CITY UNIVERSITY
LONDON

Trustworthy quantitative arguments for the safety of AVs: challenges and some modest proposals

Lorenzo Strigini
Centre for Software Reliability
City, University of London

Outline

- the problem, its components, difficulties
  - why quantitative assessment
  - merits of alternative approaches
  - difficulties of detailed modelling
- ways forward with analysis of test / operational data
  - "conservative Bayes" approaches
  - focusing on risk in operation, "bootstrapping" confidence
- tentative conclusions

Why quantitative, probabilistic assessment

- knowing how hard it is to get it right, many scoff at request for numbers: probabilities, expected numbers (of accidents, of fatalities)
- "don't make up numbers! Just invest in hazard analysis, good design, V&V"

wrong

- that investment *is of course* necessary
- but some requirements *are inevitably* quantitative
  - "kill fewer than x extra people per year", "improve road safety"

  .... you need to check on a rational basis whether that investment is likely to achieve (have achieved) the target
- especially for novel, complex systems!

Premises for this talk

The main difficulty is not random hardware faults, EMI etc

- good fault-tolerant design will cut down their contribution to a small enough level, *which we can trust* to be that low
  - much of the reasoning can assume independence between basic unwanted events

The main concern is "systematic" failures:

- due to software/design bugs, imperfect machine learning
- they happen with high probability on *specific* situations

- Although level 3 poses specific problems, most of the discussion will apply to all levels

Well-known difficulties in assessing autonomous vehicles

- vital components use machine learning: this typically undermine the very basis of "usual" verification methods
- sufficient safety is achieved through many redundancies (diverse sensors and processing; independent safety monitors)

  they reduce risk, but it is hard to quantify by how much

- system boundaries:
  - early steps to autonomy make drivers the last line of defence:
    + effectiveness harder to assess than for inanimate systems, and likely to evolve (decreasing)
    + a car navigates a society of other cars, pedestrians, cyclists, horses, ...
      how to assess risk from interactions?
      among heterogeneous, learning components in evolving ecosystem?
  - bad people will attack your computer-controlled cars

I'll ignore these latter problems. Let's walk before trying to run

What is the quantitative requirement? A range of opinions

- we'd like AVs to be no more dangerous than human drivers
  - average drivers (which includes the drunk and the crazed)?
  - some object, and propose a target 10-100 times better
  - ... or somewhat better: Kalra and Groves * estimate that introducing soon AVs that shave off 10% of current fatalities would save more lives over 30 years than waiting for AVs that save 90%

  (note: the public may dislike the risk *transfer* and *uncontrollability)*

- we'll take as reference "just as good on average": of the order of 1 fatality per 100 million miles driven, 10^{-8} fatalities/mile

- hard to demonstrate!

* N. Kalra and D.G. Groves "The Enemy of Good - Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles", RAND Corporation 2017

Usual ways of quantifying (predicting) risk / safety

... range between two extreme approaches

- bottom-up, clear box:
  - from detailed understanding of your design and its components
  - your probability of accident is some function of many parameters that describe these details.
  - great for insight to drive design, not that trustworthy for prediction with complex systems
  - requires accurate knowledge about too many things

- black box:
  operate your system, count how frequently it has failures / hazardous behaviour / accidents:
  all do this
  - good "proof of the pudding" empirical, end-to-end
  - but perhaps cannot afford driving many hundred million miles *before* you start selling
  - apart from its practical difficulties (monitoring)

Note on fault tolerance, diversity... *vs* clear-box approach

- Diverse sensors feeding into similar/diverse processing; separate safety systems (monitor/safe response); ...
- All clearly useful, essential!
- Determining *how much* they give us is hard

"Systematic" failures, due to software/design bugs, imperfect machine learning

- happen with high probability on *specific* situations ("failure regions" in the space of stimuli X states)
- for the various subsystems in a fault-tolerant system
  - we don't know the failure regions
  - we don't know how much the failure regions of different subsystems overlap
  - we don't know how often *those* stimuli randomly arise *that* strike those regions and overlaps

(we have studied this for a long time and developed some ways of helping. See www.csr.city.ac.uk/diversity)

Safety subsystems / monitor / guards

- separate and independent "safety monitors" (detecting hazardous situations and responding) are useful for safety
  - consensus opinion: cf debate yesterday, various standards and industry documents
- simple ones may be "perfect" : no systematic *false negative* failures
  - this is not certain: it depends on reasoning that may have mistakes
- the more complex the environment, the less likely is perfection
  - due to errors, necessary trade-offs with false alarms
  - in some cases, we can reason using the probability of the monitor being perfect to support some *conservative* argument/claim (see e.g. [Littlewood *et al, 2011-13-17]*
- in general the actual safety gain from the safety monitor depends on *which* hazardous states the primary control system allows/generates

Let us turn to the black box measuring approach

- detailed modelling hits some serious limitations, so we consider just looking at success (or not) in operating a vehicle (road testing, or "real" use)

- for example, we may want to support statements like "for a desired goal that this system do not cause accidents at a rate greater than [...] per mile driven; *after observing 0 accidents in [...] miles* in road testing, we have [...]% confidence that the goal is satisfied"

The 100 million miles problem

- after a car drove – say – 1 million miles without fatalities
- how do we know whether it would kill less than one person per 100 million miles? (is it just one every **10** million?)
(if you *had* fatalities, the answer is easy)
  - Kalra & Paddock at RAND reported*:
95% confidence in ~10^{-8} probab. of fatality/mile) would require 275 million miles of test driving
(12.5 years of continuous driving for 100 vehicles at 25 miles/hour)
- operational testing alone cannot give confidence of safety over longer future operation
- not news **
- implication: you need to consider all the evidence you know before you start driving ...

  and even then, it may be very hard

* "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?(Transp. Research Part A: Policy and Practice 94 (2016) 182–193)

** Littlewood & Strigini, "Validation of ultra-high dependability for software-based systems, Comm. of the ACM 36 (1993) 69–80, http://openaccess.city.ac.uk/1251

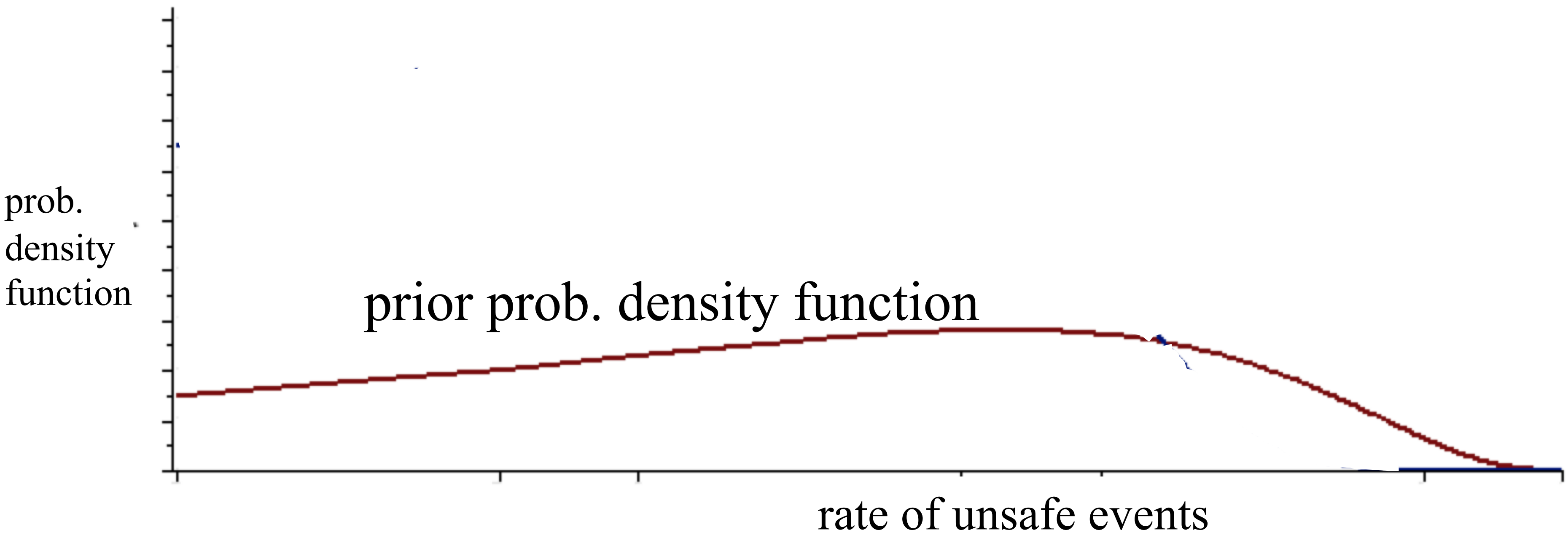
therefore...

- to account for what we knew *before* the road testing
- combining all the evidence in a rational way
- we apply *Bayesian inference*, a standard method for these goals

- to learn *how much* confidence is then justified about future operation
- and *identify gaps* that need to be filled by appropriate evidence

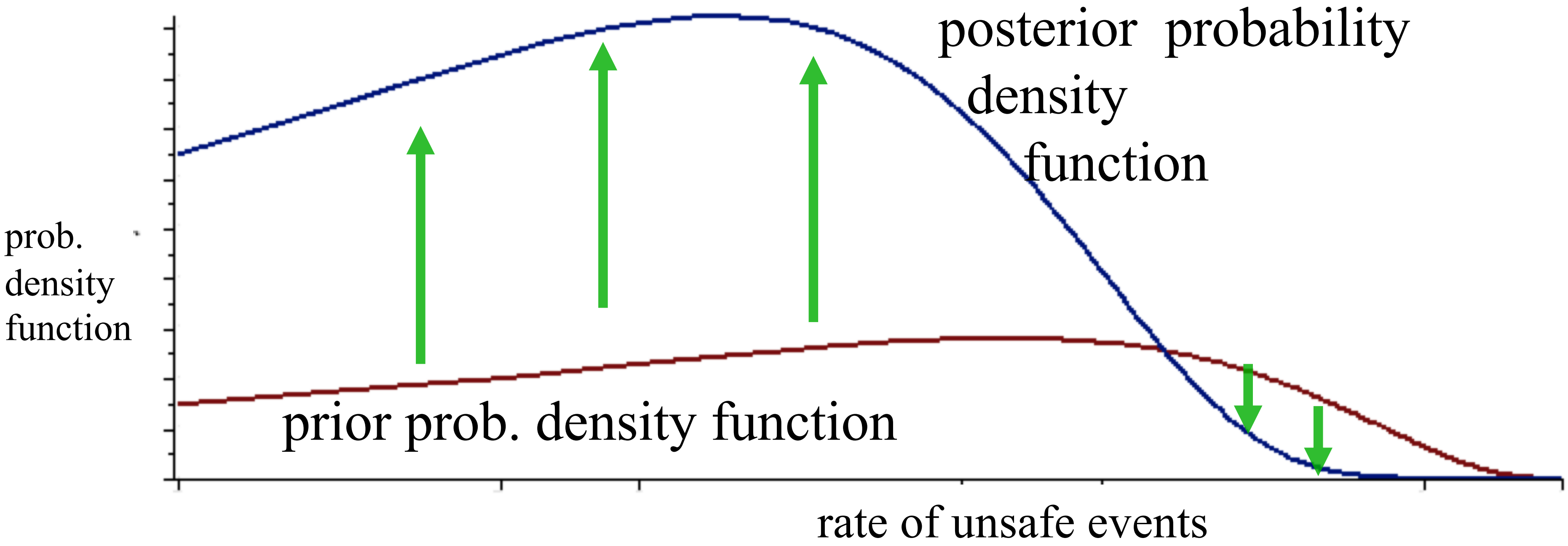
The Bayesian approach, in brief

- we see the unknown rate of failures, or accidents etc as a random variable, with a probability distribution
- development and verification support belief in a distribution for this variable (*prior distribution)*



The Bayesian approach, in brief

- we represent the rate of failures, or accidents etc as a random variable, with a probability distribution
- development and verification support belief in a distribution for this variable (*prior distribution)*
- then, observing the driving with zero/few unsafe events *changes it* ("posterior distribution")
- increasing confidence in *low* rates of unsafe events



A difficulty, and our "*conservative* Bayesian inference"

In Bayesian reasoning, the prior distribution

- is a crucial input
- to represent what we have reason to believe *before* obtaining new evidence (like road testing)
- based on quality of development, design precautions, verification activities, ...
- all important evidence, but *hard to translate* into a mathematical distribution

- common advice: use standard mathematical functions
  - ... the engineers are asked to pretend they know more than they do
  - which may produce seriously optimistic errors

- in CBI we take the opposite approach
- state less information, just what you have *really a basis (argument) for believing*, and ...
- ... we will give you the worst-case implications:
  what you can claim *conservatively*, given those *actual* prior beliefs

Example [Zhao *et al., 2019*]

Suppose

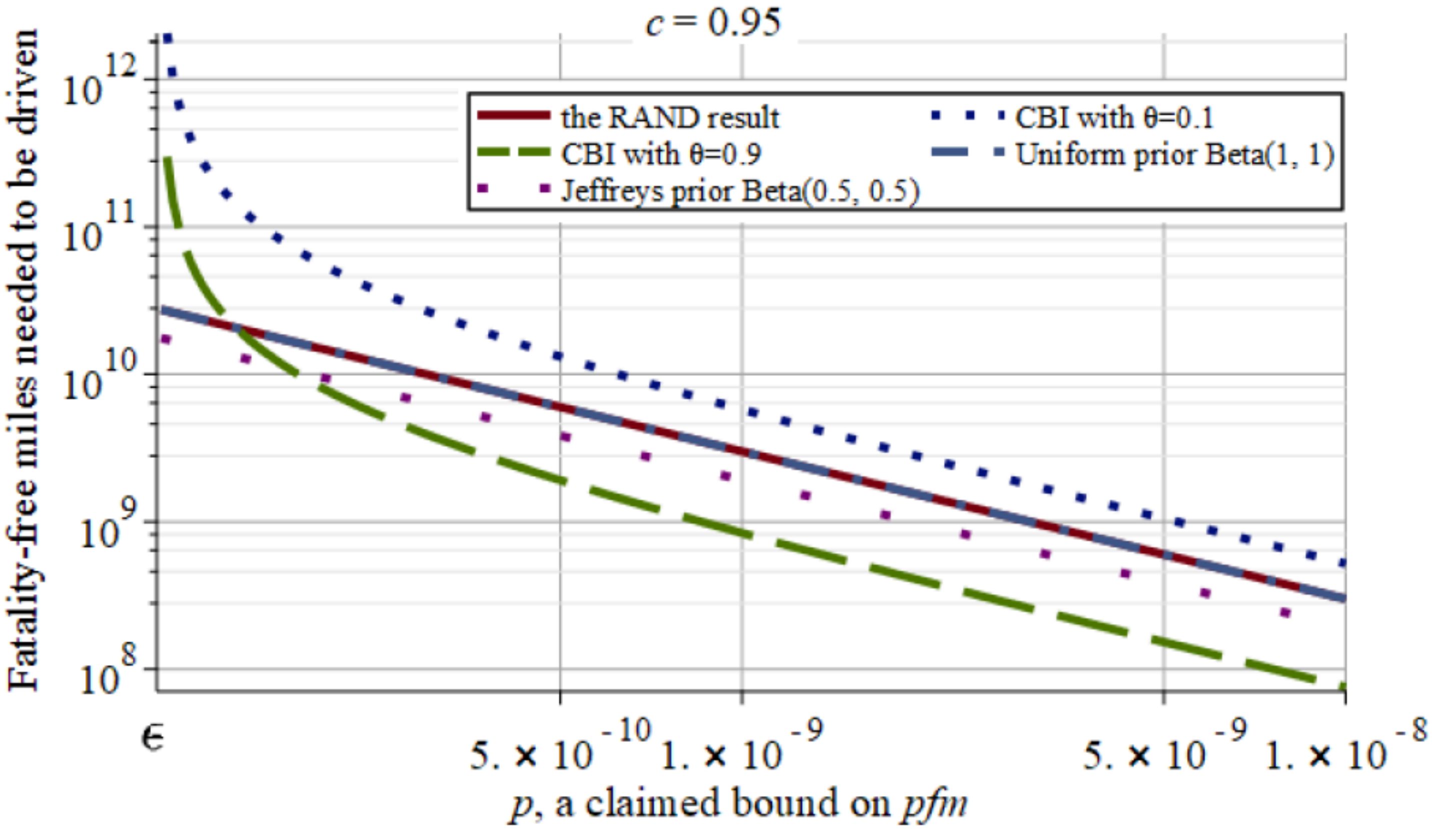
- a requirement for confidence c in "probability of fatality per mile" (*pfm)* better than a stated bound p

- design, quality of development, verification steps, historical experience give prior confidence θ that a goal "*pfm* is no more than ε" is achieved, where $\varepsilon < p$
- there is a lower bound p_l on the *pfm* considered feasible

- these bound the set of prior distributions that are possible
- so, after seeing n miles without fatalities, we can find how much confidence c , ***at least**,* can be had in *pfm* $\leq p$

Autonomous vehicles and CBI

Kalra et al paper "Driving to safety" ("RAND") vs CBI (from ISSRE 2019 paper):



- extreme claims can still be unaffordable to prove
  - but we can show how much *can* be claimed and the contribution of the other evidence: prior confidence θ of achieving the objective **does matter**
- we also addressed other questions from the "Driving to safety" paper
  - e.g.: if an accident does occur, how much accident-free driving would suffice to restore *justified* confidence?

summary ... what do we gain?

- ***probability*** that real risk ≤ target value as ***function of*** *prior* confidence *and* fatality-free test miles
  - e.g.: given 90% prior confidence θ of achieving *pfm* goal ε

    (based e.g. on simple safety guards with strong assurance, simulation testing, ...)
  - the bound *p* is demonstrated at 95% probability with ***one fourth*** the fatality-free miles driven needed to achieve 95% confidence in Rand study
  - but with only 10% prior confidence θ, *more* miles needed than in Rand study

- highlights the small print: ***sub-arguments*** required, e.g.
  - reliability arguments for the safety monitors used
  - if machine learning is allowed after deployment, arguments that it does not *reduce* effectiveness of safety guards
  - arguments for validity of results despite evolution of the driving environment

    (*cf* discussion in [Zhao et al, 2020])

What is missing?

These models assume a stationary world

- the system does not change
  - but actually manufacturers keep updating their systems

- the environment does not change
  - but it *will*
  - periodic changes (day-night, summer-winter), static differences (cities, climates, cultures), *trends* like increasing penetration of AVs

  - to account for this, we'd want to understand the kind of changes... difficult
  - some options we are exploring
    - + predictions that are robust to change (e.g. [Bishop]
    - + monitoring the operational profile for change and adjust predictions [Pietrantuono *et al* 2020]
    - + consider those simple changes that we understand, e.g. *improvements* (less harsh environment or safer system) [Zhao et al 2020]

Some changes are "probably for the better"

Example

- I used the vehicle for a long time, no accidents...
- I upload an upgrade, intended to make it safer...
- New vehicle!
  - must I consider it as having zero experience? All that operation proves nothing?
  - it seems crazy!! But *how much* does it prove?

or

- you intentionally tested in demanding environments (real / simulated)
  - so that you could deploy "with confidence" in a more benign environment
  - *how much* confidence should you derive from that "stress testing"?

- we can study this as a function of the confidence in improvement (or doubt about it) [Zhao *et al* 2020]

A better viewpoint: probability of failure in operation

so far we have seen that

- we can take into account knowledge prior to road testing
- there are gains
- but to overcome the paucity of testing compared to your extreme requirements, you need very strong claims *before* it – not commonly believable *as of now*

Let's switch viewpoint. What if

instead of demanding 10^{-8} or 10^{-10}

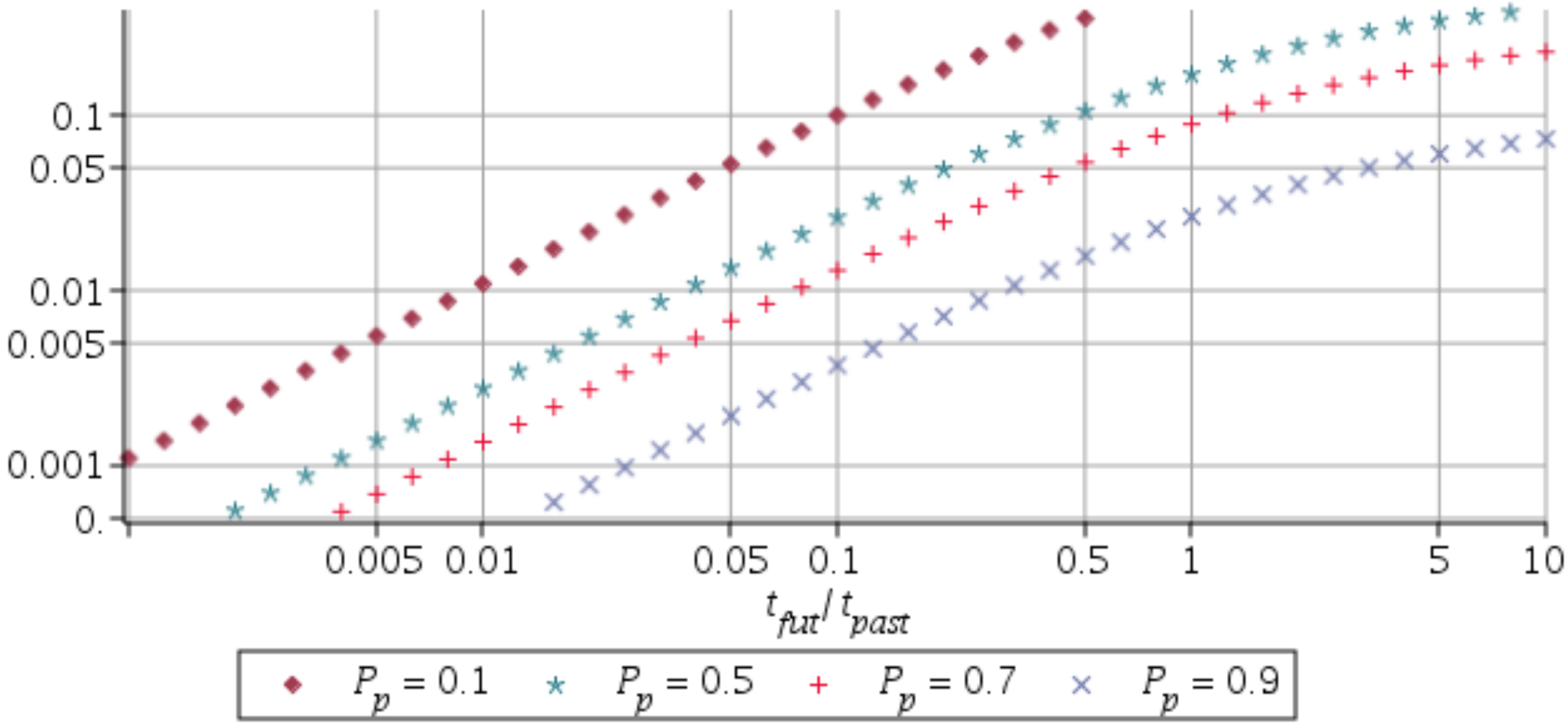
we simply ask:

how confident can we be in having no (few enough) fatalities in a reasonable period of future operation?

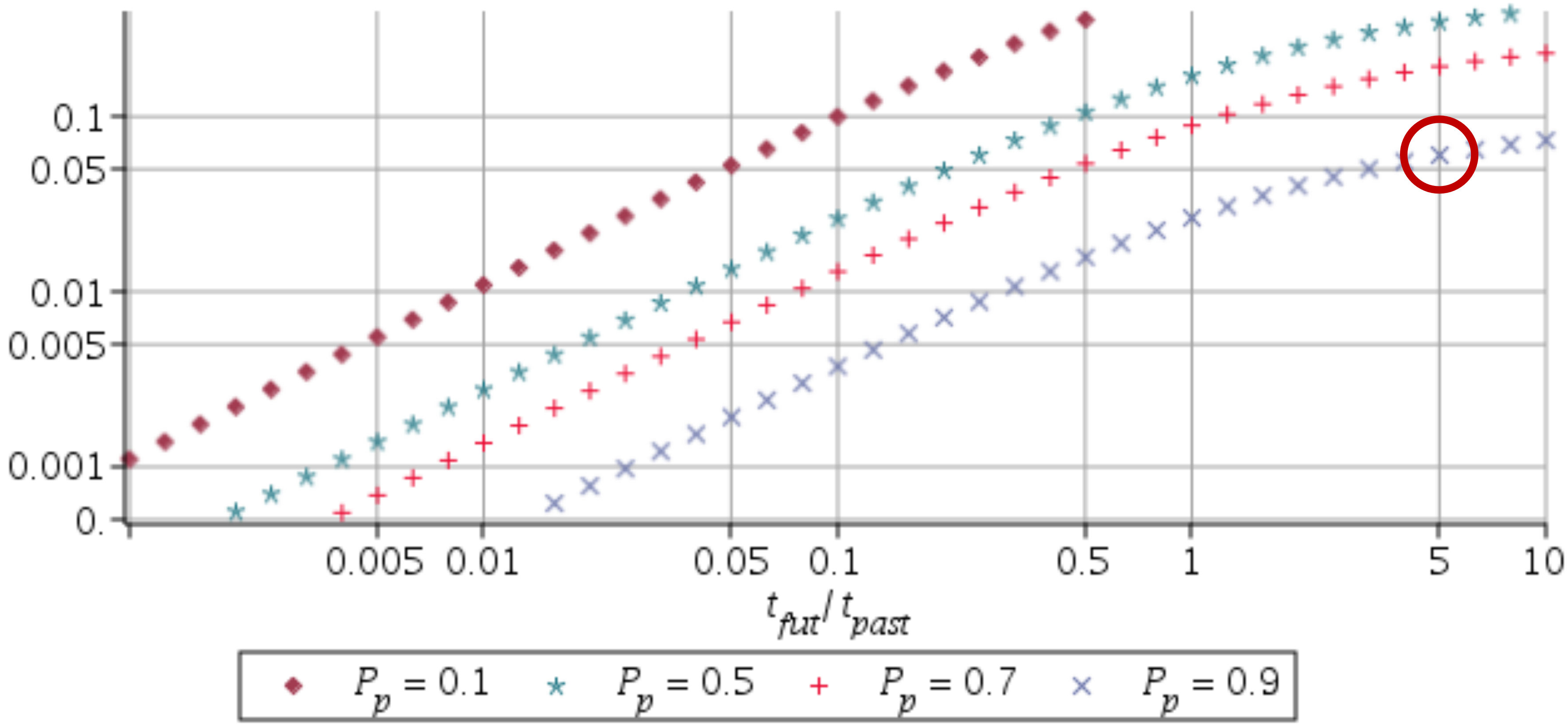
"How many mishaps" is the *real* measure of interest, after all

What we can we demonstrate about risk in operation?

- e.g. with a *prior* probability that your mishaps of interest *are* very rare by construction
- with an amount observed safe operation for a
- you forecast a *small enough* probability of mishaps over some future multiple of that amount



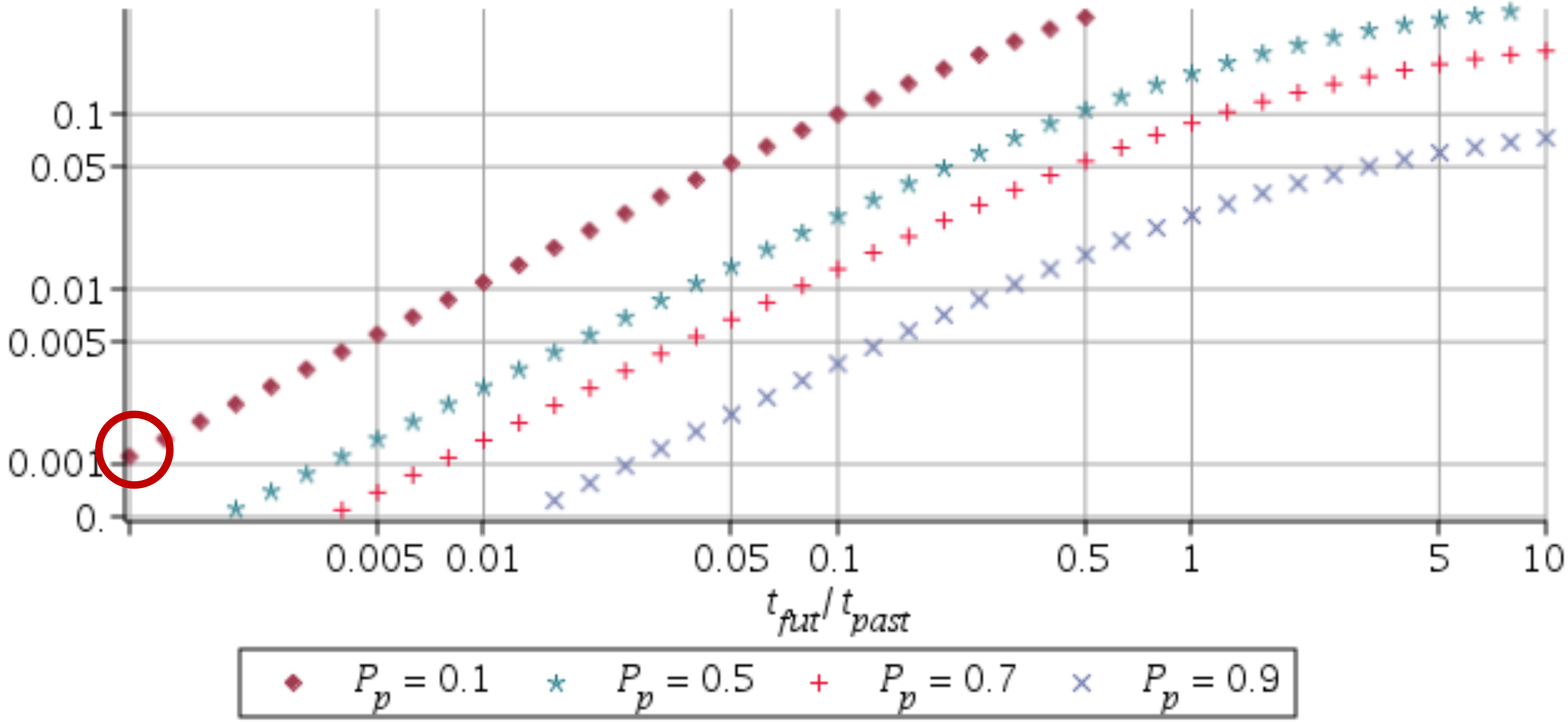
That means, for instance...



- can trust better than 94.5% probability of having no mishaps in an amount of future operation 5 times the amount of observed mishap-free operation ***if*** you have 90% prior confidence that you achieved a *much better* probability
- start with strong a priori arguments, guard against the surprise that they may be wrong

You can bootstrap your confidence by operating your systems while collecting more evidence

That meant, for instance...



- If you have driven 6.5 M miles without fatalities and seek assurance about the next 65,000 miles, 10% prior confidence that you achieved much better gives you 99.9% confidence of no mishaps

"Bootstrapping" of confidence

- suppose you start with operating, e.g., 1 vehicle for 1 year
- and at the end of the year you achieved sufficient confidence in 0 mishaps for, e.g., t_{fut}/t_{past}=5 more vehicle-years of operation

  (5 vehicles for another year or 1 for 5 years)
- after that, if all goes well, you accumulated 6 mishap-free vehicles-year: y

  can confidently run 6*5=30 more vehicle-years
- you can support constant confidence in an exponentially growing fleet
- or when fleet growth less than exponential, the accumulated experience increases your confidence and/or your time horizon

What does this "bootstrapping of confidence" give?

- a strong guarantee ("10^{-8}" or better) for the whole lifetime of a model fleet cannot be had
- but we can reason whether the risk *accepted* by operating the vehicle is *acceptable*
- allows decisions that limit risk to the public
- the vendor remains exposed – as now – to the risk of being badly wrong: "grounding", recalls

- presupposes good practices like extensive monitoring of operation, and uses their results

- it resembles the approach taken now!
- But the mathematics allows us to assess the *right* confidence to be had, given what we know or believe

Steps for application

- These methods support useful broad-brush reasoning
- Steps for use with specific industries/vehicles include
  - identifying local knowledge that supports other forms of prior beliefs
    - and extend the CBI theorems to include them
  - discuss the arguments/evidence supporting the required assumptions ("subclaims")
  - detailing the links of this quantitative reasoning to a safety case
  - all this involves use of existing practice of analyses, data collection
    - adapting the argument to match the evidence actually collected
    - or the evidence collection to help the assurance arguments
  - include *relevant* "safety indicator measures"
    - e.g., reliable counts of demands on safety monitors and their response?
  - potentially evolving a composite argument from sub-arguments
    - for subsystems
    - for regimes of operation, ODDs

Conclusions?

- Given that quantitative assessment is hard for
  - new systems
  - requirement of *high confidence* in *extreme safety*, *early on*
- Formal mathematics detects fallacies but also gives *directions for improvement*
  - focusing on shortish term "deploy or not?" decisions seems useful, even for supporting *longer term* operation
  - we demonstrated methods that seem promising and practical to extend
- The formal statistical methods have two advantages
  - they allow verification of sound reasoning
  - impose explicit statement of assumptions and the burden to argue them
- Regarding AVs now, what we have *suggests*
  - ability to argue for future operation by small increments
  - usefulness of work on supporting strong confidence prior to operational testing

Thank you...

Questions, comments?

Some references

Conservative Bayesian inference

Strigini, L. and Povyakalo, A. A. (2013). Software fault-freeness and reliability predictions. Proc *SAFECOMP 2013.* https://openaccess.city.ac.uk/id/eprint/2457/

Littlewood *et al*. (2019). On Reliability Assessment When a Software-based System Is Replaced by a Thought-to-be-Better One. Reliability Engineering & System Safety. https://openaccess.city.ac.uk/id/eprint/23238/

Zhao, X. *et al.* (2019). Assessing the Safety and Reliability of Autonomous Vehicles from Road Testing. ISSRE 2019. https://openaccess.city.ac.uk/id/eprint/22872/

Zhao, X. *et al*. (2020). Assessing Safety-Critical Systems from Operational Testing: A Study on Autonomous Vehicles. Information and Software Technology, 106393.. doi: 10.1016/j.infsof.2020.106393 , https://openaccess.city.ac.uk/id/eprint/24779/

Evolving operational profile:

Pietrantuono, R., Popov, P. T. and Russo, S. (2020). Reliability assessment of service-based software under operational profile uncertainty. Reliability Engineering & System Safety, 204, 107193.. doi: 10.1016/j.ress.2020.107193 https://openaccess.city.ac.uk/24816

Bishop, P. G. and Povyakalo, A. A. (2017). Deriving a frequentist conservative confidence bound for probability of failure per demand for systems with different operational and test profiles. Reliability Engineering & System Safety, 158, pp. 246-253. doi: 10.1016/j.ress.2016.08.019 https://openaccess.city.ac.uk/15248/

References, ctd

Primary-monitor systems:

Popov, P. T. and Strigini, L. (2010). Assessing Asymmetric Fault-Tolerant Software. ISSRE 2010. https://openaccess.city.ac.uk/id/eprint/277

Littlewood, B. and Rushby, J. (2011). Reasoning about the Reliability of Diverse Two-Channel Systems in which One Channel is "Possibly Perfect". IEEE Transactions on Software Engineering, doi: 10.1109/TSE.2011.80 https://openaccess.city.ac.uk/id/eprint/1069

Littlewood, B. and Povyakalo, A. A. (2013). Conservative reasoning about epistemic uncertainty for the probability of failure on demand of a 1-out-of-2 software-based system in which one channel is “possibly perfect”. IEEE Transactions on Software Engineering, 39(11), pp. 1521-1530. doi: 10.1109/TSE.2013.35
https://openaccess.city.ac.uk/id/eprint/2515

Zhao, X., Littlewood, B., Povyakalo, A. A., Strigini, L. and Wright, D. (2017). Modeling the probability of failure on demand (pfd) of a 1-out-of-2 system in which one channel is “quasi-perfect”. Reliability Engineering & System Safety, 158, pp. 230-245. doi: 10.1016/j.ress.2016.09.002
https://openaccess.city.ac.uk/id/eprint/15797