SAE INTERNATIONAL

WCX APRIL 9-11 2019 DETROIT™

sae.org/wcx

10 April 2019

Safety Argument Considerations for Public Road Testing of Autonomous Vehicles

Phil Koopman, Carnegie Mellon University
Beth Osyk, Edge Case Research

Carnegie Mellon University

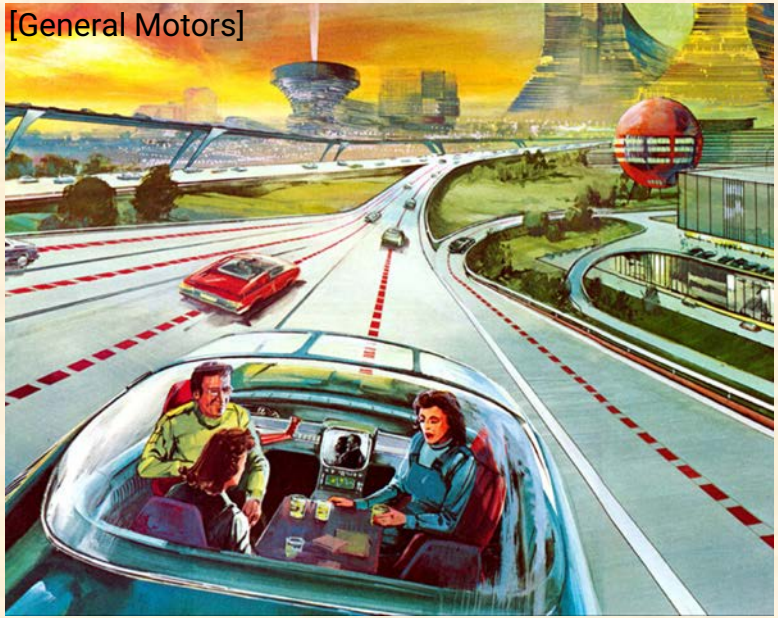
@PhilKoopman

EDGE CASE RESEARCH

Overview

EDGE CASE RESEARCH

- **Tempe AZ fatality**
  - Did we really learn the right lesson?
- **How safe is safe enough?**
  - Challenge: human supervisor effectiveness
- **Safety case for road testing:**
  - Timely human supervisor response
  - Adequate human supervisor mitigation
  - Appropriate system failure profile



We shouldn't be killing people in our haste to get to a safe future.

Tempe Arizona / March 18, 2018

EDGE CASE RESEARCH

Elaine Herzberg
Pre-impact dashcam image
Tempe Police Dept.

- **Can we avoid repeating a tragic death?**
- **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?

EDGE CASE RESEARCH

- **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?

EDGE CASE RESEARCH

- **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

- **<u>Non-occupant fatalities: 18%</u>** **about every <u>510 Million Miles</u>**
  - Motorcyclist fatalities: 14% about every 660 Million Miles
  - *Data <u>includes</u> drunk drivers, speeders, no seat belts*

➔ **Expect <u>zero</u> deaths in a 10 million mile road test campaign**

(On average, expect 0.1 fatalities, 0.02 pedestrian fatalities)

Can Humans Safely Supervise Autonomy?

EDGE CASE RESEARCH

Man reportedly caught sleeping behind the wheel of a self-driving Tesla

https://goo.gl/ZFCYzD

Sarah Whitten | @sarahwhit10

Published 11:38 AM ET Wed, 25 May 2016 | Updated 9:46 AM ET Thu, 26 May 2016

CNBC

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

Justin T. Westbrook

10/04/18 10:28am • Filed to: WAYMO

JALOPNIK

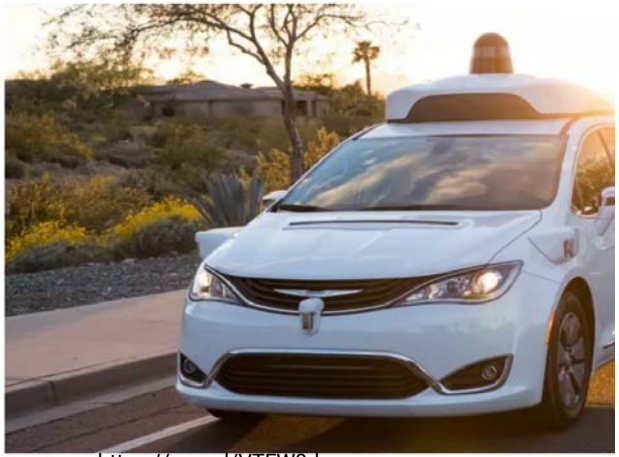


Photo: Waymo

https://goo.gl/VTFW9d

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

BUSINESS INSIDER

Graham Rapier Nov. 6, 2018, 4:20 PM

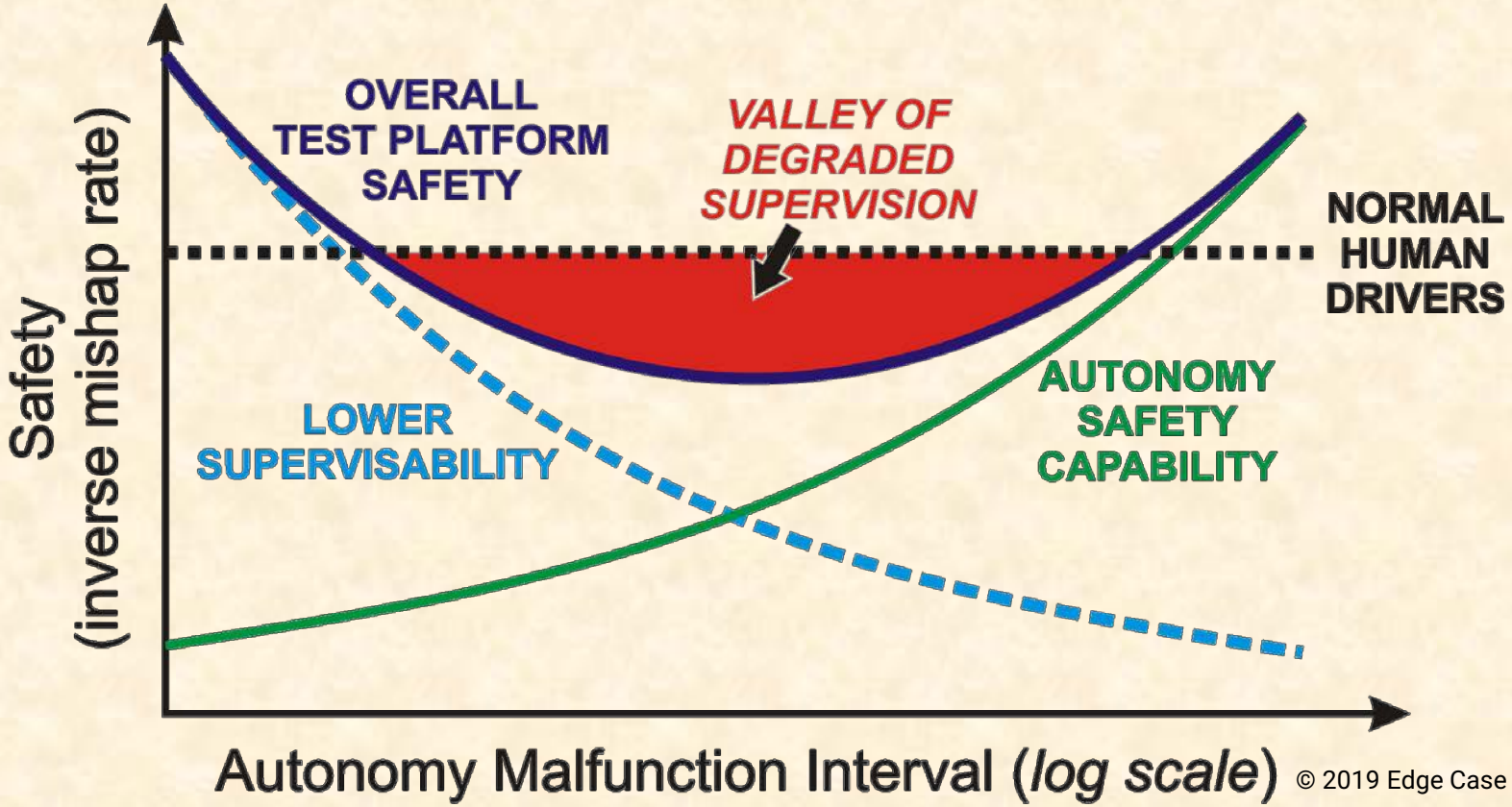
Waymo

https://goo.gl/kgRq71

Valley of Autonomy Supervisor Dropout

EDGE CASE RESEARCH

- How big and deep is this valley for a particular vehicle?



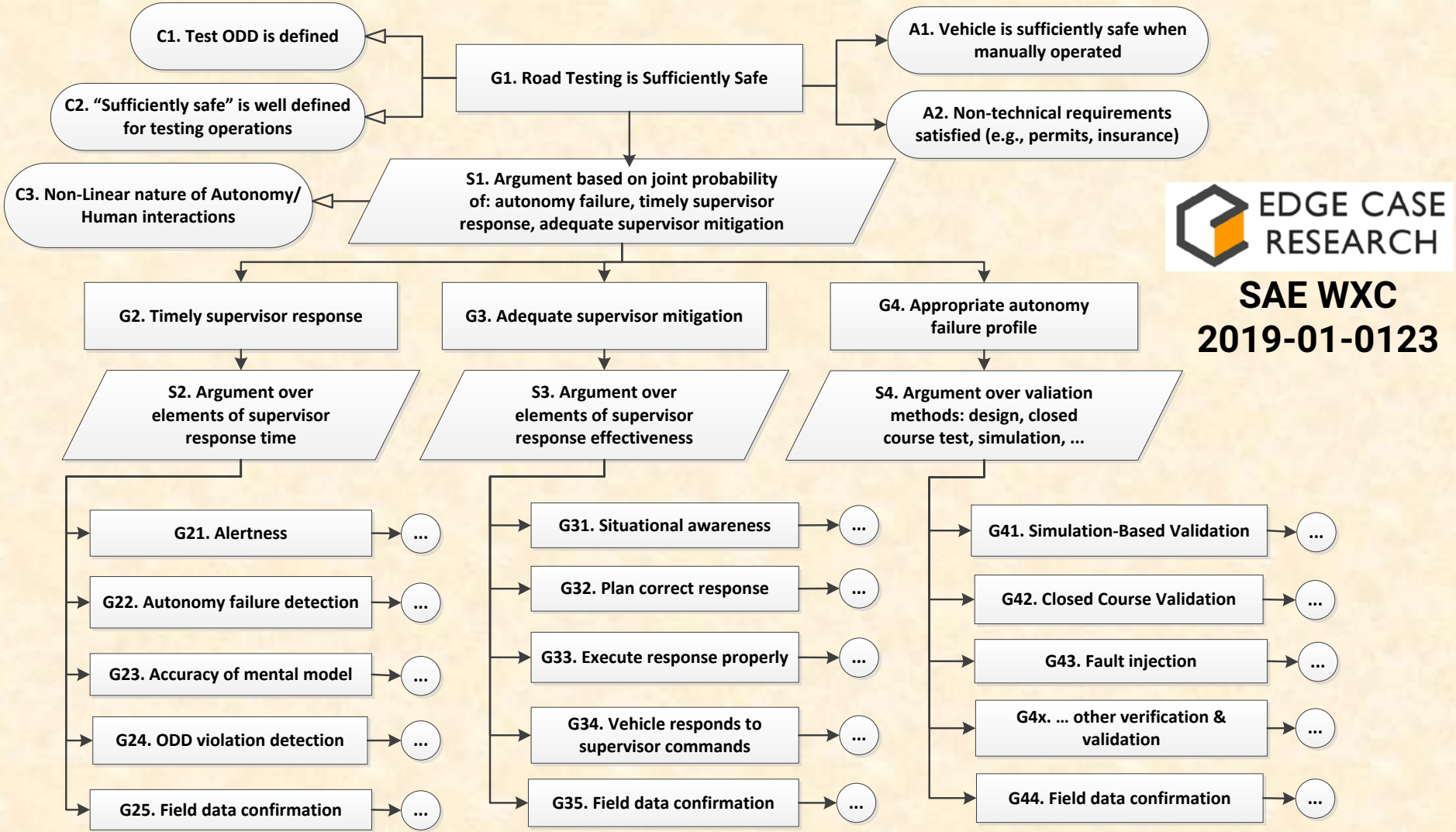
How Do You Know It's Safe Enough?

EDGE CASE RESEARCH

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

National Transportation Safety Board/Handout via REUTERS

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


C1. Test ODD is defined
C2. "Sufficiently safe" is well defined for testing operations
G1. Road Testing is Sufficiently Safe
A1. Vehicle is sufficiently safe when manually operated
A2. Non-technical requirements satisfied (e.g., permits, insurance)
C3. Non-Linear nature of Autonomy/ Human interactions
S1. Argument based on joint probability of: autonomy failure, timely supervisor response, adequate supervisor mitigation
EDGE CASE RESEARCH
SAE WXC 2019-01-0123
G2. Timely supervisor response
G3. Adequate supervisor mitigation
G4. Appropriate autonomy failure profile
S2. Argument over elements of supervisor response time
S3. Argument over elements of supervisor response effectiveness
S4. Argument over valiation methods: design, closed course test, simulation, ...
G21. Alertness
G22. Autonomy failure detection
G23. Accuracy of mental model
G24. ODD violation detection
G25. Field data confirmation
G31. Situational awareness
G32. Plan correct response
G33. Execute response properly
G34. Vehicle responds to supervisor commands
G35. Field data confirmation
G41. Simulation-Based Validation
G42. Closed Course Validation
G43. Fault injection
G4x. ... other verification & validation
G44. Field data confirmation
...

Timely Supervisor Response

EDGE CASE RESEARCH

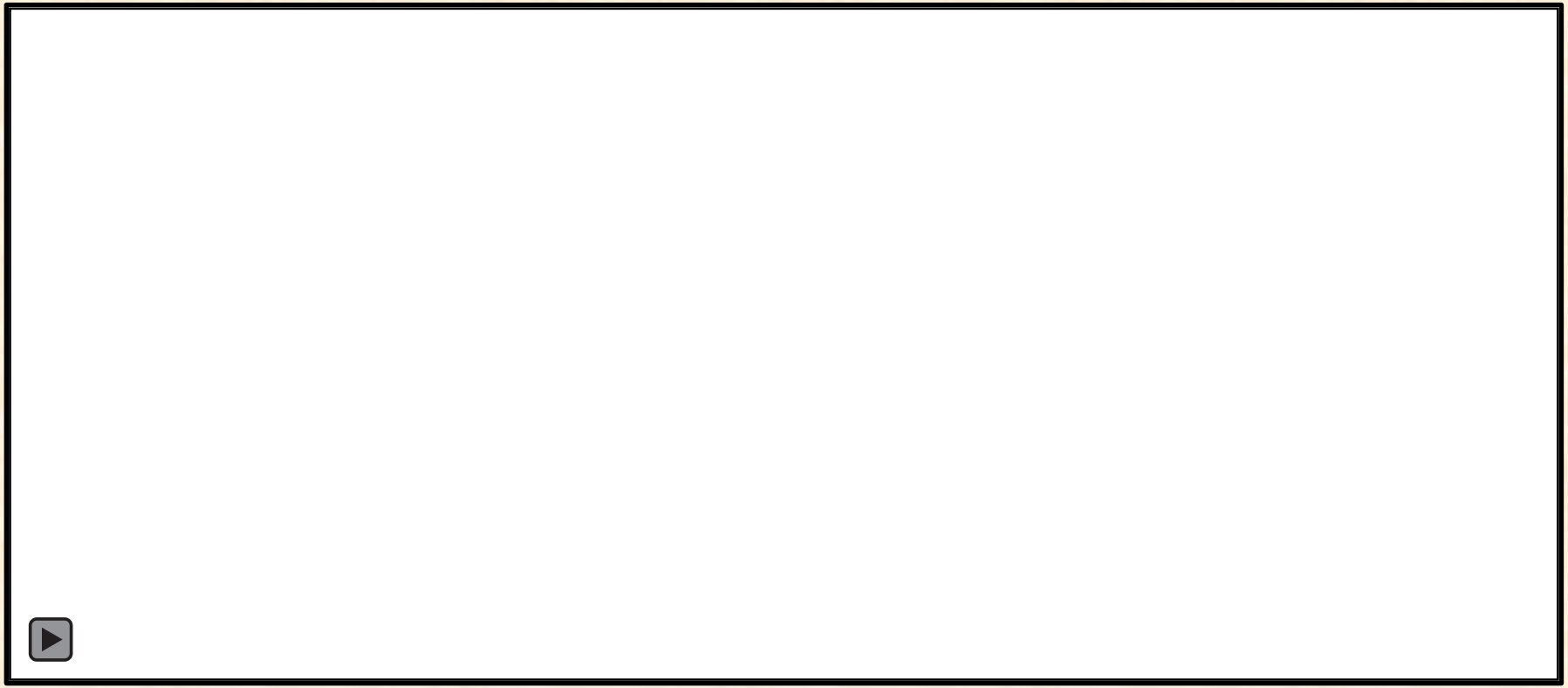
- **Human alertness**
  - Effective for only 15-30 minutes!
- **Autonomy failure detection**
  - Latency in identifying/responding
  - Risk acclimatization & false confidence
- **Accuracy of mental model**
  - How does a human supervisor model an opaque AI system?
- **ODD violation detection**
  - Does supervisor know that light haze is a problem?
- **What if autonomy leaves no error margin?**

Snooze cruise: How the drama unfolded as the two pilots 'slumbered at the controls' 2009 https://goo.gl/5htvnP

When Do You Disengage?

EDGE CASE RESEARCH

- Assume vehicle has avoided obstacles 1000+ times before



Adequate Supervisor Mitigation

EDGE CASE RESEARCH

- **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)

https://goo.gl/YUC5oU

Appropriate Autonomy Failure Profile

EDGE CASE RESEARCH

- **Humans can't provide 100% mitigation**
  - RISK = Prob(vehicle fail) * Prob(supervisor fail)
    + 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

Show Me The Data!

EDGE CASE RESEARCH

- **"Disengagements" is the wrong metric for safe testing**
  - Minimizing disengagements can incentivize unsafe testing

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

-- W. Edwards Deming

Ways To Reduce Testing Risk

EDGE CASE RESEARCH

- **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)