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
The Race To Autonomy Started With Miles

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

AV Industry
Original Plan:

Phase 1

Phase 2

Phase 3

Collect
Road Test
Miles

Profit

SouthPark S2 Ep 17
ADS Technology has come to be: Sold Based on Safety

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?”

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

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

- 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
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
AV Industry Improved Plan:

**Phase 1**

Collect Road Test & Simulation Miles

**Phase 2**

Claim Safety

**Phase 3**

Profit
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
Hypothetically: 10K miles with safety driver
- Zero safety driver interventions
- 95% confidence MTBF$_{\text{crash}}$ > 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?
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
Example: 100K miles testing with safety driver
- Zero safety driver interventions
- 95% confidence MTBF_{crash} > 33380 miles
  - (Note: automotive often does about 70% confidence)

Do 100 miles of testing with no safety driver
- \[ R(t) = e^{-\lambda t} \text{ for } \lambda = 1/33380, t=100 \implies 99.7\% \text{ no crash} \]

Now you have 100,100 miles with no crash
- 95% confidence MTBF_{crash} > 33414 miles
- Notice that 33,414 > 33,380 ... hmmm ... interesting!
  - We can bootstrap our way to proving safety!

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 $\Rightarrow$ driverless testing

Iteratively longer test cycles
- Test for X miles based on crash probability
- Each step yields bigger MTBF
- Next step can be $X + \delta$ miles due to larger MTBF
  ... math, math, math ...

Lather. Rinse. Repeat.
- Prove you are safer than a human driver

$\Rightarrow$ $\text{Profit}$$\text{Profit}$
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

- Probability of crashes is high
  - At 200M MTBF_{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?)*
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

- 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

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

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 multiple concurrent 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
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
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
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
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

➡️ 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

© 2022 Philip Koopman