Arguing safety of an improved autonomous vehicle from safe operation before the change: new results

Robab Aghazadeh Chakherlou, Kizito Salako, Lorenzo Strigini
Centre for Software Reliability
City, University of London, London, U.K.
{ Robab.Aghazadeh-Chakherlou, K.O.Salako, L.Strigini }@city.ac.uk

***Abstract*—Autonomous vehicles (AVs) are gradually appearing on the roads. However, how to demonstrate their safety is still under debate. While operational testing seems essential for building confidence in AV safety, the amount of testing required can be prohibitively expensive. Additionally, current AVs evolve continuously and are used in a changing set of environmentsRepeating substantial operational testing for each new AV version, or new use of an AV, seems unaffordable. Therefore, the idea of applying operational experience from before such a change towards claims of safety after the change is attractive. We present new results, addressing the frequent case in which a new version of the AV can be proved to be safer than a previous one, bar major errors in design or analysis assumptions. Mathematically, our new solution applies to all those scenarios in which the new version or environment is, with high probability, no less safe than the old one "no matter how safe the old one was". We call this scenario "unconditional improvement" (UI). Various previous papers addressed related scenarios in which there is some confidence that the change has improved, or at least not degraded, safety, but they solved the problem under weaker conditions: our new results substantially improve the safety claims that can be supported, especially for operation soon after the change.**

I. Introduction

Experience of safe operation or operational testing is desirable evidence for safety demonstration of any system. Such evidence alone is typically insufficient for strong statistical confidence in extreme rarity of accidents [1], [2]; yet in many contexts it plays an important role in ensuring confidence in safety. Sometimes this experience is presented as classical confidence statements; Bayesian methods allow one to combine operational evidence with "prior beliefs" derived from, e.g., quality of development and verification [1], [3], [4].

A concern is that statistical inference methods require that the measures being estimated (e.g. probability of failure per demand, *pfd*) be unchanged between the past experience from which statistical data were collected and the future use for which a safety claim is made. Thus, any change in the system or in its usage ("operational profile", "environment") is assumed to invalidate previous evidence. For instance, if we had ample experience of safe operation of a certain system in a certain environment (let's call this combination "situation A"), and wished to argue for safety in a situation B, where the system has been slightly improved or the operational environment is slightly different, this would be impossible using the generally recommended statistical methods [4], [5] for such "proven in use" evidence.

One would then need to repeat an operational testing campaign from scratch every time the situation changes (a system is updated and/or used in new environments). This would be very expensive, and yet, if the experience from A amounted to many system-years of operation, pre-operation testing for B would typically bring a much smaller amount of evidence, seemingly adding little assurance.

Therefore, statistically sound methods are needed for applying evidence collected from a previous situation of the system to demonstrating safety in the current situation [3], [6], [7].

These concerns are especially relevant for AVs, which (a) have requirements for minuscule accident rates; (b) rely on machine learning (ML), undermining conventional, non-statistical verification techniques; (c) undergo frequent changes; (d) may require complex authorisation processes for realistic testing on public roads.

For these reasons, two of the present authors (Salako and Strigini) and colleagues have proposed solutions for integrating evidence from a situation A in rigorous statistical claims for a new situation B. These address the case, frequent in practice, in which B differs from A in that the system and/or the environment of use are *likely to be safer* than those in situation A. Experience from A then improves confidence in the safety of B, and thus reduces the amount of new operational testing required before regular operation of B can be allowed to start.

This paper presents an improved method, suitable for some of the scenarios in the latest paper cited [3].

Although this approach is expected to be applicable more widely, here we limit our scope, for comparisons with the earlier work [3], [6], [7], to cases in which

- the evidence from operation is a number of discrete *demands* to which the system was subjected (e.g., trips of a vehicle), all dealt with successfully (e.g. without accidents). Since the mathematical results apply irrespective of the event whose probability is sought, we will call this event "failure", which in different situations might refer literally to failures, or to accident, or the reaching of hazardous system states, etc.;
- The objective function of interest is the (Bayesian) confidence in a required upper confidence bound p_r on the *pfd* of B, given experience of n_A failure-free demands

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

for A and n_B for B: the posterior confidence $P(X_B \leqslant p_r \mid n_A, n_B)$. We will call this objective function c for "confidence", for brevity.

Notions of confidence are important in safety assessment, as it is difficult to assert quantitative safety claims with certainty. The theory of statistical confidence deals with uncertainty arising from sampling. But much inevitable uncertainty is "epistemic", e.g., in the knowledge of the person preparing an argument for certification of a system – we will call this person "the assessor" – about their own probabilistic model of a system. With Bayesian methods, assessors can treat epistemic uncertainty via probabilities, updating uncertainty (probabilities of events of interest) by applying evidence (failure or lack of failure) from operational testing [1], [4].

Claims based on improvements over a previous system ("No worse than existing system", or *NWTES*), as inputs to, or as results of, a safety argument, are common, in various industrial sectors, with different names (e.g. "Globally at least equivalent" (GALE)). For a discussion with examples we refer the reader to an earlier paper [8].

Like the earlier work cited, we apply "Conservative Bayesian Inference" (CBI) – a form of robust Bayesian inference [9], [10]. We offer assessors a method that:

1) is Bayesian: it uses Bayes' theorem to update "prior beliefs" (probabilities of events of interest) into "posterior beliefs", based on evidence available prior to operation or operational testing. An example update is our objective function, posterior confidence c.
2) Claims based on with respect to the assessors' uncertainties about prior beliefs. Standard use of Bayesian inference often requires assessors to specify more detailed prior beliefs than they have evidence for. For example, a complete probability density function for the unknown *pfd* of a system. Assessors are offered "standard", mathematically convenient forms for such functions, with no help for verifying whether they ought to believe them. The resulting claims could therefore be more optimistic than the evidence warrants. CBI mitigates this risk by only asking for partial information about prior distributions; that is, asking for beliefs for which the assessor has credible evidence[1]. We call such statements "Prior Knowledge" statements or PKs; we give examples later. CBI considers all those prior distributions that are consistent with the assessor's PKs and seeks the infimum of the values of c that they yield, so as to return a value for c that is guaranteed to be conservative. A prior distribution that gives this most conservative posterior confidence is called *"worst-case prior distribution"*.

Our colleagues and we have produced a set of CBI results for a set of scenarios (PKs) and objective functions [3], [7], [11], [12], from which an assessor can choose, based on the evidence available before operation/testing[2].

The advance that we present here is a CBI solution for an especially interesting form of "NWTES" scenario, *Unconditional Improvement* (UI). We will first recall some limitations of earlier results [3] and show that when the UI assumption (PK) is justified, substantially stronger claims can be soundly made about a new situation B.

In the rest of this paper, II recalls earlier results; Section III explains a limitation of previous results that we begin addressing in this paper, and presents the "Unconditional Improvement" condition and some of its useful properties. Numerical results illustrate the stronger claims that UI can support. Section IV concludes the paper.

II. Previous results

The CBI approach has been studied in several papers,and applied to various objective functions, e.g. expected *pfd* [11].

In CBI extensions that use operational evidence from a *previous* system or environment [3], [6], [7], the problem of estimating a single *pfd* becomes a two-dimensional problem about the *pfd* before and after the change, seen as two random variables, X_A and X_B. Failure-free operational evidence from situations A and B is combined to estimate a pessimistic confidence bound for *pfd* in situation B. The Bayesian inference problem amounts to computing posterior confidence c:

$$\begin{aligned} c &= P(X_B \leqslant p_r \mid n_A, n_B) \\ &= \frac{\int_{[0,1]\times[0,p_r]}(1-x_A)^{n_A}(1-x_B)^{n_B}\mathrm{d}F_{AB}(x_A,x_B)}{\int_{[0,1]\times[0,1]}(1-x_A)^{n_A}(1-x_B)^{n_B}\mathrm{d}F_{AB}(x_A,x_B)} \end{aligned} \tag{1}$$

where p_r is a required bound on X_B and $F_{AB}(x_A, x_B)$ is a joint prior probability distribution for random variables X_A, X_B. Applying CBI is then an optimisation problem: determining the infimum of the values of (1) over the set of all possible prior distributions F_{AB} consistent with the assessor's constraints (PKs).

Various scenarios (different sets of PKs) studied in [3], [6], [7] are summarised in Table I and Fig. 1.

Various forms of "prior knowledge"	
PK1	$P(X_A \geqslant p_l) = P(X_B \geqslant p_l) = 1$
PK2-A	$P(X_A \leqslant \epsilon) = \theta \ , \ \epsilon < p_r$
PK2-B	$P(X_B \leqslant \epsilon) = \theta \ , \ \epsilon < p_r$
PK3	$P(X_B \leqslant X_A) = \phi$
PK4	For some intervals $I \subseteq [0,1]$ and $0 \leqslant \phi \leqslant 1$, $(1-\phi)P(X_B \leqslant X_A, X_A \in I) = \phi P(X_B > X_A, X_A \in I)$

TABLE I: Examples of "prior knowledge" statements studied in previous work [3], [6], [7].

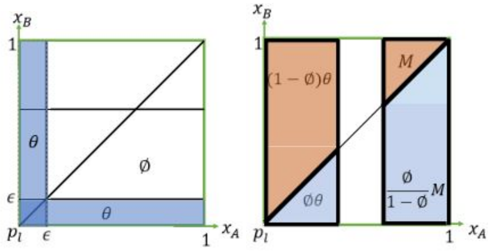
PK1 is a lower bound for *pfd*: the best *pfd* claim feasible with current technology. For simplicity, we set $p_l = 0$.

PK2 (applicable for A and/or B) means that the assessor has a certain confidence in the *pfd* satisfying a stronger bound than the required p_r. This could typically be based on experience with systems that are similar in complexity and development process. It turns out that θ gives the prior confidence, $P(X_B \leqslant p_r | n_A = 0, n_B = 0)$ in the worst-case priors.

PK3 is a "confidence in improvement" (CII) assumption: an assessor's confidence ϕ in B being no less safe than A. $\phi < 1$

[1] And can argue, e.g. with a regulatory authority, that the evidence does support them.

[2] This set does not claim to be complete. We plan to extend it according to requests from safety practitioners.



(a) PK1, PK2 and PK3 (b) PK1, PK2, PK4

Fig. 1: The PKs in Table I constrain the probabilities that F_{AB} can allocate to subsets of its domain [3]. The random variables X_A, X_B represent the unknown *pfd*s of A and B. For both figures, ϕ is the probability of the region under the diagonal, $P(X_A \leqslant X_B)$.

if the assessor knows that system B has design improvements over A, meant to improve safety. $1-\phi$ represents the probability that such improvements unwittingly harming safety. These probabilities can be based on historical experience.

PK4 will be discussed in Section III.

These papers study in detail the effect of different PKs on the posterior confidence and the amount of operational experience required for the new situation (B) to support a desired confidence c. However, an important scenario not covered previously – what we call UI in this paper – is covered in the next section.

III. Unconditional improvement (UI)

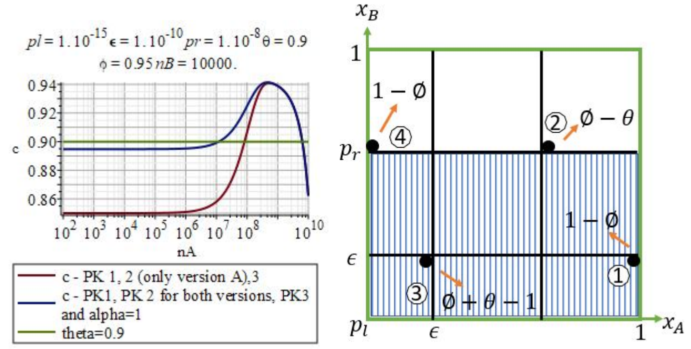
In this section we discuss UI and the worst-case posterior confidence it yields in safety after the change from A to B. This form of reasoning can greatly reinforce confidence in safety for B. Interestingly, its results are more intuitive than some results obtained in previous papers using other forms of CII. So, we take a brief detour to discuss aspects of some previous results that may appear counter-intuitive, and which do not occur in UI scenarios.

A. Two "paradoxical" effects in previous results

Fig. 2 shows examples of previous results [3]. In Fig. 2b the axes x_A and x_B represent the *pfd* of A and B, respectively.

While these are correct results of the model – the assumptions – used, we note two ways in which they contradict what a user would probably expect. Both a high n_A and a high n_B are "good news": most people would probably expect that as n_A and n_B increase, so would the posterior confidence. The plots in Fig. 2a exhibit this expected trend, up to a certain value of x_A, but also two surprising features. Indeed, the posterior confidence c: 1) at the beginning of the plot c is less than 90%, which is less than the *prior* confidence, 2) after increasing until about $n_A = 5 \times 10^8$ demands treated by A without failures, it starts decreasing – more positive evidence from A not only does not improve confidence in B any more, but actually reduces it.

These effects are explained by the shape of the worst-case priors for these scenarios, shown, e.g., in Fig 2b for the upper plot in Fig 2a [3]. The fat black points represent discrete probability masses. [3] The posterior distribution will thus have the same shape as the prior – discrete probability masses in the same points – while the probability associated to each point varies with n_A and n_B.



(a) Posterior confidence $c = P(X_B \leqslant p_r|n_A, n_B)$ for two scenarios from [3].

(b) The worst case prior distribution for the plot in Fig. 2a. $c = P(X_B \leqslant p_r)$ is the probability of the hatched region.

Fig. 2: Examples of previous results.[3]

The cause of the "paradoxical" effects now becomes clear:

1) After the first demand successfully processed in A, the event $X_A = 1$ (point 1 in Fig. 2b) becomes impossible: its probability becomes zero. The probability associated with it is redistributed between points 2, 3 and 4. c (posterior confidence) is the sum of probabilities associated to points 1 and 3: when the former becomes zero, the sum of the two drops. As n_A increases, confidence grows again until, after 10^7 demands on A (see Fig. 2a), we have regained the initial confidence in B, θ.
2) From now on, further positive experience with A improves confidence in B. However, this improvement stops at a maximum.
3) As n_A increases further, we naturally increase our confidence in A being very safe; but this implies a large probability mass at point 4: evidence of A being very safe is also evidence of B not having a *pfd* better than desired bound p_r.

This effect, however surprising, is a necessary consequence of the assumptions stated: while intuitively they suggest that good experience with A should always improve our confidence in B, mathematical analysis shows that this is not always so.

B. Understanding the inference process via odds ratios

It's useful to describe Bayesian inference in terms of posterior odds ratios. For any two events Ev_1 and Ev_2 there is a relationship between posterior odds ratio given an observation Ob and prior odds ratio (when $P(Ob\,\&\,Ev_2) \neq 0$):

$$\frac{P(Ev_1|Ob)}{P(Ev_2|Ob)} = \frac{P(Ev_1)}{P(Ev_2)} \frac{P(Ob|Ev_1)}{P(Ob|Ev_2)} \tag{2}$$

[3] Worst-case prior distributions found in CBI are usually discrete distributions, with non-zero probabilities for only a few points in the x_A, x_B plane.

The l.h.s. is the posterior odds ratio, the first term on the r.h.s. the prior odds ratio, and the second term the *likelihood ratio*.

For instance, referring to the prior distribution in Fig. 2b and considering only the effect of increasing n_A (keeping $n_B = 0$), the likelihood ratio between point 3 and point 2 is

$$(1-\epsilon)^{n_A}/(1-p_r)^{n_A} \tag{3}$$

This is an increasing function of n_A. So, in (2), increasing n_A reduces the probability at point 2 relative to that at point 3. More generally, any point will have its probability reduced relative to any other points to its left.

So (calling $P_1, \ldots, P_4$ the posterior probabilities for the points in Fig. 2b), with high enough n_A, P_1 and P_2 become negligible and P_3, P_4 grow, at first in proportion to their prior values (thus the decrease of P_2 increases P_3, i.e., c); but as n_A increases further, the likelihood ratio between points 4 and 3 grows, until P_3 starts decreasing as in Fig. 2a.

This reasoning clarifies that the applied CII (PK3) does not actually imply that all positive evidence about A would increase confidence in B. We now discuss how UI, a slightly different form of CII, avoids this.

This is why PK3 does not ensure that failure-free operation from system A (i.e. increasing n_A) increases confidence in the B system – all non-zero probability eventually accumulates at the leftmost point 4, causing confidence (1) to go to zero. To prevent this from happening, a new form of CII – called Unconditional Improvement (UI) – is defined next.

C. Unconditional improvement (UI) prior knowledge

UI is a form of CII that occurs often in practice: when evidence suggests X_B is no worse than X_A, *no matter what the true value of X_A is.* Intuitively, this guarantees that if we have evidence that A is very good, we will still believe B to be at least as good. For instance, the following quote from [3]:

> "*system B is obtained by adding to system A some safety protection: e.g., A is a safety system and B adds another independent safety monitor with authority to effect the safety action* [...]. *This way of building B ensures that the set of demands on which B fails unsafely, U_B, is a subset of those where A fails unsafely, $U_B \subseteq U_A$. Hence,*
>
> $$pfd_B = \sum_{D \in U_B} P(D) \leqslant \sum_{D \in U_A} P(D) = pfd_A \tag{4}$$

The difference between UI and previously studied scenarios as in Fig. 2a is that, under UI, we know (with probability ϕ) that system B is no less safe than system A were pfd_A "known"; in the earlier scenarios, we knew that (with probability ϕ) systems "like A" are no less safe than systems "like B". To give more concrete examples:

- a scenario that may lead to the results in Fig. 2a:
  (a) system B is a plug-in replacement for system A, built to the same requirements, but with methods recommended for more stringent safety requirements; and (b) from historical statistics, the assessor knows that were they to randomly pick, independently, an "A-type" system (developed with the weaker methods) and a "B-type" system, on a fraction ϕ of the attempts the B-type system would be found to be no less safe than the A-type system;
- a scenario in which UI applies is: (a) system B is a design improvement of A, designed to remediate known or potential safety weaknesses of A; (b) appropriate probabilistic analysisof the two designs (e.g. fault tree analysis) proves that indeed B will be safer than A. This cannot give 100% certainty: errors in the assumptions or the calculations are possible. The assessor, however, on the basis of experience judges that cases in which such "verified to be safer" changes actually harm safety happen with frequency lower than $(1-\phi)$.

The UI condition: no matter what the true value of X_A (*pfd*$_A$) is, X_B (*pfd*$_B$) is believed to be better with probability ϕ. That is, for $0 \leqslant \phi \leqslant 1$ and *all* intervals $I \subseteq [0,1]$,

$$(1-\phi)P(X_B \leqslant X_A, X_A \in I) = \phi P(X_B > X_A, X_A \in I) \tag{5}$$

UI is PK4 (from Table I) applied to *all* possible intervals in $[0,1]$); so, *any* $I = [x_{A_1}, x_{A_2}]$ for *any* distinct *pfd*s x_{A_1} and x_{A_2} (even when these are arbitrarily close in value).

Not only does (5) often represent the *NWTES* beliefs an assessor can have, but in the context we are considering, ϕ will often be larger than θ. For instance, common experience tells us that design errors such that an extra safety channel *reduces* safety are quite rare; while an ML system failing more often than is desirable is a rather common event. Therefore, we study the effects of UI when $\phi > \theta$.

Finding a worst case prior under the UI condition proved difficult, so a previous paper tried approximating it via similar, weaker conditions [3], like Fig. 1b , in which (5) applies to particular intervals of x_A. However, the results still exhibited effects similar to Fig. 2a. Theorem 1 below illustrates how UI avoids these effects, allowing confidence in *pfd*$_B$ to grow.

D. Worst-case posterior confidence under the UI assumption

The following theorem gives the least confidence in system B under the UI condition, in the limit as $n_A \to \infty$.

Theorem 1. *Let $\mathcal{D}$ be the set of all prior distributions of X_A, X_B. Using eq.* (1), *consider the optimisation problem*

$$\inf_{\mathcal{D}} P(X_B \leqslant p_r \mid n_A, n_B) \tag{6}$$

subject to PK2-A, PK2-B (Table I), and UI (eq. 5). We have

$$\lim_{n_A \to \infty} \inf_{\mathcal{D}} P(X_B \leqslant p_r \mid n_A, n_B) = \frac{\phi(1-\epsilon)^{n_B}}{\phi(1-\epsilon)^{n_B} + (1-\phi)(1-p_r)^{n_B}} \tag{7}$$

This is the expression for $P(X_B < p_r \mid n_A, n_B)$ given by the discrete prior distribution in Fig. 3.

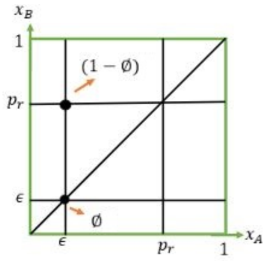
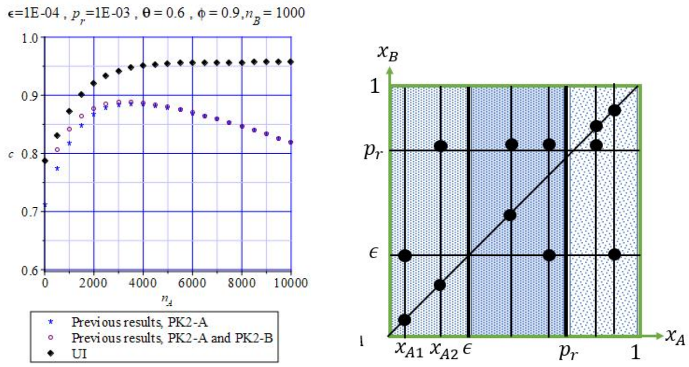


Fig. 3: A prior distribution over the unit square that gives the pessimistic (smallest) value (7) for posterior confidence (1), under the UI condition as $n_A \to \infty$. This prior assigns non-zero probability masses to only the two locations shown, and zero probability for all other locations in the square.

E. Example numerical results

Fig 4a compares CBI results with the UI assumption against results from two scenarios in the older paper [3]. With UI, the posterior confidence in B being safe enough is higher and increases monotonically with n_A: the "paradoxical" effects have disappeared.



(a) Comparison between the posterior confidence in previous results [3] and under UI.

(b) Possible patterns of discrete probability masses. Note the difference between patterns in the three regions.

Fig. 4: Effects of Unconditional Improvement assumption.

Thus with the UI assumption, often, as we said, justified in practice, e.g. in the case of intentional safety improvements to a system, positive operational experience from a previous situation A can indeed significantly boost claims about the safety of later situation B. This advantage can be substantial, depending on the parameters of the problem, $\theta, \phi, \epsilon, p_r$, but has limits: once n_A is large enough for the posterior confidence to be close to its limit in formula (7), further positive experience with A does not give any more advantage for B.

For example, Fig. 5 compares the posterior confidence $P(X_B \leqslant p_r \mid n_B)$ obtained using just prior knowledge about B – i.e. the fact that $P(X_B \leqslant p_r) = \theta$ (lower curve) – against that – $P(X_B \leqslant p_r \mid n_A, n_B)$ – obtained with UI (5) if we also have a large failure-free operational experience on A. The advantage is large while experience with B is still limited. [4] The advantage decreases as experience with B accumulates and the posterior confidence approaches 1.

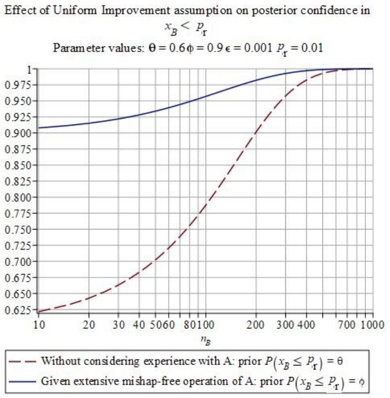


Fig. 5: Posterior $P(X_B \leqslant p_r | n_B)$ with UI and large n_A *vs* with no knowledge of A.

IV. Conclusions and Future Work

Previous publications showed that given some confidence that a change in a system or its usage will not reduce safety ("No worse than existing system" assumption, NWTES), it is possible to take advantage of the operational experience before the change for claims about safety after the change, with conservatism guaranteed via the CBI approach.

We have now presented new results that for a specific form of prior beliefs support stronger claims and remove aspects of some previous results that, although correct, could appear puzzling and discourage use of these useful methods.

Of course, the new results hold only when the "Unconditional Improvement" condition holds, which must be argued on the basis of the specific circumstances.

The large differences in assessment results that follow from apparently subtle differences in assumptions demonstrate why rigorous mathematical proof is important in these matters.

The results of the various NWTES scenarios studied so far are a promising first step towards solving the more general problem of using experience accumulated across multiple changes in a system and its environment. Future work will also need to integrate evidence about changes for which a NWTES assumption is reasonable, and changes for which it is not: e.g., an AV may be changed to improve performance, which does not imply improving safety.

Last, we note that we only presented an asymptotic result for conservative Bayesian inference using the "no worse than existing system" and "unconditional improvement" assumptions. Our planned work includes the complete solution as well as a more complete study of the practical implications, for instance about probability of surviving extensive operation without accidents rather than about confidence bounds.

[4] Experience on A could even be enough, in theory, for skipping operational testing of B: e.g., if a regulator required confidence c^* in $X_B \leqslant p_r$, and it were the case that $\phi \geqslant c^*$.

Acknowledgments

We thank the anonymous reviewers whose comments helped us to improve the manuscript.

References

[1] B. Littlewood and L. Strigini, “Validation of ultra-high dependability for software-based systems,” *CACM*, vol. 36, no. 11, pp. 69–80, 1993.

[2] 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.

[3] K. Salako, L. Strigini, and X. Zhao, “Conservative confidence bounds in safety, from generalised claims of improvement & statistical evidence,” in *51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, March 2021. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/26128/

[4] C. Atwood and et al, “Handbook of parameter estimation for probabilistic risk assessment,” U.S. Nuclear Regulatory Commission, Washington, DC, Report NUREG/CR-6823, 2003.

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

[6] X. Zhao, V. Robu, D. Flynn, K. Salako, and L. Strigini, “Assessing the Safety and Reliability of Autonomous Vehicles from Road Testing,” in *30th Int. Symp. on Software Reliability Engineering*. Berlin, Germany: IEEE, 2019, pp. 13–23.

[7] X. Zhao, K. Salako, L. Strigini, V. Robu, and D. Flynn, “Assessing safety-critical systems from operational testing: A study on autonomous vehicles,” *Information and Software Technology*, vol. 128, p. 106393, 2020.

[8] 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.

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

[10] ——, “Robust Bayesian analysis : Sensitivity to the prior,” *journal of statistical planning and inference*, vol. 25, pp. 303–328, 1990.

[11] P. Bishop, R. Bloomfield, B. Littlewood, A. Povyakalo, and D. Wright, “Toward a formalism for conservative claims about the dependability of software-based systems,” *IEEE Transactions on Software Engineering*, vol. 37, no. 5, pp. 708–717, 2011.

[12] 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.

[13] R. L. Schilling, *Measures, Integrals and Martingales*. Cambridge University Press, 2005.

Appendix: Proof of Theorem 1

Proof. Let F be any feasible prior for problem (6), i.e. a prior satisfying constraints PK2-A, PK2-B (Table I), UI (5). Denote the value of the objective function (1) given by F as $c(F)$.

1) *The feasible prior distributions can be restricted to only discrete prior distributions.* That is, for any feasible prior F and arbitrary $\delta > 0$, there exists a feasible discrete prior F^* that assigns non-zero probability masses only to finitely many points (that lie along finitely many vertical lines in the unit square), such that $|c(F) - c(F^*)| < \delta$. This follows by construction, since the integrals in the numerator and denominator of $c(F)$ can each be arbitrarily well-approximated by integrals of step functions, where these step functions converge *F-almost surely*[5] to the integrand in $c(F)$'s integrals (i.e. the *dominated convergence theorem* [13]). To create these step functions, use the main diagonal of the unit square and rectilinear grids to create a sequence of partitions of the unit square[6]. The probability masses that F assigns to each cell in a partition can be reallocated to a single point in the cell, so that probability only lies at a finite number of locations along vertical lines of the partition. By being selective about these reallocations, this creates feasible discrete priors that satisfy the constraints[7] of (6), and creates step functions that are constant over each cell (taking the value that $c(F)$'s integrand takes at the point in a cell where mass has been reallocated to). Thus, any sufficiently fine partition gives a suitable F^*.

2) *The discrete priors can be restricted further, to priors that assign probability mass to only 3 possible locations along vertical lines – the locations where a vertical line intersects the horizontal lines $\boldsymbol{x_B = \epsilon}$, $\boldsymbol{x_B = p_r}$ or the main diagonal.* For any F^*, the derivative of $c(F^*)$ along any vertical line indicates how mass must be reallocated along the line to give a smaller value for $c(F^*)$. These reallocations result in a discrete prior F^{**} that assigns probability only to locations of intersection as claimed (e.g. see Fig. 4b).

3) *For any $\boldsymbol{F^{**}}$, the value of $\boldsymbol{\lim\limits_{n_A \to \infty} c(F^{**})}$ is given by a discrete prior $\boldsymbol{F^{***}}$ that assigns probability only to locations along a single vertical line; the line itself is located at some $\boldsymbol{x_{Al} \leqslant \epsilon}$.* This follows because F^{**} must have probability masses located at points along some “left-most” vertical line at some $x_{Al} \leqslant \epsilon$, due to constraint PK2-A (I). Note that, for any $x_{Al} \leqslant \epsilon$, there exists a feasible prior F^{**} with leftmost vertical line located at x_{Al}.

Now, as $n_A \to \infty$, all probability mass must flow to this leftmost vertical line (see argument in section III-B). The UI condition ensures that $\lim\limits_{n_A \to \infty} c(F^{**})$ takes the form

$$\frac{\phi(1-x_{Al})^{n_B} + M(1-\phi)(1-\epsilon)^{n_B}}{\phi(1-x_{Al})^{n_B} + (1-\phi)\left(M(1-\epsilon)^{n_B} + (1-M)(1-p_r)^{n_B}\right)} \tag{8}$$

for some probability M. This is the value of the posterior confidence $P(X_B < p_r \mid n_A, n_B)$ given by some prior F^{***} that assigns probability only along a vertical line at x_{Al}. The derivative of (8) with respect to x_{Al} shows that (8) is a decreasing function of x_{Al} – it takes its smallest value at $x_{Al} = \epsilon$. The derivative with respect to M shows an increasing function of M – it takes its smallest value when $M = 0$. Thus, $\lim\limits_{n_A \to \infty} c(F^{**})$ is bounded below by

$$\frac{\phi(1-\epsilon)^{n_B}}{\phi(1-\epsilon)^{n_B} + (1-\phi)(1-p_r)^{n_B}} \tag{9}$$

Our initial choice of feasible prior F was arbitrary, so (9) is the infimum of the objective function's values in the limit. Thus, it is the limit of infima from (6). (9) is the value of $P(X_B < p_r \,|\, n_A, n_B)$ given by the prior distribution in Fig. 3. □

[5] The convergence occurs at every point of the unit square except, possibly, at points in subsets of the unit square that F assigns zero probability to.

[6] These partitions should have the usual properties for integration, e.g. they get finer and finer, and define a set of points that are dense in the unit square.

[7] To ensure the vertical lines satisfy the UI condition (5), it is convenient to require that a vertical line intersects the main diagonal at a point *iff* a horizontal line intersects at the same point.