Advances in Reliability, Safety and Security, Part 3 – Kołowrocki, Dąbrowska (eds.)
© 2024 Polish Safety and Reliability Association, Gdynia, ISBN 978-83-68136-15-9 (printed), ISBN 978-83-68136-02-9 (electronic)

Advances in Reliability, Safety and Security
ESREL 2024
Monograph Book Series

Using Pre-Change Operational Evidence For Predicting Post-Change Reliability, Given Prior Confidence In Fault-Freeness

Robab Aghazadeh Chakherlou, Lorenzo Strigini

City, University of London, London, United Kingdom

Abstract

For many systems, high confidence is required that they will never suffer an accident over an extended period of operation. Statistics of accident- or problem-free operation can give factual support for this confidence. But changes, to systems or to the way they are used, create problems for this part of dependability assurance. For instance, experience of safe operation before a design improvement should be still relevant for claims of safety after the improvement; but methods in current use do not show how much it should contribute to confidence in the latter. Thus quantitative assessment after changes may ignore (or instead overrate) large amounts of evidence, distorting decision making about system acceptance or evolution. To help with this problem, we extend previous work on integrating statistical evidence, from operation, with prior confidence, based on production and verification quality, that a design is free from design faults. Our extension also takes into account evidence of operation before the change, and confidence, derived from analysis, that a change did not degrade dependability. We apply "Conservative Bayesian Inference" (CBI) to allow probabilistic reasoning without specifying detailed prior distributions for the variables of interest, a serious difficulty in current use of Bayesian methods. We show: (i) that pre-change evidence can contribute substantially towards trusting the system post-change, especially while post-change experience is still limited; (ii) how this contribution depends on the strength of the analysis showing that the change improves, or does not affect, safety, and on other parameters; (iii) the limits to the advantages that pre-change evidence can bring.

Keywords: survival probability, software correctness, similarity arguments, conservative Bayesian inference, globally at least equivalent, field testing, safety critical systems, ultra-high reliability, no worse than existing system, proven in use

1. Introduction

Statistical evidence of correct, or of safe, operation of a system has an important role for demonstrating its reliability, or safety. E.g., for safety certification or licensing, for complex systems where the possibility of subtle design flaws is a serious concern, a claim of adequate safety is usually based on evidence of quality of development and verification, but is complemented by "operational testing". Even after a system is authorized to operate, the claim that it satisfies its safety requirements can be corroborated or refuted (sometimes dramatically, as, recently, for the Boeing 737 Max) by operational experience. Both forms of evidence can be combined towards quantitative safety claims via Bayesian reasoning (Atwood et al., 2003). However, for systems with stringent safety requirements, the feasible amount of operational testing is often not enough to make a strong contribution to the argument (Littlewood, 1993). We study how predictions for such difficult cases can take advantage of operational experience that is imperfectly related to the intended claim, because of intervening changes to the system or its use.

Changes create a special challenge for statistical corroboration of dependability (safety, reliability, etc.) claims. Changes may concern the system's design or configuration, or its operational environment: e.g. the practices of its users, the physical environment where it is deployed, a broader system of which this is a subsystem. For brevity, we will say that a dependability claim applies to a *context* -- a system and its environment. Statistical arguments for dependability generally use data (amount of operation and number of undesired events) collected from operation (or operational testing) in the same context about which the dependability claim is made: past observations should be a representative sample of what may happen in the future. Some standards do address change. E.g., ISO 26262 on road vehicles (ISO, 2018) (part 10) deals with certifying subsystems for use in more than one system, a case of "different contexts" according to our definition, but does not address issues with possible reuse of statistical evidence. Others like IEC 61508 (IEC, 2010) practically forbid using data from a different context. This limitation avoids many unsound claims, but at a cost. Especially if a change is minor, or arguably a change for the better, data from the old context are indeed relevant, to some extent, to the new one. Forbidding their use implies, firstly, high costs: to support strong confidence in a system, one would need to repeat abundant operational testing after each change. The cost may deter statistical arguments (denying the assessors useful factual evidence on dependability), or deter useful changes. Secondly, this constraint may cause poor decisions. Consider a scenario involving choice between two systems: S_1, which has accumulated abundant experience of safe operation, before undergoing a recent change; and S_2, which has been recently developed and has no operational history. To discard S_1's pre-change experience as irrelevant would mean that, everything else being equal, we would trust a completely new system, S_2, just as much as the one which was just updated after massive positive experience. This probably underestimates the evidence on favor of S_1. The dual problem may occur with “proven in use” arguments allowed by certain standards, e.g., part 8 of (ISO, 2018), as substitutes for a complete safety argument, potentially leading to overly optimistic assessment. Finding how much weight to give, in a claim for dependability, to safe operation in a previous context, would avoid such paradoxical conclusions, as well as reduce the cost of re-assessing or re-certifying a system after a change.

We extend here some recent literature about scenarios in which a new context is trusted, with some level of confidence, to be at least as safe as the previous one. This concept is common in reasoning about safety, e.g. in principles such as "globally at least equivalent" (GALE, or GAME in French) a requirement in some French laws for railway and road safety (Minister responsible for transport, 2021), and the claim of "substantial equivalence" (to an approved product) required by the U.S. Food and Drug Administration to allow a light-touch approval procedure for a new product (FDA, 2014). Similarly, hardware reliability claims are often based on testing under especially stressful conditions so that better reliability can be inferred for the intended conditions of use. All mentioned terms require that a changed or new system maintain or exceed the overall safety level of the previous system. The concerned agencies have guidelines (e.g., (Minister responsible for transport, 2022)) for analyses that support a claim of GAME/GALE or “essentially equivalent”. But these guidelines leave open the question of how much that claim should weigh in trusting the changed or new system, once this a *priori* analysis is complemented by the necessary observations of the new system in operational testing/experience. This paper aims to answer this question. We borrow the phrase “No Worse Than Existing System” (*NWTES*) (Littlewood et al., 2020) to refer to GALE-like statements about a post-change context. *NWTES* arguments are those that use *NWTES* claims, in situations like, for instance: a design fault has been corrected; a system has undergone a change meant to improve its safety, and/or it is deployed in its intended environment after stress-testing in a harsher environment; or a new feature has been added, with analysis showing that it has no effect on safety.

The following Section 2 defines the scope and assumptions in this paper with reference to previous literature; Section 3 specifies the mathematical minimization problem addressed, notation used, and a theorem that solves this problem. Section 4 illustrates the relevant insight gained, using numerical examples. Section 5 discusses the relevance of the results and future work.

2. Context, Related work, and contribution of this paper

Given an analysis proving *NWTES*, one might think that for the purpose of inference one can just add the number of post-change demands to the pre-change ones. But (Littlewood et al., 2020) proved this to be an error, producing over-optimistic conclusions, unless there is 100% certainty about the *NWTES* property. A set of publications, cited below, showed how the actually feasible claims for the post-change context depend on the details of prior beliefs and the exact formulation of the *NWTES* claim. To our knowledge they are the only literature to address how a probabilistic assessment should account for the uncertainty about an *NWTES* claim.

2.1. Measures of dependability; reliability-type measures

We consider systems whose operation consists of (statistically independent) demands. Their reliability or safety is characterized by a constant "probability of failure per demand" *pfd* (Bernoulli trial). We use the term "failure" as a generic term for the unwanted events to be monitored and predicted (failures, accidents, dangerous system conditions, etc., depending on the focus of the analysis), since the mathematical treatment does not depend on the specific kind of event.

A demand might be a flight, for an aircraft, a deviation of plant state from a safe envelope, for a plant safety system, a trip, for a vehicle, etc. Claims about the *pfd* of a system can never be stated with certainty: even systems developed with the best methods available may have unknown design faults; and no assessment method can give certainty. In view of this inevitable uncertainty, probabilistic requirements or claims about a system *pfd* can take various forms, including as follows.

It is often required that the *pfd* do not exceed a stated upper bound, with a certain confidence. For instance, "the *pfd* of this system in the stated operation conditions shall be at most 10^{-6} with confidence at least 90%". This common form of claim has a clear intuitive use, if "confidence" is used in the sense of "probability": it tells a decision maker that, with the stated probability, here 90%, the risk taken by operating the system is no more than it would be if the system's *pfd* were as high as the claimed upper bound, 10^{-6}. Such a requirement matches the ethical requirement not to deploy a system without strong confidence that it does not have serious dangerous flaws. Several papers (Aghazadeh Chakherlou et al., 2022; Littlewood et al., 2020; Salako et al., 2021; Zhao et al., 2019, 2020) applied different mathematical form of *NWTES* arguments to assessing a confidence level in a required confidence bound on *pfd*. But in assessing future risk, using the confidence bound as a proxy for the actual value of *pfd* may be over-optimistic. A confidence in a bound does not say what the *pfd* would be if it exceeded that bound (Littlewood et al., 1993). In the example above, the bare confidence claim says that with 10% probability, the *pfd* may have any value, up to 1: on the next demand (say, flight of an aircraft) the worst-case probability of accident is not 10^{-6} but $10^{-6} \times (90\%) + 1 \times (1 - 90\%) \approx 0.1$.

We note that we use the Bayesian meaning of the word "confidence", as a probability of the event of interest. If one instead uses the "classical" meaning ("the probability of not observing the encouraging evidence that was observed if the true *pfd* were worse than the desired bound"), a claim of confidence does not have the above useful implication. Confusing the two meanings may cause serious errors.

This major drawback is avoided by setting a requirement on the mean *pfd*, which takes into account the probability, however low, that the real *pfd* is worse than a desired or anticipated bound. An advantage of the mean *pfd* is that it gives the true estimate of the risk of operating the system for one future demand, given that the true *pfd* is unknown but a distribution for it can be estimated. Furthermore, the probability of at least one failure over a few demands, say n_{few}, is approximated conservatively (bounded above) by multiplying the expected *pfd* by n_{few}. So, for instance, an air passenger (a test pilot), deciding whether to fly in a new type of aircraft would be well served by this measure for assessing what risk he/she is accepting in doing so. A recent paper studied *NWTES* arguments for expected *pfd* (Littlewood et al., 2020). The main downside of relying on the mean *pfd* to guide decisions is that this conservatism may be excessive when making predictions over many future demands: the actual probability of failures depends on the whole probability distribution of the *pfd*; using the mean as a simple proxy could deceive a decision maker into choosing a riskier option over a less risky one (Strigini and Wright, 2014).

Sometimes, assurance is required for very large numbers of demands (amounts of operation). E.g., a decision may be needed about renewing a license to operate a system in widespread use (Bishop, 2022); for large passenger aircraft "catastrophic failure conditions" must be "so unlikely that they are not anticipated to occur during the entire operational life of all airplanes of one type" (FAA, 1988). In this case, the regulator authorizing the operation of such systems, or the companies operating them, need to assess a reliability function: the probability that the system will complete a certain amount of operation without ever suffering any accident. The contribution of this paper is to develop *NWTES* arguments for such reliability measures.

2.2. Conservative Bayesian inference (CBI)

We take a Bayesian approach: the reliability measures we calculate take into account uncertainty about the true value of the system's pfd, represented as a random variable. "Reliability" is the probability of not having any failure in a given number of future demands, as a weighted sum of the reliability values that would ensue from each possible value of the pfd. This is not the system reliability one would compute from knowing the true, unknown value of this system's pfd; it is an "expected" reliability, given what one actually knows. It is higher than (or at least equal to) the reliability that would be calculated from the mean pfd (Strigini and Wright, 2014). We apply "conservative Bayesian inference" (CBI), an approach to obviate the challenges for assessors (any individual or even group, such as regulators and experts in regulated companies, who builds or approves a Bayesian probabilistic argument.) in using Bayesian inference. Bayesian inference requires a complete "prior" distribution for the pfd, that must reflect prior beliefs justified by the evidence available. Such complete distributions are hard to construct and justify. The simplifications commonly adopted to avoid this problem may lead to dangerously optimistic predictions (Zhao et al., 2019). CBI addresses this concern by only requiring as inputs simple constraints (prior beliefs) on the prior distribution, such that they can be justified based on available evidence. These constraints determine a set of acceptable priors, and thus Bayesian inference from any given observation determines a range of predicted values for the objective function of interest. The worst value in this range is then the most pessimistic prediction compatible with the stated prior beliefs.

2.3. Prior beliefs; the case of fault-free design

We use the same constraints on the prior distribution used in some previous papers (recalled in Section 3.1): the prior beliefs for the Bayesian inference take the form of a confidence in an upper bound on the *pfd*. Specifically we study the case in which this bound is zero. A prior confidence in *pfd*=0 is realistic in a frequent and important scenario: when the concern is failures due to ***design*** faults, because (a) if there are no such faults, the *pfd is* 0, for this category of failures; and (b) this prior claim is supported by evidence of competent attempts to avoid or remove such faults. This is evidence for a probability of the product being fault-free, hence of its *pfd* being 0 (Bertolino and Strigini, 1998).

This case of prior belief in 0 *pfd* is more generally useful for illustration purposes, because it is simpler than, but for some practical scenarios a close approximation of, the more general case of a bound greater than zero (Strigini and Povyakalo, 2013). CBI with a prior belief in 0 *pfd* was previously applied to reliability-type predictions, but without change of context (Strigini and Povyakalo, 2013). We extend that previous result to take into account operational experience from before a context change, to remove the major problem mentioned, of having to discard such evidence even if extensive.

3. Mathematical formulation, notation, and worst-case prior distribution

3.1. Notation

In our scenario, an initial context A is characterized by an unknown *pfd*, represented by the random variable X_A. In A, failure-free operation has been experienced over a certain number of demands n_A. A change to A, believed to improve *pfd* or leave it unchanged (a "*NWTES*" belief), created context B, also with an unknown *pfd*, X_B. n_B demands have occurred in context B without failures. If we ignore n_A, a claim about X_B can be derived by previously published methods (Strigini and Povyakalo, 2013). But we wish to use n_A, and the *NWTES* belief, towards improving this claim. Beliefs about the values of X_A, X_B are described by a joint probability density function (*pdf*). Before it is updated on the basis of the observed n_A, n_B, this is the ***prior*** joint pdf, $f_{AB}(x_A, x_B)$. As noted earlier, specifying such detailed beliefs on a rational basis seems infeasible. We aim to help an assessor who specifies simpler, empirically justified beliefs (constraints defining a *set* of believable prior *pfd* distributions):

- prior confidence in B being "no worse than A" (*NWTES*), represented by a probability ϕ in(1):

  $$P(x_B \leq x_A) = \phi, \tag{1}$$

  where ϕ $(0.5 < \phi \leq 1)$ is the probability associate to the area – the set of (x_A, x_B) values – under, and including, the diagonal line in Figure 1. This confidence depends on the complexity of the system and of the changes, as well as the thoroughness of their analysis. The value of ϕ would be informed by experience of how often the type of analyses supporting the *NWTES* claim have proven wrong, or right, when applied to similar systems and situations. The present paper is concerned with how this confidence, however strong or weak, must affect the estimate of risk for the new context.
- prior confidence, θ $(0.5 < \theta < 1)$, that the system is fault-free (Figure 1). θ is achieved before any inference from operation of the product. E.g., safety-related standards such as (IEC, 2010) prescribe development and verification practices that depend on the required safety level: one has to conclude that the standard writers believe such practices to give some confidence of that safety level being achieved; to estimate θ one could rely on some combination of (a) historical experience of *pfd* levels achieved in systems that are comparable in complexity, functions and /or development processes; and (b) controlled studies about efficacy of verification. We consider the case that such quality evidence is the same for A and B, hence:

  $$P(x_A = 0) = \theta, \; P(x_B = 0) = \theta. \tag{2}$$

We believe the case of $\phi > \theta$ to be common, and thus worth studying first. E.g., adding an extra safety subsystem to a system usually justifies high confidence that safety has been improved: the change is simple and is analyzed with well understood techniques. It is true that historically such intended improvements have at times caused deterioration of safety, but these cases are rare. We would not be surprised if in many projects a claim of $\phi = 95\%$ could be argued. On the other hand, both the forms of evidence for θ in point 2 above are often weak. Any finite amount of operation, say n demands, is inadequate to demonstrate a $pfd \ll 1/n;$ θ is also limited by few systems being trusted to be indeed comparable to the present one; and experimental studies of the effectiveness of verification methods risk overestimating it: faults injected artificially might not be representative of real ones; real faults may be undetected in the experiment and unknown to the experimenters.

We study the posterior reliability, $R_B(n_{Bf}|n_A, n_B)$ (R_B for brevity), of the system for n_{Bf} future demands in context B, after observing runs of n_A and $\mathrm{n_B}$ failure-free demands.

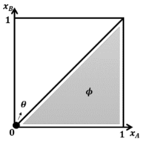

Fig. 1. General form of prior distribution based on (1), (2) over the unit square. Each point (x_A, x_B) identifies an event $(X_A = x_A, X_B = x_B)$. The probabilities θ and ϕ are associated to the point (0,0) and to the area $(x_B \leq x_A)$.

3.2. The minimization problem

In the CBI approach, we seek the worst value for an objective function (reliability, in this paper), among those allowed by the stated prior beliefs (constraints on the prior distributions). That is, Bayesian inference on each prior distribution yields a value for the posterior reliability; Bayesian inference on every prior allowed by the constraints yields a range for it. We seek the worst value from this range. Mathematically, we find, among those priors, one (it may or may not be unique) that yields the infimum (closest lower bound) for that range of posterior reliability values. For brevity we will call such a prior distribution a *worst-case prior*.

In our case of Bernoulli processes, with likelihood function $(x_A, x_B) = (1 - x_A)^{n_A}(1 - x_A)^{n_B}$, the posterior reliability is given by (3). We need to minimize expression (3), subject to the constraints in (1), (2).

$$R_B(n_{Bf}|n_A, n_B) = \frac{\int_0^1 \int_0^1 f_{AB}(x_A, x_B)(1 - x_A)^{n_A}(1 - x_B)^{n_B + n_{Bf}} dx_A dx_B}{\int_0^1 \int_0^1 f_{AB}(x_A, x_B)(1 - x_A)^{n_A}(1 - x_B)^{n_B} dx_A dx_B} \tag{3}$$

Our solution of this minimization problem relies on Theorem 1.

Theorem 1: A prior distribution that yields the infimum of the set of posterior reliability values defined in (3), among priors satisfying the constraints in constraints in (1), (2), is a discrete three-point distribution with probability masses $M_0 = \theta$, $M_1 = 1 - \phi$, and $M_2 = \phi - \theta$, for the points of coordinates (0, 0), (x_{A1}, x_{B1}), (x_{A2}, x_{B2}), as illustrated in Figure 2b. Inference from observing n_A and n_B failure-free demands updates these probabilities. The coordinates (x_{A1}, x_{B1}), (x_{A2}, x_{B2}) that give the lowest R_B (i.e., R_{BW}) depend on the values of the parameters, including n_A and n_B, and identify one of the patterns in Figures 2b to 2e. For the sake of brevity, we will sometimes call "M_1" the point with associated probability M_1, etc, and will use these abbreviations when convenient: R_B for $R_B(n_{Bf}|n_A, n_B)$ and (y_{Ai}, y_{Bi}) for $(1 - x_{Ai}, 1 - x_{Bi})$. Given Theorem 1, the worst-case posterior reliability can always be obtained from a discrete distribution as described, hence (3) takes the discrete form:

$$R_B(n_{Bf}|n_A, n_B) = \frac{\theta + (1 - \phi)y_{A1}^{n_A} y_{B1}^{n_B + n_{Bf}} + (\phi - \theta)y_{A2}^{n_A} y_{B2}^{n_B + n_{Bf}}}{\theta + (1 - \phi)y_{A1}^{n_A} y_{B1}^{n_B} + (\phi - \theta)y_{A2}^{n_A} y_{B2}^{n_B}} \tag{4}$$

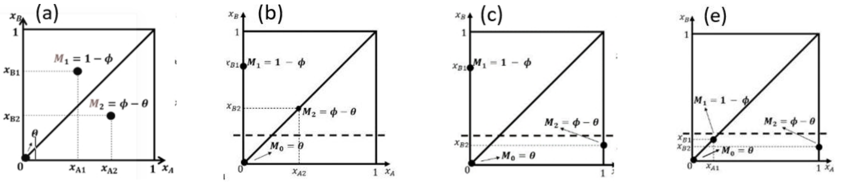

Fig. 2. (a) Discrete 3-point prior distribution satisfying prior beliefs in (1), (2); (b) to (e) show the four patterns of prior distribution (special cases of Fig. 2a) that may yield the lowest posterior reliability, R_{BW}, matching the four expressions given in the proof for R_{BW}; (b) R_{21}, (c) R_{22}, (d) R_{23}, and (e) R_{24}. The dashed lines represent hypothetical values of x, x_{B2}^*, x_{B2}^{**},which have a role in the proof outline.

3.3. Outline of proof

We recall (3) and constraints in (1), (2) as follows. The goal is to solve the optimization problem:

$$\inf_{D} \frac{\int_0^1 \int_0^1 f_{AB}(x_A, x_B)(1 - x_A)^{n_A}(1 - x_B)^{n_B + n_{Bf}} dx_A dx_B}{\int_0^1 \int_0^1 f_{AB}(x_A, x_B)(1 - x_A)^{n_A}(1 - x_B)^{n_B} dx_A dx_B}$$

subject to: $P(x_B \leq x_A) = \phi$, $\quad P(x_A = 0) = P(x_B = 0) = \theta$,

where D is the set of those prior probability distributions for variables X_A, X_B (defined over the square

$[0,1] \times [0,1]$) that satisfy the constraints in constraints in (1), (2). Within D, we seek a distribution that minimizes the posterior reliability (see (3)) for n_{Bf} future demands, given that no failures occurred over n_A and n_B demands in A and B, respectively.

We proved that a prior distribution that yields the infimum for the posterior reliability must match one of the patterns shown in Figures 2b-2e . The patterns shown in Figure 2 encompass subsets ($S_S = \{S_0, S_1, S_2\}$) of the [0, 1] × [0, 1] square: S_0 for the point (0, 0), the rest of the square being divided into S_1 above the diagonal ($x_B > x_A$), S_2 for the lower triangle ($x_B \leq x_A$). The worst reliability is found through these steps:

Step 1: We replace any given prior distribution with a discrete, 3-point distribution (Figure 2a) that yields the same posterior reliability. This determines a subset $D_d \subseteq D$ that must contain a worst-case prior. An item within D_d is actually identified by the values of four variables: $x_{A1}, x_{B1}, x_{A2}, x_{B2}$ (the co-ordinates of the points with probabilities $M_1 = 1 - \phi$, $M_2 = \phi - \theta$). For no item within D_d all partial derivatives of R_B w.r.t. all four variables are 0. Hence, the items representing the worst-case prior lies on a boundary of D_d in this 4-dimensional space.

Step 2: We restrict the pattern in Figure 2a to 4 more specific patterns with M_1, M_2 at the boundaries of the two triangles, above and below the diagonal. That is, when we optimize R_B by adjusting the positions of M_1, M_2 in Figures 2b to 2e, we obtain four possible, distinct patterns for a worst-case prior distribution, with posterior reliability given respectively by (10), (11), (12) and (13).

Step 3: We discard some of these candidate worst case patterns, by comparing the R_B values they yield. Which one is the worst depends on n_B, n_A.

Step 1: We rewrite (3) in terms of sums over $S_i \in S_s$, with $S_s = \{S_0, S_1, S_2\}$, the three subsets of the $[0,1] \times [0,1]$ square:

$$R_B = \frac{\sum_{S_i \in S_s} \iint_{S_i} f_{AB}(x_A, x_B)(1-x_A)^{n_A}(1-x_B)^{n_B+n_{Bf}} dx_A dx_B}{\sum_{S_i \in S_s} \iint_{S_i} f_{AB}(x_A, x_B)(1-x_A)^{n_A}(1-x_B)^{n_B} dx_A dx_B} \tag{5}$$

Using the "Mean Value Theorem for Integrals," we can, without changing R_B, rewrite (5) as (6) with M_i representing discrete probability masses in each region S_i of the unit square (Figure 2a). Recalling the shorthand notations $1 - x_{Ai} = y_{Ai}$, $1 - x_{Bi} = y_{Bi}$ $(i = 1,2)$ the posterior reliability can be written as:

$$R_B = \frac{M_0 + M_1 y_{A1}^{n_A} y_{B1}^{n_B+n_{Bf}} + M_2 y_{A2}^{n_A} y_{B2}^{n_B+n_{Bf}}}{M_0 + M_1 y_{A1}^{n_A} y_{B1}^{n_B} + M_2 y_{A2}^{n_A} y_{B2}^{n_B}} \tag{6}$$

where $M_0 = \theta, M_0 + M_2 = \phi, M_0 + M_1 + M_2 = 1$. Solving this equation system yields:

$$M_0 = \theta, M_1 = 1 - \phi, M_2 = \phi - \theta \tag{7}$$

Step2: Step 1 gave a 3-point prior distribution (Figure 2a). The next step is to minimize R_B with respect to the positions $(x_{A1}, x_{B1}), (x_{A2}, x_{B2})$ of the probability masses M_1, M_2. As mentioned earlier, R_{BW} occurs at a boundary of D_d. Different prior distributions result in different points on this boundary: the four patterns in Figures 2b to 2e.

Substituting probability values from (8) into (6), we obtain a modified expression, R_2.

$$R_2 = \frac{\theta + (1-\phi) y_{A1}^{n_A} y_{B1}^{n_B+n_{Bf}} + (\phi - \theta) y_{A2}^{n_A} y_{B2}^{n_B+n_{Bf}}}{\theta + (1-\phi) y_{A1}^{n_A} y_{B1}^{n_B} + (\phi - \theta) y_{A2}^{n_A} y_{B2}^{n_B}} \tag{8}$$

To minimize R_2, we study its partial derivatives w.r.t. the coordinates of M_1, M_2. Solving $\partial R_2 / \partial x_{A1}$ for x_{B1}:

$$x_{B1}^* = 1 - \left(\frac{\theta + (\phi - \theta) y_{A2}^{n_A} y_{B2}^{n_B+n_{Bf}}}{\theta + (\phi - \theta) y_{A2}^{n_A} y_{B2}^{n_B}}\right)^{\frac{1}{n_{Bf}}}$$

x_{B1}^* identifies the horizontal line in Figures 2b to 2e, which separates two cases: $\partial R_2 / \partial x_{A1} > 0$ (if $x_{B1} \geq x_{B1}^*$), for which R_2 is minimum when $x_{A1} = 0$; and $\partial R_2 / \partial x_{A1} < 0$ (if $x_{B1} < x_{B1}^*$), for which R_2 is minimum when M_1 is on the diagonal $x_{A1} = x_{B1}$. If $x_{B1} = x_{B1}^*$ then R_2 does not vary with x_{A1}. Therefore, we substitute $x_{A1} = 0$ or $x_{A1} = x_{B1}$ into (8) yielding two distinct situations. For each, we study the derivative of R_2 with respect to x_{A2}. Similarly to the process used for x_{A1}, we obtain critical values x_{B2}^* and x_{B2}^{**} for which $\partial R_2 / \partial x_{A2} = 0$. Solving all partial derivatives ($\partial R_2 / \partial x_{A1} = 0$, $\partial R_2 / \partial x_{B1} = 0$, $\partial R_2 / \partial x_{A2} = 0$, $\partial R_2 / \partial x_{B2} = 0$), we end up with four different situations:

- $\partial R_2 / \partial x_{A1} > 0$ $(x_{B1} \geq x_{B1}^*)$ and $\partial R_2 / \partial x_{A2} > 0$, $(x_{B2} > x_{B2}^*)$, with

$$x_{B2}^* = 1 - \left(\frac{\theta + (1-\phi) y_{B1}^{n_B+n_{Bf}}}{\theta + (1-\phi) y_{B1}^{n_B}}\right)^{\frac{1}{n_{Bf}}}$$

So, to minimize R_2, M_2 must be on the diagonal $x_{A2} = x_{B2}$ (Figure 2b). The reliability is R_{21} in (9).

$$R_{21} = \frac{\theta + (1-\phi)y_{B1}^{n_B+n_{Bf}} + (\phi-\theta)y_{B2}^{n_A+n_B+n_{Bf}}}{\theta + (1-\phi)y_{B1}^{n_B} + (\phi-\theta)y_{B2}^{n_A+n_B}} \tag{9}$$

- $\partial R_2/\partial x_{A1} > 0$ $(x_{B1} \geq x_{B1}^*)$ and $\partial R_2/\partial x_{A2} < 0$ $(x_{B2} < x_{B2}^*)$: minimizing R_2 requires $x_{A2} = 1$ (Figure 2c).

$$R_{22} = \frac{\theta + (1-\phi)y_{B1}^{n_B+n_{Bf}}}{\theta + (1-\phi)y_{B1}^{n_B}} \tag{10}$$

- $\partial R_2/\partial x_{A1} < 0$ $(x_{B1} < x_{B1}^*)$ and $\partial R_2/\partial x_{A2} > 0$ $(x_{B2} > x_{B2}^{**})$:

$$x_{B2}^{**} = 1 - \left(\frac{\theta + (1-\phi)y_{B1}^{n_A+n_B+n_{Bf}}}{\theta + (1-\phi)y_{B1}^{n_A+n_B}}\right)^{\frac{1}{n_{Bf}}}$$

In this case, to minimize R_2, both M_1 and M_2 must be placed on the diagonal line (see (11), Figure 2d).

$$R_{23} = \frac{\theta + (1-\phi)y_{B1}^{n_A+n_B+n_{Bf}} + (\phi-\theta)y_{B2}^{n_A+n_B+n_{Bf}}}{\theta + (1-\phi)y_{B1}^{n_A+n_B} + (\phi-\theta)y_{B2}^{n_A+n_B}} \tag{11}$$

- $\partial R_2/\partial x_{A1} < 0$ $(x_{B1} < x_{B1}^*)$ and $\partial R_2/\partial x_{A2} < 0$ $(x_{B2} < x_{B2}^{**})$: Minimizing R_2, requires $x_{A2} = 1$ (Figure 2e)

$$R_{24} = \frac{\theta + (1-\phi)y_{B1}^{n_A+n_B+n_{Bf}}}{\theta + (1-\phi)y_{B1}^{n_A+n_B}} \tag{12}$$

Step 3: Comparing these four possible expressions for R_{BW}, it is easily shown analytically that $R_{22} > R_{21}$, $R_{24} > R_{21}$, and $R_{24} > R_{23}$ for any values of the parameters. We could not prove a similar inequality between R_{23} and R_{21}: which one yields the lower R_B value, hence R_{BW}, for given parameter values, can be checked on the numerical results. As an illustration, Figure 3 compares these four candidates for R_{BW}, for one set of values of θ, ϕ, n_B, n_{Bf}.

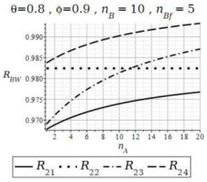

Fig.3. Comparison of four candidates for R_{BW}. For the parameter values shown, $R_{BW} = R_{21}$.

4. Results

We illustrate the results about worst-case posterior reliability with the aid of numerical examples. In these examples, the values of n_A, n_B and n_{Bf} match a decision scenario in which there may be a large amount of previous, failure-free operation in context A; assurance is sought for a future amount of operation in context B, $\mathrm{n_{Bf}}$, comparable to the amount $\mathrm{n_B}$ already observed in context B. For these ranges of parameter values the prior distributions yielding the worst posterior reliability R_{BW} have the form shown in Figure 2b, and the corresponding reliability is given by (9), in the proof outline.

4.1. Improved reliability claims from use of operational experience from previous context

Figure 4a compares the reliability R_{BW} that can be conservatively claimed for context B without the aid of the evidence $\mathrm{n_A}$ from context A and with it (the former is calculated using the earlier results (Strigini and Povyakalo, 2013), the latter from results in this paper). As expected, using n_A allows a claim of higher reliability, for any value of n_B. E.g., when not using n_A $(n_A = 0)$, for $n_B = 200$, $R_{BW} = 0.9665$, but with the same n_B, $R_{BW} = 0.9824$ for $n_A = 1000$: the probability of future failures is almost halved.

Conversely, using n_A reduces the amount n_B of operational experience in B needed for claiming a certain value of R_{BW}. Given high enough n_A and ϕ, sufficient confidence to operate in a new context can be obtained much sooner after a change. In Figure 4a, considering $\mathrm{n_A}$ halves the amount of operational experience in context B ($n_B = 200$ vs $n_B = 400$) needed to achieve $R_{BW} = 0.98$.

There may even be situations in which n_A alone supports a claim of sufficient reliability after the change, requiring no operational experience $\mathrm{n_B}$ at all. Standards would still require some testing for context B; this good practice would support the claimed arguments for the values of θ and ϕ. But the safety argument will not require operational testing meant to be statistically representative of future use in B. This decision criterion of not requiring statistics of operation in B if n_A is seen as large enough seems to be often applied informally, without probabilistic arguments, for changes considered small, or in the FDA's criterion of "substantially equivalent". In our example (Figure 4a), a requirement of 85% reliability at $n_{Bf} = 100$ is satisfied on the strength of n_A evidence alone: for $n_B = 0$, $R_{BW} \leq \theta = 0.8$ (lower curve in Figure 4a), and for $n_A = 1000$, $R_{BW} \approx 0.89$ (R_{BW} at most $\theta/(1-\phi+\theta)$). If $n_A = n_B = 0$, the worst-case probability of no failures in n_{Bf} demands is indeed θ since, with probability θ, $X_B = 0$ (failures are impossible), but otherwise, the *pfd* X_B may be as high as 1 (every demand causes failure).

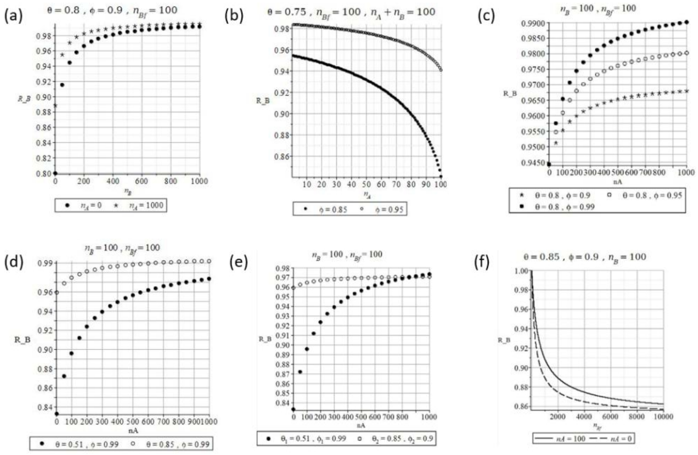

Fig. 4 Analyzing the impact of old-context evidence, n_A, of ϕ, and of θ on conservative reliability claim R_{BW} for context B. (a) R_{BW} improvement with n_B alone vs using n_A as well. (b) How R_{BW} decreases according to which fraction of the operational experience predated the change. (c) Effect of varying ϕ. (d) Effect of varying θ. (e) For high enough n_A , a high enough ϕ compensates for lower value of θ. (f) Decrease of R_{BW} with n_{Bf}. In $n_{Bf} = 10{,}000$, $(R_{BW}|n_A = 100) = 0.8625$, $(R_{BW}|n_A = 0) = 0.8572$.

We sketch an explanation of the effects seen in Figure 4a, without the mathematical detail provided in the proof outline, but accepting as proven that to find a worst-case prior it is sufficient to consider 3-point discrete distributions.

Recall that in Bayesian inference: (1) the probabilities of events (here, of (x_A, x_B) points) are updated proportionally to their likelihoods (the probability of the observation (n_A, n_B), conditional on that event); here, the likelihood of n_A failure-free demands is maximum for $X_A = 0$; the same applies to $\mathrm{n_B}$ and X_B. (2) inference can be done in steps, each step applying parts of the evidence: we can first obtain a posterior distribution for X_A, X_B on the basis of the prior and of the evidence n_A, while assuming $n_B = 0$, and then use this posterior distribution as a one-dimensional prior distribution for X_B, to obtain the posterior reliability in context B, R_B. For this second step, previous results (Strigini and Povyakalo, 2013) show that a worst-case prior has non-zero probabilities for only two values of X_B: zero, and a non-zero value (found by numerical optimization), as shown in Figure 5.

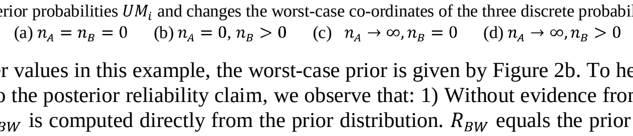
(a)
x_B
1
$M_1 = 1 - \phi$
$M_2 = \phi - \theta$
$M_0 = \theta$
0
1
x_A
(b)
x_B
1
x_B
UM_1
UM_2
UM_0
0
1
x_A
(c)
x_B
1
$UM_1 = \frac{1-\phi}{1-\phi+\theta}$
$UM_0 = \frac{\theta}{1-\phi+\theta}$
0
1
x_A
(d)
x_B
1
$UM_1 < \frac{1-\phi}{1-\phi+\theta}$
$UM_0 > \frac{\theta}{1-\phi+\theta}$
0
1
x_A

Fig. 5: Posterior probability distributions that yield lowest R_B. Inference from (n_A, n_B) updates the prior probabilities M_i of the three events to the posterior probabilities UM_i and changes the worst-case co-ordinates of the three discrete probability masses.
(a) $n_A = n_B = 0$ (b) $n_A = 0, n_B > 0$ (c) $n_A \to \infty, n_B = 0$ (d) $n_A \to \infty, n_B > 0$

For the parameter values in this example, the worst-case prior is given by Figure 2b. To help visualize how n_A and n_B contribute to the posterior reliability claim, we observe that: 1) Without evidence from operation, i.e., for $n_A = 0$, $n_B = 0$, R_{BW} is computed directly from the prior distribution. R_{BW} equals the prior probability $M_0 = \theta$ (Figure 5a); 2) With $n_A = 0$, inference from $n_B > 0$ updates M_0 to $UM_0 > M_0$, hence increases R_B for any prior, and thus also the infimum R_{BW}. This yields the lower curve in Figure 4a. The worst-case values of X_{B1}, X_{B2} are equal (Figure 5b) and decrease with increasing n_B (Strigini and Povyakalo, 2013); 3) We now consider the effect of n_A alone, with $n_B = 0$. This is complicated, but for large n_A values we can visualise it by just studying the limit behavior of R_B (with $n_B = 0$). For $n_A \to \infty,$ the posterior probability of any event with $X_A > 0$ tends to 0, and the posterior mass distribution tends to: $UM_2 = 0$, $UM_1 = (1-\phi)/(1-\phi+\theta)$, $UM_0 = \theta/(1-\phi+\theta)$ (Figure 5c). Analogously to the case of $n_A = n_B = 0$, now $R_{BW} = UM_0$, and since $UM_0 > M_0$, using n_A improved R_{BW}. Note that this UM_0 value bounds the reliability that can be claimed for B before seeing operation of B; 4) With n_A large enough for the updated probability values UM_i to be close to the limit values above, inference from $n_B > 0$ increases posterior reliability, starting with prior $UM_0 = P(X_B = 0) = \theta/(1-\phi+\theta)$ (Figure 5d); without considering n_A, it would have started with a smaller $P(X_B = 0) = \theta$. Hence the higher reliability curve in Figure 4a.

4.2. Pre-change evidence vs post-change evidence

Safety claims are sometimes based on the total amount of past successful operation, $n_A + n_B$, ignoring changes of context. But evidence of operational success in the old context A must be somewhat "discounted", compared to the direct evidence from the current context B. Figure 4b shows how a reliability claim must be reduced depending on how much of the favorable evidence comes from the old context A. This reduction is due to the limited confidence $P(X_B \le X_A) = \phi$ in B being actually no worse than A. The figure also shows how higher values of ϕ allow higher claims.

4.3. Effects of parameters ϕ and θ

Using operational experience from A to estimate reliability in B is useful, but the extent of this benefit depends on the strength of the belief, represented by the probability ϕ in context B being no worse than A.

Figure 4c shows how ϕ affects reliability for context B. The plots correspond to different value of ϕ, but the same value of θ. Higher values of ϕ lead to increased reliability in context B. This should be expected: a higher ϕ-- greater confidence that B is no worse than A-- implies that evidence from A must have a stronger effects on our beliefs about B's reliability.

Figure 4d shows instead the effect of θ, representing evidence common to both contexts. A higher θ implies higher reliability (for any given ϕ). This holds for $n_A = 0$, and remains true throughout the plot.

R_{BW} increases both with ϕ and with θ: thus, can a high enough ϕ compensate for a lower value of θ, i.e., yield a higher R_{BW} despite the lower θ? This seems reasonable: a higher ϕ means that successful operation of A contributes more to confidence in B. Figure 4e shows an example, with two plots with different values of ϕ and θ: $(\theta_1 = 0.51, \phi_1 = 0.99)$ and $(\theta_2 = 0.85, \phi_2 = 0.9)$, with $\theta_1 < \theta_2$. For $n_A = 0$, a higher θ implies higher reliability, irrespective of the value of ϕ. As n_A increases, though, the plot corresponding to lower θ ($\theta_1 = 0.51$) intersects and surpasses the plot with higher θ ($\theta_1 = 0.85$). This situation arises if $\phi_1 > (\phi_2\theta_1 - \theta_1 + \theta_2)/\theta_2$.

4.4. How posterior reliability decreases with n_{Bf}

Figure 4f illustrates some key points. Reliability of course decreases with increasing $\mathrm{n_{Bf}}$ (for any given values of n_A and n_B), but never reaches 0: it is a weighted sum of the probabilities of having no failures if $X_B = 0$ (which is 1) and if $X_B > 0$ (which tends to zero as n_{Bf} increases). The probability of $X_B = 0$ is never reduced by observing failure-free demands. So, from (4), $lim_{n_{Bf}\to\infty} R_B = \theta/(\theta + (1-\phi)y_{A1}^{n_A}y_{B1}^{n_B} + (\phi-\theta)y_{A2}^{n_A}y_{B2}^{n_B})$. The limit's worst value is θ ($\underset{n_{Bf}\to\infty}{lim} R_{BW} = \theta/(\theta + (1-\phi) + (\phi-\theta)) = \theta$), which is obtained when M_1 and M_2 move to (0, 0).

5. Discussion, conclusion and future work

We have extended a previous conservative reliability prediction method (Strigini and Povyakalo, 2013) by accounting for operational experience that comes in part from a previous context; and extended other previous work, on using pre-change evidence to assess *pfd* (Littlewood et al., 2020), to assessing reliability over multiple future demands.

Essential observations are: (1) evidence of correct/safe operation before a change does contribute to reliability type claims about correct/safe operation after the change; (2) but this contribution is less than that from operation after the change, unless there is absolute certainty ($\phi = 1$) that the change was not a change for the worse; (3) the advantage thus produced is also limited: even an infinite amount of evidence from before the change would not, alone, allow claims of perfect reliability after the change, but only improve confidence in *pfd*=0 after the change, but prior to observing operation, from θ to ϕ.

Dismissing operational evidence from before the change would be over-pessimistic, while regarding it as though it were about the context after the change would be misleadingly optimistic.

Our results may substantially improve the task of proving sufficiently low risk from systematic failures for operation of systems that: underwent changes (in the system or its use) after a long phase of successful/safe operation; require confidence of no failures occurring over large amounts of future operation; are subject to a wide range of uncertainty about their *pfd*; were thoroughly verified for absence of design faults.

Some natural extensions of this work concern studying the posterior reliability under different forms of prior beliefs, as may be justified in different practical situations. Indeed, the previous studies cited in Section 2.1 showed how the apparently simple natural language statement “context B is no worse than context A” actually allows different mathematical meanings, that apply in different real-world situations, of which here we studied one. Worthwhile extensions of this kind include: (1) the generalized scenario in which there is confidence that the *pfd* is less than some non-zero upper bound. In this paper, we assumed a prior belief of fault-freeness: there is some confidence that the *pfd* is 0. Indeed this is an important case, as we argued, but a more general case is that there is confidence that the *pfd* is less than some non-zero bound, e.g., $P(X_B < \epsilon) = \theta$. While, for some values of the other parameters, assuming $\epsilon = 0$ gives good approximate results, for others the exact solution must be computed; (2) the other generalization in which the confidence in the *pfd* not exceeding this bound differs between A and B, as addressed in a similar scenario by (Zhao et al., 2020); (3) the form of *NWTES* belief that we defined (Aghazadeh Chakherlou et al., 2022) to cover design improvements for safety or fault-tolerance where the confidence in the change being an improvement holds independently of the actual value of the *pfd* before the change; (4) cases of $\theta > \phi$: stronger confidence in the system being highly reliable than confidence in the latest change not having made it worse.

Another important extension concerns successions of changes. Some systems evolve rapidly, e.g., currently, autonomous vehicles: the amount of operation between design updates would never give strong statistical support to safety claims, yet a rational way of accounting for all these amounts of operation collectively in a safety claim would be highly desirable. Last but not least, these methods need to be extended to the case in which the evidence includes some occurrences of failures.

Acknowledgements

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

References

Aghazadeh Chakherlou, R., Salako, K., Strigini, L. 2022. Arguing safety of an improved autonomous vehicle from safe operation before the change: new results. Proceedings - 2022 IEEE International Symposium on Software Reliability Engineering Workshops, ISSREW 2022, 307–312.

Atwood, CL., LaChance, JL., Martz, HF., Anderson, DJ. 2003. Handbook of parameter estimation for probabilistic risk assessment. U.S. Nuclear Regulatory Commission.

Bertolino, A., Strigini, L. 1998. Assessing the Risk due to Software Faults: Estimates of Failure Rate versus Evidence of Perfection. Journal of Software testing , verification and reliability 8.

Bishop, P. , Povyakalo. A., Strigini. L. 2022. Bootstrapping confidence in future safety from past safe operation. 2022 IEEE 33rd Int. Symp. on Software Reliability Engineering (ISSRE), Charlotte, NC, 97–108.

FAA. 1988. Federal Aviation Administration, System design and analysis, Advisory Circular, AC 25.1309-1A.

FDA. 2014. US Food and Drug Administration: The 510(k) program: Evaluating substantial equivalence in premarket notifications [510(k)] guidance for industry and food and drug administration staff.

IEC. 2010. IEC 61508, Functional Safety of Electrical/ Electronic/Programmable Electronic Safety Related Systems, International Electrotechnical Commission.

ISO. 2018. ISO 26262, Road vehicles — Functional safety. International Organization for Standardization

Littlewood, B. 1993. Validation of Ultra-High Dependability for Software-based Systems. Communications of the ACM 36(11), 69-80, https://doi.org/10.1145/163359.163373.

Littlewood, B., Salako, K., Strigini, L., Zhao, X. 2020. On reliability assessment when a software-based system is replaced by a thought-to-be-better one. Reliability Engineering & System Safety, 197, 106752. https://doi.org/10.1016/J.RESS.2019.106752

Minister responsible for transport. (2021). French Decree on automated vehicles’ conditions of use and automated road transport systems’ commissioning.

Minister responsible for transport. 2022. Technical guide related to « GAME » demonstration for ARTS Automated Road Transport System.

Salako, K., Strigini, L., Zhao, X. 2021. Conservative Confidence Bounds in Safety, from Generalised Claims of Improvement and Statistical Evidence. Proceedings - 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2021, 451–462. https://doi.org/10.1109/DSN48987.2021.00055

Strigini, L., Povyakalo, A. A. 2013. Software fault-freeness and reliability predictions : Proceedings SAFECOMP 2013, Toulouse, France, September 24-27.

Strigini, L., Wright, D. 2014. Bounds on survival probability given mean probability of failure per demand; And the paradoxical advantages of uncertainty. Reliability Engineering and System Safety, 128, 66–83. https://doi.org/10.1016/j.ress.2014.02.004

Zhao, X., Robu, V., Flynn, D., Salako, K., Strigini, L. 2019. Assessing the Safety and Reliability of Autonomous Vehicles from Road Testing. Proceedings - International Symposium on Software Reliability Engineering, ISSRE, 2019-October, 13–23. https://doi.org/10.1109/ISSRE.2019.00012

Zhao, X., Salako, K., Strigini, L et al. 2020. Assessing Safety-Critical Systems from Operational Testing: A Study on Autonomous Vehicles. J. of Information and Software Technology 128, 106393.