Letter to the Editor

From: Bev Littlewood and Lorenzo Strigini

Subject: A Critical Response to a Recent Paper by Daniels and Tudor

It is good to see the new Safety-Critical Systems Club eJournal's first issue contributing to debate about the use of probabilistic models to evaluate the dependability of software-based systems. Such evaluation is particularly important for safety-critical systems, especially those whose failures may have massive consequences. Society needs assurances that such systems are "good enough". It is therefore appropriate that methods for providing these assurances are subject to critical examination, as Daniels and Tudor do here (Daniels and Tudor 2022) — hereafter D&T for brevity.

D&T refer to moves within a civil aviation technical community towards giving more weight, in assessment, to records of good behaviour of the product, and perhaps less weight to records of precautions in development. They do not give enough detail for others to agree or not on whether these moves would be good or bad. We thus limit our discussion to their general claims against use of statical testing and operational evidence.

Unfortunately, whilst some of the D&T claims *are* valid, we think some are not. For example, they cite the requirement for some failures of systems in aircraft to be extremely improbable. They rightly point out that the required levels cannot be assured via statistical modelling — they cite our own old paper (Littlewood and Strigini 1993) and that of others (Butler and Finelli 1993) in support. However, they go on to claim that they have identified "*an alternative way forward*" to statistical reasoning "*that does provide evidence that software is safe for its intended use before it enters service*". Both some of their opposition to statistical argument, and this latter claim, seem wrong, as we argue below. Some of our reasoning is reprised from our later paper (Littlewood and Strigini 2011).

There are some levels of reliability requirement (what we called "ultra-high" in our 1993 title) that cannot be demonstrated in advance to be satisfied. We believe this will remain true. However, statistical and probabilistic reasoning helps to determine what one can reasonably believe, and thus to answer the question, among others, whether the level of risk associated with a system is socially tolerable.

The D&T objections to statistical testing mix matters of *model misuse*, of *model validity* and of *adequacy of evidence*.

Some of their arguments are about misuse of methods. In their example of a system whose behaviour depends on the date, assessment indeed requires a proper sample of the space of all dates on which it will be used. Not procuring this, as they hypothesise, is a textbook mistake. Although they do not cite instances of such misuse occurring in practice, if it were common, it would be indeed a concern; but not a reason for giving up the advantages of statistical evidence and statistical reasoning.

"Model validity" concerns the first stage in evaluation of reliability: building a probabilistic model that represents, with acceptable approximation, the real-world properties of a stochastic process of failures. One such model is the Bernoulli Process (BP) for demand-based systems, which D&T criticise in some detail as inadequate for many real-world situations. In the BP model a system responds to a series of "demands", and each response may be either a *failure* or a *success*. The model is based on two assumptions: the outcomes of executions on successive demands are statistically independent, and the probability of failure on a demand, *p*, is constant for the successive demands. These assumptions are, of course, not always appropriate. They are, though, in many practical situations. For example, in the long association we have had with scientists and engineers in the nuclear industry, there has been wide acceptance of the BP model for reactor protection systems. We believe this is reasonable, because the two underpinning assumptions seem plausible there: demands are infrequent, so it can be expected that the outcome of a demand will be independent of the outcome of the previous demand, that might have taken place a year earlier; and there are good arguments for assuming *p* to remain approximately constant over some extent of time.

BP models are indeed not always appropriate, and for these situations there is now an extensive literature about more general models that address such issues, for example by relaxing the assumption of constant *p*. D&T do not address these, or other even simpler solutions for the problems that they cite as showing that BP models for software failures are "*fundamentally flawed*". For example, how to deal with software that has state has been publicly explained at least as far back as in our 1997 work for ESA, the European Space Agency (Strigini and Littlewood 1997).

"Adequacy of evidence" concerns the next, statistical stage of evaluation. For instance, in BP models, one uses collected evidence (successes and failures of a system, in operation or in statistical testing) to improve one's estimate of the parameter *p*, the (constant) probability of failure on demand. This will be unknown, and statistical inference about its value is needed to support probabilistic predictions about future failure behaviour of the system. With safety-critical systems, there is particular interest in the situation where there have been very many demands with no failures. This represents "best news" statistical evidence in support of a claim that a system is "good enough". See for example the Table from IEC 16508 Annex D to Part 7 which D&T reproduce[1]. D&T rightly give particular attention to this special case. They remind the reader of what many, we among them, have been pointing out: there are limits to how much evidence of failure-free execution can realistically be obtained, with consequent limits on the levels of reliability that can be assured. Thus when, for instance, a requirement of extreme improbability of certain failures is explained as "Extremely improbable failure conditions are those having a probability on the order of 1×10^{-9} or less" (as in FAA AC-1309-A, and similar documents), collecting enough evidence is currently seen as infeasible.

On the other hand, not all safety-critical systems have ultra-high reliability requirements. IEC 61508 does consider low SIL levels, for instance. The software-based primary protection system of the Sizewell B reactor was only required to have a probability of failure on demand no worse than 10^{-3} (there were other protections in the wider systems, including a hard-wired secondary system). This goal was sufficiently modest that it was eventually demonstrated with high confidence via statistical testing.

So, what can be done about systems with unassurable ultra-high reliability requirements, such as those for airplanes?

We agree with D&T that the record of in-flight operation of some critical systems, over many years, is extraordinarily impressive. So it seems that ultra-high reliability *may* have been *achieved* here, as evidenced after massive operational use. But this is rather different

[1] Note that this table from the IEC standard presents *2-sided* confidence intervals for *p*. In fact they should instead be *1-sided* confidence bounds: a user wishes to know how confident they can be, for given evidence, that *p* is *smaller* than a certain bound. They have no interest in knowing how confident they can be that *p* is larger than a bound. D&T have no comment on this.

from claiming high confidence in such ultra-high reliability *before* a system enters service. Nevertheless D&T say: "We now have nearly 30 years of service experience that satisfying the objectives of RTCA/DO-178B/C has been sufficient to address the software considerations in aircraft certification."

And go on to say: "…this paper proposes an alternative way forward that does provide evidence that *software is safe for its intended use before it enters service*." (Our italics)

We found the authors' reasoning at this point rather vague and hard to follow. The argument that 30 years of experience prove something *is* a statistical argument, but stated in a rather hand-waving manner.

D&T may mean simply that regulators agree to accept application of RTCA/DO-178B/C in lieu of evidence of satisfying 10^{-9}: a simple fact. Or they may mean that applying the standards assure that result, so that it can be taken as *evidence* of having *achieved* it. This is a bold claim. Can they actually prove it? To do so, they would need to show that the 10^{-9} objective has indeed been achieved, consistently over most systems (a hard claim to demonstrate for most of them); next, that the attainment is linked to applying RTCA/DO-178B/C prescriptions in such a way that we should consider RTCA/DO-178B/C compliance sufficient evidence. At this point they could claim a *probability* that the methods will produce satisfaction of the requirements in the next aircraft type developed. This probability, dependent on the numbers of such successes and of systems, will be a rough estimate or range, certainly, but would help to see how strong their claim is.

Evidence of good practice and diverse forms of verification is indeed a valid part of an argument for high reliability or safety. Various authors, ourselves included, have proposed ways for clarifying *how much* they can contribute to sound quantitative arguments for reliability or safety; see for example, Bishop, Bloomfield, et al, (2011), Strigini & Povyakalo (2013) and Littlewood, Salako, et al (2020). A rough estimate of probability of achieving the requirement through application of RTCA/DO-178B/C precautions would fit well in such reasoning. It would almost certainly still imply an excessive probability of catastrophic failures in a type's lifetime, yet this could be reduced with statistical evidence from testing and operation.

An especially negative effect of D&T's argument is that, while they oppose statistical reasoning for ultra-high requirements (because – paraphrasing – it will refuse to deliver the reassuring statements that one may wish to hear, albeit giving *useful* information about *how much* has actually been demonstrated), they then decry statistical methods much more generally. Yet for many systems with more modest reliability and safety requirements, quantitative assurance can indeed be effectively obtained with statistical testing and simple probability models, such as the Bernoulli Process (and its continuous time equivalent, the Poisson Process). Even safety-critical systems often fall into this category (e.g. see our earlier example of the Sizewell B nuclear reactor protection system). Thus D&T's broad opposition to reliability modelling for software runs the risk of encouraging system builders to eschew proper quantitative evaluation in favour of informal qualitative arguments: e.g. "you can trust this system to be safe enough, because we used accepted best practice in building it." Such hand-waving justification is not good enough for highly critical systems.

(Emeritus Prof) Bev Littlewood

(Prof) Lorenzo Strigini

Centre for Software Reliability

City, University of London

29 May 2022

References

Bishop, P. G., Bloomfield, R. E., Littlewood B., Povyakalo A. and Wright D. R. (2011). *Toward a Formalism for Conservative Claims about the Dependability of Software-Based Systems*. IEEE Trans Software Engineering, 37(5), pp. 708-717. doi: 10.1109/TSE.2010.67

Butler, R. W. and Finelli, G. B. (1993). *The infeasibility of quantifying the reliability of life-critical real-time software*. IEEE Trans Software Engineering 19(1): 3-12.

Daniels, D. and Tudor, N. (2022). *Software Reliability and the Misuse of Statistics*. Safety-Critical Systems eJournal 1(1).

Littlewood, B. and Strigini, L. (1993). *Validation of ultra-high dependability for software-based systems*. CACM 36(11): 69-80.

Littlewood, B. and Strigini, L. (2011). *'Validation of Ultra-High Dependability' — 20 years on*. SCSC Newsletter 20(3).

Littlewood, B., Salako, K., Strigini, L. and 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]. doi: 10.1016/j.ress.2019.106752

Strigini, L. and Littlewood, B. (1997). *Guidelines for Statistical Testing* (PASCON/WO6-CCN2/TN12). ESA/ESTEC project PASCON. https://openaccess.city.ac.uk/id/eprint/254/2/StatsTesting_TN12-3.1distrib2.pdf

Strigini, L. and Povyakalo, A. (2013). *Software fault-freeness and reliability predictions*. Proc. SAFECOMP 2013. (pp. 106-117). Cham: Springer. ISBN 978-3-642-40792-5