Why Black-Box Bayesian Safety Assessment of Autonomous Vehicles is Problematic and What Can be Done About it?

Peter Popov, Senior Member of IEEE

All models are wrong, some are useful.
George Box, a British Statistician

Abstract — **This paper deals with the Bayesian safety assessment of autonomous vehicles (AV) conducted via driving AVs on the public roads, often referred to as "driving to safety." A key safety measure is the probability of *catastrophic failure* (i.e., a road accident) per mile of driving (*pfm*), assumed a random variable.**

We argue that a Bayesian prediction based on a univariate ("black-box") probabilistic model has an *intrinsic deficiency*: it cannot accommodate the variation of *pfm* due to changing road conditions, which in turn may affect significantly the predicted *pfm* and may lead to optimistic predictions.

A multivariate probabilistic model is developed to overcome this limitation of the univariate model. Using a set of contrived examples the predictions of the multivariate model are compared with those derived with univariate models. Our results provide an intriguing insight that even when AV driving does not lead to accidents at all, the *pfm* predictions with the multivariate model may be *more pessimistic* than the assumed prior, and those derived with a black-box model, including the predictions using the recently developed "conservative Bayesian inference".

The multivariate Bayesian safety assessment can be applied to autonomous vehicles and to other complex intelligent systems such as robots, UAVs, etc., where the operating conditions vary.

Index Terms— **Autonomous vehicle, Bayesian inference, "driving to safety," Safety Assessment**

1. INTRODUCTION

Autonomous vehicles (AVs)[1] and other intelligent systems, which rely on machine learning (ML) or artificial intelligence (AI) for some of its functionality (e.g., perception, planning, etc.), have challenged many mature methods for safety assessment developed over the years for software-based cyber-physical systems (CPS). A noticeable recent example is the concept of "driving to safety", formulated in [2], which is used to assess the AV safety from data collected during driving an AV on the public roads. [2] and other related studies demonstrated that the amount of AV driving required for an AV to demonstrate levels of safety comparable with the safety of man-driven vehicles is very high (in excess of 10s of millions of miles), an observation which motivated the search for alternative methods for AV safety assessment, e.g. scenario – based testing.

The findings in [2] have been picked up by research teams in the UK, which led to the development of the "Conservative Bayesian Inference (CBI)". As the name suggests, CBI solves the "driving to safety" problem *conservatively*. Conservatism is defined precisely [3] and is summarized in section 3.2 below. The essence of the approach is that conservatism in predicting the confidence that an AV has reached a *target value* of *pfm* is sought over the space of *all* prior distributions of *pfm* (treated as a random variable) which satisfy a given prior confidence in the chosen target value of *pfm*.

1.1. Abbreviations

A number of publications, e.g., [4, 5], deal with CBI assuming that *no accidents* are observed in operation. Some other studies, e.g. [6], refined CBI for the cases with observed accidents. Being able to apply a safety assessment method to observations with accidents is important as the current AVs are far from being able to run millions of miles (kilometers) without an accident due to the limitations of the AV technology. Empirical data from AV road testing suggest that the current AVs do fail quite frequently [7]. Some more specific studies, e.g. [8], suggest that the state-of-the-art in object recognition solutions (an essential part of an AV perception) have very modest probability of successful object detection of ~80% only. Likewise, the "safety monitors", the mechanisms deployed in an AV to guarantee safety in situations when some components of the AV stack fail, are also far from being perfect [9]. Given the modest level of reliability of key components for an AV such as perception and safety monitors would suggest that one should not be surprised by regular AV accidents. Abbreviations

ODD – Operational Design Domain [3] defines the operating conditions for which an AV safety claim applies.

The manuscript was submitted on the 4th of October 2024.

Peter Popov is with the Centre for Software Reliability, at City, St. George's University of London, Northampton Square, EC1 0HB, United Kingdom. Email address: p.t.popov@city.ac.uk.

The work was supported by the Intel Collaborative Research Institute on Safety of Autonomous Vehicles (ICRI - SAVe).

[1] In this paper we adopt the term "autonomous vehicles (AV)." The theory we develop would apply to Level 4 and Level 5 defined by SAE [1] for "automated driving systems (ADS)" with a complex set of driving tasks performed in sophisticated operational environments. Autonomous vehicles are seen as a broad category of vehicles including ADS as defined in [1], but also other types of vehicles, e.g. the unmanned autonomous vehicles (UAV), robots, etc.

ODD consists of a set of explicitly defined operating conditions.

OC – an operating condition, an abstraction used to define an ODD. Typically, the operating conditions are linked to i) the AV driving conditions (e.g., on the motorway vs. in rural/urban area), and to ii) weather conditions (sunny, rainy, snow, etc.).

$\mathrm{OC_i}$ – the i-th operating condition of an ODD.

pfm – probability of failure/accident per mile of driving. A measure of safety used in the "driving to safety" approach.

pfm_i –probability of failure/accident per mile, conditional on the mile being driven in operating condition $\mathrm{OC_i}$.

$\mathrm{P(OC_i)}$ – probability of an AV driving a randomly chosen mile in $\mathrm{OC_i}$.

1.2. Notations

X – r.v. random variable

$f_x(\cdot)$ - probability density function of the r.v. X

Θ - r.v. representing *pfm*

Θ_i - r.v. representing *pfm_i*

Ψ_i - r.v. representing $\mathrm{P(OC_i)}$

$f_\theta(\cdot)$ - probability density function of Θ

$f_{\theta_i}(\cdot)$ - probability density function of Θ_i

$E[\Theta]$ - expected values of Θ

$E[\Theta_i]$ - expected value of Θ_i

$f_{\psi_i}(\cdot)$ - probability density function of Ψ_i

$E[\psi_i]$ - expected value of Ψ_i

$f_x(\cdot) * f_y(\cdot)$ - the convolution of the probability density functions of two independently distributed random variables, X and Y

$\mathrm{Dir}(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n; \mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n})$ – the Dirichlet distribution of non-negative random variables $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$

Beta(x;α, β) – a Beta distribution of the r.v. X with parameters α and β.

$L(N, r|x)$ – the likelihood of observing *r* failures in *N* miles of driving, given the values of *pfm* is *x* (i.e., *pfm* = *x*)

2. Motivation for the paper

While univariate Bayesian predictions have been extensively used in the past, e.g., to predict the probability of failure on demand of safety critical systems, including the "ultra reliable" ones [10], doubts have been voiced over the years about the adequacy of the univariate prediction models in favor of multivariate counterparts [11], [12, 13] and more recently [14]. With a univariate reliability prediction, the system under assessment is modelled as a black-box and any details about the system (e.g., its architecture) are ignored in the assessment. A multivariate inference, instead relies on a more detailed model of either the system or of its operational environment, which allows the assessor to capture important details, e.g. the use of redundancy in the system architecture, or the existence of distinctly different operating conditions which affect the system reliability and/or the reliability of some of its components, e.g. [15] and [16]. In the context of AV "driven to safety", the use of a univariate inference seems particularly problematic as it does not explicitly deal with the fact that the *N* miles processed by the Bayesian inference may be driven in quite different operating conditions. Two extreme examples of this would be: i) during road tests an AV is driven in "easy conditions" *only*, e.g., on a motorway with a light traffic, or ii) an AV is driven in difficult conditions for *atypically lengthy periods* of time. Intuitively, the likelihood of an accident in the first case seems lower, and may be much lower, than in the second case. A "driving to safety" argument based on a single measure of interest such as *pfm*, *implicitly* assumes that the risk of an accident can be assumed the *same for all miles*, irrespective of the operating conditions under which the miles are driven. This assumption is clearly *implausible*. A weaker argument in favor of a univariate inference often referred to is that the driving conditions for which the observations are collected are "typical", and that the operating conditions anticipated in the future remain *stochastically* the same as the conditions for which the observations have been collected and a safety claim – made. Such stochastic similarity, however, is merely a strong assumption, which is difficult to justify and even more difficult to enforce.

The community dealing with the AV safety assessment recognized that making a safety claim for conditions which are difficult to foresee is problematic. To confine this difficulty, the concept of the Operational Design Domain (ODD) [3] has been adopted. ODD includes a set of distinctly different operating conditions (OCs) linked to different aspects of the AV operational environment, e.g., the weather, the intensity of the traffic, time of day, etc. The terms ODD was introduced by SAE 3016J [1] for level 3 – 5 of AVs in recognition of the fact that safety is impossible to justify for an *arbitrary* operating condition; ODD defines the set of operating conditions for which an AV safety claim applies. Further details on ODD are provided in [3, 17, 18] which spell out practical ways of defining distinctly different operating conditions, mapping them to ranges of values of attributes and to constraints, defined in an ODD specification, together with mechanisms of monitoring the operating condition an AV is in at any point in time. For monitoring OpenODD defines the concept of "Current Operational Domian (COD)", seen a snapshot of all sensor readings installed in an AV. The AV is expected to update COD regularly, which will allow it to detect the OC the AV is in, including if the AV is outside the ODD boundaries. A safety claim made for a given ODD will only apply to the OCs included in the ODD. Any accidents that have occurred in conditions "outside the ODD (OoODD)" would not compromise a safety claim.

If we accept that a safety claim is based on a specific ODD, then driving to safety assessment should account for the fact that the necessary *N* miles of driving to safety will be split between the OCs included in the ODD. *Different operating conditions* OCs would imply: i) the likelihood of OCs may vary - some OCs are more likely to occur in operation than others and the OCs likelihoods may *change over time*, and ii) the likelihood of an accident may also vary between the OCs. Extensive evidence of this variation exists, e.g. from [19]: the crashes reported for AVs differ remarkably between different operating conditions linked to the weather at the time of crashes.

None of the two effects of the OCs on AV safety, however, seems accounted for in the univariate models used with "driving to safety". Accounting for these effects is the focus of this paper.

The paper makes the following contributions:

1. We develop a new method for safety assessment, based on a multivariate Bayesian assessment, which allows us to account for the uncertainties in both the "operational profile[2]" in which AV is used and the risks from accidents in the different operating conditions.
2. On contrived examples we demonstrate that the proposed Bayesian inference procedure outperforms the univariate ones and captures important aspects, which remain "invisible" for inferences based on a univariate model.
3. We demonstrate a serious deficiency of the univariate Bayesian inference used in the calculations of *pfm*. On contrived examples we show that the univariate Bayesian predictions, including those based on CBI, may be *optimistic*.

The paper is organized as follows. In section 3 a formal problem statement is provided. Section 4 presents our results – using several *contrived* examples. In section 5 we discuss the implications of our findings and threats to their validity. In section 6 we survey the related research, not mentioned in the previous sections. In Section 7 we draw conclusions and outline directions for future research.

3. Problem Statement

Now we formulate the problem of AV safety assessment as a problem of Bayesian inference.

Consider that the measure of AV safety is the *probability of catastrophic failure per mile of driving*, *pfm*, as proposed in [2]. Assume further that the probability of observing an accident within a mile is not affected by the preceding miles driven by the AV. In other words, we assume that observing successive miles of driving can be modelled mathematically as a *Bernoulli trial* [20] of *selection* of miles *at random with replacement* from the population of miles using *pfm* as the parameter of the Bernoulli trial.

Let us further assume that *pfm* is a random variable, Θ, with a probability density function, $f_\theta(\cdot)$, which captures the uncertainty about the value of *pfm*. $f_\theta(\cdot)$ is typically considered a measure of "epistemic uncertainty," related to the assessor's knowledge (belief) about the value of *pfm*.

3.1. Black-box inference

Given a prior distribution $f_\theta(\cdot)$ and observing an AV drives additional *N* miles of which *r* miles are with accidents ($r \leq N$), one can derive the posterior distribution of Θ using the Bayes formula:

$$f_\theta^{BB}(x|N,r) = \frac{f_\theta(x) \times L(N,r|x)}{\int_{x=0}^{1} f_\theta(x) \times L(N,r|x)dx} \qquad (1)$$

This is the simplest Bayesian inference, which relies on a "black-box" model about Θ. The black-box model assumes that all miles on the road are "similar," i.e., there is no reason to consider variation of *pfm* over different road conditions, an assumption, which ignores the impact of operating conditions on *pfm*.

3.2. Conservative Bayesian inference (CBI)

We briefly introduce CBI using [6]. The key idea of this *special case black-box inference* is that rather than asking a Bayesian assessor to provide a complete prior distribution for the measure(s) of interest, $f_\theta(\cdot)$, as we do in section 3.3, the assessor is asked to provide a much *less detailed info*. The minimal info required in [6] is limited to a few numbers: the prior confidence θ in a reliability/safety target ϵ, e.g. a specific value of *pfm*, and the lower bound on the reliability/safety, $p_L < \epsilon$, which is theoretically possible given the hardware unreliability of the AV. p_L may be several orders of magnitude smaller than ϵ. Then conservative posterior confidence in any value of *pmf* will be obtained using *a two-point prior distribution* as shown in Figure 1[3] below for different observations (without and with accidents).

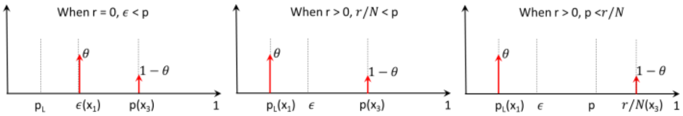


Figure 1. An illustration of CBI. A two-point prior distribution is used with the probability mass concentrated at points x_1 and x_3 shown in red in the figure. The posterior distribution will continue to be a two – point distribution with the probability mass concentrated at the same points x_1 and x_3 as in the prior.

Depending on the observations (r = 0 or r > 0 accidents) in N ($r \leq N$) miles of driving, the two – point prior distribution would guarantee a conservative prediction of $P(X \leq p|r,N)$ if the following prior is selected:

$$P(X = x) = \theta \mathbf{1}_{x>x_1} + (1-\theta)\mathbf{1}_{x>x_3} \qquad (2)$$

where $p_L \leq x_1 \leq \epsilon \leq x_3$ and the values x_1 and x_3 depend on the model parameters (p_L, ϵ, θ) and on the observations, r and N[4]. The notation $\mathbf{1}_{x>x_1}$ refers to the "unit step function" (also known as *Heaviside step function*) and denotes the fact that the function takes a value 0 for values of the argument $x \leq x_1$, and value 1 for argument $x > x_1$.

The posterior probability $P(X \leq p|r,N)$ is computed using the parameters listed above using the formula:

$$P(X \leq p|r,N) = \frac{x_1^k(1-x_1)^{N-r}\theta}{x_1^k(1-x_1)^{N-r}\theta + x_3^k(1-x_3)^{N-r}(1-\theta)} \mathbf{1}_{p>\epsilon} \qquad (3)$$

The probability mass in the second point of the posterior distribution is easy to compute as $1 - P(X \leq p|r,N)$.

3.3. White-box inference[5]

The concept of Operational Design Domain (ODD) [3], briefly introduced above, captures the idea that risks of a road

[2] Operational profile is a well-known concept in safety and reliability engineering, which we introduce formally later.

[3] All figures are included in the Appendix with much greater resolution.

[4] [6] provides further details on the relationship between x_1 and x_3, the other model parameters, and the observations, some are given in Figure 1.

[5] The term "white-box" is used often in different contexts, e.g., in software testing "white box" signifies that the tester has access to the source code of the tested software. A tester with access to the source code can plan and execute testing differently from a tester without such access who would treat the software as a "black-box." Our use of "white-box" is quite different, yet there are similarities in the sense that in both contexts "white-box"

accidents *may vary* with the operating conditions. An ODD can be seen as a *partition* of different operating (road) conditions, OCs, as follows:

OC = {OC_1, OC_2, …, OC_m} such that iff i ≠ j then $OC_i \cap OC_j = \varnothing$. An illustration of an AV driving through different OCs is given in Figure 2 below.

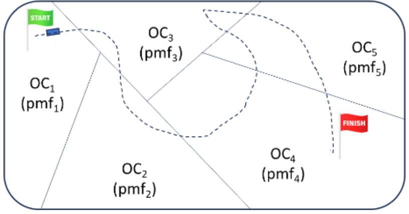


Figure 2. An illustration of ODD as a partition of operating conditions {OC$_1$, … OC$_n$}. Each OC$_i$ is characterised by the probability of accident per mile of driving, pfm$_i$, conditional on OC$_i$. A vehicle (shown as a blue rectangle on the top left of the figure) follows a "trajectory," which in the example starts in OC$_1$, crosses OC$_3$, OC$_2$, OC$_4$, OC$_5$, OC$_3$, back to OC$_5$ and finishes in OC$_4$.

Let us assume that each *OC$_i$* includes a "homogeneous" set of miles and the selection of miles can be modelled as a Bernoulli trial with a parameter *pfm$_i$* and that *pfm$_i$* may vary across OCs, (i = 1, 2, …, n).

Under the assumptions made so far, our recently developed model [15], which builds on the work by Adams [16], lands itself well to dealing with the problem at hand. Below we recall the essence of this *double-stochastic multivariate model*, which works as follows:

- Let $P(OC_i)$ be the probability of selecting at random a mile from partition *OC$_i$*. The probabilities $P(OC_i)$ define a distribution over the set OC, i.e., $\sum_{i=1}^{n} P(OC_i) = 1$.
- Let $\Theta_1, \Theta_2, \dots, \Theta_n$ be the random variables used to capture the uncertainty about the values of *pfm*$_i$ in the different OCs and $f_{\theta_i}(\cdot)$ be their respective probability density functions. Let $f_{\theta_1,\theta_2,\dots,\theta_n}(\theta_1, \theta_2, \dots, \theta_n)$ be the joint distribution of $\Theta_1, \Theta_2, \dots, \Theta_n$.

To simplify the analysis, we make an *additional* assumption that $\Theta_1, \Theta_2, \dots, \Theta_n$ are *independently distributed* random variables. In other words, we assume that changes of $f_{\theta_i}(\cdot)$ do not affect $f_{\theta_j}(\cdot), i \neq j$.

Later in the paper (Section 5, threats to validity) we discuss ways of relaxing both assumptions made above: that a Bernoulli trial should be used, and that Θ_i should be independently distributed random variables.

$P(OC_i)$, the probabilities of selecting a mile from OC_i, may vary over time or be subject to *epistemic uncertainty*, which we capture by treating $P(OC_i)$ as a random variable, Ψ_i, with a probability density function $f_{\psi_i}(\cdot)$. Since the operating conditions form a partition over the space of miles within ODD, the constraint $\sum_{i=1}^{n} \Psi_i = 1$ applies: a mile with certainty will be selected from one of the partitions. If a mile does not belong to any of the operational conditions included in an ODD, then an OoODD event should be detected.

We now express the joint distribution $f_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2,\dots, \psi_n)$, which captures the epistemic uncertainty associated with the selection of a mile from the space of all miles that belong to ODD. A suitable *multivariate distribution* which can be used here is the *Dirichlet* distribution, which satisfies the constraint $\sum_{i=1}^{n} \Psi_i = 1$. The Dirichlet distribution with *n* variates, $\Psi_i \dots, \Psi_n$ is defined as follows [21]:

$$\mathrm{Dir}(\psi_1, \psi_2,\dots, \psi_n;\boldsymbol{\alpha}) \equiv f_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2,\dots, \psi_n;a_1, \dots, a_n) = \frac{\Gamma(\sum_{i=1}^{n} a_i)}{\prod_{i=1}^{n} \Gamma(a_i)} \left[\prod_{i=1}^{n-1} {\psi_i}^{a_i-1}\right]\left[1-\sum_{i=1}^{n-1} \psi_i\right]^{a_n-1} \quad (4)$$

where $\boldsymbol{\alpha}$ is a vector $a_1, \dots, a_n$ and defines the parameters of the Dirichlet distribution.

If we denote: $A = \sum_{j=1}^{n} a_j$, then the moments of the variates of the Dirichlet distribution can be expressed as:

$$E[\Psi_i] = \frac{a_i}{A},$$

$$Var(\Psi_i) = \frac{a_i(A-a_i)}{A^2(1+A)},$$

$$Cov(\Psi_i, \Psi_j) = \frac{-a_i a_j}{A^2(1+A)}, j \neq i, \mathrm{j}$$

The marginal distribution [20] of each variate, Ψ_i, of the Dirichlet distribution is a Beta distribution, *Beta(ψ;a$_i$, A-a$_i$)*, [21].

Now, consider the case of an ODD with probabilities of the OCs known *with certainty*, i.e., $P(OC_1) = \psi_1, P(OC_2) = \psi_2, \dots P(OC_n) = \psi_n$, where ψ_i $(i = 1, \dots n)$ are known *constants*. The random variable Θ, which represents the *pfm* on a randomly chosen mile (irrespective of the operating condition it belongs to) is then the weighted sum of the random variables *Θ$_i$*, weights being the probabilities $\psi_1, \psi_2, \dots, \psi_n$, respectively.

$$\Theta_{\psi_1,\psi_2,\dots,\psi_n} = \sum_{i=1}^{n} \Theta_i \psi_i \quad (5)$$

We have already assumed that Θ_i are *independently* distributed random variables. Note that the products, $\Theta_i^{\psi_i} = \Theta_i \psi_i$, are then themselves independently distributed random variables. Let us denote the probability density functions of $\Theta_i^{\psi_i}$ as $f_{\theta\psi_i}(x)$. Then $f_{\theta\psi_i}(x)$ can be derived from $f_{\theta_i}(\cdot)$ using a standard transformation:

$$f_{\theta\psi_i}(x) = \frac{1}{|\psi_i|} f_{\theta_i}\left(\frac{x}{\psi_i}\right) \quad (6)$$

Now we can express the probability density function of $\Theta_{\psi_1,\psi_2,\dots,\psi_n}$ as follows:

$$f_{\theta|\psi_1,\psi_2,\dots\psi_n}(x|\Psi_1 = \psi_1, \Psi_2 = \psi_2, \dots, \Psi_n = \psi_n) = f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \dots * f_{\theta\psi_n}(x) \quad (7)$$

where the "*" sign in (7) above indicates a convolution of the respective probability density functions $f_{\theta\psi_i}(x)$.

Finally, we can now remove the condition that the operational profile (captured by $\Psi_1 = \psi_1, \Psi_2 = \psi_2, \dots, \Psi_n = \psi_n$) is *known with certainty* using the joint distribution $f_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2,\dots, \psi_n)$ and derive the marginal distribution:

$$f_{\theta}^{WB}(x) = \int f_{\theta|\psi_1,\psi_2,\dots\psi_n}(x|\psi_1, \psi_2, \dots, \psi_n) f_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2,\dots, \psi_n;a_1, \dots, a_n) d\psi_1 \psi x_2 \dots d\psi_n =$$

$$= \int\left[f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \dots * f_{\theta\psi_n}(x) \times f_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2,\dots, \psi_n)\right] d\psi_1 d\psi_2 \dots d\psi_n \quad (8)$$

means access to details, which are not available under "black-box" arrangements: in testing the additional details are related to the source code of software, while in the case of "white-box" model used for Bayesian inference we have access to a detailed description of the ODD assumed in the safety assessment of AV.

The integration in the last expression (8) is done with respect to all dimensions $\psi_1, \psi_2 \dots \psi_n$ of the ODD. One can see that (8) provides us with the *marginal distribution* of the *system pfm* (i.e. of the probability of an accident on a mile selected from a randomly chosen operating condition) and accounts for the *epistemic* uncertainty of both the operational profile – this is captured by the joint distribution $\mathrm{f}_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2, \dots, \psi_\mathrm{n})$ – and the distributions $f_{\theta_i}(x)$ of the conditional probabilities *pfm_i* in each partition. Clearly, $f_{\theta_i}(x)$ will affect the convolution, $f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \dots * f_{\theta\psi_n}(x)$, representing the distribution of the sum $\Theta_{\psi_1,\psi_2,\dots,\psi_n}$ expressed by (7).

We labelled (8) with "WB" ($f_\theta^{WB}(x)$) to signify the fact that this marginal distribution is derived using an inference relying on a "white-box" model accounting for both the ODD and how likely the AV is to have an accident in each of its operating conditions.

$f_\theta^{WB}(x)$, can be used differently. Apart from allowing for computing the moments, e.g., the expected value of the system *pfm*, one can compute the risk that the true probability of failure per mile can turn out to be *badly wrong* (e.g., exceed a given threshold, T), by looking at the *tail of the distribution* of system *pfm*:

$$P(\Theta \geq T) = \int_T^1 f_\theta^{WB}(x)dx$$

Let us now consider how new operational evidence from driving an AV would affect the marginal distribution $f_\theta^{WB}(x)$. Let us consider that we have collected new observations about AV performances in the form $\{(N_1, r_1), (N_2, r_2), \dots, (N_n, r_n)\}$ of miles driven, N_i, and accidents observed, r_i, $0 \leq r_i \leq N_i$, in each of the operating conditions, OC_i. Bayesian inference can be conducted in several steps:

- ***Step 1***: Update the uncertainty related to the operational profile, $\mathrm{f}_{\psi_1,\psi_2,\dots\psi_n}(\psi_1, \psi_2, \dots, \psi_\mathrm{n}|N_1, N_2, \dots, N_n)$. Note that the updated operational profile is not affected by the number of accidents, r_i, observed in OCs. The posterior distribution only depends on the number of miles driven in different OCs. If we capture the operational profile uncertainty using a Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{\alpha}) \equiv \mathrm{Dir}(\psi_1, \psi_2, \dots, \psi_\mathrm{n}; \boldsymbol{\alpha})$, then the new observations $\{(N_1, r_1), (N_2, r_2), \dots, (N_n, r_n)\}$ will lead to a new Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha^{post}})$, which is derived from $\mathrm{Dir}(\boldsymbol{\alpha})$ by a simple modification of the parameters of the Dirichlet distribution:
$$\mathrm{Dir}(\boldsymbol{\alpha^{post}}) = \mathrm{Dir}(\psi_1, \psi_2, \dots, \psi_\mathrm{n}; \alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_n + N_n) \quad (9)$$
- ***Step 2***: The conditional distributions, $f_{\theta_i}(x|N_i, r_i)$ characterizing the updated uncertainty about *pfm_i* will be updated to reflect the newly received data by conducting Bayesian inferences on the distributions of *pfm_i* in each of OC_i as follows:
$$f_{\theta_i}(x|N_i, r_i) = \frac{f_{\theta_i}(x) \times L(N_i, r_i|x)}{\int_{x=0}^1 f_{\theta_i}(x) \times L(N_i, r_i|x)dx} \quad (10)$$
(10) is identical to (1) except that we use the conditional priors $f_{\theta_i}(x)$ and the observations (N_i, r_i) collected for the respective partitions, OC_i. If the prior $f_{\theta_i}(x)$ is a Beta distribution, $Beta(x; \alpha, \beta)$, then the posterior distribution will be also a Beta distribution, $Beta(x; \alpha + r_i, \beta + N_i - r_i)$. Note that the updated conditional distribution $f_{\theta_i}(x|N_i, r_i)$ is affected by both the number of miles, N_i, and the number of failures, r_i, observed in the respective operating condition OC_i. This is the case since we have assumed that the observations in OC_i only affect Θ_i but do not affect any other Θ_j.
- ***Step 3****:* Derive $f_{\theta\psi_i}(x|N_i, r_i)$ from $f_{\theta_i}(x|N_i, r_i)$ using (6).
- ***Step 4***: Using the distributions updated in Step 1 and Step 2 above we apply (8) and derive the marginal distribution, $f_\theta^{WB_{post}}(x|N_1, r_1, N_2, r_2, \dots, N_n, r_n)$, of the system *pfm* as follows:

$$f_\theta^{WB_{post}}(x|N_1, r_1, N_2, r_2, \dots, N_n, r_n) = \int [f_{\theta\psi_1}(x|N_1, r_1) * f_{\theta\psi_2}(x|N_2, r_2) * \dots * f_{\theta\psi_n}(x|N_n, r_n)] \mathrm{Dir}(\boldsymbol{\alpha^{post}}) dx_1 dx_2 \dots dx_n \quad (11)$$

We call the last expression a "white-box" posterior distribution of the system pfm.

Steps 1 – 4 can be repeated with any new observations that become available. Consider batches of observations collected in "epochs", e_1, e_2, … e_n. The posterior distributions $Dir^{e_j}(\boldsymbol{\alpha^{post}})$ and $f_{\theta_i}^{e_j}(x|N_i, r_i)$ derived with the observations $\{(N_1^{e_j}, r_1^{e_j}), (N_n^{e_j}, r_n^{e_j}), \dots, (N_n^{e_j}, r_n^{e_j})\}$ collected within epoch e_j, will become prior(s) for the inference in epoch e_{j+1} which, in turn will use the observations $\{(N_1^{e_{j+1}}, r_1^{e_{j+1}}), (N_n^{e_{j+1}}, r_n^{e_{j+1}}), \dots, (N_n^{e_{j+1}}, r_n^{e_{j+1}})\}$, collected within epoch e_{j+1}, etc.

3.4. Black-box vs. white-box inference

In the previous subsections, 3.1 and 3.3, we defined two alternative ways of computing the posterior distribution of the system *pfm*: i) using a black-box model, which relies on (1), and ii) using a white-box model, which relies on (9), (10) and (11). The differences between the two inference procedures and of the marginal posterior distributions of the system *pfm* resulting from them are summarized in Figure 3.

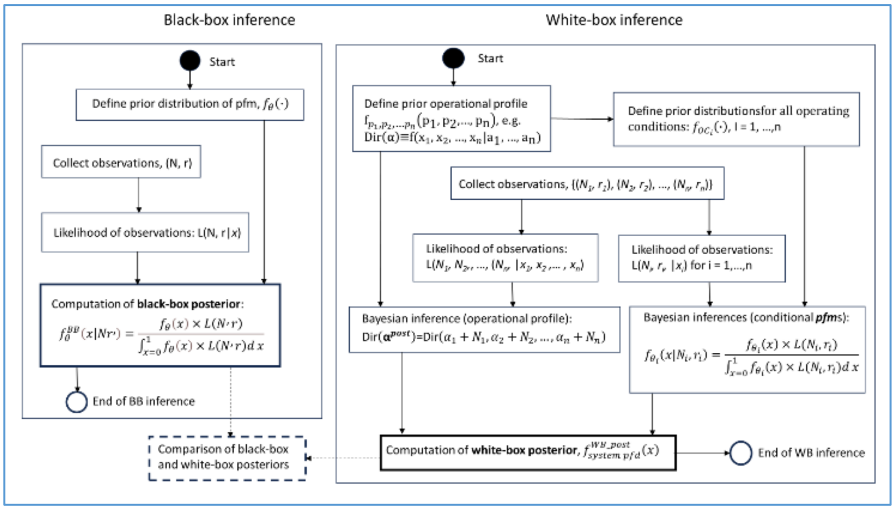

Black-box inference
White-box inference
Start
Define prior distribution of pfm, $f_\theta(\cdot)$
Collect observations, (N, r)
Likelihood of observations: L(N, r|x)
Computation of black-box posterior:
$f_\theta^{BB}(x|Nr) = \frac{f_\theta(x) \times L(N r)}{\int_{x=0}^1 f_\theta(x) \times L(N r)dx}$
End of BB inference
Start
Define prior operational profile $f_{p_1,p_2,\dots p_n}(p_1, p_2, \dots, p_n)$, e.g. $\mathrm{Dir}(\alpha) \equiv f(x_1, x_2, \dots, x_n|a_1, \dots, a_n)$
Define prior distributionsfor all operating conditions: $f_{OC_i}(\cdot)$, I = 1, …,n
Collect observations, $\{(N_1, r_1), (N_2, r_2), \dots, (N_n, r_n)\}$
Likelihood of observations: $L(N_1, N_2, \dots, (N_n \,|x_1, x_2, \dots, x_n)$
Likelihood of observations: $L(N_i, r_i \,|x_i)$ for i = 1,…,n
Bayesian inference (operational profile): $\mathrm{Dir}(\boldsymbol{\alpha^{post}}) = \mathrm{Dir}(\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_n + N_n)$
Bayesian inferences (conditional pfms):
$f_{\theta_i}(x|N_i, r_i) = \frac{f_{\theta_i}(x) \times L(N_i, r_i)}{\int_{x=0}^1 f_{\theta_i}(x) \times L(N_i, r_i)dx}$
Computation of white-box posterior, $f_{system\ pfd}^{WB\_post}(x)$
End of WB inference
Comparison of black-box and white-box posteriors

Figure 3. An illustration of the black-box and the white-box Bayesian inference, and how the posterior distributions of system pfm in both cases is derived and compared.

The black-box and the white-box inferences are clearly quite different. They require different *levels of detail* in defining the prior distributions and of collected observations. But *how significant* are the differences between the predicted marginal posteriors obtained with the two models? Can we establish systematic relationships, e.g., that the black-box predictions of system *pfm* are stochastically ordered (either pessimistic or optimistic) when compared with the white – box predictions? Answering this question is the focus of the rest of the paper.

Before we move to quantifying the difference between the predictions with the two inferences, let us discuss the conditions which will make the comparison fair.

Let us recap that we have:

- a fully defined white-box model, which includes: i) a multivariate distribution of the operational usage (e.g., captured as a Dirichlet distribution), and ii) a set of distributions $f_{\theta_i}(x)$ of the conditional *pfm*$_i$ in all operating conditions (e.g., captured as Beta distributions), and
- observations from driving an AV, which are recorded with the level of details required by the white-box model, i.e., in the format (N_1, r_1), (N_2, r_2), …, (N_n, r_n) of the number of miles driven and the number of miles with accidents observed in each operating condition, respectively.

From the white-box prior (i.e., from the multivariate distribution of the operational profile $f_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2, \ldots, \psi_n)$ and the set of marginal distributions, $f_{\theta_i}(\cdot)$ using (8) one could derive the *marginal distribution* of system *pfm*, $f_\theta^{BB}(x)$.

The detailed observations (N_1, r_1), (N_2, r_2), …, (N_n, r_n) could be aggregated in two sums: $N = \sum_{i=1}^{n} N_i$ and $r = \sum_{i=1}^{n} r_i$. Using (1) with the prior distribution $f_\theta^{BB}(x)$ and with (N, r) a black-box posterior $f_\theta^{BB_{post}}(x|N, r)$ of system *pfm (i.e., pfm*$_{BB}$) can be computed.

In parallel with the black-box inference, a white-box inference can be conducted, too. The white-box joint distribution assumed at the start of the process, is used as the prior for the white-box inference. The observations $\{(N_1, r_1), (N_2, r_2), \ldots (N_n, n)\}$ will now be used without any aggregation and will lead to a marginal posterior $f_\theta^{WB}(x|N_1, r_1, N_2, r_2, \ldots, N_n, r_n)$. In this process (9), (10) and (11) will be used. Once $f_\theta^{WB}(x|N_1, r_1, N_2, r_2, \ldots, N_n, r_n)$ is computed, it can be compared with $f_\theta^{BB_{post}}(x|N, r)$. Clearly, any difference between $f_\theta^{BB_{post}}(x|N, r)$ and $f_\theta^{WB}(x|N_1, r_1, N_2, r_2, \ldots, N_n, r_n)$ will be due *entirely* to the model used in the inference: the priors are consistent (as the marginal distributions of system *pfm* in both cases are identical) and the observations used in the inferences are identical (N_i, r_i are aggregated with the black-box inferences).

4. Contrived Examples

Let us now study the difference between the Bayesian predictions obtained with a black-box and white-box models, respectively, using several *contrived examples*.

Let us assume that an ODD is used which splits the "space of road conditions" into *five* non-overlapping operating conditions (partitions) OC$_1$, OC$_2$, OC$_3$, OC$_4$, and OC$_5$. Let us further assume that the distributions of the conditional *pfm*$_i$ are defined as *Beta* distributions with the following parameters[6]:

$f_{\theta_1}(x) \equiv Beta(\alpha = 2, \beta = 299)$, $f_{\theta_2}(x) \equiv Beta(\alpha = 2, \beta = 800)$, $f_{\theta_3}(x) \equiv Beta(\alpha = 2, \beta = 1500)$, $f_{\theta_4}(x) \equiv Beta(\alpha = 2, \beta = 1000)$, $f_{\theta_5}(x) \equiv Beta(\alpha = 1, \beta = 400)$.

The parameters of the Beta distributions are chosen to illustrate the possibility that OCs may differ both in terms of expected *pfm*$_i$ value and in terms of the *uncertainty* in the values of the conditional *pfm*$_i$ in the respective OCs.

TABLE 1 below defines three different operational profiles, OP1, OP2 and OP3, for which the probabilities P(OC$_1$), P(OC$_2$), …, P(OC$_5$) are *assumed known with certainty*. The three profiles are visibly different: Profile 1 and Profile 2 differ in the values of the probabilities P(OC$_1$) and P(OC$_3$), which are swapped. The probabilities of the other 3 OCs of Profile 1 and Profile 2 are identical. In Profile 3 all OCs are equally likely, i.e., the AV would drive an equal proportion of miles in each OC.

TABLE 1
Definition Of Three Operational Profiles, No Uncertainty.

Profile	P(OC$_1$)	P(OC$_2$)	P(OC$_3$)	P(OC$_4$)	P(OC$_5$)
Profile 1	0.1	0.2	0.4	0.25	0.05
Profile 2	0.4	0.2	0.1	0.25	0.05
Profile 3	0.2	0.2	0.2	0.2	0.2

The effect of the operational profile on the distribution of the system *pfm* is illustrated in the bottom right graph of Figure 4[7].

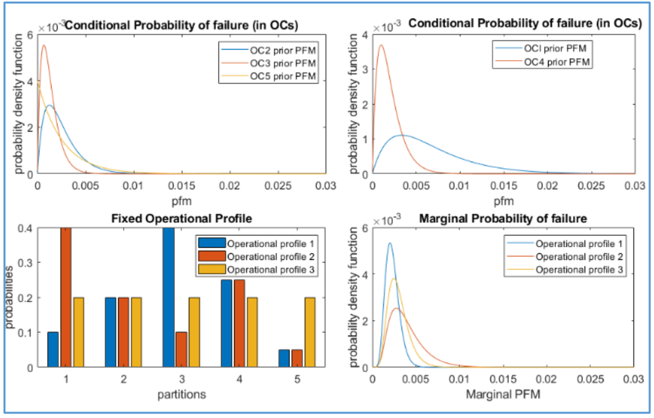


Figure 4. Illustration of the impact of the operational profile on system pfm. The same set of distributions of the conditional probabilities pfm$_i$ lead to visibly different system pfm for Profile 1, Profile 2, and Profile 3.

The system is most reliable under Profile 1 and is least

[6] Using Beta distributions is not essential for the method. If a different type of distribution is used for the conditional *pfm*$_i$ the inference will rely on numeric methods to compute the posterior distributions. Essential for the illustrations is only the assumption that the respective conditional probabilities are *independently distributed* random variables.

[7] MATLAB scripts used in the examples are available at: https://publications.city.ac.uk/viewobject.html?cid=1&id=338211

reliable under Profile 2. With Profile 3 the system reliability is in between Profile 1 and Profile 2. This ordering is not surprising – under Profile 1 the AV would spend 40% driving in OC$_1$, where the conditional probability of failure is the worst (the highest). Under Profile 2 instead the AV would spend 40% of driving in the partition where the AV is most reliable. Finally, with Profile 3 – the AV spends equal number of miles in all OCs and is less affected than under Profile 1 by the proportion of driving in OC$_1$ with worst reliability. In other words, the proposed method of deriving the distribution of the marginal *pfm* appears plausible and captures well our expectations to see the operational profile affecting visibly the system's *pfm*.

4.1. Example 1: Impact of the Operational profile uncertainty on pfm.

Let us now introduce uncertainty to the operational profile by assuming that it is captured by a Dirichlet distribution with the following parameters:

$$\mathrm{Dir}(\psi_1, \psi_2, \ldots, \psi_\mathrm{n}; \boldsymbol{\alpha}) \equiv \mathrm{Dir}(\psi_1, \psi_2, \ldots, \psi_\mathrm{n}; \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10)$$

This distribution suggests that OC$_3$ is the most likely operational condition (with expected probability of driving a mile in OC$_3$ of 0.4 (40/ (10+10+40+30+10) = 40/100). OC$_4$ is the second most likely operational condition with the expected value of the probability of driving a mile in it of 0.3 (30/100). The remaining three operating conditions – OC$_1$, OC$_2$ and OC$_5$ are equally likely with expected values of the probability of driving a mile in each of them 0.1 (10/100). Assuming that the prior distributions of the conditional *pfm$_i$* are Beta distributions, with parameters as defined at the start of section 4, we can now compute the prior distribution of the system *pfm*, shown in Figure 5.

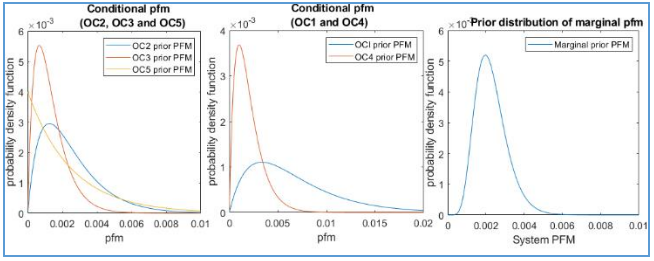


Figure 5. Prior distributions: conditional probabilities of failure, $f_{\theta_i}(x)$, and the marginal distribution of the probability of system failure under the assumed operational profile Dir($\psi_1, \psi_2, \ldots, \psi_n; \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10$).

We chose the parameters so that: i) they model OCs with different likelihoods, and ii) the likelihood of accident varies between OCs. OC$_1$ is one of the least likely operating conditions. Its impact on the probability of system failure is lower than the impact of OC$_3$ and OC$_4$. The probabilities *pfm$_i$* also vary between the driving conditions. OC$_1$ has the worst conditional probability of accident.

4.2. Example 2: Comparison of the black-box and white-box model predictions of system pfm

Now let us consider the situation with 5 identical AVs (i.e., of the same type and model of AV). Each of the AVs (AV$_1$ … AV$_5$) is assumed to have driven 100 additional miles. The observations collected by the fleet of 5 AVs, thus, are 500 miles (the sum of the miles driven by all AVs). TABLE 2 shows two similar scenarios with observations of 500 miles in total in each.

In Observation 1 none of the AVs experienced any accidents. In observation 2 AV3 observed two accidents – one in OC$_1$ and one in OC$_2$. The other vehicles (AV$_1$, AV$_2$, AV$_4$ and AV$_5$) did not observe any accidents. We chose the counts of miles driven in Observation 1 and Observation 2 to be identical for all AVs.

Now let us look at the impact of the model (black-box or white-box) used in the Bayesian inference applied by the AV *vendor* to the observations collected by the AV fleet (AV$_1$, …, AV$_5$) in different scenarios.

TABLE 2

OBSERVATIONS BY AV$_1$ … AV$_5$.

	AV ID	**N$_1$**	**r$_1$**	**N$_2$**	**r$_2$**	**N$_3$**	**r$_3$**	**N$_4$**	**r$_4$**	**N$_5$**	**r$_5$**	**Total**
Observation 1	AV1	7	0	9	0	45	0	30	0	9	0	100
	AV2	10	0	45	0	30	0	8	0	7	0	100
	AV3	45	0	30	0	7	0	9	0	9	0	100
	AV4	20	0	20	0	20	0	20	0	20	0	100
	AV5	45	0	19	0	7	0	9	0	20	0	100
	Vendor	127	0	123	0	109	0	76	0	65	0	500
Observation 2	AV1	7	0	9	0	45	0	30	0	9	0	100
	AV2	10	0	45	0	30	0	8	0	7	0	100
	AV3	45	1	30	1	7	0	9	0	9	0	100
	AV4	20	0	20	0	20	0	20	0	20	0	100
	AV5	45	0	19	0	7	0	9	0	20	0	100
	Vendor	127	1	123	1	109	0	76	0	65	0	500

The prior distributions, $f_{\theta_i}(x)$, of the conditional probabilities of failure, in OC$_1$, … OC$_5$ are as defined in the previous examples above. For the operational profile we use a Dirichlet distribution:

$$\mathrm{Dir}(\psi_1, \psi_2, \ldots, \psi_\mathrm{n}; \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10).$$

Figure 6 plots the prior and posterior distributions of the conditional probabilities of failure in OC$_1$, … OC$_5$, and of the marginal distribution of system *pfm* with the data from all 5 AVs (which in the figure are referred to as "fleet data") using a Bayesian inference with a black-box (BB) and a white-box (WB) models, respectively. The top three plots illustrate the results with Observation 1 (i.e., when no AV observed any accidents). The bottom three plots show the predictions with Observation 2 (AV$_3$ observed a failure while driven in OC$_1$ and in OC$_2$). The posterior distributions with both observations are labelled "fleet data".

The differences between the prediction with the white-box and the black-box models and *quite visible* for both Observation 1 and Observation 2. In both cases the tails of the posterior distributions obtained with the white-box model are "thicker." In other words, the white-box predictions suggest that larger values of the system *pfm* are *more likely* than the black-box predictions suggest, i.e., the white-box predictions are *more conservative*.

What is even more surprising is the comparison with the prior distributions of the system *pfm*. With both observations the black-box predictions are more optimistic than the prior, while the white-box predictions are more pessimistic than the prior. The conservatism in the white-box predictions is indeed surprising, especially for Observation 1 as in this case no failures were observed. Despite the good news with Observation 1 that none of the AVs experienced an accident, the white-box predictions by the vendor are worse than the prior. The same pattern is retained for Observation 2 (when AV3 experienced two accidents).

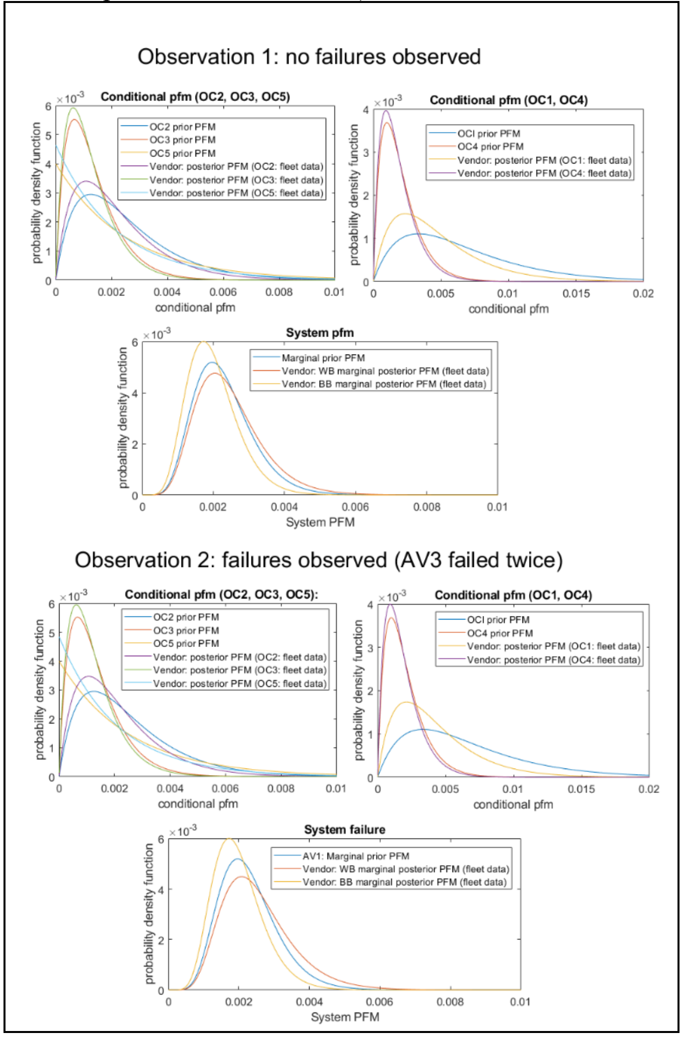


Figure 6. Effect of the model used in Bayesian inference on Vendor's predictions of the distributions of conditional pfm in OCs and of the marginal pfm.

The conservatism of the white-box predictions, even with no observed accidents, seems odd. Intuitively, we would expect the "good news" that no accidents have been observed in 500 additional miles driven by the fleet of AVs, to make us more optimistic about the AV fleet safety and to see that the posterior distribution of the system *pfm* shifts a probability mass towards the smaller values of system *pfm*. The black-box predictions support this intuition not only for Observation 1, but also for Observation 2.

The "mystery" about the white-box predictions, however, is easy to explain by looking at how the *operational profile has changed* as a result of the additional observations (of 500 miles driven by the AVs). The last rows of TABLE 2 for Observation 1 and Observation 2 show the number of additional miles driven in by all AVs in each of the OC_1, … OC_5. These are 127, 123, 109, 76 and 65, respectively, for both observations. These counts imply that the posterior "operational profile on average" (i.e., for the entire AV fleet), used in the calculations of the posterior distribution of system *pfm* (11), has changed from:

$$\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10)$$

to

$$\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\alpha_1 = 10 + 127, \alpha_2 = 10 + 123, \alpha_3 = 40 + 109, \alpha_4 = 30 + 76, \alpha_5 = 10 + 65) =$$

$$\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\alpha_1 = 137, \alpha_2 = 133, \alpha_3 = 149, \alpha_4 = 106, \alpha_5 = 75)$$

The weights of OC_1 and OC_2 in the posterior operational profile changed significantly and are now closer to the weight of OC_3. Indeed, the expected probability that a randomly chosen mile will be selected from OC_1 and OC_2 is 137/600 and 133/600, respectively, which is close to the expected probability of selecting a mile from OC_3, which would be 149/600. Prior to the additional observations OC_1, being the worst operating condition (the expected conditional pfm_1 in this operating condition was the highest among all 5 operating conditions) was quite unlikely with only 10% chance that a "randomly chosen mile" will come from OC_1. After the additional 500 miles driven, however, the weight of OC_1 has visibly increased to over 22% on average. What is the impact of the additional miles driven on the belief about the OC_1 being the worst OC, i.e., on $f_{\theta_1}(x|N_1 = 123, r_1 = 0)$? The impact on $f_{\theta_\mathrm{i}}(x|N_i, r_i)$ in OC_1,…, OC_5 is summarized in TABLE 3.

TABLE 3
DETAILED ANALYSIS OF THE IMPACT OF ADDITIONAL 500 MILES (OBSERVATION 1) DRIVEN BY THE AVS ON THE DISTRIBUTIONS OF THE CONDITIONAL PFMS IN OC_1, …, OC_5.

Operating conditions OC_i	Prior, Beta(α,β)	Prior mean $\frac{\alpha}{\alpha+\beta}$	Observations		Posterior Beta(α+r, β+N-r)	Posterior Mean $\frac{\alpha+r}{\alpha+\beta+N}$
			ID	(N, r)		
OC_1	Beta(2,299)	0.0066	1	(127,0)	Beta(2,326)	0.0047
			2	(127,1)	Beta(3,325)	0.0070
OC_2	Beta(2,800)	0.0025	1	(123,0)	Beta(2,923)	0.0022
			2	(123,1)	Beta(3,922)	0.0032
OC_3	Beta(2,1500)	0.0013	1	(109,0)	Beta(2,1609)	0.0012
			2	(109,0)	Beta(3, 1609)	0.0012
OC_4	Beta(2,1000)	0.0020	1	(76,0)	Beta(2,1076)	0.0019
			2	(76,0)	Beta(2,1076)	0.0019
OC_5	Beta(1,400)	0.0025	1	(65,0)	Beta(1,465)	0.0022
			2	(65,0)	Beta(1,465)	0.0022

It is evident from the table that the conditional pfm_1 remains the worst of all OCs: the additional miles driven reduced slightly the expected value of pfm_1 in OC_1 for Observation 1, while it gets marginally worse with Observation 2. Figure 6 provides the posterior distribution of the conditional pfm_1, $f_{\theta_1}(x|N_1 = 123, r_1 = 0)$. Thus, although the observations, and especially Observation 1 (with no accidents), are a "good news" indeed, it seems that this good news has not changed (for the better) the conditional pfm_1 *enough* to cancel out the fact that, after the additional 500 miles, OC_1, which has now become more likely, would affect more significantly the system *pfm*. The distribution of the system *pfm* is affected by both the distribution characterizing the *operational profile* and the distributions of the conditional *pfm*s in OC_1, …, OC_5. The good news of no accidents marginally changed *pfm₁* and *pfm₂* (and the other OCs, of course) but at the same time, the weight of OC_1 has increased significantly (from 10% before the additional observations to ~22.8%).

It is worth emphasizing that the black-box model would not react to changes of the operational profile at all. Under the black-box inference all miles are treated equally ignoring the impact that the observations may have on the posterior operational profile. As a result, "mysteries" like the one reported above, would remain "invisible" for predictions derived with a black-box model. This example, although contrived, raises doubts about the credibility of predictions obtained with a black-box prediction model.

4.3. Comparison with CBI

In this subsection we provide a comparison between the predictions derived with CBI and the two inferences using complete prior distributions defined in sub-section 4.2 – the black-box and the white-box predictions with the fleet data.

Let us define the CBI parameters as follows: $p_L = 10^{-5}$, $\epsilon = 10^{-3}$. The last parameter needed in the inference, θ, was computed from the marginal prior *pdf* for the examples used in sub-section 4.2. For $\epsilon = 10^{-3}$ the value of $\theta = 0.0283$. The second point of the two-point prior distribution with probability mass > 0 for Observation 1 (no failures) would be $p = 10^{-3}$. We repeated the computations also for $p = 2 \times 10^{-3}$ and for $p = 3 \times 10^{-3}$. The values of θ for these two cases are $\theta(p = 2 \times 10^{-3}) = 0.4121$, and $\theta(p = 3 \times 10^{-3}) = 0.8214$, respectively.

TABLE 4 illustrates the difference between the predictions obtained with CBI and with the full prior distributions. The confidences $P_{CBI}(X \leq p|N = 500, r = 0)$ and $P_{CBI}(X \leq p|N = 500, r = 2)$, derived for Observation 1 and Observation 2, respectively, are shown together with the corresponding confidences, $P_{BB}(X \leq p|500,0)$ and $P_{BB}(X \leq p|500,2)$, and $P_{WB}(X \leq p|N_1 = 127, r_1 = 0, N_2 = 123, r_2 = 0, N_3 = 109, r_3 = 0, N_4 = 76, r_4 = 0, N_5 = 65, r_5 = 0)$ and $P_{WB}(X \leq p|N_1 = 127, r_1 = 1, N_2 = 123, r_2 = 1, N_3 = 109, r_3 = 0, N_4 = 76, r_4 = 0, N_5 = 65, r_5 = 0)$, derived with the respective black-box and white box inferences, relying on the full prior distributions.

A couple of observations can be made about the posterior probability $P_{CBI}(X \leq p|\mathrm{N} = 500, \mathrm{r} = 0)$. Clearly it is *not* conservative when compared with the predictions derived with the other models. The CBI prediction for the target (ϵ) is more optimistic than the prediction obtained with the white-box model. The values of $P_{CBI}(X \leq p|\mathrm{N} = 500, \mathrm{r} = 0)$ equal 0.283 and equals 0.023 for the white-box prediction model. The predictions with the black-box model are significantly more optimistic (0.0537) than the CBI prediction computed for $x_1 = x_3$.

We note that CBI predictions are no worse than the prior, which is to be expected from any black-box inference including the one using a complete distribution with no failures. As we pointed out above, however, the white – box predictions for this set of observations are more conservative than the prior, due to the changed operational profile after the 500 miles of operation.

TABLE 4

COMPARISON OF CBI PREDICTIONS WITH THE PREDICTIONS DERIVED WITH FULL PRIOR DISTRIBUTIONS.

CBI model parameters	$p_L = 10^{-5}$	ϵ	θ	$P(p \leq 0.002)$	$P(p \leq 0.003)$
CBI (black-box) cdf	0	10^{-3}	0.0283	0.4121	0.8214

<table>
<tr><td colspan="5">Posterior confidence (observation 1: r_1=0, n_1 = 500), $P(X \leq p|n_2, r_2)$</td></tr>
<tr><td></td><td>ϵ (x_1)</td><td>p (x_3) = ϵ</td><td>p (x_3) = $2 \times \epsilon$</td><td>p (x_3) = $3 \times \epsilon$</td></tr>
<tr><td>CBI (black-box) cdf</td><td>10^{-3}</td><td>0.0283</td><td>0.5363</td><td>0.9260</td></tr>
<tr><td>Black-box (with full prior) cdf</td><td></td><td>0.0537</td><td>0.5603</td><td>0.9124</td></tr>
<tr><td>White-box (with full prior) cdf</td><td></td><td>0.023</td><td>0.3607</td><td>0.7633</td></tr>
<tr><td colspan="5">Posterior confidence (observation 2: r_1=2, n_1 = 500), $P(X \leq p|n_2, r_2)$</td></tr>
<tr><td></td><td>$x_1 = p_l$</td><td>$x_3 = k/n$</td><td>$p = 0.001$</td><td>$p = 0.002$</td></tr>
<tr><td>CBI (black-box) cdf</td><td>0.00001</td><td>0.004</td><td>1.333×10^{-6}</td><td>3.208×10^{-5}</td></tr>
<tr><td>Black-box (with full prior) cdf</td><td></td><td></td><td>0.0537</td><td>0.5603</td></tr>
<tr><td>White-box (with full prior) cdf</td><td></td><td></td><td>0.0219</td><td>0.3362</td></tr>
</table>

We computed also the posteriors for p = 0.002 and p = 0.003, for which θ (derived from the prior distribution, computed in previous examples) takes significantly higher values: $P(p \leq 0.002) = 0.4121$ and $P(p \leq 0.003) = 0.8214$. The predictions obtained with the white-box model are again more conservative than the ones derived with CBI. Interestingly, CBI predictions about p (x_3) $= 3 \times \epsilon$ are more optimistic than the predictions with the black-box model, although the difference is small.

Finally, we computed the CBI predictions for the case with two failures observed in 500 miles (Observation 2) for two targets $P(X \leq 0.001|500,2)$ and $P(X \leq 0.002|500,2)$. These are shown in the bottom of TABLE 4. Note how dramatically the predicted CBI probabilities changed following the observation of only two failures in 500 miles in comparison with the predictions obtained with the black – box and the white – box, respectively, both using a complete prior probability distribution. The CBI predicted probabilities are several orders of magnitude smaller (i.e., more conservative) than those with the other two models. The CBI predictions with Observation 2 are drastically different from the CBI predictions with Observation 1, even though the difference between Observation 1 and Observation 2 are only two accidents.

We can confirm that the CBI predictions for Observation 2 are indeed conservative, but the degree of conservatism is extreme (in comparison with the other models – white – box and black – box, which rely on a complete prior distribution). CBI seems extremely sensitive to the number of observed accidents, even if very few accidents are observed.

These observations, although limited, indicate that the particular CBI, based on a black-box inference: i) *does not guarantee* conservative predictions in comparison with the predictions of the white – box model described above, and ii) the degree of conservatism of CBI for cases with observed accidents may be very significant indeed. In our view, these two observations, each in their own way, raise questions about the usefulness of CBI for practical assessment.

5. Discussion and threats to validity

The results from the contrived examples demonstrate that the effect of the chosen model – black-box or white-box – on Bayesian predictions of system *pfm* may be *quite significant*. We summarize the key observations next.

The black-box (univariate) probabilistic models used for safety assessment of AV have *intrinsic limitations*, among them the fact that changes of the operational profile which may occur during the lifetime of an AV (or a fleet of AV) are not accounted for at all by black-box models. Ignoring changes of the operational profile may be justified for a large fleet of AVs where the "operational profile on average" may be expected to stay stable. The "operational profile on average" at initial stages of AV deployment, however, is *unlikely to be stable*. For instance, counting the miles of driving only during AV testing on the public roads without reference to the operating conditions these miles have been driven, leaves the possibility that the AV would be extensively tested on a *subset of operating conditions* while testing of some other operating conditions would be very limited or even non-existent (e.g. the so called "rare circumstances" [22]). If this is the case, the operational profile in testing may turn out to be quite different from the operational profile post deployment. If this is the case, a claim of satisfactory AV safety based on the testing results may not stand post deployment due to the differences in the operational profile. This situation will remain invisible with inferences based on black – box models (including CBI). This limitation of the black-box models may be fixed, of course, by using additional measurements to capture the testing profile, e.g. taking up the ideas of SPI (safety performance indicators) suggested in [23]. If SPI (or any alternative way of recording the operational profile), however, is to be used, then why not make use of these detailed data in the inference itself?

Another limitation of the black-box inference is that the inference results (i.e., safety claim) and indeed the observations the inference is based upon *cannot be extrapolated* to a different operational environment. This criticism applies to CBI predictions, too. This limitation is significant as it implies that a safety assessment based on a black-box model may need to be repeated for any new operational environment.

Multivariate probabilistic models, which account for a *variable* operational environment, bring the following advantages:

- Allow one to conduct inference, which is in tune with the needs of AV safety assessment of both a fleet of AVs or *individual* AVs.
- Force the assessors to collect operational data, which is suitable for *porting the results from a safety assessment* to a new operational profile which may differ significantly from the profile for which data has been collected and the white-box inference applied. Thus, if a new operational profile indicates that some of the operating conditions require more extensive evidence of good safety, these operating conditions will be the only ones which will require further road testing, thus reducing the cost of safety assessment for the new operational profile.
- Serve the needs of AV vendors and of the individual AVs, which are quite different. An *intriguing possibility* exists with the proposed white-box inference to be not merely applied to the individual AVs (using their own observations), but also making it possible for the observations (miles driven and accidents observed) collected by the *entire fleet* of deployed AVs to be shared among the AVs and thus allow each individual AV to learn and reduce the uncertainty about pfm_i in different OCs *much faster* (than counting on own observation only). Indeed, the volume of observations from many thousands of AV instances will be much more extensive than the observations collected by a single AV instance. This possibility to share observations among AV instances was left outside the scope of the paper and has been developed in a separate article [24].

We demonstrate that the black-box inference may lead to *more optimistic* predictions than the predictions obtained with the multivariate (white-box) model even when no accidents are observed. Our findings cast doubts about the conservatism of CBI, too, as we have shown that when no failures are observed the CBI predictions may be more optimistic than the predictions derived using the proposed white-box model. In the context of safety assessment, the optimistic predictions with a black-box model, are particularly *worrisome,* raising doubts about the suitability of black-box predictions for AV safety assessment.

Any black-box model is clearly different from the white-box model we developed in this paper. The proposed white – box model assumes that the random variables Θ_i representing the conditional pfm_i in OCs are *independently distributed* random variables. This is a strong assumption which needs validation. The black-box models, being different models, surely *implicitly imply* a form of dependence between Θ_i. What form does this dependence take? Informally, the black-box model seems to imply that Θ_i are updated *simultaneously* irrespective of the operating condition from which data has been observed. Let us assume that an AV is driven a few miles in OC_i, (N_i, r_i). Let us think of a white-box model with dependencies between Θ_i. We can further assume that this new white-box model uses (N_i, r_i) to update not only $f_{\theta_i}(\cdot|N_i, r_i)$, but also all other $f_{\theta_j}(\cdot|N_i, r_i), j \neq i$ as if (N_i, r_i) were observed in all other partitions, too. It is trivial to show that with such a form of dependence, the white – box model will have a marginal distribution of system *pfm* identical to the *pfm* predicted with a black-box inference. In other words, a black-box model implies an *extreme form of dependence* between Θ_i. While a form of dependence between Θ_i may be needed, the extreme form implied by the black-box model would be difficult to justify.

Among the *threats to validity* of our results we would like to discuss the assumptions which are essential in the proposed multivariate Bayesian inference:

- The conditional probabilities pfm_i are modelled as *independently distributed random variables*, Θ_i for i = 1,…, n. This assumption seems plausible but may in fact be difficult to justify [25]. The problem is not new and has been discussed in the past, e.g., [26] took the view that a failure/accident could be traced to a root cause (i.e., a fault), which can be triggered in more than one operating condition, thus promoting the idea that beliefs about the conditional probabilities pfm_i should be captured by *dependent* random variable. [27] looked at the impact on reliability of a software system, which consists of two parallel components, which may fail non-independently (e.g., simultaneously), and accounts for non-independence of the respective component reliability measures (treated as random variables). Technically, the independence assumption we rely upon, can be relaxed, e.g., by using suitably chosen *Copulas*[8] to capture the dependencies between the random variables Θ_i for i = 1,…, n. Scoping a credible procedure to elicit the parameters of these Copulas, however, is outside the scope of this paper. We intend to look at this problem in our future work.
- We assume that the AV operational profile is captured adequately by a *Dirichlet* distribution. Although this type of multivariate distribution has been used by many[9] in the past and, more importantly, seems quite plausible for the problem at hand, it may in some circumstances be inadequate. A promising alternative way of modelling the operational profile would be using *state-based models*, e.g., Markov and semi-Markov ones, in which the operating conditions (OC_1, …, OC_n), defined for a given ODD, appear as states of a *state-based model of the operational profile*, e.g. as we have done in own recent work [28].
- Finally, in this paper we relied on the prior work by others [2] and model the success/accident per mile of driving as a *Bernoulli* trial. Clearly the successive miles of driving may not be quite like a Bernoulli trial, although the recent work [29] provides a rationale to relax the assumption of independence between success/accidents of successive miles of driving. An alternative approach in modelling AV driving would be to consider the *duration* (in miles) in the same operating condition of ODD and model the AV driving as a trajectory via different OCs (as we have shown in Figure 2 above). We took this approach in recent studies [28, 30]. Such a model of AV driving may reveal a different insight. We intend to develop this alternative model of driving in detail in our future research.

6. Related Research

In passing we already mentioned several relevant papers. Here we discuss other examples of related research.

Bayesian inference for software reliability assessment has attracted significant interest over the years. Most of the publications take a "black-box" view, but there have been examples taking a "white-box" view in the sense we use it in this paper. We will briefly summarize these works next.

The work by Keith Miller et al. [11] is probably the first example in which the results from software "partition testing" have been used to demonstrate how software reliability can be estimated accounting for the results from partition testing. The authors focus their work on the expected value of the marginal probability of failure on demand although in the process they apply Bayesian inference to derive the distributions of the probabilities of failure conditional on partitions. This work is somewhat similar to the approach taken here, but does not account for the uncertainty in the operational profile as is done in our work.

While the work by Miller et al. [11] introduced Bayesian inference to deal with partition testing, several authors used multi-variate inference to deal with failures of component-based software. For instance, Kubal et al. [31] developed a proposal for estimating the probability of failure of a software system from the probabilities of failure of the software components used in the system. The method is based on Bayesian assessment of the probability of failure of components, assumed to *fail independently*. The implications of this strong independence assumption was later criticized in own work [32], which demonstrated that the assertion made by Kubal et al. in [31] that their method leads to "conservative" predictions of the probability of system failure are unjustified.

The problem of assessing reliability of component-based software using Bayesian inference was discussed also in [33] and a method of dealing with the complexity of multivariate inference, especially in defining a credible prior, is tackled by developing a *hierarchical model* of inference which relies on *partial views*. These views are formed by a subset of variates used in the full system model. Using views breaks the multivariate inference into manageable parts and the multivariate inference itself becomes computationally more tractable without having to rely on conjugate priors as is often done. The method relies on techniques developed by others, such as u-plot and prequential likelihood [34], to control the accuracy of the predictions as the predictions derived with the views and propagated through the inference hierarchy are subjects to prediction errors.

A two-stage Bayesian inference has attracted some interest, e.g. [35, 36], in which the rates of failure of a system (e.g. a plant) are subject to uncertainty, captured by a probability distribution, the parameters of which are also uncertain and captured by "hyperparameters". If historical data is available about the failures of the plant, the hyperparameters are gradually learnt, thus reducing the uncertainty in them. This approach is conceptually similar to the two-stage inference presented in this paper. However, there are significant differences, too. The prior works are interested in the parameter of a Poisson process – the failures are instances of this process – with a single parameter (λ), which is a random value characterized by a distribution (the second stage of the model). In our work accidents occur from

[8] Copulas are a specific way of modelling the dependence between random variables. For further details the interested reader is advised to check https://en.wikipedia.org/wiki/Copula_(probability_theory).

[9] https://en.wikipedia.org/wiki/Dirichlet_distribution#Bayesian_models

a more complex stochastic process which includes multiple "operating conditions." We are interested in the marginal probability of accident/failure not in the epistemic uncertainty of hyperparameters characterizing λ as is done in the prior work.

As mentioned in the introduction, CBI has attracted significant interest in the last few years. Most of the CBI development applies to a broader class of ultra-reliable software-based systems. CBI is a way of simplifying Bayesian reasoning without risking errors that would overestimate safety) [5, 29]. CBI builds on the long history of Bayesian reliability assessment, applied to software-based safety – critical systems, e.g., [10], [11]. A variation of the idea of CBI has been then developed for more refined AV related scenarios of practical interest, e.g., dealing with several *epochs* of observations, which may occur between different *releases* of AVs [4]: release i has seen N_i miles of operation, release $i+1$ – N_{i+1} miles, etc. The issue in such studies then becomes – how one should use in epoch $i+1$ the results from the safety assessment achieved for an AV driven up to and including epoch i.

An interesting quantitative methodology for constructing an ODD with statistical data and risk tolerance is presented in [37].

Another interesting risk decomposition methodology to derive SOTIF requirements for perception using a combination of models (Markov, Bayesian, etc.) is presented in [38].

Extensive literature exists on the use of state-based models to capture the operational profile of software subjected to operations testing. An authoritative reference is [39].

7. Conclusion and Future Research

We presented a critical review of the use of black-box (univariate) probabilistic models for Bayesian safety assessment of AVs. More specifically, we looked at Bayesian inference used to predict the distribution of the probability of accident per mile of driving using data from AV test driving on public roads. Our main result is demonstrating, via contrived examples, that *univariate models are deficient* and their use for safety assessment should probably be avoided. The limitations of predictions based on black-box models are of two kinds:

- AV safety assessment must deal explicitly with the driving conditions (operational profile) which offer different risks of road accidents. Ignoring these differences is conceptually flawed and may lead to wrong conclusions. While operational profile is routinely assumed fixed (i.e., unchanging over time) with many safety-critical systems (e.g., nuclear plants, transport systems such as railway, etc.) the operational profile of AVs is *intrinsically changeable*, especially when it comes to individual AV instances. Using predictions at initial stages of AV development to compute the miles needed to achieve high confidence in ultra-high AV reliability is particularly sensitive to the profile of AV testing. Ignoring the testing profile may lead to *unjustified conclusions*. There are two specific aspects worth emphasizing: The AV vendors may be tempted to deploy statistical methods to demonstrate that the AV is sufficiently safe. However, a testing profile which is not aligned with the anticipated operational profile may be misleading. Substantial amounts of miles in "easy" operating conditions do not necessarily demonstrate that the AV is ready to be deployed on the public roads especially if the AV has not been tested sufficiently well on "difficult road conditions" (or "rare conditions").
  - The current suggestions for "scenario-based testing" build an argument of AV safety, which does not rely on statistics, but suggests that a relatively small number of real testing scenarios[10] can be used to generate a large set of *synthetic* scenarios, which in turn would be sufficient to demonstrate AV safety. This argument, widely adopted by industry, and seemingly supported by some regulators [40], has two deficiencies: i) statistical confidence in a claim of sufficient safety seems ignored which makes it difficult/impossible to translate the safety claim to anticipated losses (including loss of human life), and ii) it seems to imply implicitly that a limited number of scenarios would be sufficient to cover the risks on the roads. Such claims are in our view simply *unjustifiable*. It seems to us that the results observed in scenario-based testing need to be interpreted using stochastic reasoning adopting the ideas presented in this paper, or proposed by others, e.g. [41].
- A side effect from our main result, which has been discussed by others in the past, including by ourselves [14], is that applying optimization to Bayesian inference aka CBI using a deficient univariate model is also problematic. The conclusions from such optimizations, too, should be taken with a degree of skepticism.

In the previous section we already identified a few areas for future development to address some of the recognized threats to the validity of our work and findings, among them: i) relaxing the reliance on Bernoulli trial as an adequate model of selecting miles from each operating condition. We will instead explore the use of stochastic state-based models [24, 30] for this purpose, ii) assuming that the distributions of the conditional *pfm*s in operating conditions are independently distributed random variables, and iii) alternative models of the operational profile, for which in this paper we use the Dirichlet distribution

The nature of the multivariate inference developed in this paper is such that it allows for the uncertainty about the operational profile and of the conditional probabilities of accident per mile of driving, *pfm*s, to be updated using *different observations*. We already highlighted in Section 5 the possibility for an *AV instance* to use own observations of miles driven to update the uncertainty about the operational profile, while the uncertainty about the conditional

[10] The number of testing scenarios needed for safety justification is rarely discussed, but if this approach to safety justification is to be more "cost - effective" than driving to safety, it seems that the expectation is that a relatively small number of test scenarios, selected smartly, will be sufficient to demonstrate safety.

probabilities of accident *pfm*s may use aggregated data (miles driven and accidents observed in operating conditions collected by a fleet of AVs of the same type. This idea has been scoped in [24] and will be developed further in our future work.

In the introduction we mentioned in passing that a safety claim, linked implicitly to a given ODD, faces additional problems: i) recognizing when an AV is getting “out-of-ODD (OoODD)”, and ii) making sure that the AV response to such an event, e.g., stopping the AV at earliest opportunity, is implemented with sufficiently high integrity so as not to compromise the overall safety claim. Both detecting OoODD itself and the implemented response to a detected OoODD, may be subject to failure. A complete safety analysis should account for failures of OoODD-related function(s), too [42]. These are important aspects of AV safety assessment, which we intend to address in the future.

Finally, we used contrived examples to illustrate the limitations of an inference based on a black-box model, and the benefits from the proposed multivariate inference. We were unable to illustrate our new theory using “field data” collected from real AV. To the best of our knowledge, such data is not available in the public domain although extensive databases with accident data are maintained by the national authorities dealing with road safety, e.g. [19]. We expect that our findings might be of interest not only to the research community but to the automotive industry, too. Such an interest may trigger effort on collecting field data with the level of details required by the proposed new multivariate inference procedure. We intend to seek actively an engagement from industry and expect in the near future to be able to apply the new inference procedure to “field data”.

8. References

1. SAE International, *J3016 : Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. 2021, SAE International: Vernier, Geneva, Switzerland. p. 41.
2. Kalra, N. and S. Paddock, *Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?* Transportation Research Part A: Policy and Practice, 2016. **94**: p. 182-193.
3. British Standards Institute (BSI), *PAS 1883:2020 Operational Design Domain (ODD) taxonomy for an automated driving system (ADS) - Specification*. 2020, BSI Standards Limited: London. UK. p. 26.
4. Bishop, P., A. Povyakalo, and L. Strigini. *Bootstrapping confidence in future safety from past safe operation*. in *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*. 2022.
5. Salako, K., L. Strigini, and X. Zhao. *Conservative Confidence Bounds in Safety, from Generalised Claims of Improvement & Statistical Evidence*. in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 2021.
6. Zhao, X., et al., *Assessing safety-critical systems from operational testing: A study on autonomous vehicles.* Information and Software Technology, 2020. **128**: p. 106393.
7. Francisco, K.F.S., *Tesla Model S causes crash on Bay Bridge*. 2022: YouTube. p. https://www.youtube.com/watch?v=xP4ldxTRPHc.
8. Zou, Z., et al., *Object Detection in 20 Years: A Survey.* Proceedings of the IEEE, 2023. **111**(3): p. 257-276.
9. Terrosi, F., L. Strigini, and A. Bondavalli. *Impact of Machine Learning on Safety Monitors*. in *SAFECOMP*. 2022. Cham: Springer International Publishing.
10. Littlewood, B. and L. Strigini, *Validation of Ultra-High Dependability for Software-based Systems.* Communications of the ACM, 1993. **36**(11): p. 69-80.
11. Miller, K.W., et al., *Estimating the Probability of Failure When Testing Reveals No Failures.* IEEE Transactions on Software Engineering, 1992. **18**(1): p. 33-43.
12. Littlewood, B., 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.
13. Popov, P. *Reliability Assessment of Legacy Safety-Critical Systems Upgraded with Off-the-Shelf Components*. in *SAFECOMP'2002*. 2002. Catania, Italy: Springer.
14. Popov, P., *Conservative reliability assessment of a 2-channel software system when one of the channels is probably perfect.* Reliability Engineering and System Safety, 2021. **216** (December): p. 23.
15. Pietrantuono, R., P. Popov, and S. Russo, *Reliability assessment of service-based software under operational profile uncertainty.* Reliability Engineering & System Safety, 2020. **204**: p. 107193.
16. Adams, T., *Total Variance Approach to Software Reliability Estimation.* IEEE Transactions on Software Engineering, 1996. **22**(9): p. 687-688.
17. ASAM e.V., *ASAM OpenODD Base Standard 1.0.0 Specification*. 2025.
18. ISO, *ISO 34503: Road Vehicles — Test scenarios for automated driving systems — Specification for operational design domain*, in *TC 22/SC 33*. 2023, ISO Copyright Office. p. 38.
19. NHTSA *SGO Incident Reports ADS January 202*. Standing General Order on Crash Reporting, 2021.
20. Trivedi, K.S., *Probability and Statistics with Reliability, Queueing, and computer science Applications (2nd edition)*. 2016, New Jersey: John Wiley & Sons, Inc.
21. Albert, I. and J.-B. Denis *Dirichlet and multinomial distributions: properties and uses in Jags*. Unité Mathématiques et Informatique Appliquées, 2012. 28.
22. Favarò, F., et al. *Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk*. 2023. 38 DOI: https://doi.org/10.48550/arXiv.2306.01917.
23. Johansson, R. and P. Koopman, *Continuous Learning Approach to Safety Engineering.* , in *CARS - Critical Automotive applications: Robustness & Safety*. 2022, HAL: Zaragoza, Spain. p. 5.
24. Popov, P., *Dynamic safety assessment of Autonomous Vehicle based on Multivariate Bayesian Inference (DyAVSA).* Journal of Reliable Intelligent Environements, 2024. **(under review)**: p. 28.
25. Klotz, J., *Statistical Inference in Bernoulli Trials with Dependence* The Annals of Statistics, 1973. **1**(2): p. 373–379.
26. May, J. and A.D. Lunn, *A model of code sharing for estimating software failure on demand probabilities.* IEEE Transactions on Software Engineering, 1995. **21**(9): p. 747-753.
27. Troffaes, M.C.M. and F.P.A. Coolen, *Applying the imprecise Dirichlet model in cases with partial observations and*

dependencies in failure data. International Journal of Approximate Reasoning, 2009. **50**(2): p. 257-268.
28. Popov, P., et al., *Stochastic Modeling of Road Hazards on Intersections and their Effect on Safety of Autonomous Vehicles*, in *arXiv:2506.02688*, ArXiv, Editor. 2025 (under review), https://arxiv.org/abs/2506.02688: Cornell U, arXiv. p. 14.
29. Salako, K. and X. Zhao, *The Unnecessity of Assuming Statistically Independent Tests in Bayesian Software Reliability Assessments.* IEEE Transactions on Software Engineering, 2023. **49**(4): p. 2829-2838.
30. Buerkle , C., et al., *Modelling road hazards and the effect on AV safety of hazardous failures*, in *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC'2022)*. 2022: Macau, China. p. 1886-1893.
31. Kuball, S., J. May, and G. Hughes. *Building a system failure rate estimator by identifying component failure rates*. in *Proceedings 10th International Symposium on Software Reliability Engineering (Cat. No.PR00443)*. 1999.
32. Popov, P. *Reliability Assessment of Legacy Safety-Critical Systems Upgraded with Off-the-Shelf Components*. in *SAFECOMP'2002*. 2002. Catania, Italy: Springer.
33. Popov, P., *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.
34. Brocklehurst, S., et al., *Recalibrating software reliability models.* IEEE Transactions on Software Engineering, 1990. **SE-16**(4): p. 458-470.
35. Bunea, C., et al., *Two-stage Bayesian models-application to ZEDB project.* Reliability Engineering and System Safety (RESS), 2005. **90**(2-3): p. 123-130.
36. Pörn, K., *The two-stage Bayesian method used for the T-Book application.* Reliability Engineering & System Safety, 1996. **51**(2): p. 169-179.
37. Lee, C.W., et al. *Identifying the Operational Design Domain for an Automated Driving System through Assessed Risk*. in *2020 IEEE Intelligent Vehicles Symposium (IV)*. 2020.
38. Yu, R., et al., *Decomposition and Quantification of SOTIF Requirements for Perception Systems of Autonomous Vehicles.* IEEE Transactions on Intelligent Transportation Systems, 2025: p. 1-14.
39. Whittaker, J.A. and M.G. Thomason, *A Markov chain model for statistical software testing.* IEEE Transactions on Software Engineering, 1994. **20**(10): p. 812-824.
40. European Commission, *COMMISSION IMPLEMENTING REGULATION (EU) 2022/1426 as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles.* Official Journal of the European Union, 2022.
41. Zhao, X., et al., *On the Need for a Statistical Foundation in Scenario-Based Testing of Autonomous Vehicles*, in *arXiv:2505.02274*, Cornell U, 2025, arXiv:2505.02274v1. p. 8.
42. Standards&Engagement, *ANSI/UL 4600: Evaluation of Autonomous Products*. 2023, ANSI.

Peter Popov (A member of IEEE) is Reader in Systems Dependability in the Centre for Software Reliability, at City, St George's University of London.

Peter holds a PhD degree in computer engineering from the National University of Ukraine "Kiev Polytechnic University." Prior to joining City, he was with Bulgarian Academy of Sciences, Sofia, Bulgaria, and prior to this worked in industry in Bulgaria.

His current research interests include probabilistic methods for rigorous safety assessment of autonomous vehicles and methods of cyber – security assessment of safety critical systems.

He is a member of the IFIP Working Group 10.4 on Dependable Computing and Fault Tolerance (https://www.dependability.org/).

APPENDIX

Figure 1

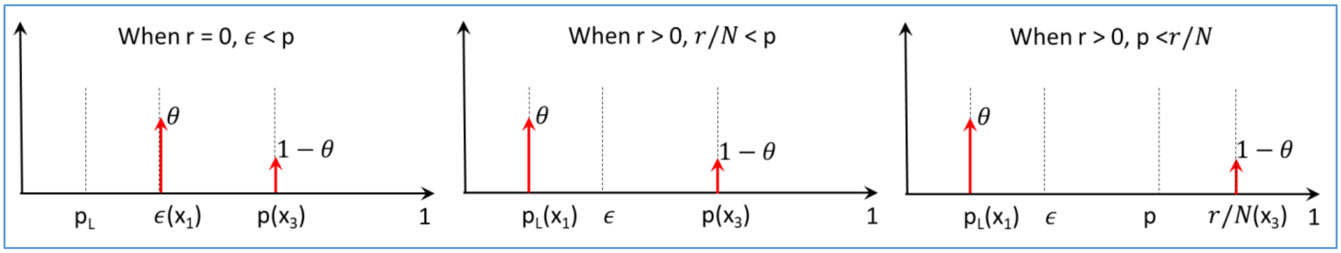


Figure 2

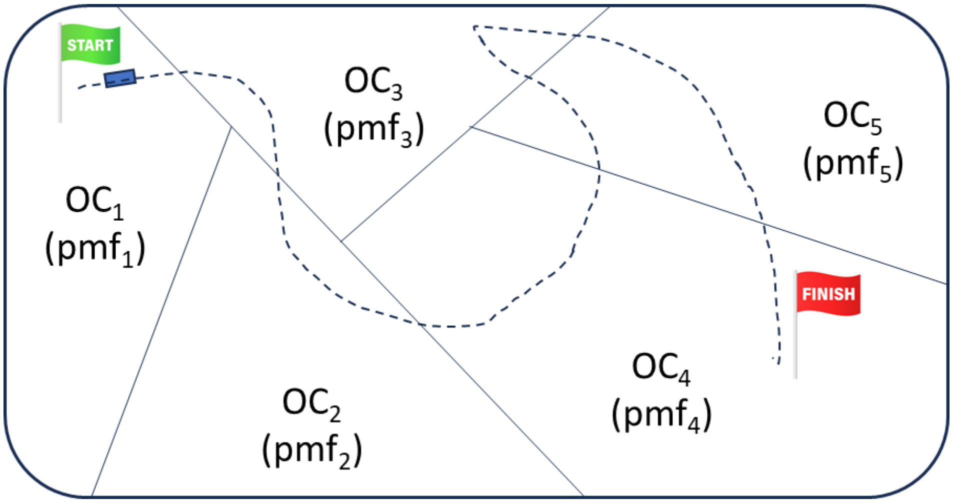


Figure 3

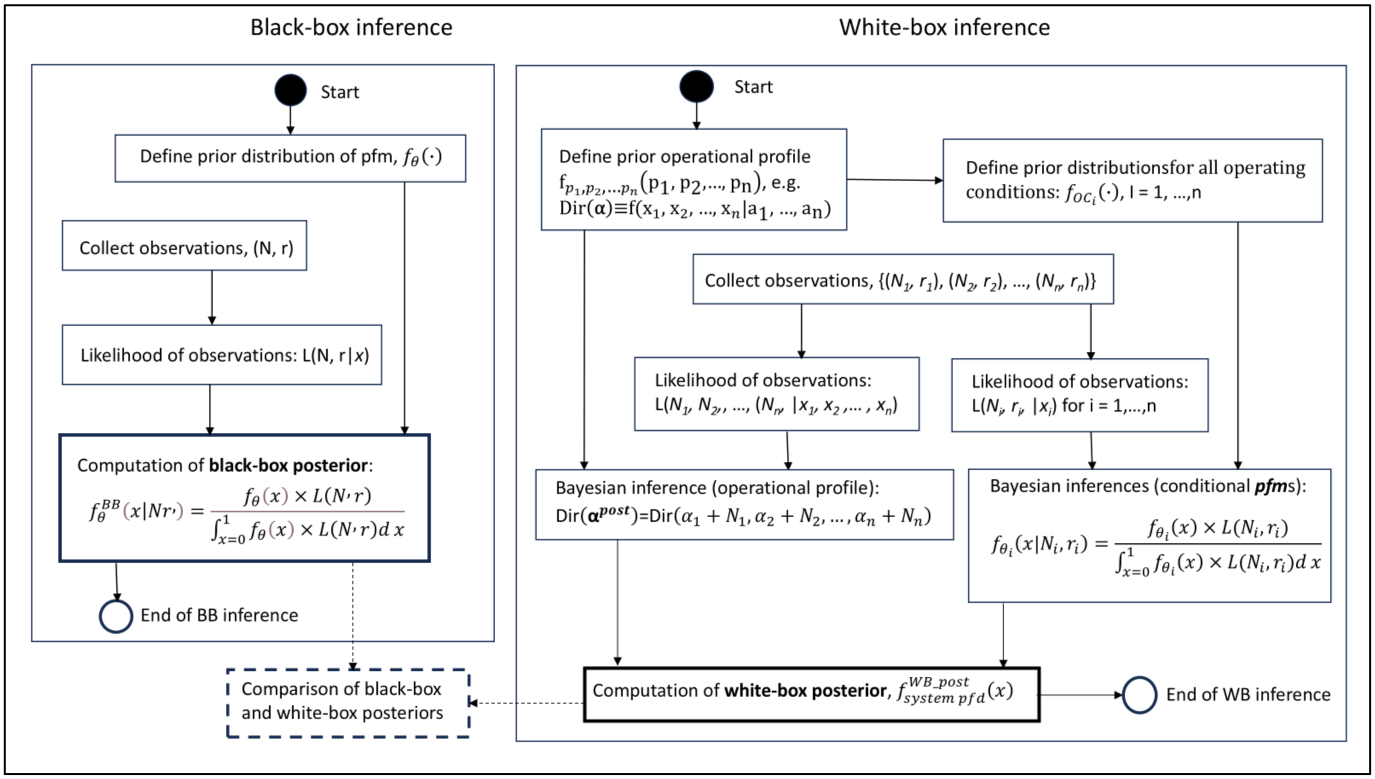

Black-box inference
White-box inference
Start
Define prior distribution of pfm, $f_\theta(\cdot)$
Collect observations, (N, r)
Likelihood of observations: L(N, r|x)
Computation of **black-box posterior**:
$f_\theta^{BB}(x|N r) = \frac{f_\theta(x) \times L(N r)}{\int_{x=0}^{1} f_\theta(x) \times L(N r) d x}$
End of BB inference
Start
Define prior operational profile $f_{p_1,p_2,\dots p_n}(p_1, p_2, \dots, p_n)$, e.g. $Dir(\alpha) \equiv f(x_1, x_2, \dots, x_n | a_1, \dots, a_n)$
Define prior distributionsfor all operating conditions: $f_{OC_i}(\cdot)$, I = 1, …,n
Collect observations, {(N_1, r_1), (N_2, r_2), …, (N_n, r_n)}
Likelihood of observations: L(N_1, N_2, …, (N_n, |x_1, x_2 ,… , x_n)
Likelihood of observations: L(N_i, r_i, |x_i) for i = 1,…,n
Bayesian inference (operational profile): $Dir(\alpha^{post})=Dir(\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_n + N_n)$
Bayesian inferences (conditional ***pfm***s):
$f_{\theta_i}(x|N_i, r_i) = \frac{f_{\theta_i}(x) \times L(N_i, r_i)}{\int_{x=0}^{1} f_{\theta_i}(x) \times L(N_i, r_i) d x}$
Comparison of black-box and white-box posteriors
Computation of **white-box posterior**, $f_{system\ pfd}^{WB\_post}(x)$
End of WB inference

Figure 4

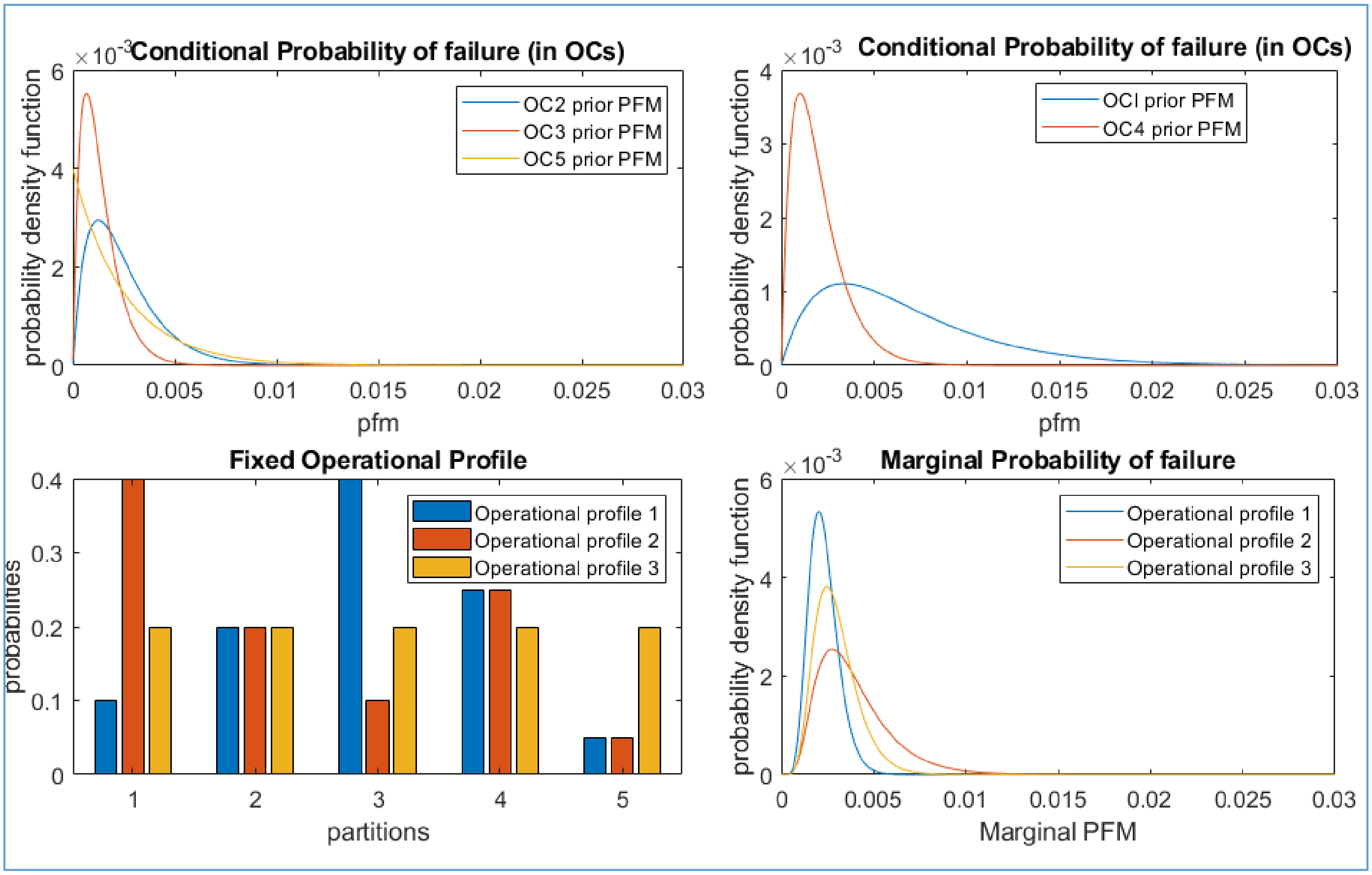


Figure 5

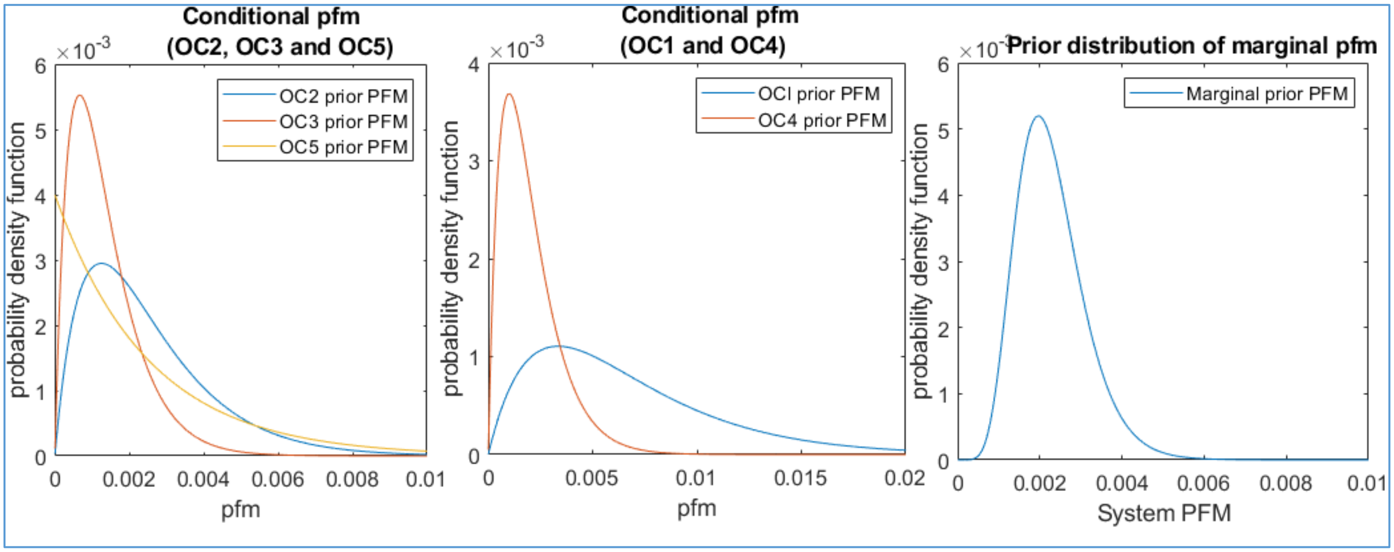
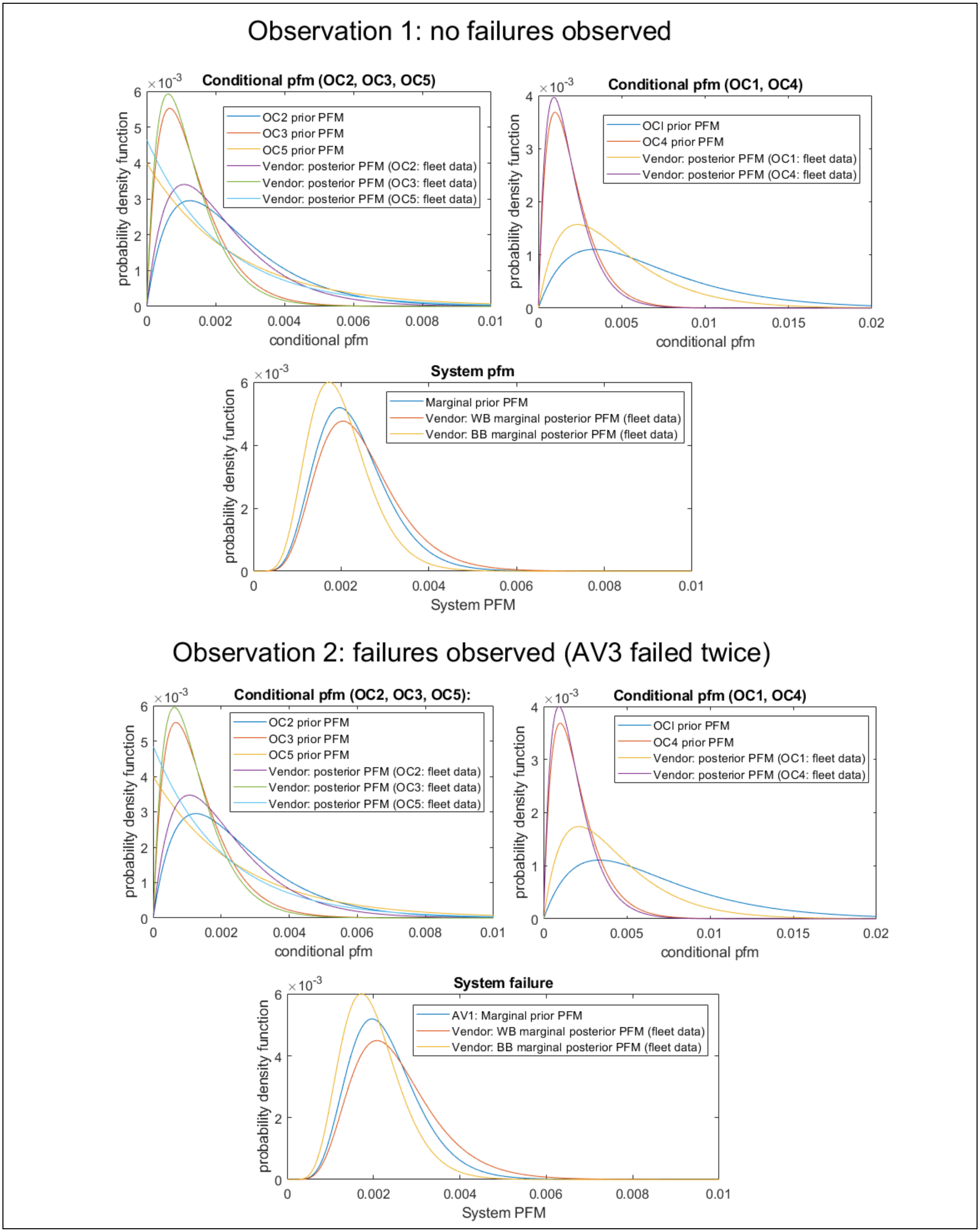


Figure 6


Observation 1: no failures observed
Conditional pfm (OC2, OC3, OC5)
×10^-3
probability density function
conditional pfm
OC2 prior PFM
OC3 prior PFM
OC5 prior PFM
Vendor: posterior PFM (OC2: fleet data)
Vendor: posterior PFM (OC3: fleet data)
Vendor: posterior PFM (OC5: fleet data)
0
0.002
0.004
0.006
0.008
0.01
1
2
3
4
5
6
Conditional pfm (OC1, OC4)
OCl prior PFM
OC4 prior PFM
Vendor: posterior PFM (OC1: fleet data)
Vendor: posterior PFM (OC4: fleet data)
0.005
0.015
0.02
System pfm
Marginal prior PFM
Vendor: WB marginal posterior PFM (fleet data)
Vendor: BB marginal posterior PFM (fleet data)
System PFM
Observation 2: failures observed (AV3 failed twice)
Conditional pfm (OC2, OC3, OC5):
Conditional pfm (OC1, OC4)
System failure
AV1: Marginal prior PFM
Vendor: WB marginal posterior PFM (fleet data)
Vendor: BB marginal posterior PFM (fleet data)