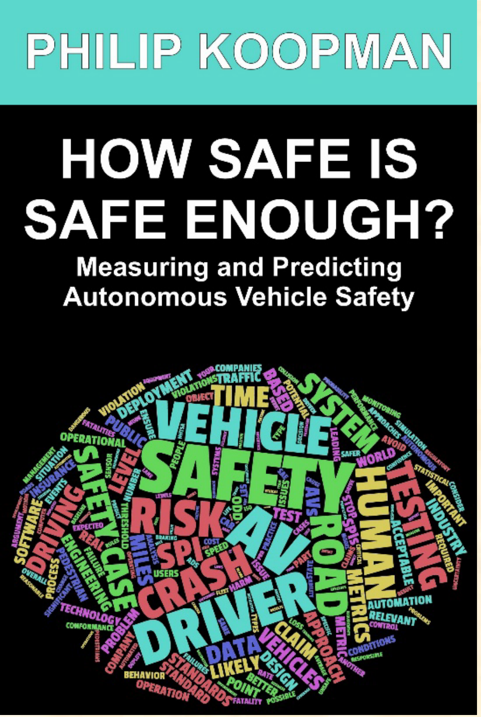


PHILIP KOOPMAN

HOW SAFE IS SAFE ENOUGH?

Measuring and Predicting Autonomous Vehicle Safety

Section 8.4 Bootstrapping

Bootstrapping Safety Assurance

(and lack thereof)

December 2022

www.Koopman.us

Prof. Philip Koopman

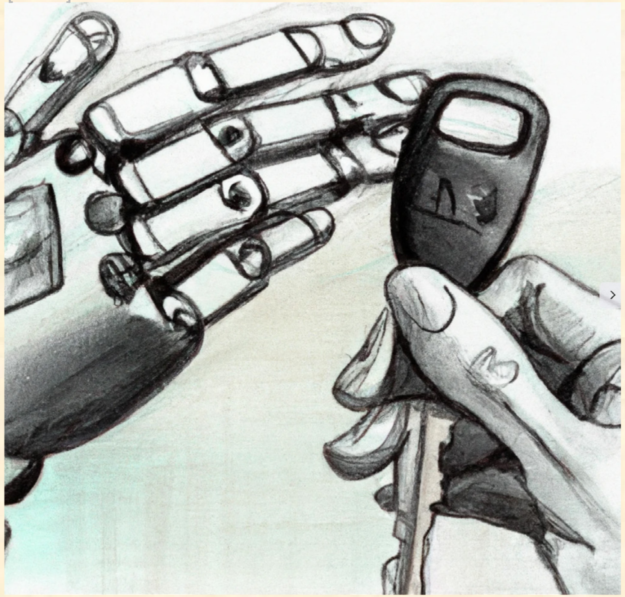
Carnegie Mellon University

Overview

Carnegie Mellon University

- Test-centric safety assurance
  - E.g., for autonomous vehicles
  - But testing alone is too expensive, so…
- Bootstrapping schemes
  - Bootstrapping by miles
  - Phased deployment
  - “Probably perfect” arguments
- Conclusion: they won’t work the way you hope they will
  - Driver-out “safety testing” is unsafe
  - Bootstrap testing won’t fix this

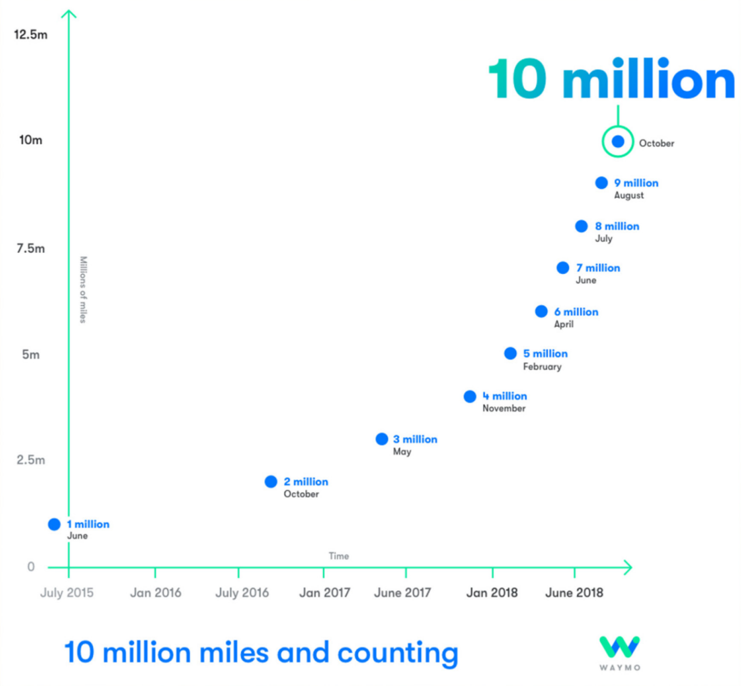
[Dall-e]



The Race To Autonomy Started With Miles

- Good for identifying common scenarios
  - Expensive; risk of a high profile crash

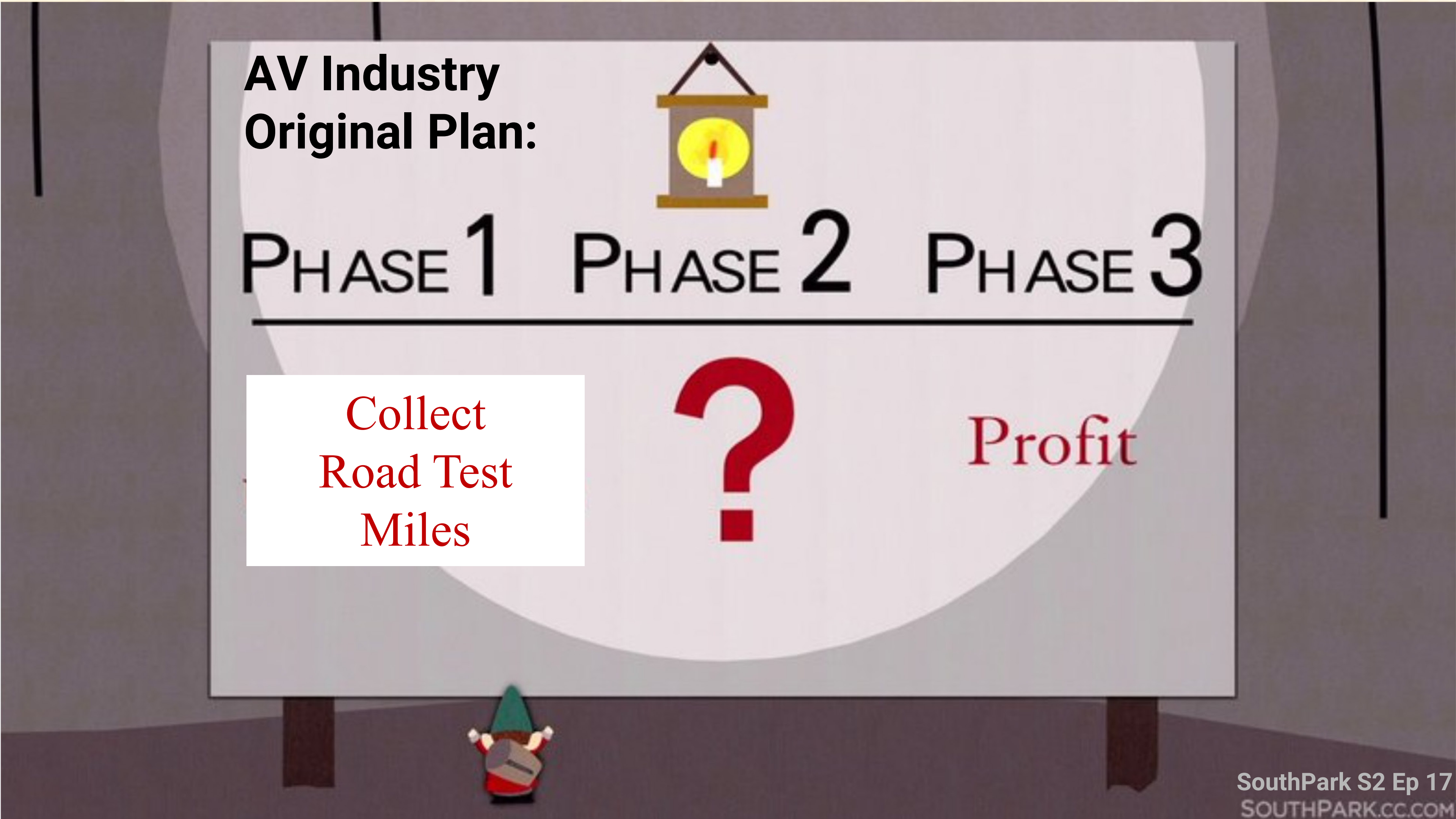
http://bit.ly/2toadfa



10 million miles and counting

https://bit.ly/3TlaPMb

AV Industry Original Plan:

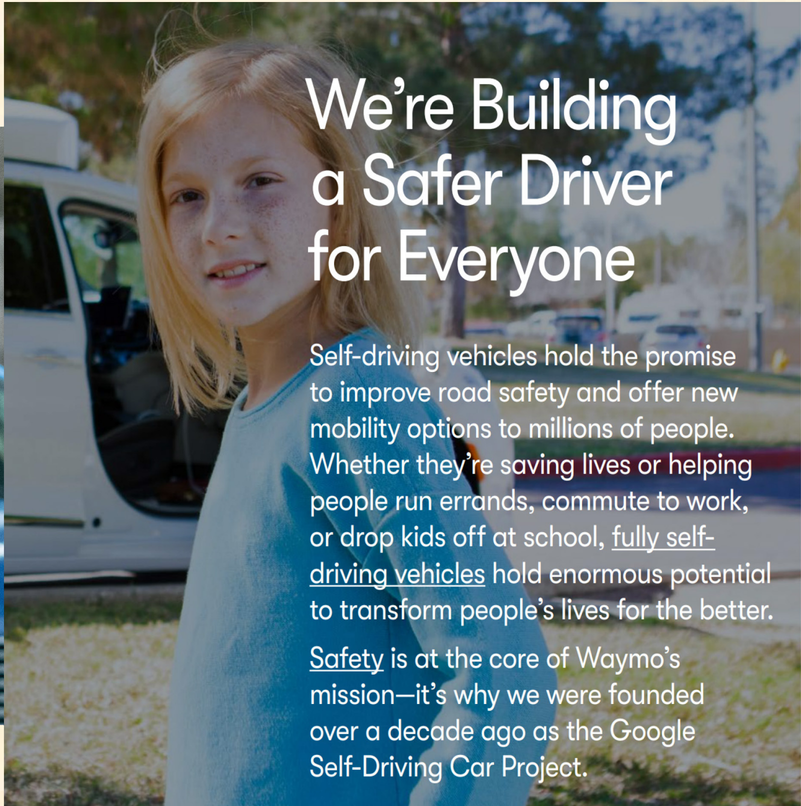


SouthPark S2 Ep 17

ADS Technology has come to be: Sold Based on Safety

Waymo VSSA https://bit.ly/2QuYhai

Ford VSSA https://bit.ly/3njionT

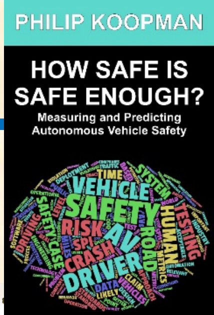


We're Building a Safer Driver for Everyone

Self-driving vehicles hold the promise to improve road safety and offer new mobility options to millions of people. Whether they're saving lives or helping people run errands, commute to work, or drop kids off at school, fully self-driving vehicles hold enormous potential to transform people's lives for the better.

Safety is at the core of Waymo's mission—it's why we were founded over a decade ago as the Google Self-Driving Car Project.

How Safe Is "Safe?"



PHILIP KOOPMAN

HOW SAFE IS SAFE ENOUGH?

Measuring and Predicting Autonomous Vehicle Safety

- ~100M miles/fatal mishap for human drivers (US)
  - 28% Alcohol impaired/Driving Under Influence
  - 26% Speed-related
  - 9% distracted driving
  - 2% drowsy …

[DOT HS 813 060 & DOT HS 813 021]

(total > 100% due to multiple factors in some mishaps)

[Dall-e]

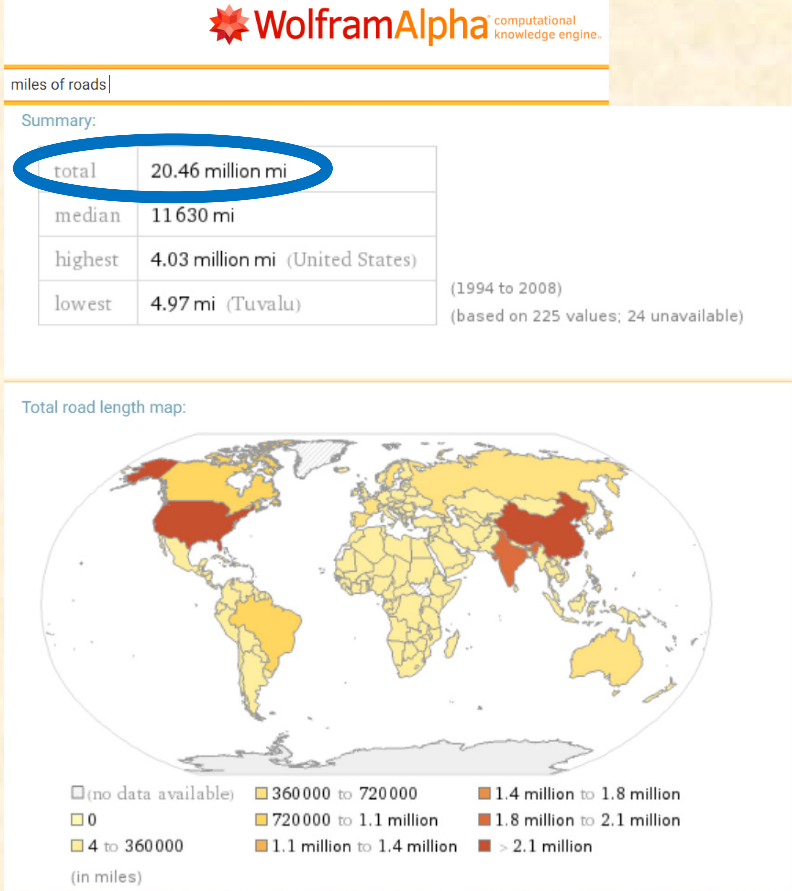
- Fully functional drivers are much better
- New AV has better safety than 10+ year old "average"car

→Better than an unimpaired, undistracted driver in new car

- ("Safe Enough" is complicated – but a different talk.)

Safety Via Brute Force Road Testing (?)

- Say 200M miles/critical mishap…
  - Test 3x–10x longer than mishap rate
    ➔ Need 2 Billion miles of testing
- That's ~50 round trips
  on every road in the world
  - With fewer than 10 critical mishaps
  - Even more testing if you find a defect and redo some testing
- Required scale is infeasible



WolframAlpha computational knowledge engine.

miles of roads

Summary:

total	20.46 million mi
median	11630 mi
highest	4.03 million mi (United States)
lowest	4.97 mi (Tuvalu)

(1994 to 2008)
(based on 225 values; 24 unavailable)

Total road length map:

□ (no data available)	□ 360000 to 720000	■ 1.4 million to 1.8 million
□ 0	□ 720000 to 1.1 million	■ 1.8 million to 2.1 million
□ 4 to 360000	■ 1.1 million to 1.4 million	■ > 2.1 million

(in miles)

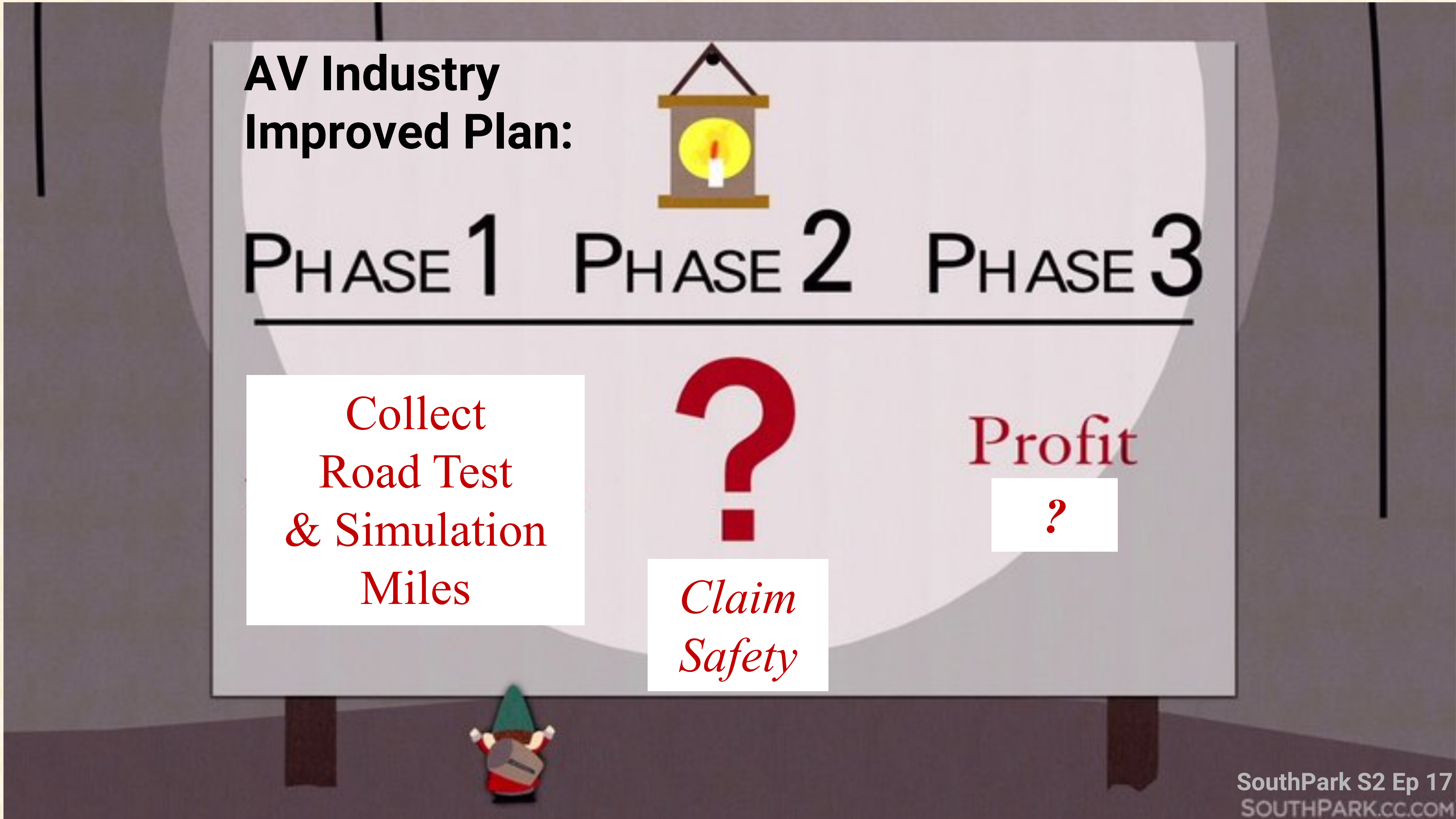
Use Simulation To Supplement Road Tests

- Highly scalable
  - “All models are wrong; some are useful.” (George Box)
  - “Simulations are doomed to succeed.”
- Still need real world miles to validate the simulations

Mairie de St FORGET

[ANSYS]

AV Industry Improved Plan:



Collect Road Test & Simulation Miles

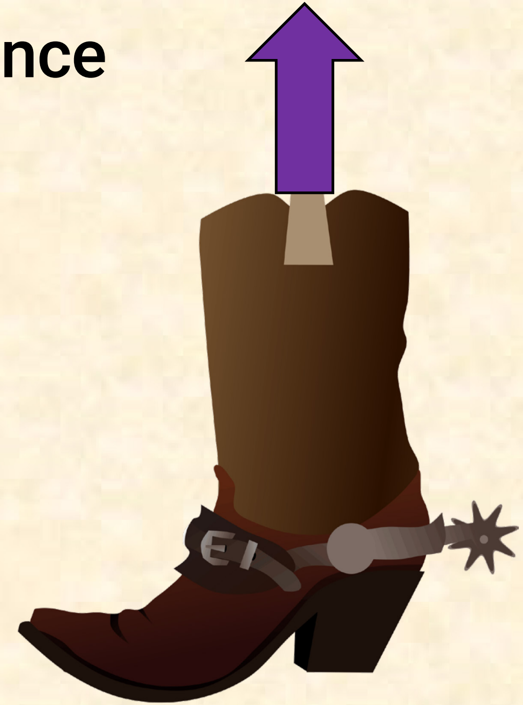
Claim Safety

?

SouthPark S2 Ep 17
SOUTHPARK.CC.COM

Bootstrapping To The Rescue (maybe)

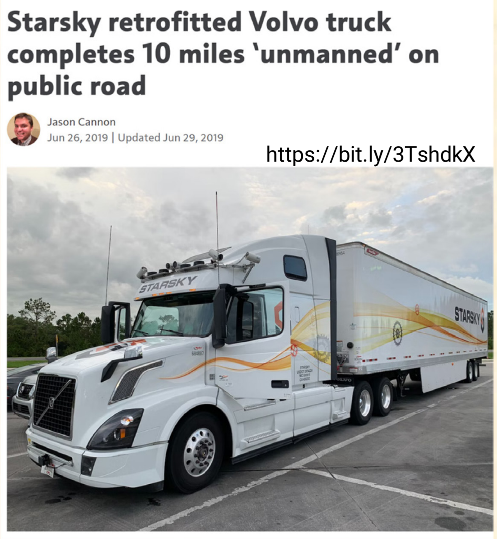
- Incremental approach to road testing assurance
  1. Observe safety driver stops intervening
  2. Remove safety driver
  3. Crash-free history predicts crash is unlikely for a small window
  4. Drive for small window with no crash
  5. Repeat Steps 3 & 4, with growing window size



- Variations
  - Pure mileage-based bootstrapping
  - Phased deployment, slow update roll-out
  - Combine with belief in probably perfect design

The Demo Question

- Hypothetically: 10K miles with safety driver
  - Zero safety driver interventions
  - 95% confidence $MTBF_{crash}$>3338 miles
- Need “driver out” demo for funding milestone
  - Demo 10 miles without driver
  - Company fails if you don’t demo on time
- What are odds of a crash on this demo?
  - $R(t) = e^{-\lambda t}$ for $\lambda = 1/3338$, $t=10$ ➔ 99.7% no crash
- Do you do the demo?
  - If there is no crash in the demo, was that safe?



Starsky retrofitted Volvo truck completes 10 miles ‘unmanned’ on public road

Jason Cannon
Jun 26, 2019 | Updated Jun 29, 2019

https://bit.ly/3TshdkX

https://reliabilityanalyticstoolkit.appspot.com/mtbf_test_calculator

One-Off Events – Safe? Or Just Lucky?

- There is a 99.7% of no crash for a demo
  - You run the demo … and … no crash
  - Claim: "therefore the demo was safe"
- What are flaws in this argument?
  - Jumped out of an undamaged airplane
    - Parachute opened, so it was perfectly safe
  - Swam with sharks … still have all limbs
- Is evading a hazard once "safe" ??
  - Getting away with taking a risk …
    is not quite the same as safety
  - Public road testing imposes risks on non-consenting road users

bit.ly/3OSqq4Q

Mileage-Based Bootstrap First Step

- Example: 100K miles testing with safety driver
  - Zero safety driver interventions
  - 95% confidence $\text{MTBF}_{\text{crash}}$>33380 miles
    - (Note: automotive often does about 70% confidence)
- Do 100 miles of testing with no safety driver
  - $R(t) = e^{-\lambda t}$ for $\lambda = 1/33380$, $t=100$ ➔ 99.7% no crash
- Now you have 100,100 miles with no crash
  - 95% confidence $\text{MTBF}_{\text{crash}}$>33414 miles
  - Notice that 33,414 > 33,380 … hmmm … interesting!
    - We can bootstrap our way to proving safety!

https://bit.ly/3VtvU98

https://reliabilityanalyticstoolkit.appspot.com/mtbf_test_calculator

Naïve Bootstrap Argument

- Start with baseline testing with safety driver
  - Perhaps 1M miles? (much less than 100M miles)
  - Then remove safety driver ➔ driverless testing
- Iteratively longer test cycles
  - Test for X miles based on crash probability
  - Each step yields bigger MTBF
  - Next step can be X+δ miles due to larger MTBF
    … math, math, math …
- Lather. Rinse. Repeat.
  - Prove you are safer than a human driver

➔ $$Profit$$

[Shutterstock]

What Happens If You Get A Crash?

- Need to test longer if there is a crash
  - For 200M miles @ 95% confidence
    - ~600M miles of testing required for no crash
    - With 1 crash: ~949M miles of testing
    - 2 crashes: ~1259M miles

      …
    - 5 crashes: ~2103M miles

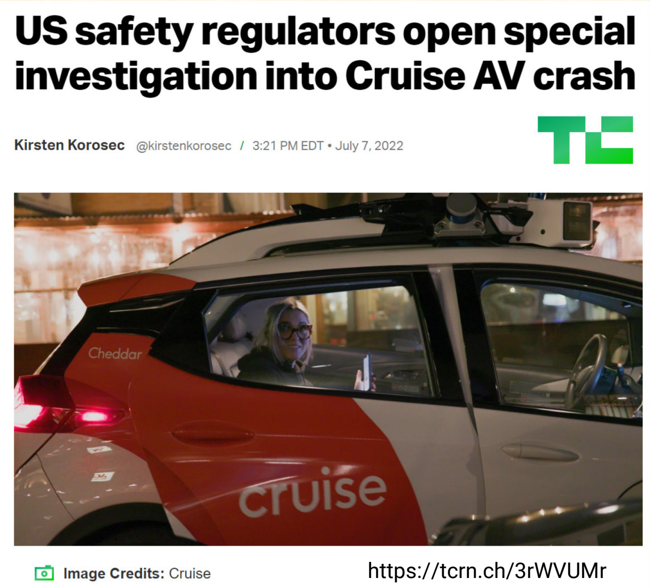
TuSimple Crash April 2022

- Probability of crashes is <u>high</u>
  - At 200M $\text{MTBF}_{\text{crash}}$, probability of crash by 600M miles is 95%
  - The math is not in your favor here … luck is required

https://reliabilityanalyticstoolkit.appspot.com/mtbf_test_calculator

Argue *That* Crash Didn't Count

- That crash does not count because {reasons}
  - It was the other driver's fault
    - A crash is still a crash
  - It was a freak/black swan occurrence
    - A crash is still a crash
  - It was a near miss instead of a crash
    - Near misses are not reported to regulators
- Argue that bug was fixed
  - Impact analysis performed
    - Do you believe in the 0% fault reinjection rate fairy?
  - Surely that was the last defect in the system. (*Really?*)



US safety regulators open special investigation into Cruise AV crash

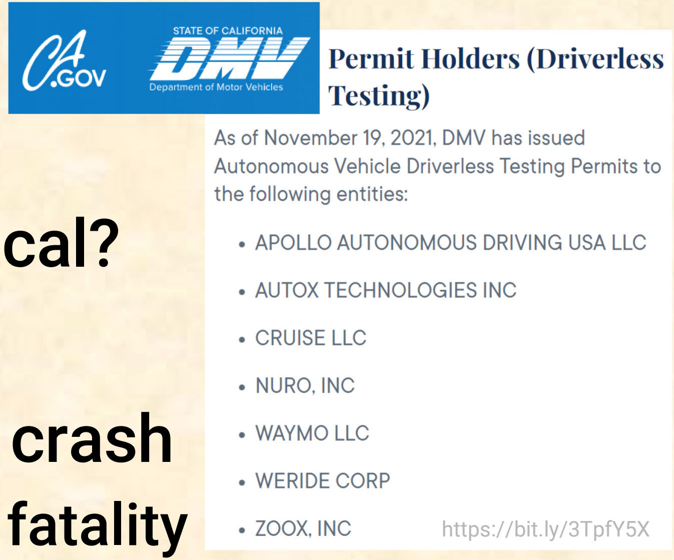
Kirsten Korosec @kirstenkorosec / 3:21 PM EDT • July 7, 2022

Image Credits: Cruise

https://tcrn.ch/3rWVUMr

Pure Bootstrap Safety Issues

- No expectation of safety up front
  - Confirms if system happens to be safe
  - Does not somehow make system safe
- Are repeated cycles of 99.7% “safe” ethical?
  - Insufficiently low bounds on mishap rate
- Find out system is unsafe is via an early crash
  - Bootstrapping in effect justifies one “free” fatality



Permit Holders (Driverless Testing)

As of November 19, 2021, DMV has issued Autonomous Vehicle Driverless Testing Permits to the following entities:

- APOLLO AUTONOMOUS DRIVING USA LLC
- AUTOX TECHNOLOGIES INC
- CRUISE LLC
- NURO, INC
- WAYMO LLC
- WERIDE CORP
- ZOOX, INC

https://bit.ly/3TpfY5X

➔ There is no such thing as uncrewed AV safety testing

- Really it is just deployment of unproven technology
  - – Pony.AI lost permit in May 2022 – empty vehicle crash

Slow-Rolling & Pilot Deployments

Carnegie Mellon University

- Introduce new versions slowly / initially operate with small pilot fleet
  - Reduces chance of large fleet having an early catastrophic failure
  - Said to be "safe" due to reduced risk
    - A variant on the one-off exposure fallacy
    - Reduces risk of <u>multiple concurrent</u> early mishaps
    - Risk reduction is not safety .. different talk
- Amounts to a bootstrap safety argument
  - Safety risk presented to individuals is unchanged
    - Loss events could still happen at unacceptable rate

[A Fish Called Wanda]

© 2022 Philip Koopman

Probably Perfect System

Bishop, Povyakalo, Strigini 2021 [https://arxiv.org/pdf/2110.10718.pdf]

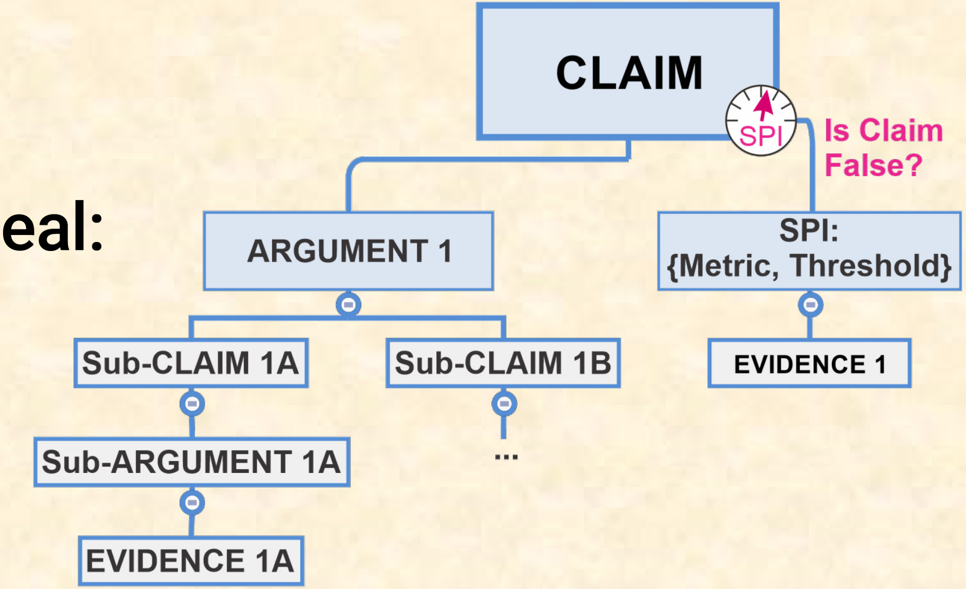
- Take credit for “probably perfect”
  - E.g., 90% probability it is safe
  - Allows faster bootstrapping
- Still might deploy unsafe system
  - E.g., 10% probability it is unsafe
    - Accumulated failure probability adds up quickly!
  - Argument destroyed along with first crash
    - Any early crash falsifies “probably perfect” hypothesis
  - Bayesian prior of “we think it is probably perfect”…

… is still an early deployment of a “possibly unsafe” system

https://bit.ly/3S5i21K

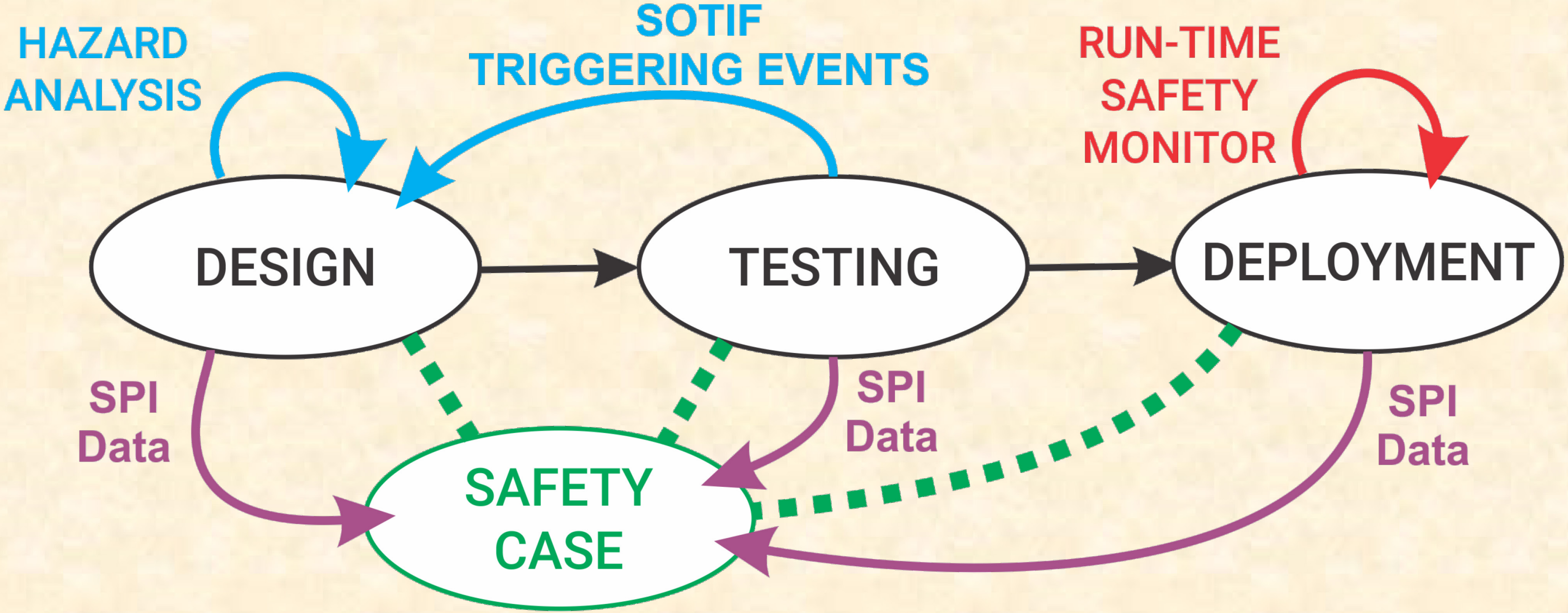
SPIs and Lifecycle Feedback

- SPI: direct measurement of safety case claim failure
  - Independent of reasoning (“claim is X … yet here is ~X”)
- A falsified safety case claim:
  - Safety case has some defect
- Root cause analysis might reveal:
  - Product or process defect
  - Invalid safety argument
  - Issue with supporting evidence
  - Assumption error
- Continual Safety case improvement



SPI-Based Feedback Approach

- Safety Case argues acceptable risk
  - SPIs monitor validity of safety case



Detailed SPI Definition

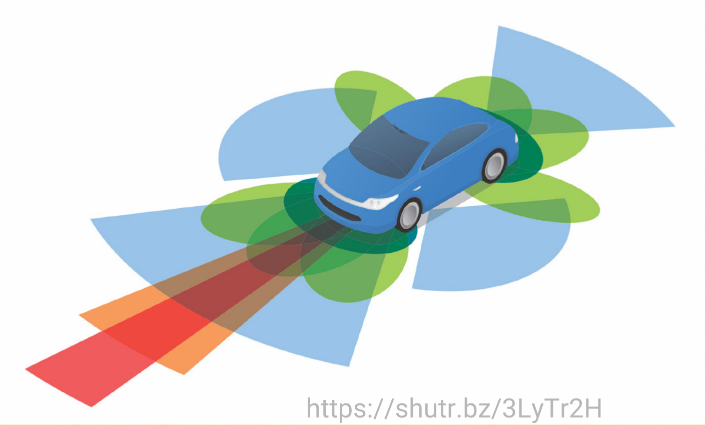
- An SPI is a metric supported by evidence that uses a threshold comparison to condition a safety case claim.
  - Metric: measurement of performance, design quality, process quality, operational procedure conformance, etc.
  - Threshold: acceptance test on metric value
    - Often statistical (e.g., fewer than X events per billion miles)
  - Evidence: data used to compute the metric
  - Condition a claim: threshold violation falsifies a specific claim
    - Argument for claim is (potentially) proven false by SPI
  - Definition ties the metric directly to the safety argument
- SPI violation: part of a safety case has been falsified

Sketch of an AV Safety Argument

Carnegie Mellon University

AV is safe enough to deploy because:

- We've followed industry safety standards & strong safety culture
- Known hazards have been mitigated
  - Residual risk is acceptable at system level
- Arrival rate of unknowns is low
  - Incidents which do not trigger runtime safing have low consequence
- Safety case has good SPI coverage
- SPIs usually detect unknowns without an actual crash
  - System is fixed to mitigate unknowns before likely reoccurrence



https://shutr.bz/3LyTr2H

➔ Idea: bootstrap on surprise arrival rates & SPI improvement

Conclusions

- Bootstrap testing is an appealing, but bad idea
  - Pure miles – safety is just a hope
  - Slow rolling – risk reduction is not safety
  - More complex approaches:
    - Maybe(?) saves the very last testing iteration if playing the odds on "probably perfect"
- Driver-out "safety testing" is unsafe
  - Keep driver in until safe enough to deploy
- Perhaps SPI bootstrapping can help
  - Bootstrap the safety case, not testing miles

https://shutr.bz/38cKv4u