Overview

- Control & Planning safety
- Perception safety
- Edge cases
- Some hard questions about AV safety
NREC: 30+ Years Of Cool Robots

1985
ARPA Demo II
NASA Dante II
Auto Harvesting

1990
NASA Lunar Rover
Auto Excavator
Auto Forklift

1995
DARPA Grand Challenge

2000
DARPA PerceptOR
DARPA LAGR

2005
DARPA UPI

2010
Army FCS

Laser Paint Removal
Auto Haulage
Auto Spraying

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

Software Safety

© 2018 Philip Koopman
Researchers evaluated 150 bugs from 11 distinct projects over 4 years [ICSE 2018]

From “RIOT Expanded Technical Brief, NAVAIR Public Release- 2016-842 'Approved for Public Release; distribution is unlimited'.
The Big Red Button era
APD (Autonomous Platform Demonstrator)
How did we make this scenario safe?

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

Approved for Public Release. TACOM Case #20247 Date: 07 OCT 2009
Traditional Validation Meets Machine Learning

- 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 over-fitting/gaming
Safety Envelope Approach to ML Deployment

- Specify unsafe regions
- Specify safe regions
  - Under-approximate to simplify
- Trigger system safety response upon transition to unsafe region
Architecting A Safety Envelope System

- **“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)
Validating an Autonomous Vehicle Pipeline

Perception presents a uniquely difficult assurance challenge
Brute Force Road Testing

- 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...
Default AV Validation: Public Road Testing

- Good for identifying “easy” cases
  - Expensive and potentially *dangerous*
  - Should concentrate on *data collection*, not debugging

http://bit.ly/2lsj6Qu
Did We Learn The Right Lesson from Tempe?

■ NOT: Blame the victim
  ● Pedestrian in road is **expected**

■ NOT: Blame the technology
  ● Immature technology under test
    - **Failures are expected!**

■ NOT: Blame the driver
  ● A solo driver drop-out is **expected**

■ The real AV testing lesson:
  ➔ **Ensure safety driver is engaged** ←
  ● Safety argument: Driver alert; time to respond; disengagement works
Can Safety Driver React In Time?

Safety Driver Tasks:
- Mental model of “normal” AV
- Detect abnormal AV behavior
- React & recover if needed

Example: obstructed lane
- Does driver know when to take over?
- Can driver brake in time?
  - Or is sudden lane change necessary?

Example: two-way traffic
- What if AV commands sudden left turn into traffic?
Closed Course Testing

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

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


[Image: Apollo](http://bit.ly/2toFdeT)
What About Edge Cases?

- You should expect the extreme, weird, unusual
  - Unusual road obstacles
  - Extreme weather
  - Strange behaviors

Edge Case are surprises
- You won’t see these in testing
  ➔ Edge cases are the stuff you didn’t think of!

© 2018 Philip Koopman
Just A Few Edge Cases

- Unusual road obstacles & obstacles
- Extreme weather
- Strange behaviors

https://dailym.ai/2K7kNS8
https://en.wikipedia.org/wiki/Magic_Roundabout_(Swindon)
https://goo.gl/J3SSya

© 2018 Philip Koopman
Why Edge Cases Matter

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)
The Real World: Heavy Tail Distribution

Common Things Seen In Testing

Random Independent Arrival Rate (exponential)

Power Law Arrival Rate (80/20 rule) (Heavy Tail Distribution)

Edge Cases Not Seen In Testing

Many Different, Infrequent Scenarios
Total Area is the same!
The Heavy Tail Testing Ceiling

- FAULT INJECTED TRAINING: HERE THERE BE DRAGONS! (UNKNOWABLE UNKNOWNS)
- DEPLOYMENT SAFETY GOAL
- UNIQUE SURPRISES (MOSTLY UNSEEN IN TRAINING)
- HEAVY-TAIL WEIGHTED TRAINING
- RARE EVENTS (SEEN ONCE IN TRAINING)
- "HARD SCENARIO" TRAINING
- UNUSUAL EVENTS
- BRUTE FORCE TRAINING
- EVERYDAY EVENTS
Malicious Image Attacks Reveal Brittleness

QuocNet:

Car Not a Car Magnified Difference

AlexNet:

Bus Magnified Difference Not a Bus


https://goo.gl/5sKnZV
ML Is Brittle To Environment Changes

Sensor data corruption experiments

Synthetic Equipment Faults

Gaussian blur

Defocus & haze are similarly a significant issue

Exploring the response of a DNN to environmental perturbations from “Robustness Testing for Perception Systems,” RIOT Project, NREC, DIST-A.
What We’re Learning With Hologram

- A scalable way to test & train on Edge Cases

Your fleet and your data lake

Hologram cluster tests your CNN

Hologram cluster trains your CNN

Your CNN becomes more robust
Edge Case Testing: Data Fuzzing

- Add small amounts of sensor noise
- Cost vs. speed tradeoff
  - Use approximate models for training
  - Use realistic models for validation
Context-Dependent Perception Failures

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

Will this pass a “vision test” for bicyclists?
~100M miles/fatal mishap for human driven road vehicles

- 28% Alcohol impaired
- 27% Speeding-related
- 28% Not wearing seat belts
- 9% distracted driving (2016 data)
- 2% drowsy
- ...

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

Unimpaired drivers are better than 100M miles

- What if AV fatality demographics change?
- Currently 19% non-occupant fatalities (at-risk road users)

Hard Question: How Safe Is Enough?

https://goo.gl/tEuoaS
Is the “94% human error” meme accurate?

- DOT HS 812 115:
  - “The critical reason was assigned to drivers in an estimated 2,046,000 crashes that comprise 94 percent of the NMVCCS crashes at the national level. However, in none of these cases was the assignment intended to blame the driver for causing the crash.”
- Low hanging fruit is 49% for AVs (https://goo.gl/2WReTj)

RAND report suggests early deployment (RR-1478-RC)

- Even 10% better than a human driver saves lives overall
- BUT – assumes no uncertainty on predicted AV fatality rate
  - Accounting for uncertainty raises the bar – probably a lot
1. Trust the car companies to self-regulate?
   - Toyota criminal sanctions. VW Dieselgate. GM Ignition Switch. Takata air bags. ... see your news feed ...
   - Numerous safety critical software recalls
   - $XX++ billion invested with incentive to deploy ASAP
   - Litigation source code analysis of machine learning?

2. Use existing automotive safety approaches?
   - FMVSS requires standardized vehicle-level road tests
   - “Driver test” must include both skills & “maturity”

3. Wait for a new safety standard?
   - SOTIF (ISO PAS 21448) addresses partial autonomy for now
   - UL 4600 Autonomy Safety standard kick-off in August 2018
Thank You!