Quick Overview

- Getting past Autonomous Vehicle (AV) safety rhetoric
- Safety Engineering in a nutshell
- Why Machine Learning (ML) breaks safety engineering
- Core ML safety technical issues
- ANSI/UL 4600 approach
- Beyond technical safety metrics

https://on.gei.co/2r2rjzg
Nobody knows when/if Autonomous Vehicles (AVs) will be safer than human drivers

- Improved safety is purely aspirational
- “AVs are safe” messaging is often propaganda

Some humans drive drunk

- On average they are still good and adaptable

But computers lack common sense

- ML is brittle when encountering novelty

Computer drivers can be imperfect even for “easy” failures

- Safety must be engineered, not assumed

Safety Engineering In A Nutshell

- Conventional vehicle safety is ~1 fatality / 100M miles (US)
  - Call it 0.00000000001 fatalities per meter
    - Including drunk, distracted drivers, etc.!
  - Testing does not prove safety
    - Too much testing needed to be practicable

- Safety comes from engineering rigor
  - Identify and mitigate hazards
  - Use engineering rigor responsive to risk presented
  - Testing validates hazard mitigation & engineering quality
  - Safety standards, e.g. ISO 26262, ANSI/UL 4600 exist…
    - ... but conformance is patchy at best; no requirement to follow these
Machine Learning Breaks Safety Engineering

- Primary safety concern: ML for perception/prediction
- Data-centric/training approach breaks safety engineering
  - Safety engineering depends on traceability
  - ML model training not traceable for safety
- Brute force simulation has limits
  - Simulation accuracy becomes life critical
  - Billions of miles real-world to validate simulated world
- ML breaks the safety certification/recall model
  - Currently a useful fiction that vehicles are “safe” when deployed
  - AVs will need lifetime monitoring and updates to maintain safety
Core ML Safety Technical Issues

- Long tail events are handled poorly by ML
  - Safety is about rare, high-consequence events
  - ML is brittle for novel events
  - ML Safety is limited by handling novel events

- Experience suggests “surprises” are heavy tail
  - Need to detect unknown relevant characteristics

- Human drivers are terrible automation supervisors
  - Approaches expecting perfect human supervision are not viable
  - Driver attention management technology needs more work
    - Common to see “moral crumple zone” strategy instead
ANSI/UL 4600 Approach

- ANSI Standard issued in 2020
  - Assessment approach to safety cases
    - Safety case: structured argument with safety claims supported by evidence
  - Autonomous vehicles: from grocery bots to trucks

- Key UL 4600 features
  - Minimum required content of safety case
  - Numerous “did you think of that?” hazard prompts
  - Quantitative measurement of safety case claims
    - Safety Performance Indicators detect falsified claims
    - Lifecycle feedback to evolve safety case as required
Beyond Technical Safety

- Engineering utilitarian approaches aren’t enough
  - Risk redistribution, fatalities as an affordable cost of business, ...
- “As safe as a human driver” has multiple interpretations
  - Technical: which driver, where, in what vehicle, which victims, etc.
    - Statistical outcome measurements; very complex
  - Legal: lack of negligent behavior
    - Compare to “reasonable” rather than “average” driver
    - Emphasize avoiding harm rather than average outcomes
- Modest proposal:
  - Any “AI” system that supplants human judgement...
    ... should be held to human standards of negligence
Resources

- Video lecture series on autonomous vehicle safety:
  - Keynote AV Safety overview video: https://youtu.be/oE_2rBxNrfc
  - Mini-course: https://users.ece.cmu.edu/~koopman/lectures/index.html#av

- “Safe Enough” book & talk video:

- UL 4600 book & talk video:

- Liability-based proposal for AV regulation & podcast
  - https://safeautonomy.blogspot.com/2023/05/a-liability-approach-for-automated.html