Conservative reliability assessment of a 2-channel software system when one of the channels is probably perfect

Peter Popov

Centre for Software Reliability, City, University of London,

Northampton Square, London, EC1V 0HB, United Kingdom

p.t.popov@city.ac.uk

ORCID: 0000-0002-3434-5272

Abstract

In this paper we subject to scrutiny some recent advances in conservative reliability assessment of 2-channel fault-tolerant software, based on the probability of perfection of one of the channels. Our approach extends the previous works by looking in detail at the implications of the assumptions made in these previous works about the relationships between the probability of failure of the channels and of the system, which have not been explored before. We demonstrate that the assumptions made by others impose significant *constraints* on the epistemic uncertainty of the probability of system failure and explore the implications of these constraints to derive new conservative bounds.

An important difference of this work from the prior works is that we use a *white-box model* of a 2-channel system, while in the previous works a *black-box system model* was used. We discuss the limitations of an assessment based on a black-box model and compare our conservative results with those, derived by others using a black-box system model.

Keywords: Bayesian inference, protection system, two-channel software system, black-box and white-box models, probability of failure on demand, probability of perfection, epistemic uncertainty.

1. Introduction

Reliability assessment of safety critical software is an essential part of developing safety-critical systems. Assessment is expected to demonstrate that critical software, as a part of a wider safety-critical system, is fit for purpose typically by showing that a software reliability target is met, i.e., that software reliability is no worse than the stated reliability target.

Conservative assessment is a widely practiced and prudent way of dealing with various uncertainties in development and in the assessment of reliability. In conservative assessment, instead of trying to establish an *accurate* estimate of software reliability, the assessor is focused on deriving an estimate, which is known *with certainty* to be pessimistic, i.e., worse than the unknown system reliability, e.g., the probability of system failure.

When a system is assessed conservatively and judged to be "good enough" (i.e., that the conservative estimate is better than the reliability target), the assessment is not merely acceptable. In this case the fact that a conservative reliability estimate is better than the target, provides *extra assurance* that the true system reliability is better than the reliability target, which is a highly desirable outcome of the reliability assessment. In some cases, however, the conservative assessment may produce estimates which are short of meeting a set reliability target. In such cases, the assessor is faced with a *dilemma*: is the system really not reliable enough or instead it is merely that the conservative reliability estimate is *too conservative*?

This paper does not offer a solution to the above dilemma and instead explores the factors which impact the conservative software reliability assessment such as the *system model* used in the assessment and *the assumptions* under which the conservative assessment is conducted.

The main contribution of this paper is highlighting the importance of scrutinising carefully *all aspects* used in a conservative reliability assessment of a 2-channel software system and in demonstrating that this effort pays-off:

- Selecting a white-box system model[1] which seems more complex in comparison with the simpler black-box

[1] In this paper the term "white-box model" is used merely to signify that the system architecture, which in this case consists of 2 separate channels, is accounted for. The "white-box" system model, therefore, is more detailed than a "black-box" system model, which ignores

model, may offer an insight that the simpler "black-box" model simply lacks.

- Scrutinising fully the assumptions made in conservative assessment may lead to the discovery of non-obvious implications of the conservatism, e.g., that the assumptions are *implausible*. This is important as then an assessor may need to trade off the benefits from a conservative assessment against its plausibility.
- The paper spells out in detail how Bayesian reliability assessment can be applied to a 2-channel software system using a *white-box system model* with a set of conservative assumptions, defined by others in the past.
- We explore the space of conservative *multivariate priors*, suitable for predictions with a white-box model, consistent with the conservative assumptions defined by others in the past and identify *the most conservative* prior, which may be used as a *lower bound* for conservative predictions.
- We offer useful *analytical results* which include:
    - How conservative assumptions are affected by the Bayesian inference for different observations of a 2-channel system in operation/testing. We confirm that the conservative assumptions made in constructing the prior are retained in the posterior for the case of "no failures", i.e., the predictions retain the properties of a conservative distribution. However, for the other possible observations – of channel and system failures – the conservatism is either *not guaranteed* (i.e., the predictions may become optimistic) or is *not useful*. These results extend the previous work on conservative Bayesian assessment.
    - Some of the analytical results reported from the white-box conservative assessment are *counterintuitive*, e.g., failures of the channel, not assumed perfect, will lead to predictions that the *system is perfect with certainty*.
- We also scrutinised the recent works by others based on a *black-box system model* and in the process make a number of contributions:
    - The *black-box* system model does not respond adequately to observations with channel failures; In some cases inadequacy may become very significant, e.g. when the channel considered possibly perfect fails.
    - The *black-box* system model may produce more optimistic predictions than a white-box model: we provide evidence of this possibility for both "no failure" observed in operation/testing and for cases when operation/testing reveals channel and system failures.

Although the paper is primarily focused on software reliability assessment, the approach we present can be applied to the broader class of systems *with design faults*, in which one or even both channels are implemented in hardware. Subtle design faults in hardware designs may lead to failures which are difficult to detect and very rare in operation. Complex hardware designs are developed with complex tools and may include "formal" proofs [3] leading to explicit or implicit statements of design perfection (e.g., "this design has been formally verified"). An example of complex hardware designs are protection systems implemented with field programmable gate arrays (FPGA). Designs may contain design faults introduced either by the respective designers in setting the high-level requirements, or due to subtle faults in the software tools used to develop the logic placed in an FPGA and to program the FPGA fabric. Conservative assessment based on the idea that one of the channels might be perfect can be used for such systems, too. We hope, therefore, that the work we present would be of interest to a wider audience, not merely to the community of colleagues interested in rigorous conservative software reliability assessment.

The paper is organised as follows: Section 2 provides the motivation for the study; in Section 3 we develop the white-box system model under a small number of conservative assumptions about the channels' reliabilities, which results in defining a 3-variate distribution consistent with the conservative assumptions made by others. In Section 4 we provide the main results of the paper – a set of results related to the conservative predictions of system reliability: i) for the special case of operation/testing when no failures occur, and ii) for the case of system operation/testing with arbitrary observations (including channel and system failures). In Section 5 we offer a contrived example illustrating the application of the approach in practice and highlight the difference between the predictions obtained with the black-box and the white-box models using conservative priors. In Section 6 we discuss the findings and the threats to their validity. In Section 7 we discuss the related research. Section 8 concludes the paper and outlines directions for future research.

details about the system architecture. Our use of the term "white-box" is consistent with [1]. In other literature sources, e.g., dealing with software testing, the term "white-box" is often used to refer to different degrees of knowledge about the tested software, e.g., whether the source code is available or not, etc. [2] makes use of the structure of the code to develop a Bayesian method of reliability assessment. Our "white-box" system model neither requires nor assumes such detailed knowledge.

2. Motivation

This work was prompted by the recent effort by colleagues at the Centre for Software Reliability, at City, University of London, who explored the idea of conservative Bayesian assessment of a 2-channel on-demand software based on the assumption of a *probable perfection* of one of the channels.

The acronyms and notations used throughout the paper are listed in Table 1.

Table 1. Acronyms and notations

pdf	Probability density function
Cdf	Cumulative distribution function
pfd	Probability of failure on demand
$f_P(\cdot)$, $F_P(\cdot)$	Probability density and cumulative distribution functions, respectively, of the random variable *P*.
P_A	Probability of failure on demand of channel A (a random variable).
P_B	Probability of failure on demand of channel B (a random variable).
P_{AB}	Probability of simultaneous failure on demand of channel A and channel B (a random variable).
$f_{P_A}(\cdot)$, $F_{P_A}(\cdot)$	Probability density and cumulative distribution functions of the probability of failure on demand of channel A.
$f_{P_B}(\cdot)$, $F_{P_B}(\cdot)$	Probability density and cumulative distribution functions of the probability of failure on demand of channel B.
$f_{P_{AB}}(\cdot)$, $F_{P_{AB}}(\cdot)$	Probability density and cumulative distribution functions of the probability of simultaneous failure on demand of channel A and channel B.
$f_{P_A,P_B}(\cdot,\cdot)$, $F_{P_A,P_B}(\cdot,\cdot)$	Joint probability density and cumulative distribution functions of the probabilities of failure on demand of channel A and channel B.
$f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot)$	Joint probability density function of the probabilities of failure on demand of channel A, of channel B and of simultaneous failure of channel A and channel B.
$f_{P_B}(\cdot\,\|P_A = x)$	Conditional probability density function of the probability of failure on demand of channel B, conditional on the probability of failure on demand of channel A being equal to *x*.
$f_{P_{AB}}(\cdot\,\|P_A = x, P_B = y)$	Conditional probability density function of the probability of simultaneous failure of channel A and channel B, conditional on the probability of failure of channel A being equal to *x* and the probability of failure of channel B being equal to *y*, respectively.
$L(n,r\|P_{AB})$	Likelihood of observing *r* simultaneous failures of channel A and channel B in *n* independent demands, conditional on the probability of simultaneous failure of channel A and channel B being P_{AB}.
$L(n,r_a,r_b,r_{ab}\|P_A,P_B,P_{AB})$	Likelihood of observing r_a failures of channel *A* only, r_b failures of channel *B* only and r_{ab} simultaneous failures of channel *A* and channel *B* in *n* independent demands, conditional on the probability of failure of channel A being P_A, the probability of channel *B* being P_B and the probability of simultaneous failures of channel A and channel B being P_{AB}.
$\delta(x)$	Dirac Delta function.
$f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot\,\|n,r_a,r_b,r_{ab})$	Joint probability density function of the probabilities of failure on demand of channel A, of channel B and of simultaneous failure of channel A and channel B, conditional on observing r_a failures of channel *A* only, r_b failures of channel *B* only and r_{ab} simultaneous failures of channel *A* and channel *B* in *n* independent demands.

2.1. Overview of the prior work: On-demand 2-channel software

We start with a brief introduction of the concept of a 2-channel on-demand software.

Consider a safety-critical system as illustrated in Figure 1. The safety of a safety-critical Plant is achieved by deploying a protection system. A deviation of the plant state from a predefined safe envelop of operation should be detected by a protection system via the “sensed plant state variables”. When a dangerous deviation of the plant state is detected the protection system tries to bring the plant to a *safe state* via a set of actuators. For many safety-critical systems the safe state might be merely shutting the plant down.

Reliability of the protection system is typically achieved via redundancy: as shown in Figure 1 the protection system consists of two functionally identical channels, channel A and channel B. Each of the two channels can bring on its own the plant to a safe state. A failure of one of the channels of the protection system, thus, is of no consequence for the safety of the plant

The case that we consider in this work is of a protection system in which the functionality of each channel is achieved by a dedicated software (SW1 and SW2, respectively) run on dedicated hardware (HW1 and HW2, respectively). Redundancy in both hardware and software is necessary so that a single point of failure is avoided. Defence against design faults in the channels is typically achieved using design diversity [4].

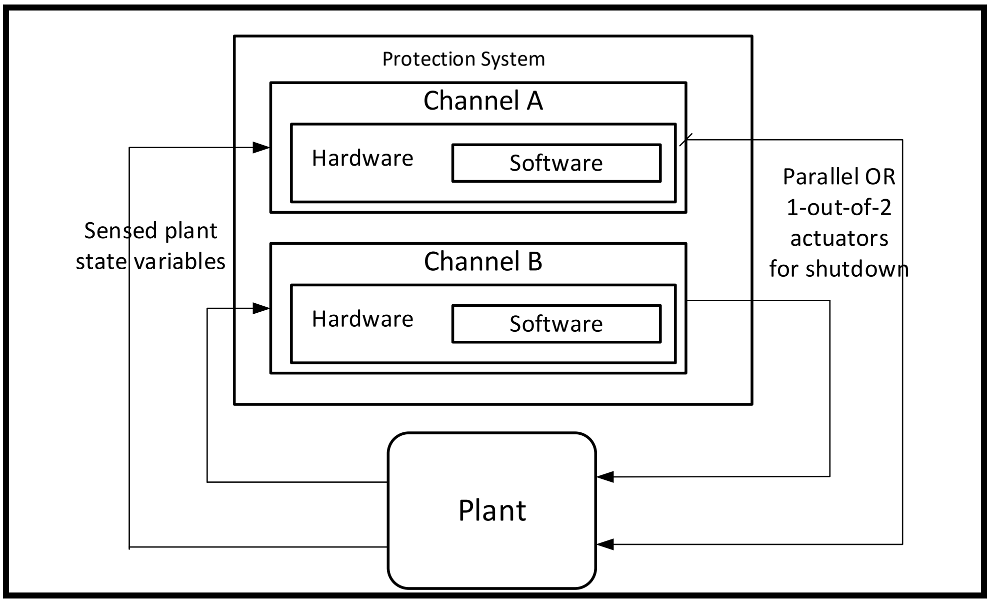
Protection System
Channel A
Hardware
Software
Channel B
Hardware
Software
Sensed plant state variables
Parallel OR 1-out-of-2 actuators for shutdown
Plant

Figure 1. A simplified diagram of a safety-critical system with a 2-channel protection system.

A catastrophic system failure of the safety-critical system shown in Figure 1 will occur when the plant enters an unsafe state and both channels of the protection system fail simultaneously to bring the plant to a safe state. This condition can be captured by a fault tree as is shown Figure 2.

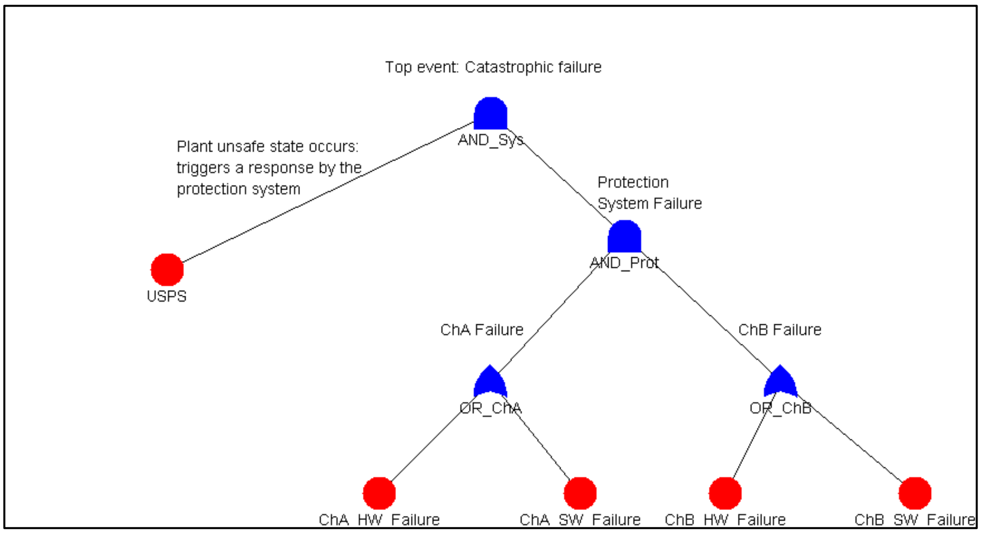
Top event: Catastrophic failure
AND_Sys
Plant unsafe state occurs: triggers a response by the protection system
USPS
Protection System Failure
AND_Prot
ChA Failure
ChB Failure
OR_ChA
OR_ChB
ChA_HW_Failure
ChA_SW_Failure
ChB_HW_Failure
ChB_SW_Failure

Figure 2. A simplified fault tree for the safety-critical system, illustrated in Figure 1.

The top event in the fault tree is the occurrence of a catastrophic system failure, which occurs when the plant enters an unsafe state and the two channels of the protection system fail simultaneously to bring the plan to a safe state. Protection channels, in turn, may fail due to either hardware or the software failures in the respective channel.

The fault-tree uses the following base events: i) unsafe state of the plant (USPS), ii) hardware failure (ChA HWF) of channel A, iii) software failure (ChA SWF) of channel A, iv) hardware failure (ChB HWF) of channel B, and v) a software (ChB SWF) failure of channel B. The top event "catastrophic system failure (CSF)" can then be expressed as:

$$\text{CSF} = \text{USPS} \wedge (\text{ChA HWF} \vee \text{ChA SWF}) \wedge (\text{ChB HWF} \vee \text{ChB SWF})$$

After an obvious transformation, the expression can be rewritten as:

$$\text{CSF} = \text{USPS} \wedge \text{ChA HWF} \wedge \text{ChB HWF} \vee \text{USPS} \wedge \text{ChA HWF} \wedge \text{ChB SWF} \vee$$

$$\text{USPS} \wedge \text{ChB SWF} \wedge \text{ChA HWF} \vee \text{USPS} \wedge \text{ChA SWF} \wedge \text{ChB SWF}.$$

The top event is expressed as a disjunction of four conjunctions of triplets of events, each of which can trigger the top event CSF: i) USPS, ChA HWF and ChB HWF, ii) USPS, ChA HWF and ChB SWF, iii) $\text{USPS} \wedge \text{ChA SWF} \wedge \text{ChB\_HWF}$ and iv) $\text{USPS} \wedge \text{ChA SWF} \wedge \text{ChB\_SWF}$. While for the first three conjunctions one can make a reasonably convincing argument that the events involved can be assumed to occur independently and hence the probability of their occurring simultaneously can be computed as a product of the marginal probabilities of occurrence of the events, such an argument cannot be made for the last conjunction term in the expression above: $\text{USPS} \wedge \text{ChA SWF} \wedge \text{ChB SWF}$. The evidence against assuming statistical independence, especially between ChA SWF and ChB SWF, accumulated over many decades of using software-based protection systems is overwhelming. This evidence suggests that even if software used in the two protection channels, as shown in Figure 1, is developed truly independently – either as a result of a bespoke development of the channels by independent teams or by deploying in the channels off-the-shelf software acquired from different (and "independent") vendors, assuming that the failures of the two software channels will occur statistically independent is not credible. In the absence of such statistical independence computing the probability of the top event CSF becomes difficult, too. This paper addresses this difficulty – analysing the attempts by others to solve the problem conservatively. Consider "on-demand" software, i.e., software, which is invoked by sending to it demands for processing as in the case of software used in the two channels of a protection system shown in Figure 1. Typical examples of on-demand software are *protection systems*, e.g., of a nuclear power plant or of any other process control plant. In such systems, the system safety is typically achieved by defining a "safe state", which the system should enter should the controlled process deviate dangerously from the intended safe envelop of operation. The sole purpose of a protection system (often implemented in software) is to react to the deviations of the plant from its safe operation.

Processing a demand by a protection system may involve a very complex sequence of inputs coming from the operational environment – from the initial signal (e.g., that the pressure in the nuclear reactor has exceeded the normal/acceptable level), followed by the software reading multiple sensors until the nature of the anomaly is established with certainty (or high degree of confidence) and then a transition to a safe state is executed as needed. We call this entire sequence of sensor reading and transition to the safe state *a demand*. Typically, a complex plant may deviate from its safe operation in many different ways, hence, many different demands on the protection system are possible. Each demand is processed by the protection system either correctly or may result in a failure.

It is common in safety-critical applications to use fault-tolerance, e.g. via design diversity [5]. 1-out-of-2 software system is commonly used[2]. The two channels of a 1-out-of-2 system process the demands independently and produce responses to them. The system is said to have processed a demand correctly if at least one of the channels has produced a *correct* response. If both channels have failed simultaneously on the same demand, then the system is said to have failed. Demands, processed correctly, will bring the plant to the predefined safe state.

	Channel A	Channel B
i)	Success	Success
ii)	Success	Fail
iii)	Fail	Success
iv)	Fail	Fail

Table 2: White-box model

	System
i)– iii)	Success
iv)	Fail

Table 3: Black-box model

With two channels there are 4 possible outcomes for the responses of the two channels on a random demand as shown in Table 2. The first 3 outcomes, i) – iii), despite the possible channel failures, represent a successful operation of the 1-out-of-2 system. If the system is modelled as a black box (Table 3), these three different outcomes become *indistinguishable* –

[2] Other architectures, such as 2-out-of-3 and even 2-out-of-4 are also used in protection systems but are outside the scope of the paper.

on all of them the system processes the demand successfully. If the system is modelled as a white-box, then clearly i) – iii) are recognised as different and may be treated differently (Table 2). The white-box model also allows one to learn about the dependence among the failure processes of the two channels. For instance, seeing many instances of outcomes ii) and iii) in operation will be evidence of negative correlation between the failures of the two channels. If instead, no instances of ii) and iii) are seen in operation would imply that failures of the channels are positively correlated. The only thing that one can learn with a black-box model is the probability of simultaneous failure of the two channels.

2.2. Black – box model vs. white-box model

Bayesian assessment can be applied to a 2-channel system using either a black-box model or a white box model. The black-box model is simpler and requires less detailed info about the system architecture and less detailed observations about the system as is evident from Table 3 above. Bayesian inference with a white-box model is computationally more complex, the observations from operation, too, have to be recorded with greater detail (as shown in Table 2 above).

The two models have been compared in the past [6] and here we summarise some of the known results:

- The white box inference makes better use of the observations as Table 2 and Table 3 demonstrate. This is not a mere technicality, but as we established in [6] may affect the predicted system reliability and confidence. Using the full observations in some cases may allow one to get confidence in a system reliability target *faster* (or slower) than if a black-box model is used.
- The black-box model only requires a description of the modelled system[3] in terms of the probability of *system* failure, characterised by the probability distribution$f_{P_{AB}}(\cdot)$, of the probability of system failure, P_{AB}, treated as a random variable. This probability (and the respective probability distribution, $f_{P_{AB}}(\cdot)$), however, *must be somehow derived* from the knowledge about the individual channels' *pfd*s (and their respective probability distributions). The channels may have been developed (or acquired) and seen some independent assessment. The outcomes of these assessments must somehow be aggregated into a form suitable for a black-box Bayesian inference with a 2-channel system. This author *is not aware of a rational way* of constructing anything meaningful about the probability of system failure, $f_{P_{AB}}(\cdot)$, without making use of the channels' assessments, and making additional assumptions about how channel reliabilities might be related (e.g., how the channel failures might be related). Assessment which only requires quantification of system reliability (i.e., about the probability of simultaneous failure of channel A and channel B) may seem appealing since the burden on domain experts to provide quantification is reduced to a single probability distribution, $f_{P_{AB}}(\cdot)$. These experts, however, still would need to deploy a "mental" model in order to get the requested minimal information *rationally and consistently* with what is known about the channels used in the 2-channel system. The advantage of a black-box inference over the white-box counterpart in terms of simplifying the system description, therefore, seems *rather questionable*.
- The spirit of Bayesian assessment is to allow for the *epistemic uncertainty* in the values of the probabilities of failures to be updated with the observations from system testing/operation. Black-box inference only allows one to learn (and update the epistemic uncertainty) about system reliability. The black-box model, however, does not allow one to learn from system testing/operation about the channel's pfd. White-box inference, on the other hand, does allow one to learn from the observations/testing about both the system and the channels' reliabilities and how their dependences evolve. This is not just a matter of taste. If an argument is built, e.g., based on assumptions that some of the channels might be perfect (i.e., free from design faults), this argument will be *entirely destroyed* if the channel assumed perfect fails in testing/operation. The black-box model may mask this sensitive outcome, e.g., if channel failures do not lead to system failures. Under the white-box model, instead, channel failures will lead naturally to setting the probability of perfection to 0. This difference between the predictions obtained using a black-box and a white-box model is examined in detail in Section 4 and illustrated in section 5.

2.3. Conservative assumptions in Littlewood – Rusby model

Littlewood and Rushby establish conservative bounds on the probability of system failure of a 2-channel system in [7] under the following *plausible* assumptions:

Assumption 1: If the probability of failure on demand of channel A, *P(A fails)* and the probability of perfection of the second channel, *P(B is perfect)*, are known *with certainty*, then the two events occur independently, i.e. the probability of the joint event "channel A fails on a randomly chosen demand and channel B is not perfect" can be computed as a product of the two probabilities:

[3] A prior distribution. We introduce the Bayesian inference formally in section 3.

$$P(A\ fails,\ channel\ B\ is\ not\ perfect) = P(A\ fails) \times (1 - P(B\ is\ perfect)) \tag{1}$$

Assumption 2: If channel B is not perfect, then it *fails with certainty* whenever channel A fails. This is a *very strong assumption* and is said to guarantee that the assessment obtained with it is *conservative*. The assumption essentially states *two things*:

- Given channel B is *not perfect*, i.e., $P_B > 0$, then channel B is *no better than channel A.* Indeed, it fails at least on all those demands where channel A fails and possibly on some other demands, too. In other words:

  $P(System\ fails \mid channel\ A\ has\ failed,\ P_B > 0) = 1$.

- The probability of system failure, $P(AB|P_A = p_a,\ P_B = p_b > 0)$ would be equal to the probability of failure of channel A. In other words, $P(AB|P_A = p_a, P_B > 0) = p_a$.

We adopt these two assumptions throughout this paper.

3. The System Model

3.1. Modelling notations

We start with a number of notations, necessary for the rest of the paper. These are consistent with the abbreviations summarised in Table 1.

We will treat the probability of failure of the channels of a 2-channel system and of the system as *random variables*: we are typically unable to tell the exact value of these probabilities, these are subject to *epistemic uncertainty*. We use capital letters to denote these random variables, P_A, P_B and P_{AB}, the probabilities of failure of channel A, channel B and of the system, respectively.

Let $f_{P_A}(\cdot)$, $f_{P_B}(\cdot)$, $f_{P_{AB}}(\cdot)$ be the probability density functions, which capture the *epistemic uncertainty*, associated with the values of P_A, P_B and P_{AB}, respectively.

The probability of perfection of a channel, e.g. of channel B, *P(B is perfect)*, is related to $f_{P_B}(\cdot)$. Non-zero probability of perfection, would imply that $f_{P_B}(P_B = 0) = \delta(x)P(B\ is\ perfect)$[4], where $\delta(x)$ is the Dirac Delta function[5]:

$$F_{P_B}(P_B = 0) = \int_0^{0+} f_{P_B}(x = 0) \times P(B\ is\ perfect)dx = \ P(B\ is\ perfect) \int_0^{0+} \delta(x)\, dx = \ P(B\ is\ perfect).$$

3.2. Bayesian Inference

Bayesian inference is based on the Bayes formula, which allows a posterior distribution to be derived from a given prior distribution and the likelihood of an observation.

For on-demand software systems, the demands submitted for processing by the system are typically assumed to be drawn *independently* (with replacement) from the space of demands according to a given demand profile, a probability distribution which defines the likelihood of a demand being selected at random from the population of all demands.

The observations – success/failure of channels and of the system – occur with different probabilities, called likelihood of the observations. For *independently selected demands*, if we observe a single channel (or the system, modelled as a black-box), the observations will be in the form (n, r), where n is the total number of demands submitted to the system and r is the number of observed failures. Clearly, $n \geq r$[6]. If we observe 2 channels, the observations will be in the form (n, r_a, r_b, r_{ab}), where n is again the number of demands submitted to the system for processing, r_a is the number of failures of channel A only, r_b is the number of failures of channel B only and r_{ab} is the number of simultaneous failures of both channels.

[4] The epistemic uncertainty about the value of *P(B is perfect)* can be captured by another distribution, e.g. with probability density function $f_{P_{B_{perfect}}}(\cdot)$.

[5] The integral $\int_{t_1}^{t_2} \delta(t)dt$ evaluates to 1, if the integration interval [t_1, t_2] includes the point t. Otherwise, the integral evaluates to 0.

[6] This inequality states the mere fact that we cannot observe more failures, r, than the total number of demands, n.

3.2.1. Black-box inference

If a black-box model is used for the 2-channel system, the random variable of interest is P_{AB}. The posterior, $f^b_{P_{AB}}(\cdot\,|data)$, can be computed as follows:

$$f^b_{P_{AB}}(x|n,r) = \frac{f_{P_{AB}}(x)L(n,r|P_{AB})}{\int_{x=0}^{1} f_{P_{AB}}(x)L(n,r|P_{AB})dx} \tag{2}$$

where, the likelihood of observing r failures in n demands ($n \geq r$), $L(n,r|P_{AB})$, is given by the binomial formula:

$$L(n,r|P_{AB}) = \binom{r}{n}(P_{AB})^r(1-P_{AB})^{n-r} \tag{3}$$

In the special case of no failures (i.e., $r = 0$), the likelihood becomes:

$$L(n,0|P_{AB}) = (1-P_{AB})^n \tag{4}$$

3.2.2. White-box inference

An inference based on a white-box model requires a tri-variate prior distribution, e.g., $f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot)$, in which the probabilities of failure of the individual channels and of the system are used as variates. The observations, e.g. from testing the 2-channel system, are used in the inference, in a similar manner to the black-box inference [6]:

$$f_{P_A,P_B,P_{AB}}(x,y,z|data) = \frac{f_{P_A,P_B,P_{AB}}(x,y,z)L(data|P_A,P_B,P_{AB})}{\int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)L(data|P_A,P_B,P_{AB})dzdydx} \tag{5}$$

where $L(data|P_A,P_B,P_{AB})$ is the likelihood of an observation (n, r_a, r_b, r_{ab}), i.e., that in n demands one observes r_a failures of channel A only, r_b failures of channel B only and r_{ab} simultaneous failures of both channels. This (multinomial) likelihood is computed as follows:

$$L(n,r_a,r_b,r_{ab}|P_A,P_B,P_{AB}) = \frac{n!}{r_a!\,r_b!\,r_{ab}!\,(n-r_a-r_b-r_{ab})!}(P_A-P_{AB})^{r_a}(P_B-P_{AB})^{r_b}(P_{AB})^{r_{ab}}(1-P_A-P_B+P_{AB})^{n-r_a-r_b-r_{ab}}$$

The special case of interest "no system failure" merely means that $r_{ab} = 0$, while r_a , r_b can both be greater than 0.

If the marginal probability of system failure is of interest, this can be derived from the 3-variate posterior by integrating out the "nuisance" parameters, P_A and P_B (i.e., the probabilities of failure of the individual channels):

$$f^w_{P_{AB}}(\cdot\,|data) = \int_{P_A=0}^{1}\int_{P_B=0}^{1} f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot\,|data)d(P_A)d(P_B) \tag{6}$$

3.2.3. Black-box vs. White-box Model

The relationship between the posteriors, $f^w_{P_{AB}}(\cdot\,|data)$ and $f^b_{P_{AB}}(\cdot\,|data)$, obtained with the two models for the same data is complex and no general law seems to exist, e.g., of stochastic ordering between the two posteriors. In our previous work [8] we recorded examples of ordering between the two predictions: in some cases the white-box inference can be more optimistic than the ones obtained with the black-box. In other cases, the ordering would be reversed. We also recorded cases with no ordering at all – the *cdf*s of the posteriors about the system *pfd* obtained with the black-box and the white-box may have a crossover point.

In general, the posteriors obtained with the two models will be different except for some special cases, some of which are summarised below:

- In [6] we reported on using a Dirichlet distribution as a 3-variate prior, related to $f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot)$. With a Dirichlet as a prior and testing on independently selected demands the marginal posteriors of the probability of system failure, $f^b_{P_{AB}}(\cdot\,|data)$ and $f^w_{P_{AB}}(\cdot\,|data)$ are *guaranteed to be identical irrespective of the observations (e.g., testing results)*, and the particular parameterisation of the prior Dirichlet distribution. The Dirichlet as a prior is the *only known case* which eliminates the predictions dependence on the model used in the inference. However, in some prior work [9, 10] a black-box model was used with a prior of the system *pfd*, which typically has the entire probability mass concentrated in a few data points. Such a marginal prior would be impossible if a white-box model were used with Dirichlet as a prior distribution. Thus, the conservative predictions established in the prior work based on a black-box model are *bound to be different* from the predictions obtained with a white-box model. In this paper we exploring these differences in detail.

- Another special case of a system prior was studied in [6], in which conservative upper bounds on the probability of failure of the channels were assumed *known with certainty*. For this case we demonstrated that using a white-box model leads to *counterintuitive* predictions: if neither of the channels fails in testing, the predictions $f^w_{P_{AB}}(\cdot\,|data)$ are guaranteed to be *worse* than the prior, $f_{P_{AB}}(\cdot)$, while the black-box model will produce predictions, $f^b_{P_{AB}}(\cdot\,|data)$, which are stochastically better than the prior. We mention this case as it points to an intriguing and important *side effect* that may result from constraining the epistemic uncertainty in pursuit of "conservatism". Our conjecture is that similar link between constrained epistemic uncertainty and Bayesian predictions might exist with other forms of conservative priors and study this possibility in this paper.

4. Results

In this section we apply Bayesian inference to predict the reliability of a 1-out-of-2 on-demand software using a white-box model and a prior distribution $f_{P_A,P_B,P_{AB}}(x,y,z)$ consistent with the conservative assumptions defined by Littlewood & Rushby [7], as summarised in section 2.3. We assume that the marginal prior distribution, $f_{P_A}(\cdot)$, characterising the epistemic uncertainty in channel A *pfd*, is *fully defined*[7]. We also assume that the probability of channel A being perfect is 0, i.e. $\int_0^{0+} f_{P_A}(x)\, d(x)$=0.

4.1. A 3-variate conservative prior

We start by constructing a joint probability distribution, $f_{P_A,P_B,P_{AB}}(\cdot,\cdot,\cdot)$, consistent with the assumptions made in section 2.3.

4.1.1. Epistemic uncertainty in *pfd* of channel B

Assumption 1 above states that there is a non-zero probability that channel B is perfect, thus P(B is not perfect) = 1 - *P(P_B*

Assumption 1 above states that there is a non-zero probability that channel B is perfect, thus P(B is not perfect) = 1 - $P(P_B = 0)$, the latter in turn is merely the value of the integral $\int_0^{0+} f_{P_B}(\cdot)\, dP_B$. For this integral to be nonzero, we need $f_{P_B}(P_B = 0) \equiv K\delta(P_B = 0)$, where *K* is the probability of perfection of channel B. In other words, if *pnp* is the true value of *P(B is not perfect)* then:

$$f_{P_B}(x) = \begin{cases} (1 - pnp) \times \delta(x), if\ x = 0 \\ g_{P_B}(x), if\ x > 0 \end{cases}. \tag{7}$$

for some non-negative function $g_{P_B}(\cdot)$ such that, $\int_{0_+}^{1} g_{P_B}(x)dx = pnp$.

Let us now consider the conditional distribution, $f_{P_B}(\cdot\,|P_A = p_a)$. Based on the assumptions from 2.3 this conditional distribution will have a shape similar to the shape shown in Figure 1:

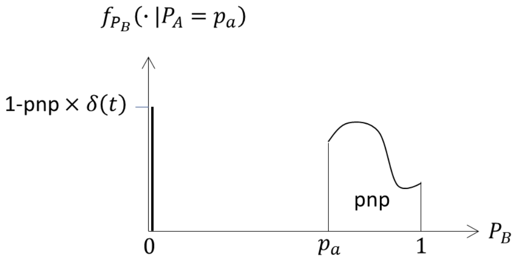

Figure 3. An illustration of the epistemic uncertainty associated with P_B, $f_{P_B}(\cdot\,|P_A = p_a)$.

A special case of the distribution will be when its entire mass on the range (p_a, 1] is concentrated at a single point, $P_b = p_a$ as is shown in Figure 2. The conditional probability distribution $f_{P_B}(\cdot\,|P_A = p_a)$ characterises the uncertainty associated with the probability of failure of channel B, conditional on $P_A = p_a$.

[7] This assumption is plausible as a black-box inference can be applied to channel A before it becomes a part of a 1-out-of-2 or other wider system, using i) the results of testing channel A on its own, or ii) the observations from real operation of channel A elsewhere. Bayesian inference with channel A would produce the prior distribution $f_{P_A}(\cdot)$.

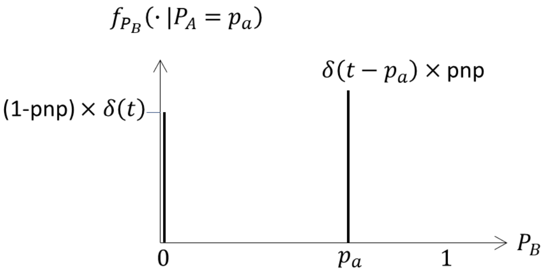

Figure 4. A special case of conservative $f_{P_B}(\cdot\,|P_A = p_a)$, when the entire mass is concentrated in two points.

Apart from the point $P_B = 0$, the assumptions of conservatism imply that $f_{P_B}(\cdot\,|P(A) = p_a)$ might be non-zero only in the interval $[p_A, 1]$, but ought to be 0 on the interval $(0, p_A)$ as is evident from Figure 1 and Figure 2. Clearly, this conditional distribution is quite *unusual* and it will be difficult to argue that it is *plausible*. The point, however, is that this distribution is guaranteed to define conservatively the epistemic uncertainty associated with P_B. And yet, Figure 1 and Figure 2 make it quite clear that one may be doubtful whether to rely on such a conservative prior distribution.

4.1.2. Joint distribution $f_{P_A,P_B}(\cdot,\cdot\,)$

Now we turn our attention to the joint distribution $f_{P_A,P_B}(\cdot,\cdot\,)$. For this we will need to specify the relationship between the probability of perfection, *P(B is perfect)* and the distribution of P_A. In the general case if these two probabilities are treated as random variables, there is no guarantee that they will be independently distributed [7]. Even, if *P(B is perfect)* is assumed known with certainty, we cannot just assume that *P(B is perfect*$|P_A = p_a$*)* is a constant and does not vary with P_A. From the point of view of the joint distribution, $f_{P_A,P_B}(\cdot,\cdot\,)$ we can have *P(B is perfect*$|P_A = p_a$*)* vary with P_A. In this case the assumed known *P(B is perfect)* would be merely the *expected value* of the random variable $P(B\ is\ perfect|P_A = p_a)$:

$$P(B\ is\ perfect) = \int_0^1 P(B\ is\ perfect|P_A = p_a) f_{P_A}(P_A) d(P_A) \tag{8}$$

Let us simplify the analysis and assume that $P(B\ is\ pefect|P_A = p_a) = P(B\ is\ perfect)$ and proceed with the construction of the joint distribution, $f_{P_A,P_B}(\cdot,\cdot\,)$. We note that Assumption 1 (see section 2.3 above) states just that – the probability of perfection of channel B does not depend on the value of the probability of failure of channel A.

4.1.3. Conditional distribution $f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = p_b \geq p_a)$

Let us now look at $f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = p_b \geq p_a)$, the conditional distribution of system *pfd*, conditional on known probabilities of failure of channel A, p_A, and of channel B, p_B.

Clearly, $f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = 0) = \delta(0)$ – perfection of channel B implies that the system is also perfect.

For any other pair of values, p_b, p_a, we should consider two cases:

- $p_a < p_b$. This case should be "impossible" under the conservative assumptions: channel B is either perfect or no better than channel A. Any definition of the conditional probability of failure, e.g., $\delta(t - p_a)$, will be acceptable.
- $p_a \geq p_b$. In this case we will have: $f_{P_{AB}}(t|P_A = p_a, P_B = p_b > 0) = \delta(t - p_a)$. Indeed, given Assumption 2 made earlier, whatever the specific value $p_b > 0$ the system is failing whenever channel A fails. P_{AB} is deterministically equal to P_A. In other words, the entire mass of the probability of system failure will be concentrated at point p_a. Not surprisingly, the particular value of the probability of failure of channel B, $p_b > 0$, does not affect $f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = p_b)$ in any way.

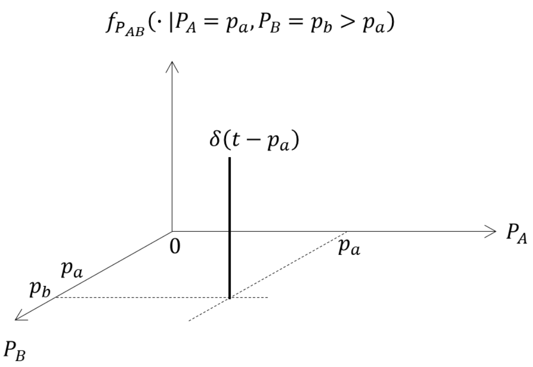

Figure 5. An illustration of the conditional distribution $\boldsymbol{f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = p_b, p_b > \; p_a)}$.

4.1.4. Epistemic uncertainty in system pfd

Having established the two conditional distributions, $f_{P_B}(y|P_A = x)$ and $f_{P_{AB}}(\cdot\,|P_A = p_a, P_B = p_b)$, we are now ready to express the full tri-variate distribution, $f_{P_A,P_B,P_{AB}}(x,y,z)$, needed for the white-box inference.

Trivially,

$$f_{P_A,P_B,P_{AB}}(x,y,z) = f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x) f_{P_A}(x) \tag{9}$$

The marginal prior distribution of the probability of system failure $f_{P_{AB}}(\cdot)$ can be derived from $f_{P_A,P_B,P_{AB}}(x,y,z)$ by integrating out the nuisance parameters.

$$f_{P_{AB}}(z) = \int_{x=0}^{1}\int_{y=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)\, d(P_B)d(P_A)=$$

$$=\int_{x=0}^{1}\int_{y=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x) f_{P_A}(x)\, d(y)d(x). \tag{10}$$

We will simplify this expression separately for $z = 0$ and for $z > 0$.

$$f_{P_{AB}}(z = 0)=\int_{x=0}^{1}\int_{y=0}^{0+} f_{P_{AB}}(z = 0|P_A = x, P_B = y = 0) f_{P_B}(y = 0|P_A = x) f_{P_A}(x)\, d(y)d(x) \tag{11}$$

and

$$f_{P_{AB}}(z > 0) = \int_{x=0}^{1}\int_{y=x}^{1} f_{P_{AB}}(z > 0|P_A = x, P_B = y) f_{P_B}(y|P_A = x) f_{P_A}(x)\, d(y)d(x) \tag{12}$$

Since, $f_{P_{AB}}(z = 0|P_A = x, P_B = y = 0) = \delta(z)$ and $f_{P_B}(y = 0|P_A = x, y > x) = \delta(y)(1 - pnp)$, (11) can be rewritten as:

$$f_{P_{AB}}(z = 0)= \delta(z)\left[\int_{y=0}^{0+} \delta(y)(1 - pnp)dy\right]\int_{x=0}^{1} f_{P_A}(x)d(x) = \delta(z)(1 - pnp) \tag{13}$$

Since, $f_{P_{AB}}(z > 0|P_A = x, P_B = y) = \; \delta(t - x)$, (12) can be rewritten as follows:

$$f_{P_{AB}}(z > 0) = \int\limits_{x=0}^{1}\int\limits_{y=x}^{1} f_{P_{AB}}(z > 0|P_A = x, P_B = y) f_{P_B}(y|P_A = x) f_{P_A}(x)\, d(y)d(x) =$$

$$\int_{x=0}^{1} \delta(t - x)\, f_{P_A}(x)d(x) \int_{y=x}^{1} f_{P_B}(y|P_A = x)d(y) \tag{14}$$

Now we note that $\int_{x=0}^{1} \delta(t - x)\, f_{P_A}(x)d(x) = f_{P_A}(x)$, which follows from the "*sifting property*" of the Dirac Delta function and that $\int_{y=x}^{1} f_{P_B}(y|P_A = x)d(y) = pnp$. Thus, we arrive at the following:

$$f_{P_{AB}}(z > 0) = \int_{x=0}^{1} \delta(t - x)\, f_{P_A}(x)d(x) \int_{y=x}^{1} f_{P_B}(y|P_A = x)d(y) = f_{P_A}(x)pnp \tag{14}$$

In summary, the marginal *pdf* of the probability of system failure becomes:

$$f_{P_{AB}}(x) = \begin{cases} \delta(x)(1 - pnp), if\ x = 0 \\ pnp \times f_{P_A}(x), if\ x > 0 \end{cases} \tag{15}$$

In other words, the marginal prior distribution of the probability of system failure (*pfd*), defined consistently with the conservative assumptions, would be *entirely* defined by the marginal *pdf* of channel A and the probability of perfection, *pnp*, of channel B. This is *not exactly surprising*. Given the set of assumptions, some of which are indeed extreme, one would have expected that the distribution of system *pfd* would be constrained. Yet, the expression of the marginal distribution of system *pfd* provides an insight about the nature of this marginal *pdf* of the system *pfd*.

It is worth noting that in the above derivation we *did not make any assumptions* about the conditional distribution $f_{P_B}(y|P_A = x)$. Its form, as shown in Figure 1, includes the special case of the entire mass of this conditional probability distribution being concentrated in a single point, $p_b = p_a$, and being 0 elsewhere (see Figure 2). Thus, the marginal distribution of system *pfd*, captured by (15), will be the same *irrespective* of the particular form that $f_{P_B}(y|P_A = x)$ takes.

Now, if one is interested in applying a black-box Bayesian inference, the prior probability density function $f_{P_{AB}}(\cdot)$ is readily available and one can use it. There is no need to ask in addition for percentiles of this distribution, as is required in [9, 10]. If one does need some percentiles, these can be easily obtained from this conservative marginal *pdf* defined by (15).

4.2. Conservative Predictions using a white-box model

Now we look at the posteriors derived with the constructed conservative priors – using a black-box or a white box system models. Are these posteriors guaranteed to be conservative in some sense? If so, in which sense? The previous works formulated the conservatism in terms of the *expected value* of the posterior system *pfd*. Can we do better and demonstrate *stronger guarantees of conservatism*, e.g., stochastic ordering of some sort between the predictions obtained with the different models?

The previous works concentrated on demonstrating conservatism with respect to a *specific observation* – testing/operation, which revealed *no system failures*. Are the predictions guaranteed to be conservative for other observations, possibly for *any observations*, e.g., for observations in which one of the channels has failed or both channels have failed, but no system failure has been observed or indeed when even a limited number of system failures (i.e., simultaneous failures of both channels) have been observed?

4.2.1. Properties of the conservative priors

First, we look at whether the assumptions stated earlier about the conservatism of the prior are retained in the posterior distribution. In other words, whether the posterior distribution derived with a white-box Bayesian inference remains conservative, in the sense of Assumption 1 and Assumption 2, stated in section 2.3.

4.2.2. Case 1: No failure observed

Consider first the case of *no channel failures* in *n* demands. Using (5) we can express the posterior probability density function as follows:

$$f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0) = \frac{f_{P_A,P_B,P_{AB}}(x,y,z)L(n,0,0,0|P_A,P_B,P_{AB})}{\int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)L(n,0,0,|P_A,P_B,P_{AB})dzdydx} \tag{19}$$

Theorem 1:

The posterior distribution $f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)$ derived with a conservative prior, consistent with Assumption 1 and 2 stated in section 2.3, and with “no failure” observed in *n* demands, retains the conservative properties captured by Assumptions 1 and 2.

Proof: Provided in Appendix A.

The next result is related to how Bayesian inference affects the assumption that the probability of perfection of channel B and the probability of failure of channel A are independently distributed random variables, which we assumed in section 4.1.2.

Theorem 2:

The assumption that the probability of perfection of channel B does not vary with the probability of failure of channel A is not guaranteed in a posterior distribution $f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)$ derived from a conservative prior consistent with Assumption 1 and 2 for an observation of “no failures” in *n* demands.

Formally, assuming that $F_{P_{AB}}(z = 0|P_A = p_{a_1}) = F_{P_{AB}}(z = 0|P_A = p_{a_2})$ for any two different values of the probability of failure of channel A, $p_{a_1} \neq p_{a_2}$, may lead to $F_{P_{AB}}(z = 0|n,0,0,0,P_A = p_{a_1}) \neq F_{P_{AB}}(z = 0|n,0,0,0,P_A = p_{a_2})$.

Proof: Provided in Appendix B.

The implication of this theorem is that the appealing assumption of independence adopted by Littlewood and Rushby model between the probability of perfection and the probability of failure of channel A may not be retained in the posterior distribution $f_{P_A,P_B,P_{AB}}(x, y, z|n, 0, 0, 0)$.

Theorem 3:

The posterior probability of system perfection derived with a conservative prior, consistent with Assumptions 1 and 2 stated in section 2.3, is greater than the prior probability of system perfection, i.e.,

$F_{P_{AB}}(z = 0|n, 0,0,0) > F_{P_{AB}}(z = 0)$.

Proof: Provided in Appendix C.

In other words, observing no failures increases the belief that the system might be perfect. This is not surprising as the evidence of "no failure" is supportive of the system being good, possibly perfect.

Finally, let us look at how the form of the prior distribution $f_{P_B}(y|P_A = x)$ impacts the predictions of the probability of system failure for the case of observing "no failures". We will compare the posterior probability of system perfection for the following forms of $f_{P_B}(y|P_A = x)$:

- Case 1: The probability mass is somehow spread over the interval [x, 1], as illustrated in Figure 1, and
- Case 2 ("opt"): The entire probability mass of $f_{P_B}(y|P_A = x, y > 0)$ is concentrated at a single point $y = x$. In this case, $f_{P_B}(y|P_A = x) = \delta(y - x) \times pnp$, as illustrated in Figure 2.

Theorem 4:

The probability of perfection, $F_{P_{AB}}^{opt}(0|n, 0,0,0)$ is smaller than $F_{P_{AB}}(z|n, 0,0,0)$, i.e.:

$$F_{P_{AB}}^{opt}(0|n, 0,0,0) < F_{P_{AB}}(z|n, 0,0,0)$$

Proof: Provided in Appendix D.

The implication of this theorem is quite clear: the "opt" case of prior distribution is the *most conservative prior* in the sense of minimising the posterior probability of system perfection. Although intriguing, this result *falls short* of establishing *stochastic ordering* between the posterior distributions derived using the "opt" prior and the other form of prior distribution of $f_{P_B}(y|P_A = x, y > 0)$.

We note that the result from Theorem 2 applies to both $F_{P_{AB}}^{opt}(0|n, 0,0,0)$ and $F_{P_{AB}}(0|n, 0,0,0)$. Thus, combining Theorem 2 and 4, leads to the following relationship between the marginal probabilities of system perfection:

$$F_{P_{AB}}(0) < F_{P_{AB}}^{opt}(0|n, 0,0,0) < F_{P_{AB}}(0|n, 0,0,0) \tag{20}$$

4.2.3. Case 2: Failures observed

We now briefly look at the cases, in which failures are observed in operation/testing. Such observations were not studied in the prior work [7, 9, 10].

We are particularly interested to find observations, for which the posterior violates Assumptions 1 and 2 used in constructing a conservative prior distribution. We are also interested to find out the price of using the conservative prior, e.g., to establish observations, which lead to *counterintuitive* predictions about the reliability of the system and its channels.

The general form of the joint posterior distribution was already provided earlier by equation (5). Below we list the likelihood for several observations including channel or system failures:

- Failures of channel A only. Consider n tests/operational demands in which $r_a > 0$ failures of channel A are observed and no failures of channel B ($r_b = 0, r_{ab} = 0$). The likelihood of this observation is:

$$L(n, r_a > 0, r_b = 0, r_{ab} = 0 | P_A, P_B, P_{AB}) = \frac{n!}{r_a!\,(n - r_a)!}(P_A - P_{AB})^{r_a}(1 - P_A - P_B + P_{AB})^{n - r_a}$$

- Failures of channel B only. Consider n tests/operational demands in which $r_b > 0$ failures of channel B are observed, and no failures of channel A ($r_a = 0, r_{ab} = 0$). The likelihood of this observation is:

$$L(n, r_a = 0, r_b > 0, r_{ab} = 0 | P_A, P_B, P_{AB}) = \frac{n!}{r_b!\,(n - r_b)!}(P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n - r_b}$$

- System failures only. Consider n tests/operational demands in which $r_{ab} > 0$ simultaneous failures of channel A and channel B are observed, but none of the channels fails on its own ($r_a = 0, r_b = 0$). The likelihood of this observation is:
  $L(n, r_a = 0, r_b = 0, r_{ab} > 0 | P_A, P_B, P_{AB}) =$
  $$= \frac{n!}{r_{ab}!\,(n - r_{ab})!}(P_{AB})^{r_{ab}}(1 - P_A - P_B + P_{AB})^{n - r_{ab}}$$
- Finally, we could have a combination of channel only failures, $r_a > 0, r_b > 0$, respectively. In this case the likelihood of the observation is:

$$L(n, r_a > 0, r_b > 0, r_{ab} = 0 | P_A, P_B, P_{AB})$$
$$= \frac{n!}{r_a!\,r_b!\,(n - r_a - r_b)!}(P_A - P_{AB})^{r_a}(P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n - r_a - r_b}$$

We analyse each of the first three cases listed above next. The fourth case is a combination of observing channel failures. Its implications, therefore, can be easily derived from the cases describing the implications of failure of each of the channels on their own.

4.2.3.1. Case 2a: Channel B failure only are observed in operation/testing

This case illustrates an observation $r_a = 0, r_b > 0, r_{ab} = 0$ in n demands/tests ($n \geq r_b$). Intuitively, a failure of channel B, considered perfect with non-zero probability is straightforward. Such an observation destroys the hypothesis that channel B *might be perfect*. Hence, we expect the posterior distribution to evolve to a form, in which the probability of perfection is set to 0. The following theorem confirms that observing failures of channel B indeed leads to a posterior with probability of failure of channel B set to 0.

Theorem 5:

Posterior distribution $f_{P_A, P_B, P_{AB}}(x, y, z | n, r_a = 0, r_b > 0, r_{ab} = 0)$ derived from a conservative prior consistent with Assumptions 1 and 2, stated in section 2.3 and with observations $r_a = 0, r_b > 0, r_{ab} = 0$ in n demands ($n \geq r_b$), will be such that the probability of perfection of channel B is set to 0:

$F_{P_B}(y = 0 | n, r_a = 0, r_b > 0, r_{ab} = 0) = 0.$

Proof: Provided in Appendix E.

Note that the result does not depend on the number of failures of channel B. Even if a *single failure* of channel B is observed, the belief is destroyed that channel B might be perfect. At this point, the reliability benefits from channel B are lost – the system becomes as reliable as channel A (Assumption 2), i.e., channel B is assumed failing deterministically, whenever channel A fails.

Note also that using a black-box inference *will overlook this very significant change* of the probability of channel B perfection. With a black-box model, a demand which leads to a failure of channel B only, will be considered a successfully processed demand and thus the black-box inference will lead to a posterior predicting improvement of the probability of system perfection.

4.2.3.2. Case 2b: Channel A failure only observed in operation/testing

Now consider the case of observing failures of only channel A in testing/operation, i.e., observing $r_a > 0, r_b = 0, r_{ab} = 0$ in n demands/tests ($n \geq r_a$). An observation with failures of channel A only in operation should not be surprising and should be anticipated. The prior works, however, have not analysed the response of a conservative prior to such observations. The next theorem establishes a very *surprising* result:

Theorem 6:

Posterior distribution $f_{P_A,P_B,P_{AB}}(x, y, z|n, r_a > 0, r_b = 0, r_{ab} = 0)$ derived from a conservative prior consistent with Assumptions 1 and 2, stated in section 2.3, and after observations $r_a > 0, r_b = 0, r_{ab} = 0$ in n demands ($n \geq r_a$), will be such that the probability of perfection of channel B is set to 1.

$F_{P_B}(y = 0|n, r_a > 0, r_b = 0, r_{ab} = 0) = 1.$

Proof: Provided in Appendix F.

The implications of seeing failures of channel A are indeed very dramatic and surprising: even a single failure of channel A, which itself should not be surprising, leads to a posterior, in which the channel B and the system are now believed to be *perfect with certainty*. Such a prediction is *counterintuitive*, a clear sign that the particular *conservative model is problematic* as it implies a *very strong dependence* between the uncertainties in the probabilities of failure of the channels. Details of why this "side effect" occurs are provided in Appendix F, but intuitively the cause are the constraints we have imposed on the prior, $f_{P_B}(y|P_A = x, y \geq x)$, (as illustrated in Figure 1 and Figure 2). Given that channel A fails, the constrained distribution $f_{P_B}(y|P_A = x, y \geq x)$ changes dramatically. The probability mass attached to $F_{P_B}(y > x|P_A = x)$ in the prior, in the posterior will be multiplied by $(P_A - P_{AB})^{r_a}$. However, for values of $P_B > 0$ Assumption 2 implies $P_A = P_{AB}$, thus leading to $f_{P_A,P_B,P_{AB}}(x, y, z|n, r_a > 0, r_b = 0, r_{ab} = 0) = 0$. P_A and P_{AB}, however, do not have to be equal! They are so merely due to Assumption 2, made in adopting a conservative prior of a *particular form*. As a result, the entire probability mass of the conditional distribution $F_{P_B}(y|P_A = x)$ will be moved to point $P_B = 0$. As a result, channel B in the posterior will be believed to be perfect with certainty, which in turn, implies that the system itself will be believed to be perfect with certainty.

We note that this dramatic and counterintuitive change of the epistemic uncertainty will remain invisible with a black-box inference. A demand which leads to a failure of channel A will be considered processed correctly by the system and will lead to an increase of the confidence in system's perfection, but not as dramatic as the inference with the white-box model implies.

4.2.3.3. Case 2c: System failure only observed

Now we look at another possible observation – observing *simultaneous failures* of the two channels, i.e., $r_a = 0, r_b = 0, r_{ab} > 0$ in n demands/tests ($n \geq r_{ab}$). Such observation would be evidence of a strong correlation between the failures of the two channels.

Intuitively, this case is a special case of Case2a – channel B fails – and we expect to see again (as in Case 2a) that the belief about a possible perfection of channel B in the posterior is destroyed. This is indeed confirmed by the following:

Theorem 7:

Posterior distribution $f_{P_A,P_B,P_{AB}}(x, y, z|n, r_a = 0, r_b = 0, r_{ab} > 0)$ derived from a conservative prior consistent with Assumptions 1 and 2, stated in section 2.3, and after observations $r_a = 0, r_b = 0, r_{ab} > 0$, will be such that the probability of perfection of channel B is set to 0:

$F_{P_B}(y = 0|n, r_a = 0, r_b = 0, r_{ab} > 0) = 0.$

Proof: Provided in Appendix G.

In summary, when we observe failures – either of the individual channels or of the system – the conservative prior aligned with Assumption 1 and 2 stated in section 2.3 leads to predictions with "side effects":

- A failure of channel A makes us believe that channel B is perfect, which is an extreme form of *negative dependence* between the epistemic uncertainties in the pfds of channel A and B.
- When either channel B or the system fail, we lose confidence in channel B perfection, which in turn, makes the *conservative model useless*: the conservative model assumes that if channel B is not perfect the system is merely as reliable as channel A. In this case, there is no reliability gain from using channel B.

It seems that the conservative model built with Assumption 1 and 2 and captured by the prior distribution is quite *brittle* with implausible "side effects". Although no such "side effects" have been detected with "no failures" observations, the brittleness of the predictions, which is entirely due to the assumptions on which the prior is built, poses doubts about the suitability of the conservative prior for practical assessment.

We stress that the two cases, Case 2a and Case 2b, will be indistinguishable for the black-box model under any parametrisation – conservative or not. This is due to the intrinsic nature of the black-box model. With the white-box model, however, the two cases lead to two extreme changes of the epistemic uncertainty related to channel B's perfection: with Case 2a any hope of perfection of channel B is lost; with Case 2b – all doubts in perfection of channel B (hence in the system) are removed and the system is believed to be perfect with certainty. The difference could not have been greater.

The attentive reader may have noticed that we did not consider a combination of the above 3 cases – combination of channel A and channel B failures and of system failures (i.e., two or more of the failure counts are greater than 0: $r_a > 0, r_b > 0, r_{ab} > 0$). Such cases can be handled using the 3 cases described above. For instance, an observation (n, r_a, r_b, r_{ab}), which contains a combination of non-zero failure counts, can be split suitably into batches such that at most one of the (r_a, r_b, r_{ab}) is non-zero. A batch can include any number of "no failures" observations. As an example, consider that batch$_1$ consists of $n_1 > 0$ "no failure" observations followed by a failure of channel A, batch$_2$ consists of $n_2 > 0$ "no failure" observations, followed by a simultaneous failure of both channels, batch$_3$ consists of $n_3 > 0$ "no failure" observations followed by a failure of channel B, etc. Processing the first batch will use the conservative prior and will produce posterior$_1$, which in turn will become a prior for the inference with batch$_2$. The posterior derived with the observations in batch$_2$ will then become the prior for the inference with batch$_3$, etc. Having processed all batches one will arrive at a posterior prediction which takes into account the full set of observations grouped in batches. Although technically, such an iterative procedure is sound, the value that one will get using it seems questionable. Clearly, after the first batch with a channel failure the predictions become "questionable":

i) after the first failure of channel A we will conclude that the system is perfect with certainty. Note that such a conclusion does not guaranteed that failures of channel B are impossible in the future. The certainty in channel B (hence the system) perfection is a mere consequence of using a constrained prior based on assumptions 1 and 2.

ii) after the first failure of channel B on its own or simultaneously with channel A, we will conclude that the system is as good as channel A on its own. At this point one will question the value of a conservative prior based on Assumptions 1 and 2 and will probably abandon processing the further batches of observations.

5. A contrived example

In this section we illustrate the work of the white-box conservative model on a contrived example of a 2-channel system, which is close to the example used in [7]. The system model is defined using Assumption 1 and 2 for constructing the prior[8]. Channel A's *pfd* is represented by a truncated Beta distribution in the range [0, 0.01] with parameters α=1, β = 10, i.e., the expected value of $P_A \approx 10^{-3}$ with an upper bound on P_A of 0.01 ($P_A \leq 0.01$).

For channel B's we construct the conditional *pdf* of its *pfd* using a truncated Beta distribution with the same parameters (α=1, β = 10) but the probability mass is assigned according to:

- Figure 1: $f_{P_B}(y = 0 | P_A = x) = \delta(y)(1 - pnp)$ and the rest of the probability mass is assigned within the interval $[x, x + 0.01]$, or
- Figure 2: the probability mass for the conditional probability distribution $f_{P_B}(y | P_A = x)$ is concentrated in two points: $f_{P_B}(y = 0 | P_A = x) = \delta(y)(1 - pnp)$ and $f_{P_B}(y = x | P_A = x) = \delta(y - x)pnp$, respectively.

The non-zero probability of perfection of channel B is set to *pnp* = 0.5 and is independent of the probability of failure of channel A (i.e., does not vary with P_A).

Now we illustrate the inference for a selected number of observations:

Case 1: n = 5000 tests with no failures observed (r_a=0, r_b = 0, r_{ab} = 0).

Case 2: n = 5000 tests with a single failure of channel A (r_a = 1, r_b = 0, r_{ab} = 0).

Case 3: n = 5000 tests with a single failure of channel B (r_a = 0, r_b = 1, r_{ab} = 0).

Case 4: n = 5000 tests with a single system failure (i.e., a simultaneous failure of channel A and channel B) (r_a = 0, r_b = 0, r_{ab} = 1).

[8] The calculations were conducted using a bespoke MATLAB script, available from the author on demand.

We constructed the joint prior using the parameterisation above and then derived from it the marginal distribution of the probability of system failure, $f_{P_{AB}}(x)$. This marginal distribution was then used to derive the "black-box" posterior distribution $f^b_{P_{AB}}(x|n, r_{ab})$ for all cases listed above using (2). Clearly, for the black-box inference, Case 1, Case 2 and Case 3 are indistinguishable as in all three cases the system does not fail.

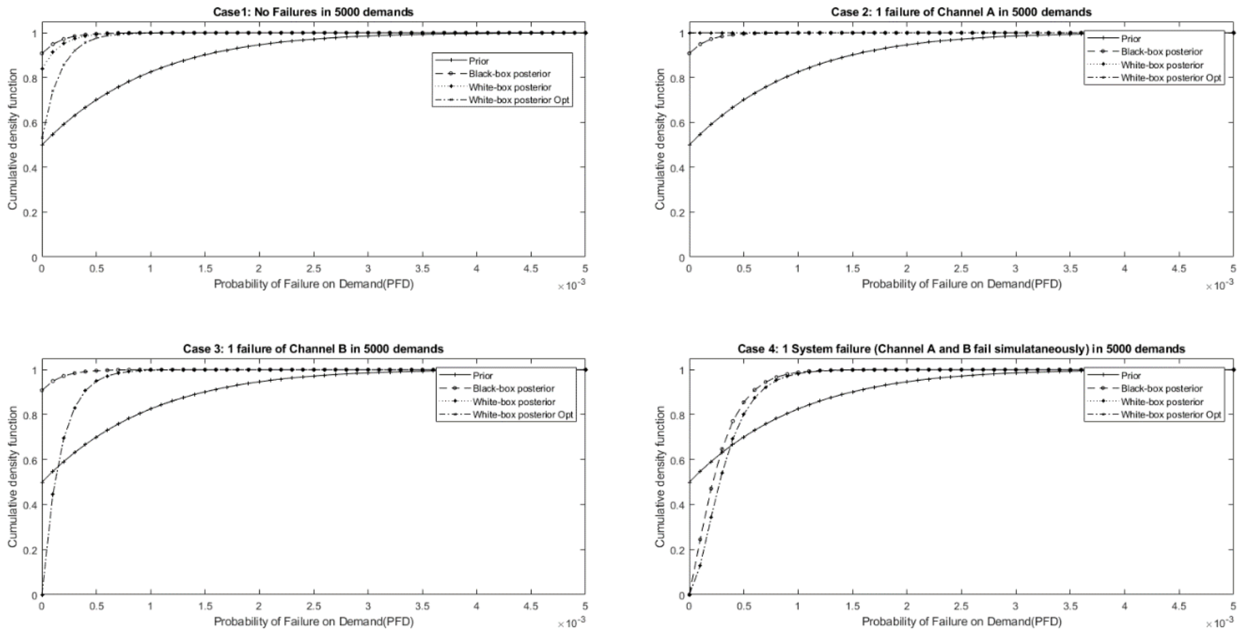

Figure 6. Conservative inference in action: an illustration of the difference between white-box and black-inference results for different observation.

We apply also a "white-box" inference using the results described in section 4. For the white-box model inference we define two priors whose $f_{P_B}(y|P_A = x)$ differ - according to Figure 1 or Figure 2, respectively. From the joint posteriors we derive $f^w_{P_{AB}}(\cdot\,|data)$, and plot them in Figure 4.

The curves representing the white box posterior distribution of the probability of system failure are labelled as follows:

- "White-box posterior" for the posterior derived from a joint prior in which $f_{P_B}(y|P_A = x)$ is defined as shown in Figure 1 (referred to as the non-"opt " case in the commentary below), and
- "White-box posterior Opt" for the posterior derived from a joint prior in which $f_{P_B}(y|P_A = x)$ is defined as shown in Figure 2 (referred to as the "opt" case in the commentary below).

The difference between the white-box posteriors ("opt" and non-"opt") and the black-box posteriors are shown in Figure 4 above.

The case of "no-failures" in operation, studied in [7], is illustrated in the left-top plot of Figure 4. One can see clearly that the three posterior distributions for the probability of system failure *are different*: there is a stochastic ordering between them. The most *optimistic* prediction is obtained with the *black-box model*. The white box "opt" prediction is the most conservative and the white-box (non-"opt") is in between. The illustration not merely confirms the ordering between the "opt" and non-"opt" white box predictions of the probability of system perfection, which we established with inequality (20), but a stochastic ordering is in place between the distributions, too. All three predictions are more optimistic than the prior.

The other 3 plots illustrate the findings that we established in section 4 for the observations with some failures:

- A single failure of channel A (the top right plot in Figure 4) leads to a white-box posterior whereby the system appears *perfect with certainty* – the entire mass is concentrates at point $P_{AB} = 0$. The prediction is not affected by the choice of the prior ("opt" or non-"opt"). Both white-box predictions are identical and imply that the system is now predicted as perfect with certainty. This is a dramatic shift from the prior where we assumed 50% doubts that channel B is perfect, which in turn let to 50% doubt that the system itself is perfect.

- When channel B fails (the bottom left plot in Figure 4) the black-box predictions and the two white-box predictions *differ dramatically*: the black-box prediction is stochastically more optimistic than the prior. This is not surprising as the evidence (i.e., the observations) presented to the black-box model is all "supportive": no system failures are observed in 5000 demands/tests. We have a clear demonstration of the limitations of the black-box model here: since the failure of channel B is masked in the black-box model, the black-box model does not respond to such failures and "interprets" the evidence as supportive. The situation is, of course, quite different for the white-box models. As the plot shows, the white box predictions for low values of the probability of system failure are worse than the prior. The two *cdf* curves cross between 10^{-4} and 2×10^{-4}. The white-box predictions indicate that there is no chance for system perfection as we have assumed with the chosen prior. The posteriors *cdf* curves from the "opt" and non-"opt" white-box models are indistinguishable (and cannot be seen as separate curves on the plot). The white-box predictions are *stochastically worse* (i.e., more conservative) than the black-box predictions. Such white-box predictions are plausible: a failure of channel B implies that channel B has now been demonstrated to be imperfect with certainty. According to the conservative assumptions built-in the prior, if channel B is not perfect, then the system is merely as reliable as channel A, i.e., there is no reliability gain from using channel B.
- When a single system failure is observed (i.e., channel A and channel B fail simultaneously on the same demand), we observe difference between the black-box and the white-box posterior distributions. The two white-box predictions are again indistinguishable – the values of the two *cdf*s are very close and therefore the corresponding curves cannot be seen in the plot. The black-box predictions respond to the observation of a single system failure plausibly: the probability mass that the prior placed at $P_B = 0$ is now gone: indeed, the system has failed (albeit once in 5000 demands), which rules out the possibility of perfection of the system. Again, the white-box predictions turn out to be more conservative than the black-box predictions: there is a stochastic ordering between the *cdfs* representing the black-box and the white-box posterior distributions. Not surprisingly, both the black-box and the white-box *cdfs* have crossover points with the prior *cdf*, which indicates that there is no ordering between the prior distribution (constructed under the assumption that channel B might be perfect with probability of 0.5) and the posterior distributions, for which the system was demonstrated not to be perfect. Despite the system failure, however, after 5000 demands the confidence that the system reliability is better than 0.0005 is higher than it was in the prior for all predictions.

6. Discussion and Threats to Validity

We established analytically the critical importance of two decisions: i) what model is selected for conservative reliability assessment of fault tolerant software, and ii) once a model is chosen what assumptions it relies upon. Our findings suggest that scrutinising the implications of these decisions is an essential part of making the assessment credible, i.e., of gaining confidence in the assessment results.

More specifically we demonstrated that achieving conservatism by restrictions of the *epistemic uncertainty* in the probability of failure on demand (*pfd*) of the channels of a 1-out-of-2 fault-tolerant software and of their dependencies, may have surprising *counterintuitive consequences*. Making restrictive assumptions about channels' *pfd*s is convenient for an assessor as it simplifies the problem of quantifying the dependence between the failures of the channels used in a 1-out-of-2 system. Our results suggest that *convenience comes at a significant price of raising doubts about predictions' adequacy*!

We confirm that white-box predictions are conservative in the sense that they retain the essential properties used to construct a conservative prior for the case of "no-failure" in test/operation.

For observations with failures of channel A, the channel assumed not to be perfect, the predictions obtained with a white-box model are *not conservative at all*! Even a single failure of channel A will lead to predictions that channel B – hence the system as a whole – are perfect with certainty, which is clearly implausible.

The white-box inference based on the particular conservative prior deals naturally with observations, whereby channel B assumed likely to be perfect fails – either on its own or simultaneously with the second channel. For these cases, however, which are not included in the prior work [7, 9, 10], the predictions obtained with the white-box model are *extremely conservative* as in such cases the system becomes merely as reliable as the second channel (channel A). Such extreme conservatism does not seem of any practical interest whatsoever.

Although the black-box inference was not the focus of this work, a couple of observations from the numerical examples presented in section 5 seem important.

- The numerical results in section 5 suggest that for plausible cases the black-box predictions, expected to be conservative, may be *more optimistic* than those obtained using a white-box model. If one is really interested in obtaining conservative predictions one should consider using an inference based on a white-box model. Establishing rigorously ordering between the black-box and white box predictions is outside the scope of the paper.
- A black-box inference does not deal adequately with channel failures. This is true for the black-box inference in general, irrespective of the prior. With the particular conservative prior, however, constructed under the assumption of possible perfection of one of the channels, ignoring the failures of channel B assumed possibly perfect has dramatic consequences. The essential point here is that once the conservative prior is constructed one should not use a black-box model which ignores evidence (e.g., channel failures), which may potentially destroy the assumptions built in the prior. From this point of view, the white box inference has a clear edge – it will respond adequately to observations, which include channel and system failures.

The paper looked at the simplest architecture of a fault tolerant protection system, a 1-out-of-2 system. We are aware that other, more complex architectures, are used in practice, too, e.g., 2-out-of-3, 2-out-of-4, etc. We chose to work with a 1-out-of-2 for several reasons:

- The prior work [7], which we scrutinised in this paper, refers to a 1-out-of-2 system. We thought natural to base the analysis on a system with the same architecture. This architecture is used in protection systems, e.g. [11, 12], but also in some other contexts.
- An essential part of this work and of [7] is the assumption that one of the channels is "probably" perfect. This idea has been proposed and studied for 1-out-of-2 systems. To the best of our knowledge no prior work applies similar ideas to more complex architectures. Perfection is typically justified with references to "simplicity", possibly using "formal" methods, such as theorem provers, in software development, etc. While it is clear that these techniques can be applied to the channels of a 1-out-out-of-2 or the channels of more complex architectures, it is unclear whether the channels of more complex architectures are likely to be considered "possibly perfect".
- Finally, there is a methodological aspect in the choice of a 1-out-of-2 over the more complex architectures. The main contribution of this work is about the importance of: i) choosing the "right" model, and ii) parameterising the model well. Using a simple architecture in the analysis seems important so that excessive complexity in various mathematical transformations can be minimised. Dealing with more complex architectures is an area for future research.

Among threats to validity of the results from this work we acknowledge:

- We illustrate the approach on several contrived examples in section 5 (see Figure 4) for which we had to make additional assumptions about the shape of the prior distribution. For the conditional probabilities of failure of channel B, conditional on a particular value of reliability of channel A we used a truncated Beta distribution. Although using Beta distribution is common in Bayesian assessment, we acknowledge that the results presented in Figure 4 may be affected by the choice of parameters used to define the tri-variate prior distribution. For instance, the stochastic ordering between the black-box and the white-box posterior distributions, goes well beyond the theoretical results presented in section 4 and may not extend beyond the parameterisation used in the contrived examples. It is worth pointing out, however, that the observations made in section 5 are merely illustrations of the general theory presented in the paper and its feasibility for practical assessment. Validity of the theory is not affected.
- In developing the theory, we relied on some general theorems (e.g., the properties of Dirac Delta function). One may argue that priors, which include Dirac Delta function are extreme and should be avoided. We agree with such a viewpoint but would like to point out that "possible perfection" of channel B is an is essential concept in conservative assessment, which naturally leads to the use of Dirac Delta function. We merely explored the implications of the assumptions to demonstrate the "side effects" of predictions based on the particular conservative priors.

7. Related Research

The most relevant sources have already been referenced earlier in the paper. In this section we briefly outline a few additional sources.

The first formulation of the approach to assessment based on possible correctness (i.e., perfection) is due to Littlewood [13, 14]. The importance of the idea was discussed in [15] in the context of reliability assessment of high integrity systems. Further insight is provided in [16] introducing the concepts of "quasi" perfection.

[9] is an important paper which formulated the problem of conservative reliability assessment using a black-box system model. The conservative Bayesian assessment relies on a small number of quantiles, rather than the entire prior distribution, for a conservative prediction. While the approach is interesting for the rigor in dealing with the problem, the paper does not address the issue of how the required quantiles can be elicited in practice. The guarantees of conservatism are only valid for the case of "no failures" observed in operations/testing.

A number of recent papers provide incremental extensions on the idea formulated in [9], among them [10, 17-19]. They all refine the concept of conservative assessment by relying on more quantiles from the distribution of the probability of system failure, treating the system as a black-box. As in [9], the focus of these works is the "no failure" observation and the rigor in the mathematical treatment of the problem. How a practitioner is expected to derive the parameters of interest (i.e., the required quantiles), however, is outside the scope of these papers.

An interesting example of applying a conservative assessment is also [20] in which a black-box Bayesian assessment is applied to establish conservatively the number of miles needed for an assessor to be able to claim with sufficient confidence that an autonomous vehicle is sufficiently safe (i.e., "driving to safety"). The paper addresses an important application problem and is a straightforward application of the conservative Bayesian assessment based on a black-box model developed by some of the same authors earlier.

A somewhat related work is [19], which solves a slightly different problem: Bayesian assessment when legacy software is replaced by a new product believed to be better. The focus of the paper is on spelling out the problem including the sources of uncertainty and analysing the implications of these for the predictions.

Another relevant work is [2], in which the author takes into account the structure of software code (as a "flow network structure") with a claim that this method is a significant improvement over the alternatives, which disregard the control flow. A similar work was conducted by May et al., e.g. [21]. The focus of these works, however, is a single-channel safety-critical software. We acknowledge that the approach is interesting and possibly applicable to a 2-channel software but is not concerned with a conservative assessment.

In a recently published paper [22] Bishop and Povyakalo offer a method for estimating the probability of failure on demand of a software system built with components from testing the system and offer a conservative method of assessing the confidence in this probability. The method uses the "structure function" defined for the system under study and assumes that none of the software components fails in testing. The conservatism is based on the observation that if an input leads to a failure of a replicated component, all replicas of this component used in the system will fail simultaneously. The method uses classical confidence bounds and does not rely on Bayesian inference at all.

Another paper dealing with conservative reliability assessment is [23]. This is an example of conservative reliability assessment in manufacturing. Despite the significant difference between the application domains – manufacturing vs. software – the work also looks to establish a conservative prior for the probability of failure. The authors use an iterative approach and Monte Carlo simulation to construct a conservative prior of the probability of failure. The simulations are used to compensate the lack of sufficient empirical observations. The findings in the paper are further elaborated in a Doctoral Thesis [24], supervised by one of the authors of [23].

In own work, an approximate method of Bayesian inference for systems with complex "structure function" was developed, [25]. The main idea of the method is that the system is modelled by a set of views, whereby each view captures a part of the system structure. The views are linked via random variables (typically the probabilities of failure of software components) which appear in more than one view (a "common random variable"): The marginal distribution of the common random variables is propagated from a more detailed view to a view operating at a higher level of abstraction ("propagation chain"), thus allowing for the marginal system *pfd* to be expressed using the propagation chain without having to define joint probability distribution with an excessive number of variates. The method also computes the prediction error in comparison with alternative less detailed inference models (including the black-box system model) and always selects the model which for the given observations provides the most accurate prediction. The philosophy of this method is quite different from the work presented in this paper: it is not seeking to obtain conservative predictions. Instead, the method is trying to maximise the accuracy of the predictions by comparison of the predictions of the available approximate models. This work, however, seems quite relevant if the method of conservative assessment proposed in [7] and extended in this paper is to be applied to more complex architectures, such as 2-out-of-3 and 2-out-of-4.

8. Conclusions and future research

This paper provides an insight into the effects of a *particular form of conservative Bayesian prediction* of the probability of system failure of a 1-out-of-2 software system, in which one of the channels may be perfect with non-negligible probability.

The work presents a few important innovations in comparison with the prior work:

- The inference is based on a "white-box model". This model is more sensitive to the full range of possible observations, including the observations with channel and system failures.
- Although the work is extensively based on the conservative assumptions defined by others in [7], we spell out the implications for the shape of the 3-variate prior, consistent with these informal conservative assumptions. The approach that we take to construct a tri-variate prior suitable for a white-box inference seems applicable to a large class of conservative priors, constrained in different ways. We then:
  - Scrutinise the implications of the specific conservative assumptions for the Bayesian predictions and report on a *counterintuitive* result: any number of failures of the channel, assumed with certainty *not to be perfect*, leads to predictions that the system itself is perfect. This implausible prediction is entirely due to the particular form of conservatism, which imposes constraints on the epistemic uncertainty in channels' and system's *pfds*.
  - We also study how the shape of the prior distribution $f_{P_B}(y|P_A = x)$ affects the predictions about the system probability of perfection, $F_{P_{AB}}(0|n, 0,0,0)$, and establish the shape of the prior, which minimises it. Although this result is important as it established an important *lower bound* on the probability of system perfection, it falls short of demonstrating that this shape of prior will lead to the most conservative predictions of *system pfd*, which is the primary concern when a 2-channel system is deployed. The existence of stochastic ordering between the most conservative predictions and the predictions obtained with other plausible priors has been established numerically.
- We demonstrate that the assumptions made in [7] are sufficient for the marginal *pdf* of the system *pfd*, $f_{P_{AB}}(\cdot)$, to be derived in full from the assumed known *pdf* of the channels' *pfd*s, thus making a black-box inference easily applicable without having to ask, as some prior works [9, 10] suggest, for quantiles on the system *pfd*. Such quantiles can be easily obtained from $f_{P_{AB}}(\cdot)$.

The work presented in this paper is essentially a detailed scrutiny of the conservative assumptions presented in [7] which, as our analysis shows, lead to *quite unusual prior distribution*. Justifying such a distribution may be very difficult, indeed. The conservatism of the assessment under the particular conservative assumptions is, thus, achieved at the price of raising doubts about the credibility of the assessment based on such assumptions.

Reasoning about reliability of fault tolerant software based on "possible" perfection of some of the channels is quite appealing. The concept of perfection is well supported by the use of well-established techniques of formal verification. Being conservative in the assessment of software used in safety-critical applications is widely-adopted. Our work points to a number of deficiencies with the particular form of combining "perfection" with conservative assessment. These difficulties, however, *do not seem intrinsic and unavoidable*. Finding alternative forms of combining perfection and conservatism in software assessment, which are free from the problems we have identified, seems and important direction for future research.

Finally, another important aspect worth addressing in the future is how arguments based on software tools perfection can be extended and applied to critical systems, in which complex logic/computations are implemented in silicon (e.g., FPGA and Systems on Chip). I am grateful to one of the reviewers for pointing out the proliferation of 1-out-of-2 architectures at chip and board level. Many semiconductor manufacturers provide SIL 3 level microcontrollers, e.g., TI's TMS570 and RM serial functional safety controller. In FPGA applications, these architectures are also widely used, e.g., NewTec's SafeFlex - Functional Safety Development Kit (FSDK). Verification of such solutions is likely to face issues similar to those discussed here. Hence, effort to establish the benefits (or otherwise) of a conservative assessment for systems implemented in silicon seems highly desirable.

Acknowledgment

The author would like to thank his colleagues at the Centre for Software Reliability, at City, University London, Prof. Bev Littlewood who commented on an earlier version of the paper and Dr Andrey Povyakalo and Dr Kizito Salako, who pointed out the elegant way of proving inequality C6 in Appendix C.

The author is also grateful to the anonymous reviewers and the associate editor for their insightful comments on the previous versions of the paper.

The work has been supported by the Intel Collaborative Research Institute - Safe Automated Vehicles (ICRI-SAVe).

References

1. P. Bourque and R.E. Fairley, *SWEBOK (v.3.0): Guide to Software Engineering Body of Knowledge*, P. Bourque and R.E. Fairley, Editors. 2014, IEEE Computer Society. p. 335.
2. Yaguang Yang, *Test based safety-critical software reliability estimation using Bayesian method and flow network structure.* Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2019. **233**(5): p. 847-856.
3. Christoph Kern and Mark R. Greenstreet, *Formal verification in hardware design: a survey.* ACM Trans. Des. Autom. Electron. Syst., 1999. **4**(2): p. 123–193.
4. G. G. Preckshot, *Method for Performing Diversity and Defense-in-Depth Analyses of Reactor Protection Systems*, in *NUREG/CR-6303*, U.S.N.R. Commission, Editor. 1994, Lawrence Livermore National Laboratory. p. 52.
5. Bev Littlewood, Peter Popov, and Lorenzo Strigini, *Modelling software design diversity - a review.* ACM Computing Surveys, 2001. **33**(2): p. 177 - 208.
6. B. Littlewood, P. Popov, and L. Strigini. *Assessment of the Reliability of Fault-Tolerant Software: a Bayesian Approach*. in *19th International Conference on Computer Safety, Reliability and Security, SAFECOMP'2000*. 2000. Rotterdam, the Netherlands: Springer.
7. B. Littlewood and J. Rushby, *Reasoning about the reliability of diverse two-channel systems in which one channel is 'possibly perfect'.* IEEE Trans Software Engineering, 2012. **38**(5): p. 1178–94.
8. P. Popov. *Reliability Assessment of Legacy Safety-Critical Systems Upgraded with Off-the-Shelf Components*. in *SAFECOMP'2002*. 2002. Catania, Italy: Springer.
9. B. Littlewood and A. Povyakalo, *Conservative reasoning about the probability of failure on demand of a 1-out-of-2 software-based system in which one channel is 'possibly perfect'.* IEEE Trans Software Engineering, 2013. **39**(11): p. 1521–30.
10. X. Zhao, et al., *Modeling the probability of failure on demand (pfd) of a 1-out-of-2 system in which one channel is "quasi-perfect".* Reliability Engineering & System Safety, 2017. **158** ( ): p. 230-245.
11. N.E. Buttery, *The use of probabilistic safety analysis in design and operation —Lessons learned from Sizewell B, Annex 14*. 2002, IAEE. p. 14.
12. Health and Safety Executive *Sizewell B nuclear powerstation: The findings of NII's assessment of British Energy's periodicsafety review*. 22.
13. B. Littlewood, *The use of proof in diversity arguments.* IEEE Transactions on Software Engineering, 2000. **26**(10): p. 1022-1023.
14. P. Bishop, et al., *Toward a Formalism for Conservative Claims about the Dependability of Software-Based Systems.* IEEE Transactions on Software Engineering, 2011. **37**(5): p. 708 - 717.
15. J. Rushby. *Software Verification and System Assurance*. in *2009 Seventh IEEE International Conference on Software Engineering and Formal Methods*. 2009.
16. L. Strigini and A. A. Povyakalo. *Software Fault-Freeness and Reliability Predictions*. in *International Conference on Computer Safety, Reliability, and Security (SAFECOMP'2013)*. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.
17. B. Littlewood, et al., *On Reliability Assessment When a Software-based System Is Replaced by a Thought-to-be-Better One.* Reliability Engineering & System Safety, 2019. **197**(May, 2020).
18. X. Zhao, et al., *Conservative Claims for the Probability of Perfection of a Software-based System Using Operational Experience of Previous Similar Systems.* . Reliability Engineering and System Safety, 2018. **175**: p. 265-282.
19. B. Littlewood and Rushby. J., *Reasoning about the reliability of diverse two-channel systems in which one channel is 'possibly perfect'.* . IEEE Trans Software Engineering., 2012. **38**(5): p. 1178–94.
20. X. Zhao, et al., *Assessing Safety-Critical Systems from Operational Testing: A Study on Autonomous Vehicles.* Information and Software Technology, 2020. **128**(December, 2020.).

21. J. May, G. Hughes, and A.D. Lunn, *Reliability estimation from appropriate testing of plant protection software.* Software Engineering Journal, 1995. **10**(6): p. 206-218.
22. P. Bishop and A. A. Povyakalo, *A conservative confidence bound for the probability of failure on demand of a software-based system based on failure-free tests of its components.* Reliability Engineering & System Safety, 2020. **203**.
23. H. Cho, et al., *Conservative reliability-based design optimization method with insufficient input data.* Structural and Multidisciplinary Optimization, 2016. **54**(6): p. 1609-1630.
24. M.-Y. Moon, *Confidence-based model validation for reliability assessment and Confidence-based model validation for reliability assessment and its integration with reliability-based design optimization its integration with reliability-based design optimization* in *Mechanical Engineering*. 2017, University of Iowa, USA Iowa City. p. 158.
25. P. Popov, *Bayesian reliability assessment of legacy safety-critical systems upgraded with fault-tolerant off-the-shelf software.* Reliability Engineering & System Safety, 2013. **117**: p. 98-113.

Appendix A

The general expression of the posterior 3-variate distribution for observing no failures in *n* demands can be written as:

$$f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0) = \frac{f_{P_A,P_B,P_{AB}}(x,y,z)(1-P_A-P_B+P_{AB})^n}{K3}$$

where $K3 = \int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)\, L(n,0,0,|P_A,P_B,P_{AB})dzdydx$.

As we have seen in section 3, there are 3 areas in the hypercube defined by the 3-variate distribution:

- Area 1, where $f_{P_A,P_B,P_{AB}}(x,y,z) = 0$.
- Area 2, where $f_{P_A,P_B,P_{AB}}(x,y,z) = K\delta(s)$ or $f_{P_A,P_B,P_{AB}}(x,y,z) = \delta(s-x)$.
- Area 3, where $f_{P_A,P_B,P_{AB}}(x,y,z) > 0$.

One can easily establish from (19) that the inference does not change these areas – the boundaries of the areas in the posterior distributions match exactly the boundaries in the prior. Indeed, the posterior density, $f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)$, is a product of the prior density $f_{P_A,P_B,P_{AB}}(x,y,z)$ at a given point of the hypercube multiplied by $(1-x-y+z)^n$, a constant dependent on the specific values of the probabilities of failure of the channels and of the system at the particular point of the hypercube of the three-variate distribution, divided by a normalising constant, K3. The values of *x*, *y* and *z* of *practical interest* are very small numbers, typically in the order of less than 10^{-2}, for which $(1-x-y+z)^n$ would be a positive number.

The implications of this observation are that the posterior will preserve the properties that we established in section 4.1 for the prior distribution.

QED.

Appendix B

We now establish whether the Bayesian inference for an observation "no failures" in *n* demands will affect the assumed constant probability of perfection $P(P_B = 0|P_A) = const$ for all values of P_A.

We will compare $P(P_B = 0|n,0,0,0,P_A = p_{a1})$ and $P(P_B = 0|n,0,0,0,P_A = p_{a2})$ for two different values of the probability of failure of channel A, $p_{a1} \neq p_{a2}$, respectively.

Clearly, the conditional probability of interest can be derived from the conditional distribution, $f_{P_B}(y|n,0,0,0,P_A = x)$, which can be expressed as:

$$f_{P_B}(y|n,0,0,0,P_A = x) = \frac{f_{P_A,P_B}(x,y|n,0,0,0)}{f_{P_A}(x|n,0,0,0)}. \tag{B1}$$

We derive $f_{P_A,P_B}(x,0|n,0,0,0)$ and $f_{P_A}(x|n,0,0,0)$ from the joint posterior distribution $f_{P_A,P_B,P_{AB}}(x,0,z|n,0,0,0)$ next.

B1.1. Posterior joint distribution $f_{P_A,P_B}(x,0|n,0,0,0)$

$$f_{P_A,P_B}(x,0|n,0,0,0) = \frac{\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,0,z|n,0,0,0)dz}{K3}$$

$$= \frac{\int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = 0)\, f_{P_B}(y = 0|P_A = x) f_{P_A}(x)(1-x-y+z)^n dz}{K3}$$

$$= \frac{f_{P_B}(y = 0|P_A = x)}{K3} f_{P_A}(x) \int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = 0)\,(1-x-y+z)^n dz$$

$$= \frac{\delta(0)(1-pnp)}{K3} f_{P_A}(x) \int_{z=0}^{1} \delta(0)\,(1-x-0+z)^n dz = \frac{\delta(0)(1-pnp)}{K3} f_{P_A}(x)(1-x)^n \tag{B2}$$

Thus:

$$f_{P_A,P_B}(x = p_{a1}, 0|n,0,0,0) = \frac{\delta(0)(1-pnp)}{K3} f_{P_A}(x)(1-p_{a1})^n \tag{B3}$$

$$f_{P_A,P_B}(x = p_{a2}, 0|n,0,0,0) = \frac{\delta(0)(1-pnp)}{K3} f_{P_A}(x)(1-p_{a2})^n \tag{B4}$$

B1.2. Posterior joint distribution $f_{P_A}(x|n,0,0,0)$

The marginal *pdf*, $f_{P_A}(x|n,0,0,0)$ can be derived from the joint *pdf* $f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)$ by integrating out the nuisance parameter, P_B and P_{AB}:

$$f_{P_A}(x|n,0,0,0) = \int_{y=0}^{1}\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}\,(x,y,z|n,0,0,0)d(z)d(y)$$

$$= \frac{\int_{y=0}^{1}\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)(1-x-y+z)^n d(z)d(y)}{K3}$$

$$= \frac{1}{K3}\int_{y=0}^{1}\int_{z=0}^{1} f_{P_B,P_{AB}}(y,z|P_A = x) f_{P_A}(x)(1-x-y+z)^n d(z)d(y)$$

$$= \frac{f_{P_A}(x)}{K3}\int_{y=0}^{1}\int_{z=0}^{1} f_{P_B,P_{AB}}(y,z|P_A = x)(1-x-y+z)^n d(z)d(y)$$

$$= \frac{f_{P_A}(x)}{K3}\int_{z=0}^{1}\left[\int_{y=0}^{0+} f_{P_B,P_{AB}}(0,z|P_A = x)(1-x-0+z)^n d(y)\right.$$

$$\left. + \int_{y=0+}^{1} f_{P_B,P_{AB}}(y,z|P_A = x)(1-x-y+z)^n d(y)\right] d(z) =$$

$$= \frac{f_{P_A}(x)}{K3}\int_{z=0}^{1}\left[\int_{y=0}^{0+} f_{P_{AB}}(z|P_A = x, P_B = 0) f_{P_B}(0|P_A = x)(1-x+z)^n d(y)\right.$$

$$\left. + \int_{y=0+}^{1} f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x)(1-x-y+z)^n d(y)\right] d(z)$$

Taking into account that $f_{P_{AB}}(z|P_A = x, P_B = 0) = \delta(0)$ and that $f_{P_{AB}}(z|P_A = x, P_B = y > x) = \delta(z-x)$, the expression above becomes:

$$
\begin{aligned}
f_{P_A}(x|n,0,0,0) &= \frac{f_{P_A}(x)}{K3}\int_{z=0}^{1}\left[\left[\int_{y=0}^{0+} f_{P_{AB}}(z|P_A=x,P_B=0)f_{P_B}(0|P_A=x)(1-x-y+z)^n d(y)\right.\right.\\
&\left.+\int_{y=0+}^{1} f_{P_{AB}}(z|P_A=x,P_B=y)f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right] d(z)\\
&= \frac{f_{P_A}(x)}{K3}\int_{z=0}^{1}\left[\left[\int_{y=0}^{0+}\delta(z=0)\delta(y=0)(1-pnp)(1-x-y+z)^n d(y)\right.\right.\\
&\left.+\int_{y=x}^{1}\delta(z-x)f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right] d(z)=\\
&= \frac{f_{P_A}(x)}{K3}\int_{z=0}^{1}\left[\delta(z=0)(1-pnp)(1-x+z)^n\right.\\
&\left.+\delta(z-x)\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right] d(z)\\
&= \frac{f_{P_A}(x)}{K3}\left[\left[\int_{z=0}^{1}\delta(z=0)(1-pnp)(1-x+z)^n d(z)\right.\right.\\
&\left.+\int_{z=0}^{1}\delta(z-x)\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\, d(z)\right]\\
&= \frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n+\int_{z=0}^{1}\delta(z-x)\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\, d(z)\right]
\end{aligned}
$$

Now we change the order of integration for the second summand in the expression above, which leads to the following:

$$
\begin{aligned}
&f_{P_A}(x|n,0,0,0) =\\
&\frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n+\int_{z=0}^{1}\delta(z-x)\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\, d(z)\right]\\
&\qquad= \frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n+\int_{y=x}^{1} f_{P_B}(y|P_A=x)\int_{z=0}^{1}\delta(z-x)\,(1-x-y+z)^n d(z)d(y)\right]\\
&\qquad= \frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n+\int_{y=x}^{1} f_{P_B}(x|P_A=x)(1-y)^n d(y)\right]
\end{aligned}
$$

We can now express the value of the posterior $f_{P_A}(x|n,0,0,0)$ for different values of x:

$$f_{P_A}(p_{a1}|n,0,0,0) = \frac{f_{P_A}(p_{a1})}{K3}\left[(1-pnp)(1-p_{a1})^n+\int_{y=p_{a1}}^{1} f_{P_B}(p_{a1}|P_A=p_{a1})(1-y)^n d(y)\right] \quad \text{(B5)}$$

$$f_{P_A}(p_{a2}|n,0,0,0) = \frac{f_{P_A}(p_{a2})}{K3}\left[(1-pnp)(1-p_{a2})^n+\int_{y=p_{a2}}^{1} f_{P_B}(p_{a2}|P_A=p_{a2})(1-y)^n d(y)\right] \quad \text{(B6)}$$

Now using (B1) we derive:

$$f_{P_B}(0|n,0,0,0,P_A=p_{a1}) = \frac{f_{P_A,P_B}(x=p_{a1},0|n,0,0,0)}{f_{P_A}(p_{a1}|n,0,0,0)}$$

$$= \frac{\dfrac{\delta(0)(1-pnp)}{K3} f_{P_A}(p_{a1})(1-p_{a1})^n}{\dfrac{f_{P_A}(p_{a1})}{K3}\left[(1-pnp)(1-p_{a1})^n+\int_{y=p_{a1}}^{1} f_{P_B}(p_{a1}|P_A=p_{a1})(1-y)^n d(y)\right]}$$

$$= \frac{\delta(0)(1-pnp)(1-p_{a1})^n}{\left[(1-pnp)(1-p_{a1})^n+\int_{y=p_{a1}}^{1} f_{P_B}(p_{a1}|P_A=p_{a1})(1-y)^n d(y)\right]} \quad \text{(B7)}$$

and

$$f_{P_B}(0|n,0,0,0,P_A=p_{a2}) = \frac{f_{P_A,P_B}(x=p_{a2},0|n,0,0,0)}{f_{P_A}(p_{a2}|n,0,0,0)} = \frac{\delta(0)(1-pnp)(1-p_{a2})^n}{\left[(1-pnp)(1-p_{a2})^n+\int_{y=p_{a2}}^{1} f_{P_B}(p_{a2}|P_A=p_{a2})(1-y)^n d(y)\right]} \quad \text{(B8)}$$

The ratio of the two probability densities of the probability of failure of channel B becomes:

$$\frac{f_{P_B}(0|n,0,0,0,P_A = p_{a1})}{f_{P_B}(0|n,0,0,0,P_A = p_{a2})} = \frac{\frac{\delta(0)(1-pnp)(1-p_{a1})^n}{\left[(1-pnp)(1-p_{a1})^n + \int_{y=p_{a1}}^{1} f_{P_B}(p_{a1}|P_A = p_{a1})(1-y)^n d(y)\right]}}{\frac{\delta(0)(1-pnp)(1-p_{a2})^n}{\left[(1-pnp)(1-p_{a2})^n + \int_{y=p_{a2}}^{1} f_{P_B}(p_{a2}|P_A = p_{a2})(1-y)^n d(y)\right]}} =$$

$$\frac{(1-p_{a1})^n}{(1-p_{a2})^n}\frac{(1-pnp)(1-p_{a2})^n + \int_{y=p_{a2}}^{1} f_{P_B}(p_{a2}|P_A=p_{a2})(1-y)^n d(y)}{(1-pnp)(1-p_{a1})^n + \int_{y=p_{a1}}^{1} f_{P_B}(p_{a1}|P_A=p_{a1})(1-y)^n d(y)} \qquad \text{(B9)}$$

In the general case we cannot tell whether the ratio (B9) is greater or smaller than 1 – the value will depend on the integrals involved, which in turn will depend on the form of $f_{P_B}(\cdot\,|P_A)$.

We can look for more clarity by comparing $f_{P_B}(0|n,0,0,0,P_A = p_{a1})$ and $f_{P_B}(0|n,0,0,0,P_A = p_{a2})$ for the *special case* when the entire mass of $f_{P_B}(y > 0|P_A = x)$ is concentrated at *x*, i.e. $f_{P_B}(y > 0|P_A = x) = \delta(y - x)pnp$. In this case further simplification can be achieved:

$$\begin{aligned} f_{P_A}(x|n,0,0,0) &= \frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n + \int_{y=p_x}^{1} f_{P_B}(x|P_A = p_x)(1-y)^n d(y)\right] \\ &= \frac{f_{P_A}(x)}{K3}\left[(1-pnp)(1-x)^n + pnp\int_{y=p_x}^{1} \delta(y-x)(1-y)^n d(y)\right] \\ &= \frac{f_{P_A}(x)}{K3}[(1-pnp)(1-x)^n + pnp(1-x)^n] = \frac{f_{P_A}(x)}{K3}(1-x)^n \end{aligned}$$

The ratio of the two probability density functions (B9) then becomes:

$$\frac{f_{P_B}(0|n,0,0,0,P_A=p_{a1})}{f_{P_B}(0|n,0,0,0,P_A=p_{a2})} = \frac{f_{P_A}(p_{a1})\frac{\delta(0)(1-pnp)(1-p_{a1})^n}{(1-p_{a1})^n}}{f_{P_A}(p_{a2})\frac{\delta(0)(1-pnp)(1-p_{a2})^n}{(1-p_{a2})^n}} = \frac{f_{P_A}(p_{a1})}{f_{P_A}(p_{a2})} \qquad \text{(B10)}$$

It turns out that even if we constrain the conditional distribution of the probability of failure of channel B to a form where the entire probability mass is concentrated in two points (0 and *y* = *x*) the posterior probability of perfection may still vary with the probability of failure of channel A unless we assume that $f_{P_A}(\cdot)$ is constant (i.e. in our prior belief about the value of the probability of failure P_A we are indifferent between the values P_A can take) for all values of P_A of interest. Such indifference seems implausible, e.g., if channel A has seen a significant operational exposure. If $f_{P_A}(\cdot)$ being constant for different values of channel A's *pfd* is ruled out, then (B10) suggests that the constancy of $f_{P_B}(0|P_A = x)$, which we assumed in the prior will be lost, too, after a finite amount of failure-free operation. Interestingly, the ratio (B10) does not seem to change with the number of observations, *n*, as all terms in (B10) dependent on the number of demands *n* are cancelled out.

QED.

Appendix C

In this appendix we establish a relationship between the prior and the posterior probability of perfection of the system, i.e., between $F_{P_{AB}}(z = 0)$ and $F_{P_{AB}}(z = 0|n,0,0,0)$, respectively.

We will first express the posterior marginal distribution of system *pfd*, $f_{P_{AB}}(z|n,0,0,0)$, in the case of observing no failures in operation/testing and then from it we will look at the probability mass $F_{P_{AB}}(0|n,0,0,0)$.

$f_{P_{AB}}(z|n,0,0,0)$ can be derived from the 3-variate *pdf*, $f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)$, by integrating out the nuisance parameters, P_A and P_B, i.e.:

$$\begin{aligned}
f_{P_{AB}}(z|n,0,0,0) &= \int_{x=0}^{1}\int_{y=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z|n,0,0,0)d(y)\,d(x) \\
&= \frac{\int_{x=0}^{1}\int_{y=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)(1-x-y+z)^n d(y)\,d(x)}{K_p} \\
&= \frac{\int_{x=0}^{1}\int_{y=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x) f_{P_A}(x)(1-x-y+z)^n d(y)\,d(x)}{K_p} \\
&= \frac{\int_{x=0}^{1} f_{P_A}(x)\left[\int_{y=0}^{1} f_{P_{AB}}(z|P_A=x,P_B=y) f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right]d(x)}{K_p} \qquad \text{(C1)}
\end{aligned}$$

where $K_p = \int_{x=0}^{1}\left\{\int_{y=0}^{1}\left[\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z)(1-x-y+z)^n\,d(z)\right]d(y)\right\}d(x)$.

We split the integration over y (the part of expression C1 shown in square brackets above) into two parts – around $y = 0$ and for $y > 0$. The internal integral, shown in square brackets above, becomes:

$$\int_{y=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x)(1-x-y+z)^n d(y) =$$

$$\int_{y=0}^{0+} f_{P_{AB}}(z|P_A = x, P_B = 0) f_{P_B}(y|P_A = x)(1-x-0+z)^n d(y)$$

$$+ \int_{y=0+}^{1} f_{P_{AB}}(z|P_A = x, P_B = y, y > 0) f_{P_B}(y|P_A = x, y > 0)(1-x-y+z)^n d(y) \qquad \text{(C2)}$$

Taking into account that $f_{P_{AB}}(z|P_A = x, P_B = 0) = \delta(z)$[9] and $f_{P_B}(0|P_A = x) = \delta(y)(1-pnp)$, the first summand above can be simplified as follows:

$$\int_{y=0}^{0+} f_{P_{AB}}(z|P_A = x, P_B = 0) f_{P_B}(0|P_A = x)(1-x-0+z)^n d(y) =$$

$$\int_{y=0}^{0+} \delta(z=0)\delta(y)(1-pnp)(1-x-y+z)^n d(y) =$$

$$\delta(z=0)(1-pnp)(1-x+z)^n =$$

$$\delta(0)(1-pnp)(1-x+z)^n \qquad \text{(C3)}$$

For the second summand of (C2), we note that $f_{P_B}(y|P_A = x, y > 0) = 0$ for any $y < x$ and that Assumption 2 (see section 2.3) implies that $f_{P_{AB}}(z|P_A = x, P_B = y) = \delta(z-x)$. Thus, without loss of generality, we can change the bounds of integrations and simplify the integral, as follows:

$$\int_{y=0+}^{1} f_{P_{AB}}(z|P_A = x, P_B = y, y > 0) f_{P_B}(y|P_A = x)(1-x-y+z)^n d(y) =$$

$$\int_{y=x}^{1} \delta(z-x) f_{P_B}(y|P_A = x)(1-x-y+z)^n d(y) \qquad \text{(C4)}$$

Thus, expression (C1) becomes:

[9] The Dirac Delta function, $\delta(z)$, takes value 0 everywhere except at point 0, where its value is infinity. $\delta(z-a)$ implies that the spike is at z=a. The fundamental property of the Dirac Delta function is that $\int_{x=a-}^{a+} \delta(x-a)f(x)d(x) = f(a)$. Also $\delta(ax) = \frac{\delta(x)}{|a|}$.

$$
\begin{aligned}
f_{P_{AB}}(z|n,0{,}0{,}0) &= \frac{1}{K_p}\int_{x=0}^{1} f_{P_A}(x)\Bigg[\delta(0)(1-pnp)(1-x+z)^n \\
&\quad + \int_{y=x}^{1}\delta(z-x)f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\Bigg]d(x) \\
&= \frac{1}{K_p}\Bigg\{\delta(0)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x+z)^n d(x) \\
&\quad + \int_{x=0}^{1} f_{P_A}(x)\delta(z-x)\left[\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right]d(x)\Bigg\}
\end{aligned}
$$

Now after well-known transformations (see http://mathworld.wolfram.com/DeltaFunction.html) and changing the order of integration, the second summand of the expression above becomes:

$$
\begin{aligned}
&\int_{x=0}^{1} f_{P_A}(x)\delta(x-z)\left[\int_{y=x}^{1} f_{P_B}(y|P_A=x)(1-x-y+z)^n d(y)\right]d(x) \\
&\quad = \int_{y=x}^{1}\left[\int_{x=0}^{1}\delta(x-z)\, f_{P_B}(y|P_A=x)f_{P_A}(x)(1-x-y+z)^n dx\right]dy \\
&\quad = \int_{y=x}^{1} f_{P_B}(y|P_A=z)f_{P_A}(z)(1-y)^n d(y) = f_{P_A}(z)\int_{y=x}^{1} f_{P_B}(y|P_A=z)(1-y)^n d(y)
\end{aligned}
$$

Thus:

$$
f_{P_{AB}}(z|n,0{,}0{,}0) = \frac{1}{K_p}\left[\delta(z)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x+z)^n d(x) + f_{P_A}(z)\int_{y=x}^{1} f_{P_B}(y|P_A=z)(1-y)^n d(y)\right]
$$

Clearly, the first summand will be equal to 0 for any $z \neq 0$ and will be infinity for $z = 0$ due to the Dirac Delta function, $\delta(z)$. The contribution of the second term for $z = 0$, thus, becomes negligible. We can rewrite the posterior *pdf* as follows:

$$
f_{P_{AB}}(z|n,0{,}0{,}0) = \begin{cases} \dfrac{\delta(0)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x)^n d(x)}{K_p}, z = 0 \\ \dfrac{f_{P_A}(z)\int_{y=x}^{1} f_{P_B}(y|P_A=z)(1-y)^n d(y)}{K_p}, z > 0 \end{cases} \tag{C5}
$$

One can see that if we set $n = 0$ (which will also set $K_p = 1$), (C5) will be reduced to an expression consistent with (15), as we would expect.

Now, let us compare the posterior and the prior probability of system perfection, i.e., $f_{P_{AB}}(z = 0|n, 0{,}0{,}0)$ and $f_{P_{AB}}(z = 0)$. The former is expressed by (C5) and the latter – by (15). Consider the ratio:

$$
\frac{f_{P_{AB}}(z=0|n,0{,}0{,}0)}{f_{P_{AB}}(z=0)} = \frac{\frac{\delta(0)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x)^n d(x)}{K_p}}{\delta(0)(1-pnp)} = \frac{\int_{x=0}^{1} f_{P_A}(x)(1-x)^n d(x)}{K_p} > 1 \tag{C6}
$$

The inequality follows from the observation that the normalising constants used with posteriors represent the expected values of the likelihood of the events: "no failure" for the black-box and the white-box models, $E_{no\ failure}^{black-box}$ and $E_{no\ failure}^{white-box}$, respectively. The event "no failure" for the black-box model, $E_{no\ failure}^{black-box}$, is a superset of the event $E_{no\ failure}^{white-box}$, i.e. $E_{no\ failure}^{black-box} \subseteq E_{no\ failure}^{white-box}$, from which using the axioms of probabilities we conclude that $P\left(E_{no\ failure}^{black-box}\right) \geq P\left(E_{no\ failure}^{white-box}\right)$ irrespective of the probabilistic measure used.

QED.

Appendix D

In this appendix we look at how the form of the prior distribution $f_{P_B}(y|P_A = x)$ impacts the predictions of the probability of system failure. We compare the posterior probability of perfection of the system for the cases of $f_{P_B}(y|P_A = x)$ illustrated in Figure 1 and Figure 2:

- Case 1: The probability mass is somehow spread in the interval [x, 1] (see Figure 1) , and
- Case 2 (“opt”): The entire mass of $f_{P_B}(y|P_A = x, y > 0)$ is concentrated at a single point $y = x$. In this case, $f_{P_B}(y|P_A = x) = \delta(y - x) \times pnp$ (see Figure 2).

Recall the normalising term, K_p, which we introduced in (C1).

$$\begin{aligned}K_p &= \int_{x=0}^{1} \left\{ \int_{y=0}^{1} \left[\int_{z=0}^{1} f_{P_A,P_B,P_{AB}}(x, y, z)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} d(x) \\ &= \int_{x=0}^{1} \left\{ \int_{y=0}^{1} \left[\int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y,) f_{P_B}(y|P_A = x) f_{P_A}(x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} d(x) \\ &= \int_{x=0}^{1} f_{P_A}(x) \left\{ \int_{y=0}^{1} \left[\int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y) f_{P_B}(y|P_A = x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} d(x)\end{aligned}$$

Let K_{p1} and K_{p2} denote the value of K_p for the two cases listed above, respectively.

For both cases $f_{P_{AB}}(z|P_A = x, P_B = y, y > 0) = \delta(z - x)$. The difference between K_{p1} and K_{p2} is due to the different forms $f_{P_B}(y|P_A = x, y > x)$ can take, which for the “opt” case becomes $f_{P_B}^{opt}(y|P_A = x, y > x) = \delta(y - x)$. Let us express the part of the integrals for the two cases shown in curly the brackets in the last expression above:

$$\begin{aligned}Inner^{case\ 1} &= \left\{ \int_{y=x}^{1} \left[\int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y,) f_{P_B}(y|P_A = x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \left\{ \int_{y=x}^{1} \left[\int_{z=0}^{1} \delta(z - x) f_{P_B}(y|P_A = x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \left\{ \int_{y=x}^{1} f_{P_B}(y|P_A = x) \left[\int_{z=0}^{1} \delta(z - x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \left\{ \int_{y=x}^{1} f_{P_B}(y|P_A = x)(1 - x - y + x)^n \, d(y) \right\} = \left\{ \int_{y=x}^{1} f_{P_B}(y|P_A = x)(1 - y)^n \, d(y) \right\} \\ &< \left\{ \int_{y=x}^{1} f_{P_B}(y|P_A = x)(1 - x)^n \, d(y) \right\} = (1 - x)^n \int_{y=x}^{1} f_{P_B}(y|P_A = x) \, d(y) \\ &= (1 - x)^n \times pnp \qquad \text{(D1)}\end{aligned}$$

A similar expression for case 2 (“opt”) would lead to:

$$\begin{aligned}Inner^{case\ 2} &= \left\{ \int_{y=x}^{1} \left[\int_{z=0}^{1} f_{P_{AB}}(z|P_A = x, P_B = y,) f_{P_B}^{opt}(y|P_A = x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \left\{ \int_{y=x}^{1} \left[\int_{z=0}^{1} \delta(z - x) f_{P_B}^{opt}(y|P_A = x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \left\{ \int_{y=x}^{1} f_{P_B}^{opt}(y|P_A = x) \left[\int_{z=0}^{1} \delta(z - x)(1 - x - y + z)^n \, d(z) \right] d(y) \right\} \\ &= \int_{y=x}^{1} f_{P_B}^{opt}(y|P_A = x)(1 - y)^n \, d(y) = \int_{y=x}^{1} \delta(y - x) pnp (1 - y)^n \, d(y) \\ &= pnp \int_{y=x}^{1} \delta(y - x)(1 - y)^n \, d(y) \\ &= pnp \times (1 - x)^n \qquad \text{(D2)}\end{aligned}$$

From (D1) and (D2) it is clear that: $Inner^{case\ 1} < \ Inner^{case\ 2}$, which in turns implies that:

$$K_{p1} < K_{p2} \qquad \text{(D3)}$$

Now using (C5) and (D3) we can evaluate the ratio of $f_{P_{AB}}(0|n, 0,0,0)$ and $f_{P_{AB}}^{opt}(0|n, 0,0,0)$:

$$\frac{f_{P_{AB}}(0|n,0,0,0)}{f^{opt}_{P_{AB}}(0|n,0,0,0)} = \frac{\dfrac{\delta(0)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x)^n d(x)}{K_{p1}}}{\dfrac{\delta(0)(1-pnp)\int_{x=0}^{1} f_{P_A}(x)(1-x)^n d(x)}{K_{p2}}} = \frac{K_{p2}}{K_{p1}} > 1$$

which can be rewritten as:

$$f_{P_{AB}}(0|n,0,0,0) = \frac{K_{p2}}{K_{p1}} f^{opt}_{P_{AB}}(0|n,0,0,0) \qquad \text{(D4)}$$

Clearly, using (D4) we can express the posterior probability of perfection of channel B, $F_{P_{AB}}(0|n,0,0,0)$, as follows:

$$F_{P_{AB}}(0|n,0,0,0) = \int_0^{0+} f_{P_{AB}}(z|n,0,0,0)dz =$$

$$\int_0^{0+} \frac{K_{p2}}{K_{p1}} f^{opt}_{P_{AB}}(z|n,0,0,0)dz = \frac{K_{p2}}{K_{p1}} \int_0^{0+} f^{opt}_{P_{AB}}(z|n,0,0,0)dz = \frac{K_{p2}}{K_{p1}} F^{opt}_{P_{AB}}(0|n,0,0,0) > F^{opt}_{P_{AB}}(0|n,0,0,0)$$

QED.

Appendix E

Recall that the posterior distribution for any observations is expressed by (5), which for the observation in question ($r_a = 0, r_b > 0, r_{ab} = 0$) in *n* demands becomes:

$$\begin{aligned}
f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a=0,r_b>0,r_{ab}=0) &= \frac{f_{P_A,P_B,P_{AB}}(x,y,z)L(n,r_a=0,r_b>0,r_{ab}=0|x,y,z)}{\int_0^1\int_0^1\int_0^1 f_{P_A,P_B,P_{AB}}(x,y,z)L(n,r_a=0,r_b>0,r_{ab}=0|x,y,z)} \\
&\propto f_{P_A,P_B,P_{AB}}(x,y,z) \times (P_B - P_{AB})^{r_b} \times (1 - P_A - P_B + P_{AB})^{n-r_b} \\
&= f_{P_{AB}}(z|P_A = x, P_B = y) \times f_{P_B}(y|P_A = x) \times f_{P_A}(x) \times (P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n-r_b}
\end{aligned}$$

The normalising coefficient $K = \int_{z=0}^{1}\int_{y=0}^{1}\int_{x=0}^{1} f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a=0,r_b>0,r_{ab}=0)dx\,dy\,dz$ is a positive number. This can be demonstrated using transformations similar to those used in Appendix D above (to derive D1 and D2).

Clearly, $f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a=0,r_b>0,r_{ab}=0)$ can be analysed by looking at the following two cases:

- $P_B = y = 0$. In this case $f_{P_{AB}}(z|P_A = x, P_B = y = 0) = \delta(z)$, $f_{P_B}(y|P_A = x) = \delta(y) \times (1 - pnp)$, and
- $P_B = y > 0$. In this case $f_{P_{AB}}(z|P_A = x, P_B = y > 0) = \delta(t - x)$, $f_{P_B}(y|P_A = x)$ is somehow spread between [x, 1], as shown in Figure 1 or as shown in Figure 2. This case, $P_B = y > 0$, does not affect the probability of perfection of channel B.

Let us look at the case, $P_B = 0$:

$$f_{P_A,P_B,P_{AB}}(x,y=0,z|n,r_a=0,r_b>0,r_{ab}=0)$$

$$\begin{aligned}
&\propto f_{P_{AB}}(z|P_A = x, P_B = y = 0) \times f_{P_B}(y = 0|P_A = x) \times f_{P_A}(x) \times (P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n-r_b} \\
&= \delta(z) \times (1 - pnp) \times \delta(y) \times f_{P_A}(x) \times (P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n-r_b}
\end{aligned}$$

We note that in this case $P_B = P_{AB} = 0$. The expression above, thus, becomes:

$$\begin{aligned}
&f_{P_A,P_B,P_{AB}}(x,y=0,z|n,r_a=0,r_b>0,r_{ab}=0) \\
&\qquad \propto \delta(z) \times (1 - pnp) \times \delta(y) \times f_{P_A}(x) \times (P_B - P_{AB})^{r_b}(1 - P_A - P_B + P_{AB})^{n-r_b} \\
&\qquad = \delta(z) \times \delta(y) \times (0)^{r_b}(1 - pnp) \times f_{P_A}(x)(1 - P_A - 0 + 0)^{n-r_b} \\
&\qquad = \delta(z) \times \delta(y) \times (0)^{r_b}(1 - pnp) \times f_{P_A}(x)(1 - P_A)^{n-r_b}.
\end{aligned}$$

Integrating out the nuisance parameters, P_A and P_{AB}, we derive $f_{P_B}(y = 0|n, r_a = 0, r_b > 0, r_{ab} = 0)$ as follows:

$$f_{P_B}(y=0|n,r_a=0,r_b>0,r_{ab}=0)=\int_{z=0}^{1}\int_{x=0}^{1} f_{P_A,P_B,P_{AB}}(x,y=0,z|n,r_a=0,r_b>0,r_{ab}=0)dx\,dz$$

$$=\frac{\int_{z=0}^{1}\int_{x=0}^{1}\delta(z)\times\delta(y)\times(0)^{r_b}(1-pnp)\times f_{P_A}(x)(1-x)^{n-r_b}dx\,dz}{K}$$

$$=\frac{(0)^{r_b}\times(1-pnp)\delta(y)}{K}\int_{z=0}^{1}\int_{x=0}^{1}\delta(z)\times f_{P_A}(x)(1-x)^{n-r_b}dx\,dz$$

$$=\frac{(0)^{r_b}\times(1-pnp)\delta(y)}{K}\int_{z=0}^{1}\delta(z)dz\int_{x=0}^{1}f_{P_A}(x)(1-x)^{n-r_b}dx$$

$$=\frac{(0)^{r_b}\times(1-pnp)\delta(y)}{K}\int_{x=0}^{1}f_{P_A}(x)(1-x)^{n-r_b}dx$$

Clearly, the integral $\int_{x=0}^{1}f_{P_A}(x)(1-x)^{n-r_b}dx$ evaluates to a positive number, lets denote is as K$_a$. Thus, we can derive the probability of perfection of channel B, $F_{P_B}(y=0|n,r_a=0,r_b>0,r_{ab}=0)$ as follows:

$$F_{P_B}(y=0|n,r_a=0,r_b>0,r_{ab}=0)=\frac{\int_{y=0}^{0+}K_a\times(0)^{r_b}\times(1-pnp)\delta(y)dy}{K}=$$

$$\frac{K_a\times(0)^{r_b}\times(1-pnp)}{K}\int_{y=0}^{0+}\delta(y)dy=\frac{K_a}{K}\times(0)^{r_b}\times(1-pnp)=0$$

Clearly, the term $(0)^{r_b}$ is the reason for the mass to be set to 0.

QED.

The implication of this theorem is intuitively straightforward: a failure of channel B destroys the hope that the channel is perfect.

Appendix F

Here we look at the posterior of the probability of perfection of channel B derived from a conservative prior, consistent with Assumptions 1 and 2 (section 2.3) and for observations: $r_a>0,r_b=0,r_{ab}=0$ observed in n demands ($n\geq r_a$).

For the given conservative prior and the specific observations (5) leads to the following posterior:

$$f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a>0,r_b=0,r_{ab}=0)=\frac{f_{P_A,P_B,P_{AB}}(x,y,z)L(n,r_a>0,r_b=0,r_{ab}=0|x,y,z)}{\int_0^1\int_0^1\int_0^1 f_{P_A,P_B,P_{AB}}(x,y,z)L(n,r_a=0,r_b>0,r_{ab}=0|x,y,z)}$$

$$\propto f_{P_A,P_B,P_{AB}}(x,y,z)(P_A-P_{AB})^{r_a}(1-P_A-P_B+P_{AB})^{n-r_a}$$

$$=f_{P_{AB}}(z|P_A=x,P_B=y)f_{P_B}(y|P_A=x)f_{P_A}(x)(P_A-P_{AB})^{r_a}(1-P_A-P_B+P_{AB})^{n-r_a}$$

We start by noting that in this case the normalising coefficient K is positive (see Appendix D for further details):

$$K=\int_0^1\int_0^1\int_0^1 f_{P_A,P_B,P_{AB}}(x,y,z)L(n,r_a>0,r_b\geq0,r_{ab}=0|x,y,z)$$

As before, we can split the analysis of $f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a>0,r_b=0,r_{ab}=0)$ into two cases:

- $f_{P_{AB}}(z|P_A=x,P_B=y=0)$. In this case, $P_B=0$, $P_{AB}=0$, as well. Hence, $f_{P_{AB}}(z|P_A=x,P_B=y=0)=\delta(z)$, $f_{P_B}(y=0|P_A=x)=\delta(z)(1-pnp)$. We have:
  $f_{P_A,P_B,P_{AB}}(x,y,z|n,r_a>0,r_b=0,r_{ab}=0)$
  $$\propto f_{P_{AB}}(z|P_A=x,P_B=y=0)f_{P_B}(y|P_A=x)f_{P_A}(x)(P_A-P_{AB})^{r_a}(1-P_A-P_B+P_{AB})^{n-r_a}$$
  $$=\delta(z)\times\delta(z)\times(1-pnp)(P_A-P_{AB})^{r_a}(1-P_A-P_B+P_{AB})^{n-r_a}$$

  There are two distinct sub-cases:

- ○ $P_A = 0$. As the values of all probabilities P_A, P_B and P_{AB} in this sub-case become all equal to 0, it follows that $P_A - P_{AB} = 0$ and $1 - P_A - P_B + P_{AB} = 1$. Thus:

$$\begin{aligned} &f_{P_A,P_B,P_{AB}}(x = 0, y = 0, z = 0|n, r_a > 0, r_b = 0, r_{ab} = 0) \\ &\qquad \propto f_{P_{AB}}(z|P_A = x > 0, P_B = y = 0)f_{P_B}(y|P_A = x)f_{P_A}(x)(0 - 0)^{r_a}(1 - 0 - 0 + 0)^{n-r_a} \\ &\qquad = \delta(z)\delta(z)(1 - pnp)(0)^{r_a}(1)^{n-r_a} = [\delta(z) \times z]\delta(z)(0)^{r_a}(1)^{n-r_a}(1 - pnp) = 0 \end{aligned}$$

  It turns out that this posterior probability density, $f_{P_A,P_B,P_{AB}}(x = 0, y = 0, z = 0|n, r_a > 0, r_b = 0, r_{ab} = 0)$ will be set to 0.

- ○ $P_A > 0$. It is easy to see that for this sub-case: $P_A - P_{AB} = P_A > 0$ and $1 - P_A - P_B + P_{AB} = 1 - P_A$. Thus, the posterior probability density becomes:

$$\begin{aligned} &f_{P_A,P_B,P_{AB}}(x = p, y = 0, z = 0|n, r_a > 0, r_b = 0, r_{ab} = 0) \\ &\qquad \propto f_{P_{AB}}(z|P_A = x = 0, P_B = y = 0)f_{P_B}(y|P_A = x)f_{P_A}(x)(P_A)^{r_a}(1 - P_A)^{n-r_a} \\ &\qquad = \delta(z)\delta(y)(1 - pnp)(P_A)^{r_a}(1 - P_A)^{n-r_a} > 0 \end{aligned}$$

  It turns out that a non-zero density will be retained[10]. The values assigned to this density will, of course, be dependent on the *normalising* coefficient.

- $f_{P_{AB}}(z|P_A = x, P_B = y > 0)$. In this case, $f_{P_{AB}}(z|P_A = x, P_B = y > 0) = \delta(t - x)$. This case does not affect directly the probability of perfection of channel B but affects it *indirectly*. The *pdf* of the probability of failure of channel B is distinctly different depending on how the values *x* and *y* are related:
  - ○ $f_{P_B}(y > 0|P_A = x, y < x) = 0$
  - ○ $f_{P_B}(y > 0|P_A = x, y \geq x)$ may be greater than 0 or equal to 0.

  If $f_{P_B}(y > 0|P_A = x) = 0$, the inference cannot change that. Let us concentrate on the cases where $f_{P_B}(y > 0|P_A = x, y \geq x)$. Clearly, according to assumption 2 made in section 2.3, it follows that $P_{AB} = P_A$, hence $P_A - P_{AB} = 0$. Further, $1 - P_A - P_B + P_{AB} = 1 - P_B$. Thus, we can establish that:

$$\begin{aligned} &f_{P_A,P_B,P_{AB}}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0) \\ &\qquad = \frac{f_{P_A,P_B,P_{AB}}(x, y \geq x, z)(P_A - P_{AB})^{r_a}(1 - P_A - P_B + P_{AB})^{n-r_a}}{K} \\ &\qquad = \frac{f_{P_{AB}}(z|P_A = x, P_B = y \geq x)f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)(0)^{r_a}(1 - P_B)^{n-r_a}}{K} \\ &\qquad = \frac{\delta(t - x)f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)(0)^{r_a}(1 - P_B)^{n-r_a}}{K} \end{aligned}$$

We can now derive the posterior probability density function, $f_{P_B}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0)$ by integrating out the nuisance parameters, *x* and *z*:

$$\begin{aligned} f_{P_B}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0) &= \int_{z=0}^{1}\int_{x=0}^{1} f_{P_A,P_B,P_{AB}}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0)dx\,dz \\ &= \int_{z=0}^{1}\int_{x=0}^{1} \frac{\delta(z - x)f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)(0)^{r_a}(1 - P_B)^{n-r_a}}{K}dx\,dz \\ &= \frac{(0)^{r_a}(1 - P_B)^{n-r_a}}{K}\int_{z=0}^{1}\int_{x=0}^{1} \delta(z - x)f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)dx\,dz \end{aligned}$$

Changing the order of integration – over z and x will allow us to simplify the expression as follows:

$$\begin{aligned} f_{P_B}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0) &= \frac{(0)^{r_a}(1 - P_B)^{n-r_a}}{K}\int_{x=0}^{1} f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)\left[\int_{z=0}^{1} \delta(z - x)dz\right]dx \\ &= \frac{(0)^{r_a}(1 - P_B)^{n-r_a}}{K}\int_{x=0}^{1} f_{P_B}(y|P_A = x, y \geq x)f_{P_A}(x)dx \end{aligned}$$

[10] For brevity, we omit the steps of how from the *pdf*, $f_{P_A,P_B,P_{AB}}(x = p, y = 0, z = 0|n, r_a > 0, r_b = 0, r_{ab} = 0) > 0$ we arrive at $F_{P_B}(y > 0|P_A = x) > 0$. These steps are identical to those shown in Appendix E, but the difference is that the posterior *pdf* is positive, e.g., does not contain a term which sets it to 0.

Given the definition of $f_{P_B}(y|P_A = x, y \geq x)$ and $f_{P_A}(x)$, it is clear that $\int_{x=0}^{1} f_{P_B}(y|P_A = x, y \geq x) f_{P_A}(x)dx > 0$, which leads to the conclusion that:

$$f_{P_B}(x, y > x, z|n, r_a > 0, r_b = 0, r_{ab} = 0) = \frac{(0)^{r_a}(1 - P_B)^{n-r_a}}{K} \int_{x=0}^{1} f_{P_B}(y|P_A = x, y \geq x) f_{P_A}(x)dx = 0$$

for values $P_B = y > 0$ and $P_A = x, y \geq x$. The reason is the term $(P_A - P_{AB})^{r_a} = (0)^{r_a} = 0$. This, however, is exactly the area in the prior, for which the probability density function, $f_{P_B}(x, y > x, z)$ might be non-zero except the points in the prior, which capture the probable perfection of channel B: $f_{P_A,P_B,P_{AB}}(x, y = 0, z)$. Thus, it appears that a failure of channel A will make the entire probability mass in the posterior $f_{P_B}(x, y, z|n, r_a > 0, r_b = 0, r_{ab} = 0)$ *shift* from values of $P_B > 0$ to $P_B = 0$. This shift will occur irrespective of the values of P_A. In other words, a failure of channel A and no failures of channel B will make us believe that *channel B is perfect with certainty*!

QED.

Appendix G

In this appendix we look at the posterior probability of perfection of channel B derived with a conservative prior consistent with Assumptions 1 and 2 and for observations of system failures only (i.e., simultaneous failures of channel A and B): $r_a = 0, r_b = 0, r_{ab} > 0$ observed in n demands ($n \geq r_a$).

As in the previous cases we start with (5), which for the particular observation becomes:

$$\begin{aligned} f_{P_A,P_B,P_{AB}}(x, y, z|n, r_a = 0, r_b = 0, r_{ab} > 0) &\propto f_{P_A,P_B,P_{AB}}(x, y \geq x, z) \times (P_{AB})^{r_{ab}}(1 - P_A - P_B + P_{AB})^{n-r_{ab}} \\ &= f_{P_{AB}}(z|P_A = x, P_B = y) \times f_{P_B}(y|P_A = x,) \times f_{P_A}(x) \times (P_{AB})^{r_{ab}}(1 - P_A - P_B + P_{AB})^{n-r_{ab}} \end{aligned}$$

As in the previous cases, let us denote as K the normalising coefficient:

$$K = \int_0^1 \int_0^1 \int_0^1 f_{P_A,P_B,P_{AB}}(x, y, z) L(n, r_a = 0, r_b = 0, r_{ab} > 0|x, y, z)$$

Using transformation similar to those used in Appendix D to derive D1 and D2 it can be shown that K > 0. We split the analysis of the posterior density function $f_{P_A,P_B,P_{AB}}(x, y, z|n, r_a = 0, r_b = 0, r_{ab} > 0)$ into two sub-cases:

- $P_B = y = 0$. Clearly, in this case $P_{AB} = 0$.

$$\begin{aligned} &f_{P_A,P_B,P_{AB}}(x, y = 0, z|n, r_a = 0, r_b = 0, r_{ab} > 0) \\ &\quad = \frac{f_{P_{AB}}(z|P_A = x, P_B = y = 0) \times f_{P_B}(y = 0|P_A = x) \times f_{P_A}(x) \times (0)^{r_{ab}}(1 - P_A)^{n-r_{ab}}}{K} \\ &\quad = \delta(z) \times \delta(y) \times (1 - pnp) \times f_{P_A}(x) \times (0)^{r_{ab}} \times (1 - P_A)^{n-r_{ab}} \\ &\quad = [\delta(z)] \times \delta(y) \times (1 - pnp) \times f_{P_A}(x) \times (0)^{r_{ab}} \times (1 - P_A)^{n-r_{ab}} \end{aligned}$$

By integrating out the nuisance parameters, P_A and P_{AB} we can express the posterior marginal probability density function, $f_{P_B}(x, y = 0, z|n, r_a = 0, r_b = 0, r_{ab} > 0)$ as follows:

$$\begin{aligned} f_{P_B}(y = 0|n, r_a = 0, r_b = 0, r_{ab} > 0) &= \frac{\int_{z=0}^{1} \int_{x=0}^{1} f_{P_A,P_B,P_{AB}}(x, y = 0, z|n, r_a = 0, r_b = 0, r_{ab} > 0)dx\,dz}{K} \\ &= \frac{(0)^{r_{ab}} \times \delta(y) \times (1 - pnp)}{K} \int_{z=0}^{1} \int_{x=0}^{1} [\delta(z)] \times f_{P_A}(x) \times (1 - x)^{n-r_{ab}} dx\,dz \\ &= \frac{(0)^{r_{ab}} \times \delta(y) \times (1 - pnp)}{K} \int_{z=0}^{1} [\delta(z)]dz \int_{x=0}^{1} f_{P_A}(x) \times (1 - x)^{n-r_{ab}} dx \\ &= \frac{(0)^{r_{ab}} \times \delta(y) \times (1 - pnp)}{K} \int_{x=0}^{1} f_{P_A}(x) \times (1 - x)^{n-r_{ab}} dx \end{aligned}$$

Given the definition of $f_{P_A}(x)$, the integral $K_a = \int_{x=0}^{1} f_{P_A}(x) \times (1 - x)^{n-r_{ab}} dx > 0$.

Now we can express the probability of perfection of channel B as:

$$
\begin{aligned}
F_{P_B}(y=0|n,r_a=0,r_b=0,r_{ab}>0) &= \int_{z=0}^{0+} f_{P_B}(x,y=0,z|n,r_a=0,r_b=0,r_{ab}>0)dz \\
&= \int_{z=0}^{0+} (0)^{r_{ab}} \frac{K_a \times \delta(y) \times (1-pnp)}{K} dz = (0)^{r_{ab}} \frac{K_a \times (1-pnp)}{K} \int_{z=0}^{0+} \delta(y) dz \\
&= (0)^{r_{ab}} \frac{K_a \times (1-pnp)}{K} = 0
\end{aligned}
$$

As in Appendix E we have a term $(P_{AB})^{r_{ab}}$, which sets the probability of channel B perfection $F_{P_B}(y=0|n,r_a=0,r_b=0,r_{ab}>0)$ to 0. This conclusion is quite plausible: if channel B has failed (in this case simultaneously with channel A) it cannot be considered perfect any longer. The case is a special case of observing channel B failing on its own, which we analysed in Appendix E above.

- $P_B = y > 0$. This case does not affect the probability of perfection of channel B and the analysis is omitted.

QED.