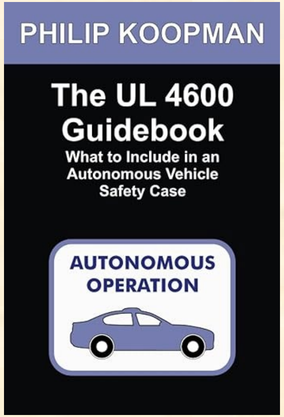
Automated Vehicle Safety Cases: Scope & Structure

Prof. Philip Koopman

Carnegie Mellon University

July 31, 2024
TRB/ARTS24

www.Koopman.us

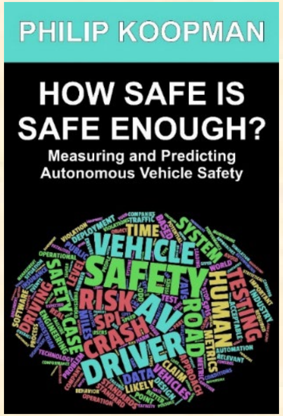


PHILIP KOOPMAN

The UL 4600 Guidebook

What to Include in an Autonomous Vehicle Safety Case

AUTONOMOUS OPERATION



PHILIP KOOPMAN

HOW SAFE IS SAFE ENOUGH?

Measuring and Predicting Autonomous Vehicle Safety

Assurance Arguments To Support Safety

Carnegie Mellon University

- Safety case:
  - Logical argument + Evidence → Safety Claim
- Scope:
  - What do you mean by acceptably “safe”?
  - Why do you think you are safe?
  - Why do <u>you</u> believe your argument?
  - Why should <u>we</u> believe your argument?
- There is no “One True Safety Case” structure

[Dall-e]

Scope of "Acceptably Safe" Claim

Carnegie Mellon University

- Net statistical safety (safer than average driver?)
  - Establishing a baseline is very complex!
- Tolerance for risk transfer
  - What if pedestrian risk doubles? (etc.)
- Tolerance for negligent behavior
  - What if breaking a traffic rule results in harm?
- Fine-grain absence of unreasonable risk
  - Recalls tend to be for specific behaviors
- Ethical behavior & equity concerns
  - Consequences of testing & deployment decisions

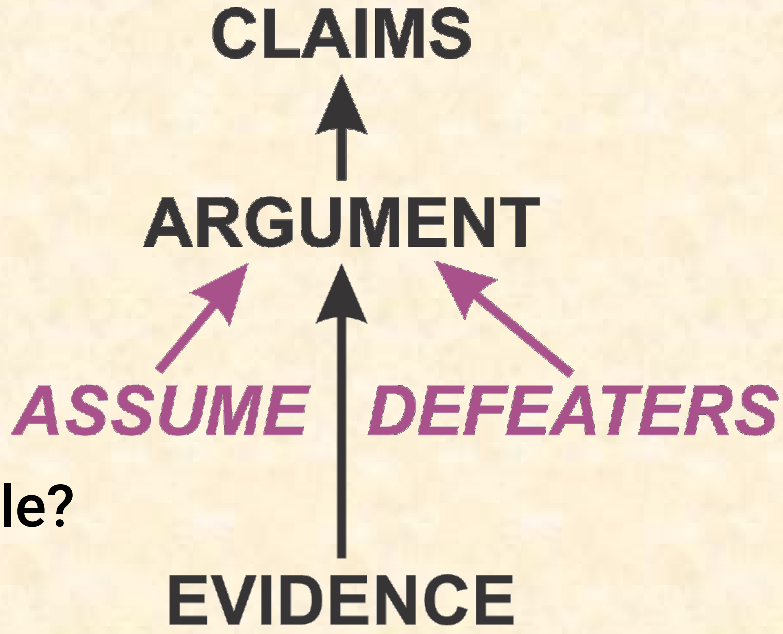
https://bit.ly/3KO9PPe

Reference: Redefining Safety for AVs https://arxiv.org/abs/2404.16768

Why Do You Think You Are Safe?

- Claims + well reasoned argument
  - Claim true because A and B and C
  - No rhetoric allowed
- Potential defeaters considered
  - Why might this argument be false?
- Identify assumptions
  - Why are these assumptions reasonable?
- Supported by evidence
  - Engineering rigor, simulations, test

Reference: UL 4600 Chapter 5

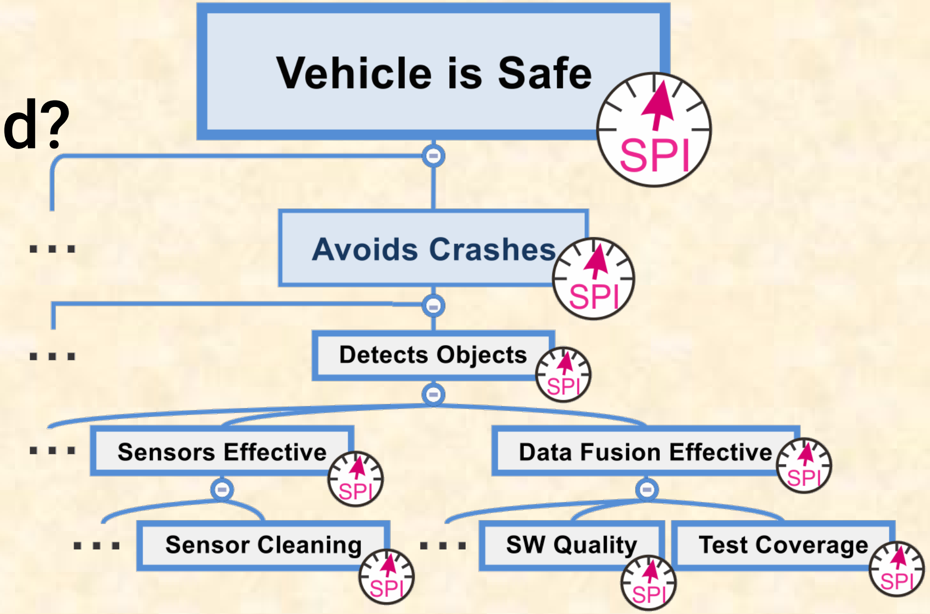


Safety Case

Why Do *YOU* Believe Your Argument?

Carnegie Mellon University

- Safety case review
  - Tool checks for consistency, no loose ends
  - Peer review by internal teams
- What if the argument is unsound?
  - Safety Performance Indicators
    ➔Instrument safety case claims
- Reviewer independence
  - What happens to a safety reviewer who says “no”?



Reference: UL 4600 Chapters 16 & 17

Why Should *WE* Believe Your Argument?

Carnegie Mellon University

- Credibility of safety case
  - What exactly are the claims?
  - Expose some of the safety case
  - Integrity of independent review process
- Public SPI metrics
  - How do they trace to your safety case?
- Conformance to UL 4600
  - A standard for *assessing safety cases*
  - #DidYouThinkofThat?
    - Argument completeness, validity

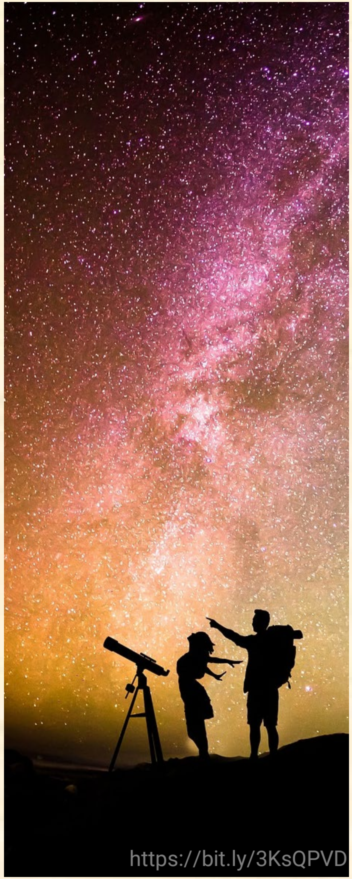
SPI: Safety Performance Indicator

Searching For The One True Safety Case

Carnegie Mellon University

There is no One True Safety Case!

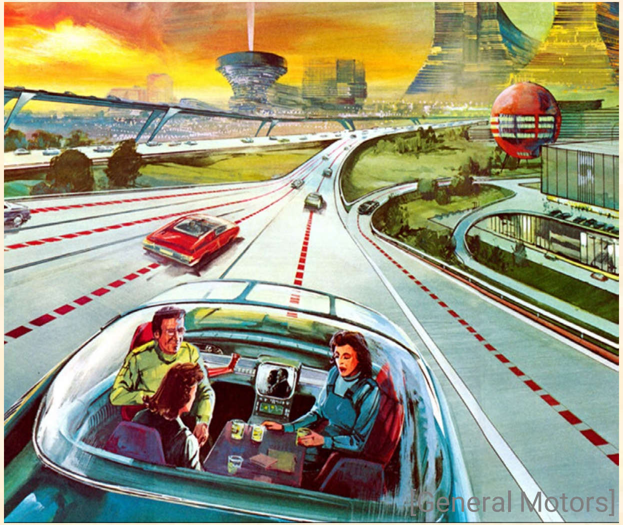
- Claims might vary by operational concept
- Argument strategies vary
  - Operational environment, role of remote support
  - System architecture & development strategy
  - Depth / assumption scope will vary
  - Notation approach will vary (graphical vs. textual)
- Evidentiary needs vary by argument strategy
  - SPI instrumentation enables broader assumptions
- The act of creating the case has significant value



https://bit.ly/3KsQPVD

Summary

- Safety case:
  - Logical argument + Evidence → Safety Claim
- Scope:
  - What do you mean by acceptably "safe"?
  - Why do you think you are safe?
  - Why do <u>you</u> believe your argument?
  - Why should <u>we</u> believe your argument?
- Structure
  - Quality of argument matters, not notation



[General Motors]

FREE view UL 4600 launch page: https://bit.ly/ul4600