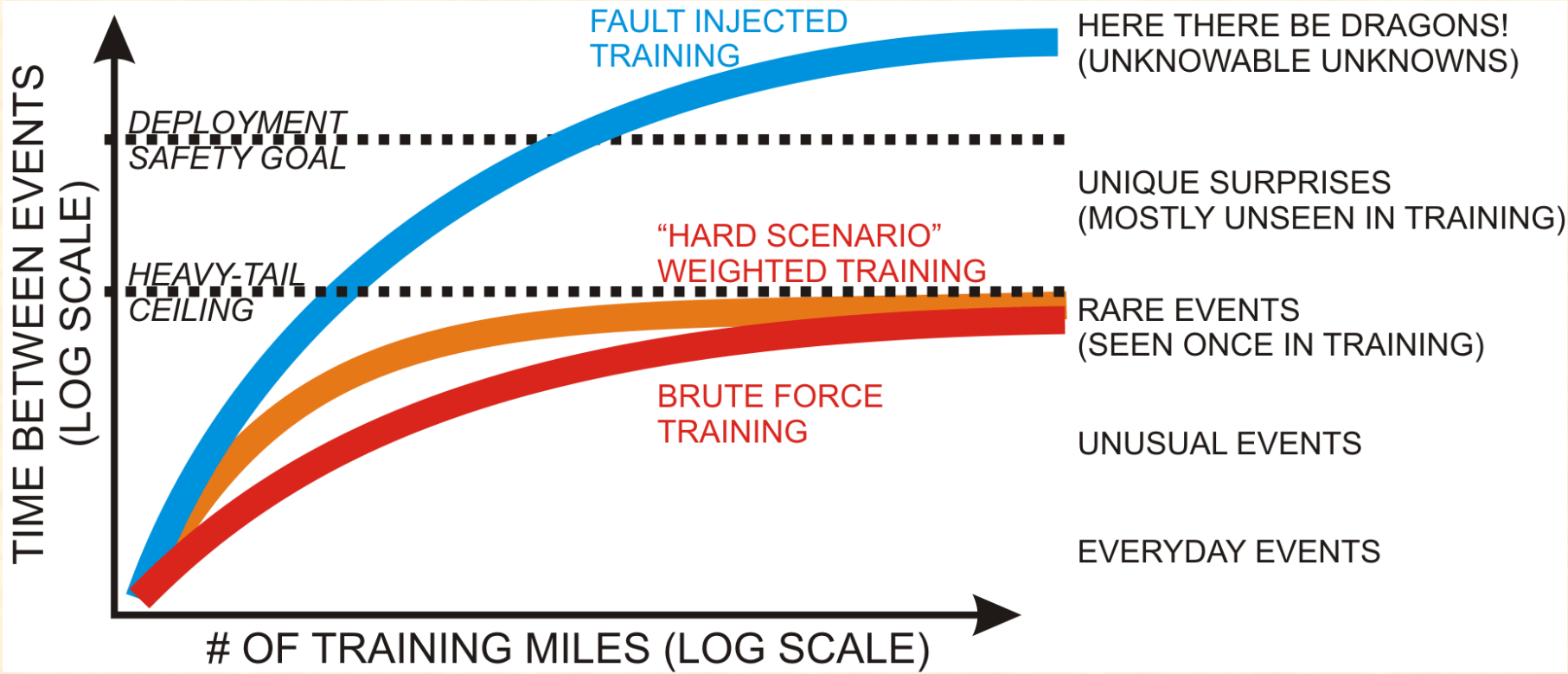
EDGE CASE RESEARCH



The Heavy Tail Testing Ceiling

EDGE CASE RESEARCH

- **Need to find "Triggering Events" to inject into sims/testing**

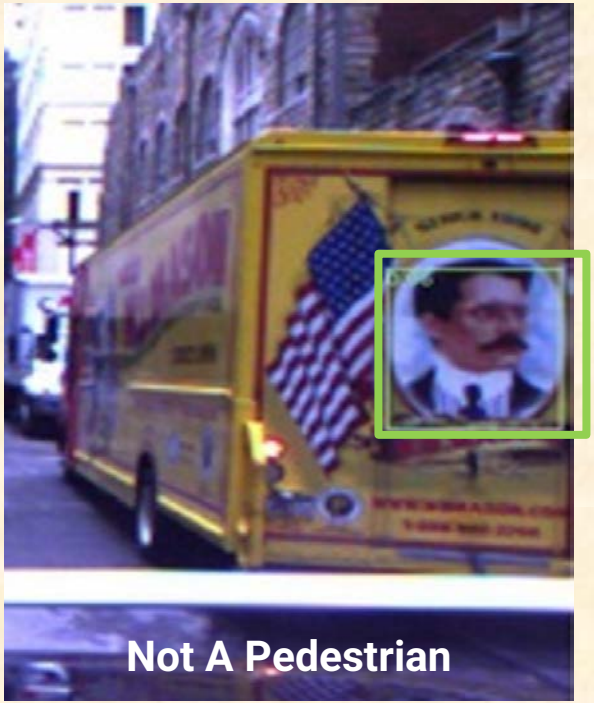


Edge Cases Pt. 1: Triggering Event Zoo

EDGE CASE RESEARCH

- **Need to collect surprises**
  - Novel objects
  - Novel operational conditions

- **Issue:**
  novel for person ≠ novel for Machine Learning
  - ML can have “edges” in unexpected places
  - ML might train on features that seem irrelevant to people



Edge Cases Part 2: Brittleness

Carnegie Mellon University

- **Sensor data corruption experiments**

Defocus & haze are a significant issue

Synthetic Equipment Faults

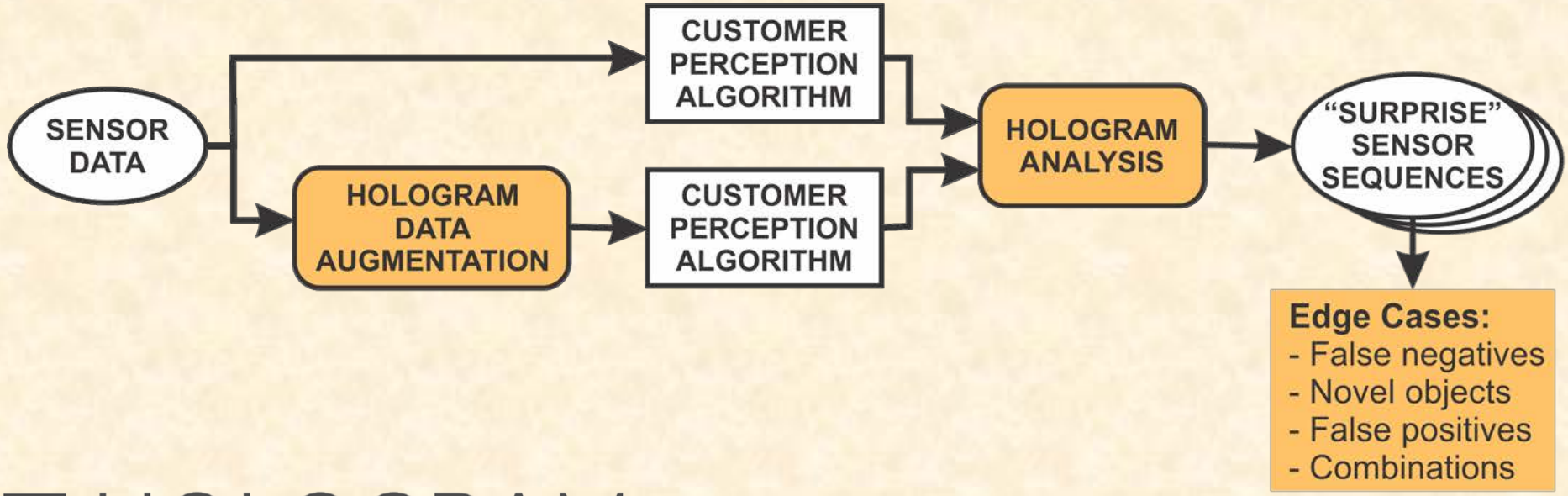
Gaussian Blur & Gaussian Noise cause similar failures

Exploring the response of a DNN to environmental perturbations from “Robustness Testing for Perception Systems,” RIOT Project, NREC, DIST-A.

Hologram Detects Edge Cases

EDGE CASE RESEARCH

- **Brittle perception behavior indicates Edge Cases**
  - Can uncover false negatives and detect novel objects

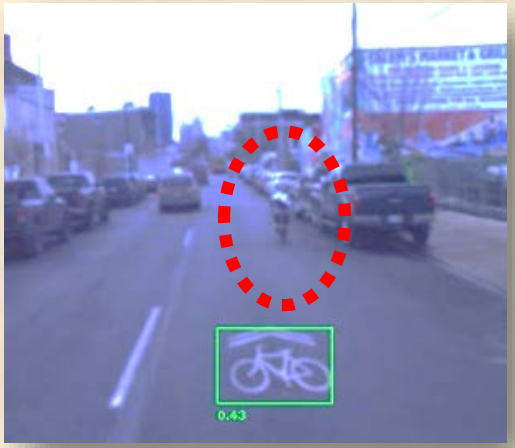


HOLOGRAM

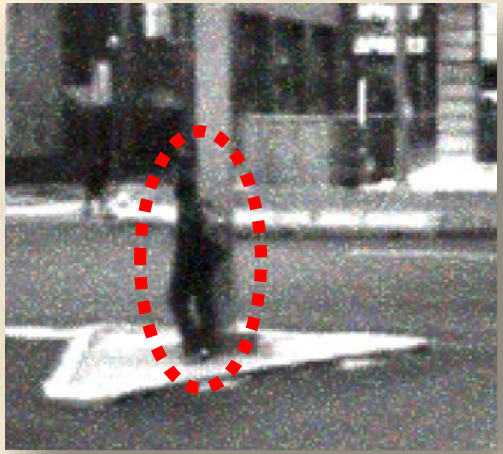
Context-Dependent Perception Failures

EDGE CASE RESEARCH

- **Perception failures are often context-dependent**
  - False positives and false negatives are both a problem



False positive on lane marking
False negative real bicyclist



False negative when
person next to light pole

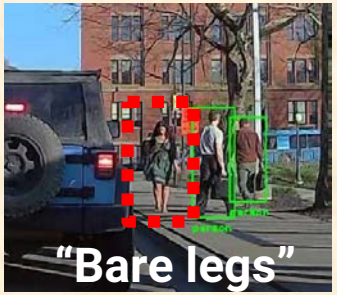
False negative when
in front of dark vehicle

Will this pass a "vision test" for bicyclists?

Example Triggering Events via Hologram

EDGE CASE RESEARCH

- **Mask-R CNN: examples of systemic problems we found**



Notes: These are baseline, un-augmented images // Your mileage may vary.

HOLOGRAM

HOLOGRAM


PITTSBURGH
PARAMEDIC RESCUE
AMBULANCE
27
For Lease
MiVue 420
2018/10/30 13:18:43
N 40°26'19.7" W 80°00'8.62"


DO NOT
BLOCK
INTERSECTION
WARD
MiVue 420
2018/11/05 12:50:47
N 40°26'43.5" W 79°56'55.9"
HOLOGRAM

Operations & Human Interactions

EDGE CASE RESEARCH

- **Drivers do more than just drive**
  - Occupant behavior, passenger safety
  - Detecting and managing equipment faults
- **Operational limitations & situations**
  - System exits Operational Design Domain
  - Vehicle fire or catastrophic failure
  - Post-crash passenger evacuation
- **Interacting with non-drivers**
  - Pedestrians, passengers
  - Police, emergency responders

Lifecycle Issues

EDGE CASE RESEARCH

- **Handling updates**
  - Fully recertify after every weekly update?
  - Security in general
- **Vehicle maintenance**
  - Pre-flight checks, cleaning
  - Corrective maintenance
- **Supply chain issues**
  - Quality fade
  - Supply chain faults

https://bit.ly/2IKlZJ9

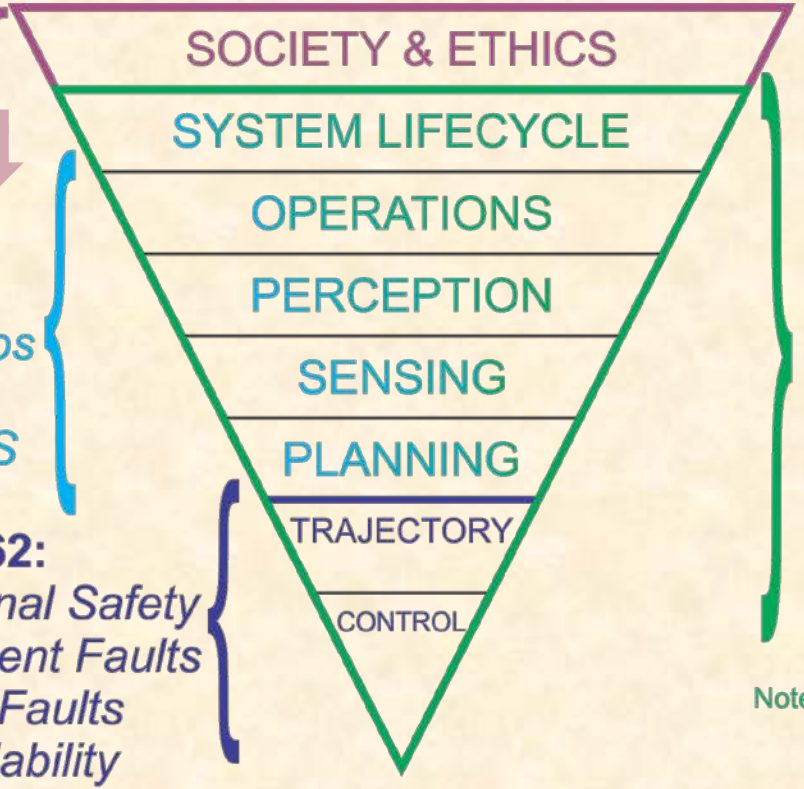

https://bit.ly/2VavsjM

Is windshield cleaning fluid life critical?

Safety Standard Landscape

EDGE CASE RESEARCH

IEEE P700x
- *Ethical Life-Cycle*



ISO 21448 (SOTIF):
- *Enumerate Scenarios*
- *Triggering Events*
- *Historically for ADAS*

ISO 26262:
- *Functional Safety*
- *Equipment Faults*
- *Design Faults*
- *Controllability*

Note: ISO 26262 covers functional safety lifecycle activities.

UL 4600: (Draft Scope)
- *Machine Learning Faults*
- *External Operational Faults*
- *Non-Driver Human Faults*
- *Nondeterminism*
- *High Residual Unknowns*
- *Field Feedback and Updates*
- *Lack of Human Oversight*
- *System-Level Safety Metrics*
- *Full Autonomous Driving: Excludes Human Supervision*

Note: Provides significantly more detailed treatment regarding autonomous systems for those topics also mentioned in ISO 21448, ISO 26262. Intended to supplement, not replace these.

Ways To Improve AV Safety

EDGE CASE RESEARCH

- **More safety transparency**
  - Independent safety assessments
  - Industry collaboration on safety
- **Minimum performance standards**
  - "Driver test" is necessary -- but not sufficient
    - How do you measure maturity?
- **Autonomy software safety standards**
  - ISO 26262/21448 + UL 4600 + IEEE P700x
  - Dealing with uncertainty and brittleness

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

EDGE CASE RESEARCH
MAKING AUTONOMY SAFER
info@ecr.guru