



Prof. Philip Koopman

**Carnegie  
Mellon  
University**

# Redefining Safety for Autonomous Vehicles

SafeComp  
September 20, 2024

[PhilKoopman.Substack.com](http://PhilKoopman.Substack.com)

PHILIP KOOPMAN

**HOW SAFE IS  
SAFE ENOUGH?**

Measuring and Predicting  
Autonomous Vehicle Safety



- Time for computer-based system safety to evolve
  - Autonomous systems show us where to improve
- Definitional build-up:
  - Loss
  - Risk
  - Safety Constraint
  - Safety Engineering
  - Safety Case
  - Acceptable safety
- Viewpoint: safety as constraint satisfaction rather than risk optimization



# Is “Safety Case” Definition Broken?

- DefStan 00-56: “... in a given operating environment”
  - AVs: open, incompletely defined environments
  - Unexpected obstacles, vehicle types, etc.

## Crash into utility pole



## Crash into articulated bus



# Is “Risk” Definition Broken?

- Typically: combination of probability and severity
  - See also *Positive Risk Balance* (“safer than human driver”)
  - What about risk redistribution onto vulnerable populations?
- 11 of 74 SF Fire Dept. robotaxi incidents in Tenderloin District
  - Economically distressed
  - High drug use
- Mishaps at edge of Tenderloin:
  - Cruise fire truck crash
  - Cruise pedestrian dragging



# Key Insights:

- Robots can fail even if they do not drive drunk
  - Is negligent driving OK?
  - Is uneven risk distribution OK?
  - Should losses due to “rare” events be OK?
- No human operator to blame
  - Who is responsible for negligent behavior?
  - Who/what monitors “for a given environment”?
  - Social interactions are in-scope for technology
- Let’s explore revising safety terminology



# Definition of Loss

- ISO 26262 Harm: physical injury to people
  - But what about other incidents?
- Loss: an adverse outcome, including damage to the system itself, negative societal externalities, damage to property, damage to the environment, injury or death to animals, and injury or death to people

## Autonomous Waymo car runs over dog in San Francisco

The vehicle was in autonomous mode with a safety driver present in a 25 mph zone.

RON AMADEO - 6/7/2023, 2:24 PM

<https://bit.ly/4cLX2s4>



# Definition of Risk

## ■ Classical risk: combination of probability and severity

- ISO 26262 includes controllability
- But, we see recalls for *patterns* of losses

Federal regulator finds Tesla Autopilot has 'critical safety gap' linked to hundreds of collisions

<https://bit.ly/3SXklHr>

The NHTSA report comes as Tesla signals it is betting its future on autonomous driving.



## ■ NHTSA EA22002 / Recall 23V838

- 956 Tesla crashes/ 29 fatalities <https://bit.ly/4cChQ4z>
- Avoidable crashes, loss of yaw control
- Inadvertent AutoSteer override

## ■ Risk: combination of the probability of occurrence of a loss, or pattern of losses, and the importance to stakeholders of the associated consequences

# Definition of Safety Constraint

- Is safety net minimizing the sum of risks?
  - Near zero probability \* catastrophic consequence = ???
- Risk due to negative externalities
  - How does design team assign consequence to blocking a fire truck?
  - Rules & regulations help here
- Safety constraint: a limitation imposed on risk or other aspects of the system by stakeholder requirements

San Francisco's fire chief is fed up with robotaxis that mess with her firetrucks. And L.A. is next

<https://bit.ly/3Wc3bXA>



San Francisco Fire Chief Jeanine Nicholson says state regulators are moving too fast on robotaxi expansion, jeopardizing public safety. Meanwhile, Waymo and Motional are planning to begin robotaxi service in Los Angeles. (Lea Suzuki / San Francisco Chronicle via Associated Press)



# Definition of Safety Engineering

- Testing alone does not create safe software

## One Million Driverless Miles

But ... arguing safety via brute force testing is a pervasive narrative



- Safety engineering: a methodical process of ensuring a system meets all its safety constraints throughout its lifecycle, including at least hazard analysis, risk assessment, risk mitigation, validation, and field engineering feedback

# Definition of a Safety Case

- Safety case: ... “given application in given environment”
  - What about open environment? (Cruise crash into bus)
  - What about enforcing operational limits?

**Tesla driver arrested for DUI after allegedly using self-driving option while drunk then passing out**

Police in California forced electric car to stop automatically by pulling in front of it <https://bit.ly/3zODCUW>

- Safety case: structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and sound argument that safety engineering efforts have ensured a system meets a comprehensive set of safety constraints

# Definition of Acceptable Safety

- More to safety than positive risk balance
  - Meet ethical constraints (e.g., risk distribution)
  - Non-negligent driving (e.g., justifiable road rule violation)
  - No recallable behaviors (even if net risk is OK)
  - Meet legal restrictions (e.g., passenger drop-off)
- Net acceptability across all stakeholders
  - Auto industry, insurance industry
  - Regulators, legislators
  - Road users, consumer advocates



[Dall-e]

Acceptable: meets all safety constraints as shown by a safety case

# Summary

## ■ New definitions needed; No person to handle:

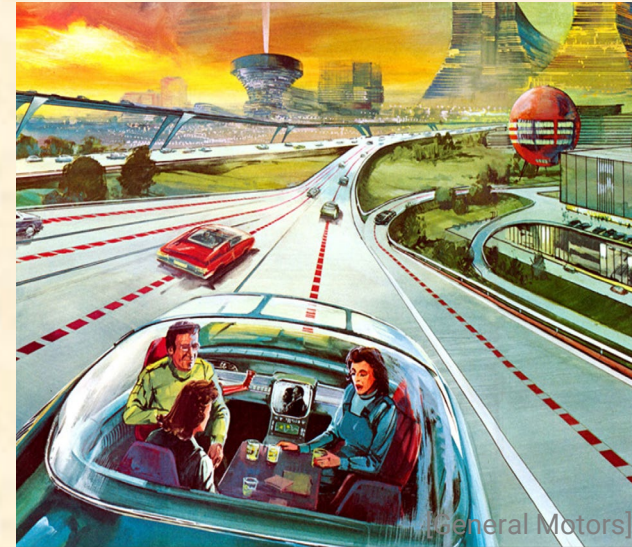
- Surprises in environment
- Enforcement of operational limits
- “Do the right thing” rule interpretation
- Legal and ethical constraints

## ■ Can't we just re-interpret existing terms?

- Minimal-compliance organizations are likely to fall short
- Terms should say what they mean

## ■ Extended paper addresses specific safety standards

<https://arxiv.org/abs/2404.16768>



- **Loss:** an adverse outcome, including damage to the system itself, negative societal externalities, damage to property, damage to the environment, injury or death to animals, and injury or death to people
- **Risk:** combination of the probability of occurrence of a loss, or pattern of losses, and the importance to stakeholders of the associated consequences
- **Safety constraint:** a limitation imposed on risk or other aspects of the system by stakeholder requirements
- **Safety engineering:** a methodical process of ensuring a system meets all its safety constraints throughout its lifecycle, including at least hazard analysis, risk assessment, risk mitigation, validation, and field engineering feedback
- **Safety case:** structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and sound argument that safety engineering efforts have ensured a system meets a comprehensive set of safety constraints
- **Acceptable:** meets all safety constraints as shown by a safety case