Bootstrapping confidence in future safety from past safe operation

Peter Bishop[1,2], Andrey Povyakalo[1], Lorenzo Strigini[1]
[1]City, University of London, London, U.K.
{P.Bishop, A.A.Povyakalo, L.Strigini }@city.ac.uk
[2] Adelard LLP, London, United Kingdom
pgb@adelard.com

***Abstract*—We propose an approach to managing the rolling out of a new system type so as to contain the risk of mishaps in its operation to an acceptable level, while using the evidence of safe operation to support confidence for extending the scale of operation. This cautious approach of "bootstrapping" confidence in the safety of a system is now widely applied to autonomous vehicle (AVs), our example. AVs are subject to extreme safety requirements; a major concern is the inability to give meaningful quantitative assurance of safety of an AV type, to the extent required by society, before it is used extensively.**

We exploit a previously published approach to achieving more moderate, but useful, assurance, e.g. about low enough probability of causing accidents in a limited period of operation; and demonstrate how this approach supports choice of production/deployment strategies, so as to manage the growth of a fleet of AVs for a given accepted level of risk. Via a formal mathematical description of "confidence bootstrapping", we show: (1) that it is a substantially sound approach in the right circumstances, and useful for deciding about the early deployment phase for a new system; (2) *how much* confidence can be rightly derived from such a "cautious deployment" approach, avoiding over-optimism; (3) under which conditions our sound formulas for future confidence are applicable; (4) thus, which analyses of the concrete situations, and/or constraints on practice, are needed in order to enjoy the advantages of provably correct confidence in adequate future safety over a definite time ("confidence horizon").

I. Introduction

It is common practice to start operation of a system on a limited scale, based on reasonable confidence that the system is safe enough for this, so as to collect more evidence to support this belief in its safety and thus justify larger-scale operation[1].

This paper studies how to base such decisions on a sound quantitative basis. It is motivated by the prominent current example of autonomous vehicles (AVs). A major concern for developers of autonomous vehicle, authorities responsible for authorising their use, and end users, is how to achieve confidence that they will be safe enough in operation. The safety objectives are inevitably, at least in part, quantitative: we do not want accidents to be too frequent. This might be stated as a specific maximum frequency of accidents or fatalities per mile, or a requirement that these frequencies be no greater than some fraction of the average for human drivers, or comparable to the better human drivers, etc.. It is important to have quantitative arguments that properly demonstrate what level of confidence we can have in these requirements being satisfied. Given the reliance of AVs on machine learning, arguments that they are safe enough hinge by necessity on empirical evidence, from test driving: first of simulated vehicles in simulated environments, then of real vehicles on the road and in traffic.

However, a major difficulty arises when spelling out these quantitative requirements for road vehicles on public roads: when these reasonable requirements are translated into numerical targets about, for instance, a low enough "probability of fatal accidents per mile driven", demonstrating that the requirement is satisfied by just operating the vehicles and collecting statistical evidence is impractical. The cost and practical difficulties in building such confidence statistically seem insurmountable [1]–[3].

An alternative, more feasible quantitative claim can be made about the probability of safety over a finite amount of future operation [4]. For example, the claim could be that for a fleet of cars there is a 90% probability that it will not suffer accidents due to its self-driving functions, over the next year of operation. Such a statement could be used by vendors or regulators to decide whether to allow operation for one year, or by insurers or investors to assess the risk or insurance premium for a company developing such an advanced system.

This paper's main contribution is to extend the approach in [4] so as to support rational deployment decisions about the rolling out of a new system type over the lifetime of the fleet.

Such an ability is especially needed when, as with AVs, much uncertainty surrounds the degree of safety in operation until demonstrated by operation itself. The practice of cautious small-scale operation to gain some confidence in safety, followed by gradual increases in the scale of operation (e.g. number of vehicles in operation) is a natural and common response to uncertainty: observing safe small-scale operation naturally gives confidence that these systems will keep op-

This work was supported in part by ICRI-SAVe, the Intel Collaborative Research Institute on Safe Automated Vehicles.

[1]We call this process informally "confidence bootstrapping": the initial, limited amount of confidence allows the limited amount of operation that bolsters confidence, allowing more operation, and so on, in a positive feedback loop. We only use the word "bootstrapping" as a common English language metaphor; this paper does not deal with the technical statistical concept of "bootstrapping" as a form of resampling.

erating safely. But *how much* confidence should one derive from – say – 1000 vehicle-kms of safe operation? Should it be, for instance, 90% confidence that the vehicle will not cause accidents for the next 100 vehicle-kms? 1000 kms? How quickly can we safely extend pilot operation? When will we attain adequate confidence for commercial operation? Intuitive judgement is notoriously unreliable for this kind of problem.

The novelty of this paper is in proposing a mathematically sound method for these decisions. The method depends on assessing the probability of experiencing no mishaps over a stated future amount of operation; or equivalently, if we want the probability of a future mishap to be lower than a certain acceptable bound, assessing how much more the vehicle can operate before reaching that bound. This calculations yield a "confidence horizon", until which the required risk level is not exceeded. We show, in precise mathematical terms, how experiencing mishap-free operation extends this horizon. But for decision making like authorising a period of operation of a fleet of vehicles, the confidence horizon should be translated into calendar time, e.g. months or years. We show how this translation is affected by the changes in the size of the operational fleet, with examples of: pilot operation of a constant-size test fleet; linearly increasing fleet size; increases in production rate; and whole-lifetime evolution including progressive retirement of the type. For a growing fleet, the confidence horizon in calendar time grows more slowly than for a constant-size fleet. Our method allows a decision maker to control the fleet's growth rate so as to keep the associated risk within desired limits. An additional merit of this method, compared to quantitative safety assessment in common use, is reduces reliance on arbitrary mathematical assumptions, thanks to its "conservative Bayesian inference" approach [4].

Many different kinds of AVs are in development or in use, creating many different scenarios for this problem of gaining confidence in their safety. We will refer to two types of AVs that are at opposite extremes of the range, from the viewpoint of how easy or hard it is to accumulate confidence in their future safety from progressively extended periods of operation. At one end of the spectrum, we will call "A-type" vehicles those "SAE level 5" AVs meant to be sold to many millions of ordinary consumers and to transport them over public roads, in complex and quite unpredictable environments. At the opposite end, we will call "Z-type" those AVs that are deployed in small numbers to perform well-understood, limited tasks in constrained and protected environments, e.g., a self-moving crane in a factory or a heavy truck in an open-face mine, or a self-driving metro train. Z-type AVs may still have extreme safety requirements (e.g. if they transport dangerous material in a chemical plant); but the simpler environment and less pressure for high performance and fast evolution reduce difficulties in both their development and their assessment.

In the following Section II we address some relevant aspects of the state of the art and previous research on quantitative safety assessment for software and for AVs; in Section III, we recall the method presented in [4] (III-A,III-B), and introduce our mathematical approach (III-C), and in Section IV we present some numerical examples; Section V discusses implications, limitations and direct extensions of the method, followed by conclusions and future work in Section VI.

II. Background and related work

A. Statistics for assessing extreme reliability/safety levels

Statistical assessment of reliability or safety is a long-established discipline with a colossal corpus of literature.

The basic principles are intuitively simple: logging failures and a "time" variable (time in operation, number of demands or similar variables depending on the kind of system) and estimating a failure rate as the ratio between the two measures. Complications like variable failure rates to account for physical ageing of a system have long been covered in textbooks and engineering handbooks.

Extensions to *software* reliability have also been well-established for a long time [5]. These address predictions about probability of system failures due to design, rather than random physical, faults. In other words, scenarios in which whether a system fails or not at a certain point may be a deterministic function of the stimuli to which it has been subjected until then, but probabilistic reasoning is needed to accounts for uncertainty about both which sequences of stimuli will actually occur and which ones would actually cause failure (since the system has unknown "design faults", often called "bugs").

With software receiving increasingly critical roles in systems with safety relevance, attention grew in the 1980s and 1990s on two areas that concern the present work:

1) software in theory might never fail. By definition, correct (in other words, fault-free) software could not fail. Advances in formal proof of correctness seemed to promise the practical possibility of fault-free critical software. At the same time, discovery of critical bugs in critical software brought into relief the fact that complex systems fail due to design faults, not just physical failures of components; and that demonstrably fault-free software was generally unattainable, in practice, at the time. Safety critical industries and their regulators had to deal with the implications;
2) reliability requirements on the software might be "ultra-high" (as dubbed in a paper co-authored by one of us [2]): so stringent that they could not be convincingly shown to be satisfied by just observing a system operate, without failures, through a feasible period of operational testing. That is, requirements such that the system could be worse than required and yet good enough to pass any affordable amount of operational testing.

The issue of "ultra-high" reliability requirements[2] was highlighted in the 1990s in civil aviation. The requirement for "catastrophic failures" to be "so unlikely that they are not

[2]We use the word "reliability" for consistency with relevant literature, although the main concern in this paper is safety. Having requirements about safety rather than reliability does not change the terms of the problem. It means requirements about the improbability of certain system failures - accidents – rather than of any system failure.

anticipated to occur during the entire operational life of [...] one type" [6]) translated into quantitative terms as a probability of 10^{-9} or less for one flight hour. This requirement still appears in the current version of these rules by the U.S. Federal Aviation Administration and the European Union Aviation Safety Agency (EASA). Various authors argued that strong confidence of such a requirement being satisfied, before an aircraft could be certified for commercial operation, was unfeasible [2], [3]: (1) software could not be convincingly demonstrated to be fault-free; (2) demonstrating from statistical evidence that the faults, potentially present, would not cause catastrophic failures too frequently, would require unaffordable amounts of operational testing.

A minor difficulty here was more complex mathematics. Reasonably unreliable systems will sometimes be seen to fail. If they fail for instance n_f times in n_d demands, n_f/n_d is a practical and intuitive estimator of their probability of failure per demand. More advanced statistical calculations are only needed if we require confidence bounds around this estimate. But if a system never failed, the n_f/n_d estimator would be 0, and thus likely optimistic. To avoid this, it is appropriate to estimate instead one-sided confidence bounds: statements of interest will have the form "this system's *pfd* under this usage is less than *this bound*, with *this confidence*", a less intuitive statement than "it is about n_f/n_d".

The problem remains of demonstrating satisfaction of requirements of very low *pfd* or failure rate. The solutions differ in different industries. Civil aviation regulators in the West choose simply *not* to require a demonstration: for design faults – in software or other complex design artefacts – they do not require statistical or probabilistic arguments, but just demonstration that high quality practices were followed; compliance with guidelines like (currently) the joint EUROCAE ED-12C/RTCA DO-178C documents is sufficient for certification.

A different (we would say more realistic) approach is taken, e.g., by the U.K. Office for Nuclear Regulation (ONR), which asks [7] for safety arguments to include two "legs", "Production Excellence" and "Independent confidence–building measures", the latter including statistical testing. It must be said that this approach has been usually applied to systems, within a nuclear power plant, that individually do not have "ultra-high" requirements: they contribute to the plant's safety as parts of redundant and diverse layers of defence.

For ultra-high requirements, the statistics of failures (or, rather, absence of failures) and operation will not give enough confidence that a target like the above 10^{-9} has been attained, before commercial operation begins. It is necessary to combine this statistical evidence with what was known before the statistical observation, which can be done rigorously via a Bayesian approach, as recommended by researchers [2], [8] and already accepted in regulatory practice, regarding non-design faults, in various industrial sectors [9].

In Bayesian inference, "prior beliefs", described as probabilities of events and probability distributions of random variables, are updated in view of the evidence observed – e.g., failure-free operation. E.g., given a prior probability of catastrophic failures being rare enough, observing a period of operation without such failures will yield a higher "posterior" probability.

A major difficulty is specifying these prior probabilities and probability distributions. A textbook application of Bayesian inference to assessing a system's *pfd* from observing failure-free operation would require an assessor to translate the evidence prior to operational testing – e.g., precautions taken in software development, forms of verification applied, qualification of staff in a project – into a complete probability distribution for the *pfd* of the system considered. To simplify this daunting task, shortcuts are often suggested – even in regulatory documents – like choosing intuitively plausible or mathematically convenient functions. This approach risks implicitly forcing the assessors to build flawed mathematical arguments: internally correct but based on premises – the prior distributions – that they cannot justify.

Concern with the sensitivity of conclusions to hard-to-justify priors has produced extensive research in "robust Bayesian inference" [10]. In this direction, we and our colleagues have developed an approach of "conservative Bayesian inference" (CBI): we only require an assessor to specify *constraints* on the prior distribution; and then prove what is the *most conservative* conclusion that can be drawn from combining a prior in the set thus defined with the observed new evidence (records of operation and failures or lack thereof) [11]–[13].

B. Assessment, certification, licensing of autonomous vehicles

AVs have attracted huge investments and political interest, creating intense pressure to authorise their operation. However, how to demonstrate that they are safe enough for general use on public highways is a classical "ultra-high reliability" problem. The safety level of human drivers is indicated by statistics like, in the U.S., a rate of about 1 fatality in 10^8 miles driven. Simple classical confidence calculations have been published [1] to show how expensive it would be to gain from road testing strong confidence that an AV type equals even this average performance (an average among all drivers, good or bad).

E.g., given a safety target that an AV type must deliver no more fatalities/mile than this average U.S. driver, the classical analysis requires 275 million fatality-free miles [1] in order to achieve 95% confidence in the target being attained. A Bayesian (CBI) analysis [13], [14] only confirms how hard the goal is. For instance, aiming at 95% posterior probability for the same target being achieved, this amount of road testing could be reduced to 69 million fatality-free miles, if one had 90% prior confidence that the AVs are two orders of magnitude safer than the 10^{-8} target.

The ultra-high reliability problem is akin to that faced in aviation, but a major difference is that key functions in AVs are performed by machine learning-based systems, as opposed to conventional software. This undermines mature verification methods that exist for the latter, to check that the software (when operating on non-faulty hardware) would satisfy specific requirements.

The AV industry has naturally embraced use of independent safety monitors (see e.g. [15], [16]) to detect and resolve hazardous situations. However, this does not solve the problem of demonstrating that a type A AV will be as safe as required. Firstly, hazardous situations may be subtle to detect and hard to resolve; more importantly, it does not change the cost of statistical demonstration of safety, because the effectiveness of a monitor for a particular AV cannot be estimated independently of the specific AV [2], [17], [18].

Two relevant research reports from the RAND Corporation [19], [20] address the interconnected issues of interpreting societal demand for "sufficient safety", demonstrating its attainment via measurement, and risk communication between vendors, government and the public. These reports have popularised in the debate the terms "lagging indicators" and "leading indicators" for safety. "Lagging indicators" designates aspects of actual, measured safety of operation, like numbers and severity of accidents. "Leading indicators" are proxy measures that suggest how safe the vehicle will be. They may range from measures of events in operation (e.g. near misses or "disengagements" – handovers to a human driver) to characteristics of the design or the development process. The major concern at this stage is identifying how such indicators may be related to actual safety, as measured in e.g. accidents per vehicle-km.

Koopman et al [21], [22] advocate an ongoing safety assurance approach where leading and lagging indicators are used to estimate achieved safety and also provide opportunities for safety improvement. This strategy is also incorporated into the ANSI/UL 4600 safety assurance guidance for AVs [23] where the safety case is no longer static, but is updated using feedback from actual operation [3].

The work we present here fits in this pattern of demonstrating safety via realistic accumulation of evidence and its integration into a safety argument. The novelty of this paper is that we offer a mathematically formal way of representing such incremental safety arguments.

III. Basic model and results

We show the reasoning for the case that the process of mishaps occurring[4] can be modelled as "Bernoulli trials": the system is subjected to a series of *demands*, and mishaps on different demands are independent events with the same probability ("probability of failure per demand", *pfd*). Bernoulli processes are a common model for failure processes [9], [26]. One could call "demand" a single trip; or, as in [1], [13], driving a mile or a kilometre. If mishaps are rare, the Bernoulli process should be a tolerable approximation of reality, despite successive demands, if defined this way, not being independent [27].

We use Bayesian inference: the uncertain values of interest are considered random variables; a "prior" distribution that represents the state of knowledge and uncertainty about their values before new evidence is observed, and is updated on the basis of this new evidence. In the present context, the new evidence is that some amount of operation was completed with no mishaps. In particular, we use here the approach of "conservative Bayesian inference" (CBI for brevity) to avoid the need to specify the prior distribution in full, and instead depend on specifying only some characteristics of it that one *can* trust to have good reasons for believing.

We use here a specific version of CBI [4] where, based on

- a very limited partial description of the prior distribution;
- the observation of a period of mishap-free operation;

a rigorous lower bound can be obtained on the probability of the system continuing to operate without mishap for a specific amount of future operation.

We follow this approach because, as we argued [4]:

- the real question of interest is whether the probability of having *any* mishap *over a certain period of operation* (e.g., the whole lifetime of the system; or the next year of operation) is acceptably low (in other cases, whether the probability of *too many* mishaps is); estimating a *pfd* is just a mathematical detail of how one can answer this question;
- for many systems, there is strong confidence, *before* we start operating them, that they are reasonably safe: that for some short period of operation, they are unlikely to cause mishaps. Indeed, the decision to start even small-scale operation of such systems would not happen if the decision makers (vendors, operators, and/or regulators) did not have this confidence. The source of the confidence is typically in that these systems were developed following good quality practice, the code was extensively verified, etc. How much confidence this evidence *should* really generate is a separate problem.

This prior confidence is affected by uncertainty, of course. Expensively developed and verified systems have been put into operation despite failure modes with astonishing high probabilities: historically, e.g., the early Space Shuttle software had a probability of 1 in 67 per flight of failing to start properly [28]; the initial version of the Ariane 5's control system, a 100% probability of destroying the rocket [29]. In general, a thorough interrogation of what one knows should indicate some estimate of a probability (less than 100%) that the *pfd* is acceptably low. We note that this is a Bayesian probability: the system's *pfd*, given the way the system will be used and the world around it, is a specific number, but is unknown. This probability of acceptable *pfd* describes our "epistemic"

[3]This is similar in principle to the mandatory aviation incident reporting schemes, where severe incidents and accidents have to be reported and analysed so that safety can be improved [24], [25]. This approach has been operating for many decades and accident rates have fallen continuously over this time.

[4]We use the umbrella term "mishap" for the negative event about which one wants to give predictions. Some may want to reason about probability of a deadly accident as in [1], [13], or of any accident, or of *potential* accidents, e.g. violations of an assigned safety envelope, or even failures that analyses reveal *could* cause an accident. In safety, one would typically wish to use this last definition (*potential* accident); but to have data for statistical inference, it may be necessary to reason just about serious accidents, those that are reported and will appear in logs.
We retain instead the word "failure" in "probability of failure on demand", as a standard term in the probabilistic treatment of safety.

uncertainty about what the *pfd*'s real value is; this uncertainty is a crucial factor in our decision whether to take the "gamble" of operating the system.

In Bayesian terms, the unknown *pfd* is a random variable – we will call it Q for brevity – and our uncertainty about it is described by a probability distribution, say a probability density function, $f_Q(q)$.

A. Argument based on probability of pfd = 0

The simplest version of our conservative form of reasoning [4] applies for systems so simple that one has some substantial confidence that they are *free* from safety-relevant faults, i.e., that their *pfd* is zero.[5]

Of course one never has 100% confidence of this. Thus a parameter of this kind of argument is the probability P_p of the statement "*pfd* $= 0$" being true. So, with probability P_p the system has zero probability of mishap per demand: we could operate it for an infinite amount of time and a mishap would never happen (remember that we are talking about mishaps due to the design of the self-driving function, not due to physical failures, or to fatally reckless behaviour of other drivers, "acts of God", etc.: with this restrictive focus, if there is no fault in the system, it will never fail so as to generate a mishap). With probability $(1 - P_p)$, thus, the system does have design faults and will – sooner or later – fail: experience a mishap. The Bayesian description of the problem is that the *pfd* may have any value, with different probabilities: using the notation we used earlier [4], we call Q the unknown *pfd*. Q is a random variable, with a *prior* probability density function $f_Q(q)$, which in this case is.

$$f_Q(q) = P_p\,\delta(q) + (1 - P_p)\,f_{Qn}(q) \tag{1}$$

where $\delta(q)$ is Dirac's delta function[6] and $f_{Qn}(q)$ is the probability density function (*pdf*) for the system *pfd* conditional on *pfd*> 0.

If the future period of operation for which we wish to know that the system is safe enough is made up of T_{fut} demands, the probability of surviving it without mishap (a *reliability* function) is:

$$\begin{aligned} R(T_{fut}) &= \int_0^1 (1-q)^{T_{fut}} f_Q(q)\,dq \\ &= \int_0^1 (1-q)^{T_{fut}} \left(P_p\,\delta(q) + (1-P_p)\,f_{Qn}(q)\right)dq \\ &= \int_0^1 (1-q)^{T_{fut}} P_p\,\delta(q)\,dq \\ &\qquad + \int_0^1 (1-q)^{T_{fut}} (1-P_p)\,f_{Qn}(q))dq \\ &\qquad = P_p + (1-P_p)\int_{0+}^1 (1-q)^{T_{fut}} f_{Q_N}(q)\,dq \end{aligned} \tag{2}$$

[5]This case is convenient for the purpose of presentation: its results yield simpler plots than the general case, which we introduce in section V-A2, of imperfect confidence in the *pfd* being less than a small non-zero value.

[6]Defined as: $\delta(x) = 0$ for any $x \neq 0$ and $\int_{-\infty}^{+\infty} \delta(x)dx = 1$.

One can observe that the reliability will stay higher than P_p for any duration of future operation; the risk we take in the "gamble" of operating this system is a weighted sum between zero risk (if indeed *pfd* $= 0$) and the risk due the potential defects in the system. With the common approach, instead, of only seeking a confidence bound on *pfd*, and using it for a conservative prediction of future probability of mishap, one ignores the fact that with a certain probability those defects are absent and thus pose no risk in operation. This would be needlessly pessimistic. Once we start operating the system, operation experience feeds new evidence about how much we should trust the hypothesis that *pfd* $= 0$.

B. How mishap-free operation extends the confidence horizon

After T_{past} independent demands without mishaps, the *posterior* probability of T_{fut} further mishap-free demands is (from Bayes' theorem):

$$R(T_{fut}|T_{past}) = \frac{\int_0^1 (1-q)^{T_{past}+T_{fut}} f_Q(q)\,dq}{\int_0^1 (1-q)^{T_{past}} f_Q(q)\,dq} \tag{3}$$

As mentioned earlier, so that the results can be trusted not to be an artefact of unjustified details of the prior distribution, we applied a "conservative Bayes" method [4]: given a value of P_p, and assuming we do not know the rest of the prior distribution, $f_{Q_N}(q)$, this method obtains the most pessimistic posterior reliability, for any pair $\{T_{past},\ T_{fut}\}$, compatible with that P_p [4]. Thus our confidence in future mishap-free operation is the most pessimistic given these inputs. Clearly one could study "confidence bootstrapping" with any other choice of prior distribution. The advantage of this approach is the guaranteed conservatism with respect to uncertainties about the detailed prior distribution of the *pfd*.

Figure 1 exemplifies the results of the conservative inference. In this scenario where the prior distribution is described only via its P_p value, this probability is a function of the ratio T_{fut}/T_{past}, so these bi-dimensional plots are sufficient to describe, for any values of P_p and T_{past}, the probability of having any mishaps over a certain T_{fut}. Our confidence level is 1 minus that value: the probability of *no* mishap happening in the next T_{fut} amount of operation. So, given the P_p that one trusts, and the confidence level one desires in future operation without mishaps, one can see how much such future operation can be. We will call the value of T_{fut} for which this confidence holds the "confidence horizon" that the experience T_{past} supports for that required confidence level.

E.g., for a desired confidence 95% that no mishaps will occur (i.e., $1 - 0.95 = 0.05$ on the vertical axis in Fig. 1), the plot shows that if I have $P_p = 0.9$ the confidence horizon is about 5 times T_{past} (the exact value of the probability for $T_{fut} = 5\ T_{past}$ is 0.94). It is useful to visualise the confidence horizon as a multiple of $Tpast$. We will write $T_{hor} = kT_{past}$, so in this case $k = 5$; we will use this value in our examples below.

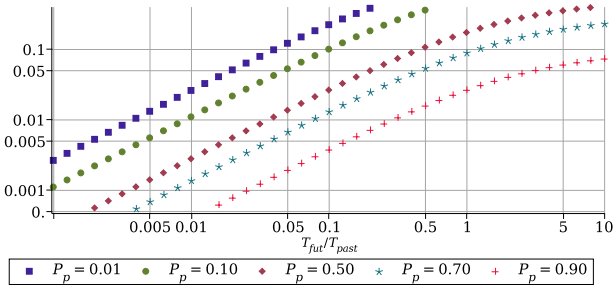

Fig. 1: Probability of one or more mishaps (i.e., 1 minus confidence that no mishap will occur) over T_{fut} future demands, supported by given P_p and T_{past} amount of past mishap-free operation (from [4]).

C. Bootstrapping confidence; confidence horizon measured in amount of operation and in calendar time

Our scenario is now as follows. A required probability of no mishaps occurring has been chosen. For this required confidence level, the current P_p value gives a "confidence extension coefficient" k that determines the confidence horizon, $T_{hor} = kT_{past}$. On this basis, a fleet of vehicles is allowed to operate. If the fleet size is constant, and the way it is operated (environments in which they run, times of operation, kinds of trips, how many trips of each kind per day or month) does not change, then the formulas for the confidence horizon, although written in terms of future number of demands, could just as well be written in terms of time on the road, or calendar time: each one of these measures is proportional to the others. If we use lowercase 't' for calendar time, while keeping uppercase 'T' for amounts of operation (e.g., vehicle-days, or total trips) we can write that $t \propto T$. We define an average rate of operation, o_{avg} (measured in demands/vehicle/year, i.e., kms/vehicle/year, or trips/vehicle/year, etc, depending on the definition of "demand" adopted for measurement), and assume that every vehicle operates at this rate – an approximation that will become good enough once enough vehicles are on the road (we will discuss its limitations later). So, at any time, we have clear indication of what our confidence horizon is, and this is growing longer; alternatively, if our main interest is the probability of avoiding mishaps over a fixed horizon, this probability is growing constantly.

But suppose that the vehicle is in production: the fleet size is increasing. A given confidence horizon $T_{hor} = k{*}T_{past}$, stated e.g. as a number of vehicle-months, has to be spread over operation of a growing number of vehicles: in calendar time, our confidence extension coefficient will be less than k. In other words, $t_{hor}/t_{past} < T_{hor}/T_{past}$. The confidence horizon is a necessary time buffer for decision makers: could the growth of fleet size reduce it too much? Will it progressively shrink to nothing? We study these questions below.

1) Case 1: constant number of vehicles: At any time t, with a constant number n of vehicles, $T_{past} = t_{past} \times n \times o_{avg}$ and $t_{hor} = kt_{past}$.

As time passes, the future confidence horizon expands, proportionally to the time elapsed. This may be a quite satisfactory situation, e.g. if I am in early pilot operation with a fixed number of vehicles, and want to run this pilot phase for e.g. one year, then, in our example of *k*=5, after 2 months in operation I will have sufficient confidence of mishap-free operation for the remaining 10 months; and this confidence will increase towards 100%, as is natural, as more time passes without mishaps; or if I am concerned about a fixed population (type Z vehicles, e.g. the set of autonomous bespoke heavy load vehicles in a certain mine), after 1 year I will be confident enough for the next 5 years, and if the planned operating life is - say - 30 years, after 5 mishap-free years we will be confident enough of safe operation for the rest of the operating life.

While no mishaps occur, the confidence horizon expands constantly; if we were interested instead in absence of mishap for a constant period into the future, a constant t_{fut}, say one year, or 5 years, our confidence in this outcome would keep increasing. This will be true for all scenarios that we examine.

2) Case 2: linear growth of fleet in operation: Suppose now that after a pilot period of operation, a production line is activated that delivers vehicles at a constant rate r_v. Let us count time from this moment. The pilot period accumulated a total amount of operation (demands) T_p, so that we have "accumulated confidence" for a future amount of operation kT_p, considered sufficient to decide to start production. As the fleet grows, we can soon ignore this initial "capital" of confidence: it becomes negligible, compared to that accumulated through operation of the mass-produced vehicles. We assume for the sake of simplicity that every vehicle starts operation as soon as produced, and all operate at the same rate o_{avg}.

We want to calculate the confidence horizon after a time t_{past} has elapsed from the start of mass production (and mass operation) of the AV type. The amount of operation until then will be

$$T_{past} = \int_0^t (t-\tau)\, r_v\, o_{avg}\, \mathrm{d}\tau = \frac{r_v o_{avg} t_{past}^2}{2} \tag{4}$$

giving a confidence horizon

$$T_{hor} = k\, T_{past} \tag{5}$$

To translate this into calendar time, we consider that

$$T_{past} + T_{hor} = (k+1)\, T_{past} \tag{6}$$

hence, substituting from (4):

$$T_{past} + T_{hor} = \frac{(k+1)\, r_v o_{avg} t_{past}^2}{2} \tag{7}$$

and observing that, by analogy with (4):

$$T_{past} + T_{hor} = \frac{r_v o_{avg}\, (t_{past} + t_{hor})^2}{2} \tag{8}$$

one sees from equating (7) and (8) that:

$$t_{hor} = \left(\sqrt{k+1} - 1\right) t_{past} \tag{9}$$

e.g., for our example of *k*=5, $t_{hor} \approx 1.45\, t_{past}$.

That is, after 1 year of operation, we have gained the required confidence in future mishap-free operation for approximately another year and a half.

We can call the coefficient on the right hand side

$$k_{linear} = \sqrt{k+1} - 1 \tag{10}$$

We note that:

- this result does not depend on the production rate, but only on it being constant;
- though always smaller than k, for a high enough k this coefficient k_{linear}, of the order of $\sqrt{(k)}$, can still be high, e.g. for $k = 10$, $k_{linear} \approx 2.32$.

A warning is appropriate: while k is a ratio between operation times in vehicle-months, which applies no matter how the fleet size evolves over time, k_{linear} is a ratio between interval lengths in calendar time, and only applies if vehicles enter operation at a constant rate. For each law of fleet size change as a function of time, a different law of evolution of the confidence horizon in calendar time applies. In practice any such evolution law can be easily studied by simulation, as we do for the examples we present later in this paper.

Readers may observe that the row for $k = 100$ in Table I contains P_p values that are only marginally smaller than the desired confidence of mishap free operation for a time $k\,t_{past}$. Indeed, if we want assurance of safety for an amount of future operation much greater than we have observed, our confidence in it needs to rely mostly on the confidence we had prior to that observation [2]. The way of reasoning we propose does not create evidence out of thin air; it does instead indicate clearly what can be trusted based on the evidence itself.

The good news from this table is that for moderate amounts of future operation, the required prior confidence P_p is markedly lower than the posterior confidence sought. That is, with this approach a decision maker can trust that operation for a finite future interval is safe, and during that interval, additional positive evidence can be collected, if indeed the system is as safe as intended. One can manage incremental deployment, the only sure way of building up, gradually, evidence of sufficient safety, with an understanding of the amount of risk it entails, and can thus contain that risk at a level considered acceptable.

3) Increasing the rate of growth of the fleet in operation: If the production rate increases, the confidence horizon will decrease accordingly (or confidence will decrease if we want the same horizon). E.g., if, in the previous example, after five years of operation we suddenly double the production rate, by opening another identical production line, the amount of past operation becomes, for any $t_{past} = t > 5$, the sum of that due to vehicles from the older factory and vehicles from the new one:

$$T_{past} = \frac{r_v o_{avg} t^2}{2} + \frac{r_v o_{avg} (t-5)^2}{2} \tag{11}$$

		Desired P(no mishap)		
		90%	95%	99%
k	k_{linear}	P_p	P_p	P_p
100	9.05	0.89	0.95	0.99
10	2.32	0.86	0.93	0.99
5	1.45	0.84	0.92	0.98
3	1	0.8	0.9	0.98
2	0.732	0.77	0.88	0.97
1	0.414	0.67	0.82	0.96
0.6	0.265	0.56	0.75	0.94
0.5	0.225	0.52	0.72	0.93
0.2	0.0954	0.26	0.5	0.86
0.04	0.0198	0.0092	0.071	0.53

TABLE I: P_p values required for different levels of confidence in further mishap-free operation for the static fleet case (k) and the linear fleet increase case (k_{linear}). Each P_p column shows the value of $P_p = P(pfd = 0)$ that would support the level of posterior confidence listed above the column, for the k value on that row.

and

$$T_{past} + T_{hor} = \frac{r_v o_{avg} (t + t_{hor})^2}{2} + \frac{r_v o_{avg} (t + t_{hor} - 5)^2}{2} \tag{12}$$

but it is also true that

$$T_{past} + T_{hor} = (k+1)\, T_{past} \tag{13}$$

So, substituting (11) and (12) in (13) and solving, we obtain, for our example with $k = 5$:

$$t_{hor} = -t + 5/2 + \frac{\sqrt{24\,t^2 - 120\,t + 275}}{2} \tag{14}$$

which at the time the production increases, has the value 5.8 (units of time): the confidence horizon for the combined output of the two factories is now substantially lower than that calculated without taking into account the new factory, which was 7.2.

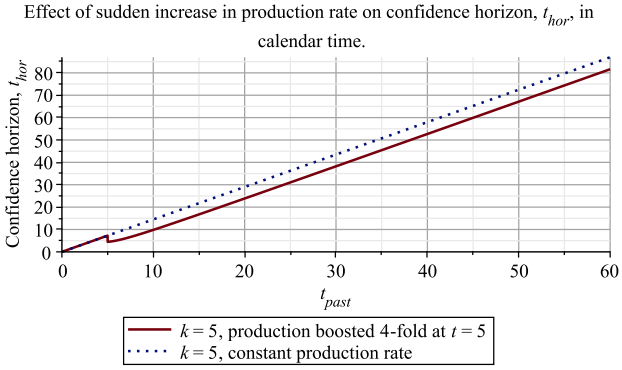

Fig. 2: Growth of the confidence horizon (for 95% confidence in no mishaps occurring), in the case of continuous growth of the fleet, and effects of step increase in production rate. After the increase at time $t = 5$, the confidence horizon drops from 7.2 to 4.5, but recovers to the previous value by time $t = 8$.

We skip for reasons of space the general solution for this problem, and show an example of these effects in Figs 2 and 3, for a greater increment (four-fold) in production rate and

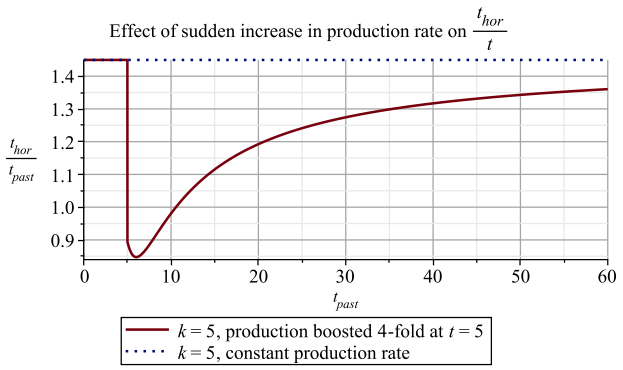

Fig. 3: Ratio between confidence horizon and t_{past}, constant with continuous growth of the fleet, and the dip caused by a step increase in production rate.

thus a larger dip in the confidence horizon. Note that (1) the effects of the dip tend to disappear, in relative terms (Fig. 3), in the long run; (2) if the increase in production rate is assumed known before it happens, this dip would start, gradually, earlier than shown, as soon as the confidence horizon for those AVs that were produced before the increase reaches the moment at which the increase will take place.

4) Maturity and retirement : If production continues for a long time, older vehicles will start to be retired, hence the confidence extension coefficient will gradually rise back from k_{linear}, to reach k when a constant-fleet size equilibrium is reached (new vehicles are added at the same rate at which older ones are retired). When production is eventually scaled down and ends, as the fleet size dwindles the confidence horizon will increase even faster, but by then it will generally already extend beyond the end of the life of the fleet.

IV. Examples

1) Example: type Z vehicle, for restricted environments: Figs. 4 and 5 show a hypothetical history of fleet size evolution for a type-Z AV, and the corresponding evolution of the confidence horizon. Very satisfactorily, the desired 95% confidence of no mishaps in all future operation is reached early in the fleet lifetime.

But it is more interesting to observe the evolution of the confidence horizon in the early times, when operational experience is still limited and decisions are needed whether wider-scale operation should be considered safe enough to

Fig. 4: A possible fleet evolution pattern for type Z vehicle. Pilot operation proceeds with 5 vehicles for 24 months. Series production starts at month 24, with 10 vehicles starting operation per month. Production ends after 16 years. Vehicles are retired at age 25 years.

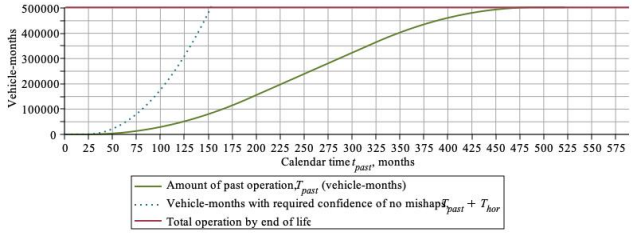

Fig. 5: Evolution of confidence horizon for type Z fleet of Fig. 4. The confidence horizon, in vehicle-months of operation, is the vertical distance between the bottom solid line (accumulated vehicle-months of operation) and the dotted line. The prior confidence P_p is assumed 0.92, giving a "confidence extension coefficient" $k = 5$.

proceed. Fig. 6 zooms in on these early days. The confidence horizon grows quickly while operating a constant-size fleet but dips when linear increase starts. If this dip were thought

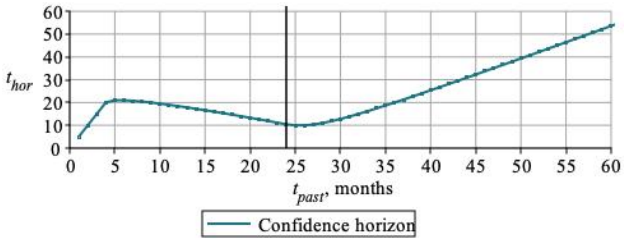

Fig. 6: Evolution of confidence horizon in the early days for the type Z fleet of Fig. 4. At month 24 (marked by the vertical line), new vehicles start operation at a rate of 10 vehicles/month. The future period for which we have 95% confidence of no mishaps (confidence horizon) thus drops to 10 months, before increasing again as mishap-free operation accumulates.

excessive, one can phase the production increment: e.g. in Fig. 7 at month 24, production is started at a rate of only 3 vehicles/month, then at month 36 production is increased to the desired 10 vehicles/month.

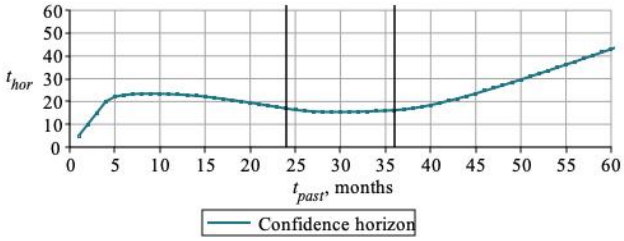

Fig. 7: Evolution of confidence horizon during early operation for the type Z fleet of Fig. 4, changed to reach the introduction rate of 10 vehicles/month in two stages: 3 vehicles/month at month 24, increased to 10 at month 36. This limits the dip in the confidence horizon to 15.4 months

2) Example: type A vehicle: For type A vehicles, we do not expect substantial prior confidence that *pfd* $= 0$. However, before starting test on public roads, a decision maker will require some confidence of acceptable probability of mishaps, obtained from e.g. verification of the machine

learning platform and of the safety monitors, lab testing of the machine learning components, extensive simulated road testing and testing of the vehicle on restricted-access test circuits, comparisons with results seen in other, similar systems, etc.

We discuss in section V-A how the method can be extended to deal with this different input. Here, instead, we only consider an approximation valid for a short initial amount of operation, during which a small bound on *pfd* has an effect on predicted probability of mishap approximately equivalent to that of having *pfd* $= 0$, and thus introduces a small error in the calculation of the confidence horizon described so far.

Suppose for instance that the evidence available, mentioned earlier, yields for the decision maker an estimated bound on *pfd* such as to give accident rate no more than $10^{-8}\,\mathrm{km}^{-1}$. This is near the average for human drivers, and thus not quite adequate for type A vehicles. Furthermore, given the limitations of the evidence (e.g. uncertainty on the realism of simulations), our decision maker sets a modest 80% confidence in this bound. If they require, for instance, 90% confidence in mishap-free operation in the test period, this yields, from Table I, $k = 3$.

This approximation is only useful for an early period of operation, presumably of test vehicles only. If we assume e.g., test vehicles running 10,000 kms/month, and given the aforementioned required 90% probability of no mishaps in this phase, the approximation will be acceptable for the first 100 vehicle-months of operation: $10^{-8} \times 100 \times 10^4 = 0.01 \ll 1 - 0.9$.

Thus for instance, if operations starts with one test vehicle only, after one month of operation there will be confidence in operating $k = 3$ vehicles for another month with probability of mishap at most 90%, etc. If for instance the test fleet expansion is stopped after month 3, with 9 vehicles and $T_{past} = 13$ vehicle-months, the confidence horizon in vehicle- months will be $T_{hor} = 3 \times 13$. In calendar months, it will be another $t_{hor} = T_{hor}/9 = 4.3$ months of test operation.

In the next section, we discuss issues raised in this example: simple mathematical extensions and practical implications, e.g. for continuing validation activity.

V. Discussion

The discussion until now has assumed that the chosen mathematical model of the real world processes is correct. We believe that it would be correct in many situations with type-Z AVs (as we defined them in the introduction), and in much more restricted cases for type-A AVs. We discuss here the various limits to its applicability, what extra evidence or research would be needed to extend these limits, and general insight that can be derived from this study.

A. The probability of pfd $= 0$ versus "effective fault-freeness"

The form of conservative Bayesian inference we presented relies on a prior probability P_p of the system having 0 *pfd*, although our earlier paper [4] also outlined other forms. This one has advantages of simplicity when presenting "confidence bootstrapping" concepts, but it is not the only option. We now discuss these options in more detail.

1) Probability of pfd $= 0$ *:* A non-negligible probability that the system *pfd* is zero may be plausible for very simple systems, but not generally for the control functions of AVs, which tend to be complex and depend heavily on machine learning. However, practically all AV manufacturers embrace the use of "safety monitor" subsystems, which monitor the vehicle's situation, checking for any violation of preset safety constraints, and have authority for taking emergency action. These are potentially very simple: for instance, detecting a fixed obstacle via lidar and braking if it is approaching too quickly is a simple function (in principle). For type-Z vehicles, for which the environment is simple and safety is paramount, these safety subsystems could plausibly be so simple and well verified to have a high probability of 0 *pfd*, making our method applicable. P_p would never be 1, since subtle misunderstandings of hazards, or errors in verification, may happen, but it would be high enough to support useful levels of confidence in safety. We would not claim that this is possible for class-A vehicles, because: (1) the dangerous situations that can occur in traffic may be very complex and not amenable to simple detection or simple accident avoidance decisions; (2) manufacturers have to pursue a difficult balance between safety and performance (e.g., delivering passengers to destination in times that are comparable to human-driven vehicles). However, some manufacturer might try this route of making the safety monitor subsystems more thoroughly verifiable for the whole set of potential hazardous situations.

2) "Effective fault-freeness": If the system is complex, so that P_p is very small, high quality development and verification could still support a prior confidence about the *pfd* being, if not 0, at least smaller (better) than a very low bound, q_S. Some scenarios in which this could apply are discussed in our earlier paper [4], showing that the prior confidence has then the form

$$\int_0^{q_S} f_Q(q)\,dq = P_S \tag{15}$$

leading to a probability of operating over T_{fut} without mishap:

$$\begin{aligned} R(T_{fut}) &= \int_0^1 (1-q)^{T_{fut}} f_Q(q)\,dq \\ &= \int_0^{q_S} (1-q)^{T_{fut}} f_Q(q)\,dq + \int_{q_S}^1 (1-q)^{T_{fut}} f_Q(q)\,dq \\ &\geq P_S (1-q_S)^{T_{fut}} + \int_{q_S}^1 (1-q)^{T_{fut}} f_Q(q)\,dq \end{aligned} \tag{16}$$

Note that if q_S is such that over the time horizon of interest its contribution to risk is negligible (or, equivalently, if we limit our predictions to T_{fut} values for which this is true):

$$(1-q_S)^{T_{fut}} \approx 1 \tag{17}$$

then the right-hand side of equation (16) is approximated by (2) with negligible effect on the confidence in mishap-free operation over T_{fut}, and ability to improve confidence

as mishap-free operation accumulates. When $q_S = 0$, (16) reduces to (2).

This more general model offers a way forward. However, we have to discuss its limits and possible difficulties. For type-A AVs, condition (17) would probably apply initially only for very short T_{fut}. We can instead develop an exact worst-case prediction based on (15) and the bound in (16) and inference from mishap-free operation. Some extra difficulty arises, which may or may not be surmountable depending on the circumstances. The pessimism implied by using the bound in inequality (16) implies that the predicted probability of safe future operation would never exceed $(1 - q_S)^{T_{fut}}$: as though the true *pfd* were q_S. E.g., if one initially chose $q_S = 10^{-3}$, the inference would indicate that the true *pfd* is no better than 10^{-3}, even after seeing 10^6 or 10^9 mishap-free demands. This pessimistic prediction would be neither useful nor credible.

To avoid this, one needs to accumulate other evidence, in parallel with mishap-free operation, to establish a lower P_S or a less coarse description of the distribution $f_Q(q)$ for q close to zero. This should be possible. Indeed, AV manufacturers have ongoing programmes of accompanying road testing and operation of their AVs with continuing verification activities in the form of, e.g., much more extensive simulated driving, pursuing static verification of some safety properties, etc. These activities could progressively improve both P_S (increasing it) and q_S (reducing it). Last but not least, manufacturers would respond to both accidents, and violations of safe behaviour envelopes that do not cause accidents, with improvements. To what extent all these activities could allow this form of "confidence bootstrapping" to be be deployed, as the operation amounts involved increase, remains to be studied on the basis of the details of the activities and evidence about their efficacy.

B. Conditions for validity of the model

The "Bernoulli trials" model we used assumes that the *pfd* will not change – between the past from which we draw inferences and the future, or during future operation. This is only true with certainty if neither the vehicle, nor its mode of use, change. These conditions may well hold for type-Z AVs, meant for well-understood tasks in controlled environments (factories, mines), for which the traditional rule is followed of avoiding change, as far as possible, for critical functions of safety-critical systems. This rule indeed exists because change undermines the confidence that has been built through expensive verification work. For type-A AVs, meant for future mass sale to consumers, this condition does not seem to hold at the present state of aggressive development pace. The Bernoulli trial model would only apply over short periods of time between upgrades; although, in the future, maturity of designs and the need to satisfy authorities and consumers about safety might at some point make changes in self-driving functions much less frequent.

Given frequent upgrading of A-type AVs, can "confidence bootstrapping" work? A sufficient condition for this model still to help is to convincingly demonstrate that the changes do not reduce (or actually improve) safety. Then, one could use evidence of safe operation of previous versions of a vehicle as though it concerned the latest version, and only err in the direction of pessimism. How to formalise mathematically this kind of argument, taking into account the uncertainty that may affect them, has been studied for various scenarios [13], [30]–[32]. We need still to study how these results could be combined with those we present here.

These methods for accounting for "changes for the better" would also apply to changes of the environment of use. For type-A AVs using public roads, the environment will certainly be changing, if nothing else due to the increasing presence of AVs. This may well mean that the environment will gradually become more benign, allowing the method we presented to be extended along the lines of the papers cited above. Other ways to account for a changing environment are being studied, including monitoring the changes so as to update predictions accordingly [33] and making predictions robust by accepting extra conservatism [34].

In summary, some extensions to deal with changing environment and evolving AVs are available for specific scenarios, and suggest that research may deliver extensions to a broader range of scenarios.

Irrespective of evolution of the vehicle and the traffic environment, a reader may question whether mishaps on different demands (e.g., different trips) are indeed independent and equiprobable events. We mentioned (Sec. III) that if mishaps are rare enough events, assuming independence is an acceptable approximation. More extensive discussion and references are available in [13], [30]. An essential requirement is that the demands from whose outcomes we infer probabilities must have the same probability distribution (of characteristics that may affect whether a mishap occurs – like position and velocity of vehicles, temperature, visibility etc.) as the future demands to which those probabilities are meant to apply. Thus in any statistical learning (not just with our method) care is required, e.g., not to use operation limited to Summer as a suitable sample for prediction about winter conditions. Concerns like this will especially apply while the scale of operations is still so small that fair sampling cannot be ensured: e.g., a single test vehicle cannot practically sample all the regions in which the future fleet is meant to operate.

We should note that even when assumptions do not strictly hold and the numerical results are not trusted, we expect the model to help qualitative understanding of the effects to be expected from expansion of operation.

C. Long-term operation and mishaps

We have described a way of "bootstrapping confidence" on the basis of operating an AV without any mishaps. With this approach, even a single mishap would completely undermine confidence in future operation. Other studies have addressed, for different contexts, how confidence would grow again with subsequent mishap-free operation has been addressed [14], [35]. Regarding our current context, it is reasonable to demand that an acceptably safe AV should not suffer mishaps in early operation: if the target is of the order of less than one serious

accident in 10^8, or more, kms, one such mishap in the first – say – 10^5 kms would be a very strong alarm signal. It is thus reasonable for our method to respond with a total loss of confidence.

So, the method we have described is suitable for a type-Z AV (expected to have no mishaps over its whole lifetime) or the early period of adoption of a type-A AV.

It becomes inadequate later, when a type-A AV, even if acceptably safe, would inevitably start to suffer mishaps. The fact is that even if an AV is acceptably safe, in that its *pfd* is as required (say, 10^{-r}), as it goes through extensive operation it will still reach a stage when mishaps are bound to happen. After the first 10^{r-1} demands, there is a non-negligible 10% probability of having had at least 1 mishap. After $(0.7\ 10^r)$ demands, this probability becomes 50%. Thus a mishap in the first 10^{r-1} demands poses the question whether this AV has acceptable *pfd*$\leq 10^{-r}$, but hit that 10% probability of an early mishap, or instead possibly has an unacceptable *pfd*=10^{-r+1}, or even worse *pfd*, but hit a lucky, but not wildly improbable, mishap-free run. To answer these questions, the method we have presented is as yet inadequate. We plan to study extensions in this direction. In the much longer run (after a number of demands much larger than 10^r), the problem disappears: if the rate of mishaps is stable, it is easily assessed with standard statistical methods, to confirm that the vehicle is acceptably safe (or that it is not).

Accidents, or even near misses, that appear due to defects of a system, are likely to trigger attempts to diagnose and remove the defects that cause them, thus – it is hoped – improving the *pfd* in subsequent operation. It has been shown [36] that if these events were due to defects that will be removed, with some probability, following any accidents or near misses, the total number of accidents over the system lifetime will also be bounded. This bound is affected by the number of defects, and by how effective the safety monitors and other monitoring arrangements are in causing a defect to be detected before it causes an accident. With machine learning systems, just as with mature, very complex conventional systems, it cannot be taken for granted that attempts at removing defects will reduce the *pfd*. One cannot exclude that such "repairs" will not only be subject to a law of diminishing returns, but possibly just cause the system *pfd* to oscillate up and down without a definite decreasing trend. However, extending the above model to describe these situations may bring additional insight for this kind of scenarios.

D. Risk criteria

We have reasoned so far about scenarios in which the main concern is whether there will be any mishap (due to the self-driving functions) in operation. This is one of the possible concerns, appropriate, it would seem, for a public authority that only wants to authorise operation if there is high confidence that the system will not cause harm; or for a manufacturer fearing that any accident during the early life of a new model might turn the public off buying it: accidents would risk all the expected returns on the massive investment made in development. However, there are other possible viewpoints. For instance, accidents may entail compensation costs after each accident; or recalls after each accident. That is, in some circumstances the dominant concern may not be *whether* there will be accidents, but *how many*. This requires an extension to the model presented here, using other published results [12].

Last, while we have focused on how the confidence horizon progressively extends into the future, there will be situations in which the main concern is absence of mishaps over a fixed term into the future. For these circumstances, the model we have presented is very satisfactory, as it shows this confidence increasing steadily as experience of mishap-free operation accumulates.

VI. Conclusions

"Bootstrapping" confidence in a system, by operating it on a gradually increasing scale, with the next increase being deemed safe enough on the basis of safe operation in the previous increments, is common practice. With systems that are based on machine learning, not amenable to some of the standard ways of gaining confidence in a software-based system before operation, "bootstrapping" is even more important.

But to trust that we are deriving correct decisions from this encouraging evidence, we need to know *how much* our confidence should grow thanks to mishap-free operation. We have presented a formal mathematical way for a sound (conservative) derivation of how much confidence one can really have on the basis of a certain amount of mishap-free operation.

This method applies well for managing growing scale of operation in some scenarios: specifically, giving assurance in the short term for early operation of what we called type-A AVs, those for which assurance is hardest; and probably whole-life assurance for type-Z ones (corresponding to some "SAE level 4" forms of autonomy), those built for extremely safe operation in constrained and controlled environments.

For other scenarios, our present study is an encouraging indication that similar solutions may be developed, although they may require new research: developing the mathematical methods, but also demonstrating empirically whether the assumptions of these methods hold in practice (within some acceptable degree of approximation), or devising variations in design practices (e.g. regarding safety monitors), or in data collection practices, that would allow the assumptions to be proved valid and thus grant the benefit of a sounder basis for confidence in future safety. Case studies are also needed on choosing model parameter values of the models on the basis of evidence prior to operation. There are thus various areas for future work. On the mathematical side, the most urgent ones probably concern extending our methods to cover the case in which mishaps do occur (albeit few of them), so filling the gap identified in section V-C between the early, no-mishap days and the very long term phase in which mishaps are rare but numerous enough to make statistical analysis straightforward; and extending the efforts mentioned in V-B for taking into account changes of the vehicle and/or its environment of use.

The proposed form of argument can be applied to early deployment of any system. For systems with "ultra-high" dependability requirements, where strong beliefs that the system will live up to its required level of safety cannot be empirically justified before extensive operation, this way of reasoning supports responsible management of incremental deployment so as to keep risk at an acceptable level all along.

This is a more modest goal than certifying, as is required, at least in theory, e.g., in aviation, that risk is acceptably low for the whole lifetime of a type; but its very modesty allows it to be argued and scrutinised rigorously.

Our earlier paper [4] proposed a form of sound argument for such modest, but practically useful, predictions of adequate safety over some future period of interest. The present paper shows how that approach can be used both to manage risk via controlling the growth of the operational fleet and also, in some cases, to know when the more extreme, initially unverifiable claim will become believable with high confidence.

ACKNOWLEDGMENT

We are grateful to Philip Koopman, John Rushby and the anonymous reviewers for thought-provoking comments on previous versions of this work.

REFERENCES

[1] N. Kalra and S. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transp. Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[2] B. Littlewood and L. Strigini, "Validation of ultra-high dependability for software-based systems," *Comm. of the ACM*, vol. 36, pp. 69–80, 1993.

[3] R. Butler and G. Finelli, "The infeasibility of quantifying the reliability of life-critical real-time software," *IEEE Trans. Softw. Eng.*, vol. 19, no. 1, pp. 3–12, 1993.

[4] L. Strigini and A. Povyakalo, "Software fault-freeness and reliability predictions," in *Computer Safety, Reliability, and Security*, F. Bitsch, J. Guiochet, and M. Kaâniche, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 106–117.

[5] M. R. Lyu, Ed., *Handbook of Software Reliability Engineering*. IEEE Computer Society Press and McGraw-Hill, 1996.

[6] FAA, "Sytem design and analysis," Federal Aviation Administration, Advisory Circular AC 25.1309-1A, 1988.

[7] *Safety Assessment Principles – 2014 edition (Revision 1, 2020)*, Office for Nuclear Regulation (ONR). [Online]. Available: "https://www.onr.org.uk/saps/"

[8] K. W. Miller, L. J. Morell, R. E. Noonan *et al.*, "Estimating the probability of failure when testing reveals no failures," *IEEE Transactions on Software Engineering*, vol. 18, no. 1, pp. 33–43, 1992.

[9] C. Atwood, J. LaChance, H. Martz *et al.*, "Handbook of parameter estimation for probabilistic risk assessment," U.S. Nuclear Regulatory Commission, Washington, DC, Report NUREG/CR-6823, 2003.

[10] J. O. Berger, "An overview of robust Bayesian analysis," *Test*, vol. 3, no. 1, pp. 5–124, 1994.

[11] P. Bishop, L. Strigini, A. Povyakalo, and X. Zhao, "Using Conservative Bayesian Inference for reliability modelling," CSR, City, University of London, DISPO Project Technical Report PB-DISPO2018-02, 2018.

[12] P. Bishop, R. Bloomfield, B. Littlewood *et al.*, "Toward a formalism for conservative claims about the dependability of software-based systems," *IEEE Trans. Softw. Eng.*, vol. 37, no. 5, pp. 708–717, 2011.

[13] X. Zhao, K. Salako, L. Strigini *et al.*, "Assessing safety-critical systems from operational testing: A study on autonomous vehicles," *Information and Software Technology*, vol. 128, p. 106393, 2020.

[14] ——, "Assessing the safety and reliability of autonomous vehicles from road testing," in *IEEE 30th Intern. Symp. on Software Reliability Engineering (ISSRE 2019)*, 2019, pp. 13–23.

[15] M. Wood, P. Robbel, M. Maass *et al.* (2019) Safety first for automated driving. Several companies. [Online]. Available: https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driving.pdf

[16] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a Formal Model of Safe and Scalable Self-driving Cars," *arXiv e-prints*, p. arXiv:1708.06374, Aug. 2017.

[17] P. Popov and L. Strigini, "Assessing asymmetric fault-tolerant software," in *Proc. of the 21st Int. Symp. on Software Reliability Engineering*. San Jose, CA, USA: IEEE Computer Society Press, 2010, pp. 41–50.

[18] F. Terrosi, L. Strigini, and A. Bondavalli, "Impact of machine learning on safety monitors," in *SAFECOMP 2022, 41st International Conference on Computer Safety, Reliability and Security*. Munich, Germany: Springer, 2022, in press.

[19] L. Fraade-Blanar, M. S. Blumenthal, J. M. Anderson, and N. Kalra, *Measuring Automated Vehicle Safety - Forging a Framework*, RAND Corporation, 2018. [Online]. Available: https://www.rand.org/pubs/research_reports/RR2662.html

[20] M. S. Blumenthal, L. Fraade-Blanar, R. Best, and J. L. Irwin, *Safe Enough-Approaches to Assessing Acceptable Safety for Automated Vehicles*, RAND Corporation, 2020. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA569-1.html

[21] P. Koopman and B. Osyk, "Safety argument considerations for public road testing of autonomous vehicles," *SAE International Journal of Advances and Current Practices in Mobility*, vol. 1, no. 2019-01-0123, pp. 512–523, 2019.

[22] P. Koopman and M. Wagner, "Positive trust balance for self-driving car deployment," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2020, pp. 351–357.

[23] ANSI/UL, *Evaluation of Autonomous Products, UL 4600, Edition 2*, 2022.

[24] FAA, *Accident and Incident Data*, Federal Aviation Administration. [Online]. Available: https://www.faa.gov/data_research/accident_incident

[25] EASA, *Aviation Safety Reporting*. [Online]. Available: https://www.easa.europa.eu/domains/safety-management/aviation-safety-reporting

[26] *IEC 61508, Functional Safety of Electrical/ Electronic/Programmable Electronic Safety Related Systems*, International Electrotechnical Commission, 2010. [Online]. Available: https://webstore.iec.ch/publication/22273

[27] L. Strigini, "On testing process control software for reliability assessment: the effects of correlation between successive failures," *Software Testing, Verification and Reliability*, vol. 6, no. 1, pp. 33–48, 1996.

[28] J. Garman, "The 'bug' heard round the world," *ACM SIGSOFT Software Engineering Notes*, vol. 6, no. 5, pp. 3–10, 1981.

[29] J. L. Lions, "Report by the Inquiry Board on the Ariane 5 Flight 501 failure," ESA/CNES, Tech. Rep., 19 July 1996.

[30] B. Littlewood, K. Salako, L. Strigini, and X. Zhao, "On reliability assessment when a software-based system is replaced by a thought-to-be-better one," *Reliability Engineering & System Safety*, vol. 197, p. 106752, 2020.

[31] K. Salako, L. Strigini, and X. Zhao, "Conservative confidence bounds in safety, from generalised claims of improvement & statistical evidence," in *51st IEEE/IFIP Intern. Conf. on Dependable Systems and Networks (DSN 2021)*, 2021.

[32] R. A. Chakherlou, K. Salako, and L. Strigini, "Arguing safety of an improved autonomous vehicle from safe operation before the change: new results," in *RAIS 2022, International Workshop on Reliability of Autonomous Intelligent Systems*. Charlotte, NC, U.S.A.: IEEE Computer Society Press, 2022, in press.

[33] R. Pietrantuono, P. Popov, and S. Russo, "Reliability assessment of service-based software under operational profile uncertainty," *Reliability Engineering & System Safety*, vol. 204, p. 107193, 2020.

[34] P. Bishop and A. Povyakalo, "Deriving a frequentist conservative confidence bound for probability of failure per demand for systems with different operational and test profiles," *Reliability Engineering & System Safety*, vol. 158, pp. 246–253, 2017.

[35] B. Littlewood and D. Wright, "Stopping rules for the operational testing of safety-critical software," in *25th IEEE Annual International Symposium on Fault -Tolerant Computing (FTCS-25)*. Pasadena: IEEE Computer Society Press, 1995, pp. 444–451.

[36] P. Bishop, "Does software have to be ultra reliable in safety critical systems?" in *Computer Safety, Reliability, and Security*, F. Bitsch, J. Guiochet, and M. Kaâniche, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 118–129.