Autonomous Vehicle Testing & Safety

Tech.AD, November 16, 2018

Prof. Philip Koopman

Carnegie Mellon University

Edge Case Research
Overview

- **Tempe AZ fatality**
  - Did we really learn the right lesson?

- **Safety case for road testing:**
  - Timely human supervisor response
  - Adequate human supervisor mitigation
  - Appropriate system failure profile

- **Challenges:**
  - Monitoring human effectiveness
  - Valley of Supervisor Dropout

---

We shouldn’t be killing people in our haste to get to a safe future.
A tragic death has occurred
- How can we mitigate risk in the future?

Activities that do NOT improve safety of autonomous vehicle (AV) testing:
- Arguing that delaying deployment costs lives
- Deciding which human was at fault
- Finding out why autonomy failed (surprise!)

The issue is safe AV testing platforms
- AV testing platform = autonomy + safety driver + safety support + test procedures
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 supervisor
  - Solo human drop-out is **expected**

The real AV testing safety lesson:

⇒ Ensure human supervisor is effective ⇐
- If human safety driver is unsafe, you are doing unsafe testing
How Safe Is Safe Enough?

2016 Police-reported crashes
- 3,174,000,000,000 miles
- 34,439 fatal crashes (0.5%) every 92 Million Miles
- 2,177,000 injury crashes (29.9%) every 1.5 Million Miles
- 7,277,000 property damage (69.6%) every 0.6 Million Miles

Non-occupant fatalities: 18% about every 510 Million Miles
- Motorcyclist fatalities: 14% about every 660 Million Miles
- Note: data includes drunk drivers, speeders, no seat belts

Expect zero deaths in a 10 million mile road test campaign
(Probability of any fatality is 10%; pedestrian fatality is 2%)
Can Humans Supervise Autonomy Safely?

Man reportedly caught sleeping behind the wheel of a self-driving Tesla  
[Link](https://goo.gl/ZFCYzD)

Google's Waymo Self-Driving Car Crashed After Driver Dozed Off Back in June

[Image]

A Waymo self-driving car sent a motorcyclist to the hospital — but the human driver was at fault

[Image]
How Do You Know It’s Safe Enough?

Safety Case:
A structured written argument, supported by evidence, justifying system is acceptably safe for intended use.

Example structure:
- Timely Supervisor Response / sub-claims & evidence
- Adequate Supervisor Mitigation / sub-claims & evidence
- Appropriate Autonomy Failure Profile / sub-claims & evidence
Timely Supervisor Response

- **Human alertness**
  - Effective for only 15-30 minutes!

- **Autonomy failure detection**
  - Latency in identifying/responding
  - Risk acclimatization & false confidence

- **Accuracy of mental model**
  - How does a human supervisor model an opaque AI system?

- **ODD violation detection**
  - Does supervisor know that light haze is a problem?

- **What if autonomy leaves no error margin?**
When Do You Disengage?

Assume vehicle has avoided obstacles 1000+ times before
Adequate Supervisor Mitigation

- **Situational awareness**
  - Surrounding traffic; environment

- **Plan correct response**
  - Takes time for driver to re-engage
  - Stop? Swerve? Hit?

- **Execute response properly**
  - Risk of incorrect startle response to emergency

- **Vehicle responds to supervisor commands**
  - Disengagement should be natural
  - Does disengagement really work? (conform to ISO 26262)
How big is this valley for a particular vehicle?
Humans can’t provide 100% mitigation
- \( \text{RISK} = \text{Prob(vehicle fail)} \times \text{Prob(supervisor fail)} + \text{Prob(supervisor mistake)} \)
- NON-LINEAR effect of supervisor dropout

 Surprise!
Supervising good autonomy is more difficult!

Need to understand likely vehicle failure rate
- Simulation-based & closed course validation, etc.

Need to understand supervisor performance
- Supervisor training, test plan, vehicle failures

https://goo.gl/YUC5oU
Show Me The Data!

“Disengagements” is the wrong metric for testing safety
- Minimizing disengagements can incentivize unsafe testing

Data collection based on safety argumentation
- Timely supervisor response
- Adequate supervisor mitigation
- Appropriate autonomy failure profile

Show Me The Data!

--- W. Edwards Deming

ALL OTHERS BRING DATA
Ways To Reduce Testing Risk

- It’s all about testing safely
  - “Human at fault” is still unsafe testing!

- Create a testing safety case
  - Timely Supervisor Response
  - Adequate Supervisor Mitigation
  - Appropriate Autonomy Failure Profile

- Reduce road testing exposure
  - More simulation
  - Validate instead of debug on public roads
  - Collect road data instead of testing
  - Test below 20 mph (reduced pedestrian lethality)

https://goo.gl/dBdSDM