Dynamic safety assessment of Autonomous Vehicle based on Multivariate Bayesian Inference (DyAVSA)

Peter Popov [1][0000-0002-3434-5272]

[1] Centre for Software Reliability, City St George's, University of London, Northampton Square, London EC1V 0HB, United Kingdom
p.t.popov@city.ac.uk

Abstract. This paper deals with the Bayesian safety assessment of autonomous vehicles (AV) using as a key safety measure the probability of catastrophic failure per mile of driving (*pfm*), assumed a random variable.

The paper takes the view that *pfm* may (and typically will) vary due to changing road driving conditions. Accommodating this variation in a Bayesian inference on *pfm* requires one to use a multivariate probabilistic model whereby the changeable *pfm* is captured explicitly for the different driving conditions. The model that we use in this work is derived from our prior work and accounts for the uncertainties in both – the operational profile (i.e., the likelihood of the different driving conditions) and the *pfm*s, conditional on the respective operating conditions.

The concept of the "dynamic AV safety assessment (DyAVSA)" is presented in the work, too, whereby the Bayesian predictions used *at run time* rely on the operational data collected by a fleet of AVs. DyAVSA benefits both: i) the AV vendors, for monitoring the safety changes of the entire AV fleet; ii) the owners/users of individual AVs, whose safety assessment is personalized and different from the assessment of the AV fleet.

DyAVSA thus offers a *major change* in the AV safety management than is currently the case. It allows the AV users/owners to benefit from the aggregated safety relevant data collected from a fleet of AVs. Our findings show that the benefits from DyAVSA for the owners/operators of the individual AV instances may be significant: the safety predictions they can make by using the data collected by *the entire fleet of* AV instances and shared among them, may differ considerably from the predictions the AV instances would be able to make relying on own observations only. Sharing data would lead to a much *more rapid reduction of uncertainty* in the *pfm*s than would be the case if the AV instances relied on their own observations only.

The presented DyAVSA, based on a multivariate Bayesian safety assessment, can be applied to other complex intelligent systems such as robots, UAVs, etc.

Keywords: Autonomous vehicle, Safety Assessment, Bayesian inference, "driving to safety".

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 the data collected during driving an AV on the public roads. [2], and other related studies, e.g., [3], 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.

While there is an ongoing active debate as to how AV safety can be assured cost effectively, it is clear that it will take years for vendors to demonstrate adequate AV safety to the regulators [4] and more importantly to convince the public that AVs are safe to be used on the public roads [5].

Whatever form the AV safety assurance/certification takes [6] this would be a *pre-deployment* safety assurance. Once the national authority grants permission for the use of a particular AV brand on the public roads, the vendors will enter a *post-deployment* period of data collection from the deployed AVs, which will be used to improve further the AV functionality and, of course, to improve the AV safety. This post-deployment cycle is *not unique* to AVs. It is routinely followed in many safety – critical domains (nuclear [7], aviation [8] to name a few). The periodic safety reviews are an opportunity for the operators and/or vendors of the critical systems to review the safety claim made for the particular system in light of the *new evidence* that will have been collected from the installed systems since the previous periodic review. Should the review discover that the new evidence is not supportive of the safety claims, corrective actions will follow, which in turn may trigger a new cycle of certification[2].

How the post-deployment AV safety reviews will be shaped in the future is yet to be seen. Waymo, a leading AV manufacturer, acted decisively after an accident of their robotaxi in Phenix, Arizona and recalled voluntary the entire fleet of robotaxis. Some suggest that the transition from pre- to post-deployment safety assessment should be more gradual whereby the AV vendors should be allowed to deploy a limited fleets of

---

[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.

[2] In some cases, the boundary between the pre- and post-deployment safety assurance may be less clear. For instance, the two crashes of Boeing 737 MAX in 2018 and 2019, which led to the death of several hundred of passengers, triggered the grounding of all MAX planes worldwide for almost two years. A high – profile investigation was triggered shortly after the second crash, followed by a scrutiny of how the MAX safety was assessed. A detailed account is available at: https://en.wikipedia.org/wiki/Boeing_737_MAX_groundings#2020.

AVs after a preliminary safety assurance and use the fleet of deployed AVs to collect operational data and *gradually improve* the AV safety on the roads [9]. This proposal for "continuous safety assessment" is appealing, especially for AVs of high level of automation (e.g. Level 4 and Level 5 according to [1]) given the great uncertainty about the operating conditions the AVs may be used in.

Some AVs may find themselves used mostly in "easy" operating conditions (e.g., on roads with light traffic). Some other AVs may be used in difficult operating conditions (e.g., in heavy traffic in urban areas and extreme weather). Foreseeing all possible operating conditions is difficult as the evidence from the national authorities collecting statistics on AV accidents suggest [10].

It is well known in safety engineering that safety is not only a property of the system under assessment, but also of the operating conditions in which the system is used. AVs of Level 4 and 5 are examples of systems used in a *highly changeable* operational environment, which makes the safety assessment very difficult indeed. In recognition of this difficulty, the AV community adopted an approach to safety assurance based on constraining the operational environment by introducing the concept of Operational Design Domain (ODD) [11]. ODD defines an "envelop" on the operating conditions by restricting them to a subset of all operating conditions that an AV may find itself in (e.g., the intensity of the road traffic, the weather, the type of the road, etc.). A safety claim linked to an ODD would apply only to the operating conditions within the stated ODD. Any accident which takes place outside the ODD will not affect adversely the AV safety claim since the AV is not assured as safe outside the ODD. A detection of "outside-of-ODD (out-OOD)" should trigger a transition to a "safe state" (e.g., by stopping the AV at earliest opportunity when it is safe to do so). Clearly both detecting "out-ODD" and responding to it by taking the AV to a safe state, may be subject to failure. Dealing with the implications of imperfection in detecting and responding to the event "out-ODD" is outside the scope of this paper. The implications of "out-ODD" for safety assessment, however, are briefly discussed later in the paper.

The contributions of this paper are:

- We propose and *develop in detail* a method for "continuous safety assessment", we called Dynamic AV Safety assessment (DyAVSA) based on a multivariate Bayesian inference procedure, which we developed in [12] and recently adapted to the needs of AV safety assessment [13]. To the best of our knowledge, this approach to continuous safety assessment is innovative and has not been applied before. A key element of the method is that operational data (on miles driven and accidents encountered) collected by *all* deployed AV instances of a particular type of AV is *shared* with all other deployed AV instances, thus allowing each instance to benefit from the operational "experience" of the entire AV fleet. This process could be facilitated by the AV vendor who may serve as a collector of all operational data and subsequently share it with the entire AV fleet.

- We demonstrate on a set of contrived examples[3] the *benefits* from the proposed DyAVSA procedure by comparing the outcomes from applying the multivariate Bayesian inference procedure differently: i) by the vendor to the data collected by the entire fleet of deployed AV instances and using an operational profile "on average" (i.e., accounting for the data collected from the entire AV fleet); ii) by the individual AV instances to their "own data" only, i.e., to the data about the miles driven and the accidents observed by the respective AV instance only. In this case, the AV instances are not aware of the operational data collected by the fleet of deployed AVs of the same type, and iii) DyAVSA, i.e., by the AV instances using own operational data to estimate their own "operational profile" and using the fleet operational data (e.g., shared by the AV vendor) to estimate the conditional probabilities of accident per mile of driving in each of the operating conditions defined by a given ODD.

Our findings demonstrate that DyAVSA can bring about significant benefits for continuous post-deployment AV safety assessment. The fleet data will allow for a much faster reduction of the uncertainty about the conditional probabilities of accident in different operating conditions than each of the AV instances can achieve by counting on own observations only. The AV instances can, therefore, benefit from the shared operational data among the fleet of AVs and conduct *individualized* AV safety assessments. These may indicate that some AV instances are driven in an operational profile, for which the safety claim may be (or has been) violated. Such a targeted safety assessment (via the use of DyAVSA) will allow the AV vendor to change *significantly* the policy of issuing advisories (i.e., AV recalls) to only those AV instances driven in operational environments leading to a safety claim violation.

1.1 Abbreviations

ODD – Operational Design Domain [11]. ODD consists of a set of 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.).

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, proposed in [2].

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

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

1.2 Notations

X – r.v. random variable

[3] The attempt to identify a suitable "field dataset" on which DyAVSA could be demonstrated unfortunately was unsuccessful as the analysed databases of road accidents (in USA, UK and Germany) do not seem to provide the level of details required by DyAVSA.

$f_x(\cdot)$ - 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 $P(OC_i)$
$f_\theta(\cdot)$ - 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.
$\mathrm{Dir}(X_1, X_2, \ldots, X_n|a_1, \ldots, a_n)$ – the Dirichlet distribution of non-negative random variables $X_1, X_2, \ldots, X_n$
$\mathrm{Beta}(X|\alpha, \beta)$ – 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

The traditional approach to post-deployment safety assessment relies on *periodic safety reviews*. These are used in many safety-critical industries. The approach, however, has limitations, which in the context of AV safety assessment are significant. For instance, the period of safety review (i.e., of operational data collection) is typically quite long (a year or longer). Responding to the indicators of potentially unsafe AV operation with such a long delay is itself a risk as actual accidents may result unless one acts upon early indications of unsafe operation. There is also an ongoing active debate as to what indicators of unsafe operation one should use with AVs: "near misses", safety performance indicators (SPI) [9] and "surrogate safety measures" (see Section 6 for further details) are only a few noticeable examples. In these circumstances it is not obvious how one should apply periodic safety reviews.
A safety claim may take different forms. The current view with AV safety is that a safety claim must be linked to an ODD, but there are different views on how this link should be applied. One view would be that a safety claim must hold *true in all operating conditions* inside the defined ODD. Such an approach will require an extensive safety assessment even for conditions which are *very unlikely to occur* in real operation. "Driving to safety", proposed by [2], takes a different approach and adopts as a measure of safety the probability of accident (i.e., catastrophic failure) per mile of driving (*pfm*), expected to be below a given threshold, e.g., lower than the value of the same measure computed for human drivers. *Pfm* is by definition a measure "on average", aggregated over *all* operating conditions, inside a given ODD. It is clear, that a sufficiently low *pfm* value can be achieved if the AV is driven mostly in "easy" conditions (with low *pfm*) and very rarely (if at all) in road conditions with high risk of accident (i.e., high *pfm*). A rational assessor in such circumstances would be tempted to achieve a safety

target based on *pfm* by reducing the risk of accident in those operating conditions which an AV will spend *most* of its driving and possibly allocate less effort on reducing the risk in conditions which are (very) unlikely to occur in operation[4]. Such an interpretation of a safety claim is fine as long as the likelihood of the operating conditions is stable (i.e., it does not change much over time and across different AV instances). However, in reality, the operating conditions vary. They may vary considerably. This leads to the possibility that a safety claim established by the AV vendor pre-deployment for an assumed *mix of operating conditions*[5], may be violated if/when the mix changes. Different AV instances are very likely to be used in different operating conditions, which may lead to violations of the safety claim for some AVs. This, in turn, would put the passengers of the affected AV instances at unacceptably high risk of road accidents. In summary, given the operational conditions in which different AV instances are used and the *intrinsic* variability of these conditions, it seems essential that a suitable monitoring procedure is put in place which allows the AV instances to conduct continuous run-time safety assessment and evaluate the impact of their current operational profile on the safety claim. Should the safety claim be violated, the affected AV instances should be stopped or at least notified of the increased risk from a road accident.
In a recent study into "driving to safety" [13], we scrutinized the role of the model used in a Bayesian inference and demonstrated that a *univariate* model, as proposed by [2], has a fundamental weakness – it cannot account for the *variation* of the likelihood of accident in highly dynamic operating conditions. We developed a multi-variate Bayesian inference, aligned well with the concept of ODD and variable operating conditions. We demonstrated that an inference procedure based on the proposed multivariate model is superior to any univariate Bayesian inferences, including the "conservative Bayesian inference" [15].

The newly developed multivariate Bayesian inference accounts for the uncertainty in both: i) the operational profile an AV is driven in, which may change significantly over time, and ii) the uncertainty about the *pfm* conditional on the operating conditions included in an ODD. The inference is split into two inference parts: i) learning about the evolving AV operational profile, and ii) learning about the conditional *pfm*s in the operating conditions included in the ODD. These two stages of the Bayesian inference provide an interesting *possibility* of updating the predicted operational profile using *one set of operational data*, e.g., in line with the AV own observations, and of updating the conditional *pfm*s using a different set of operational data. As a result of this flexibility,

---

[4] Similar judgements of discarding rare events (i.e. that can occur in unlikely operating conditions) are not unusual in safety assessment. An extreme example is the Fukushima nuclear plant disaster, where the impact of a tsunami was considered in the risk-assessment of the nuclear plant, but the likelihood of extreme tsunamis was considered very low, hence building a very high seawall - unnecessary. Shortly before the earthquake in 2011, the error in seawall calculations (and the assumptions made) was discovered by the national nuclear regulator, but the operator did not rush to implement adequately high seawalls [14].

[5] The mix of operating conditions, e.g. defined by an ODD, together with a *probability distribution* on the set of distinctly different operating conditions is known in software reliability engineering (e.g., in software testing) as "operational profile", which we introduce formally in Section 3.

the vendor and the AV instances can use the developed multivariate inference differently. The vendor can use the observations from the entire AV fleet for both – to update the estimated operational profile and the conditional *pfm*s. The individual AVs can use the multivariate procedure with the own observations only for both – to update the own operational profile and the estimates of the conditional *pfm*s. The AV instances, however, can use the own observations to update the own operational profile, but for the conditional *pfm*s the AV can use the observations of the entire AV fleet, which will consist of much more extensive operational evidence about the safe operation of the *AV brand* in each of the operating conditions of a given ODD. We called the latter option of applying the multivariate Bayesian inference relying on the fleet data "Dynamic AV Safety Assessment (DyAVSA)".

This flexibility with the data which can be fed into the inference procedure is due to the *nature* of the multivariate Bayesian inferences proposed in [13], which consists of two relatively separate inference steps. In this paper we study the difference between the predictions derived with the multivariate Bayesian inference by the vendor and by the individual AVs using different mixes of operational data (own and/or from the AV fleet).

3 The system model

3.1 Multivariate Bayesian inference

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* (i.e., an accident) per unit of distance, e.g., per mile (or kilometer), of driving following the proposal in [2]. Assume further that the probability of observing a failure 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 (each resulting in a success of an accident) can be modelled mathematically as a *Bernoulli trial* of miles of driving *selected at random* from the population of all miles with a probability of failure/accident per mile (*pfm*). 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 called a measure of "epistemic uncertainty", related to the assessor's knowledge (belief) about the value of *pfm*.

The concept of Operational Design Domain (ODD) [11], informally introduced earlier, captures the idea that risks of road accident *may vary* with the operating conditions. An ODD is typically defined as a *partition* of different operating (road) conditions, OCs, as follows:

OC = $\{OC_1, OC_2, \ldots, OC_m\}$ such that iff i ≠ j then $OC_i \cap OC_j = \varnothing$. This is illustrated below in **Fig. 1**.

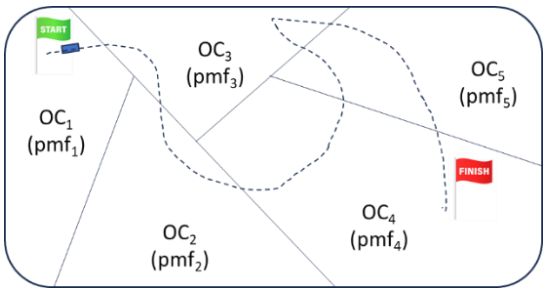

Fig. 1. A model of ODD as a partition of operating conditions {OC_1, … OC_n}. Each OC_i is characterized by a probability of failure (accident) per mile of driving, pfm_i. A vehicle (shown as a blue rectangle on the left) is shown as driving along a "trajectory" (shown as a dashed curve), which starts in OC_1, crosses OC_3, OC_2, OC_4, OC_5, OC_3, OC_5 and finishes in OC_4.

Let us assume that each OC_i includes a "homogeneous" set of miles in the sense that for each of the miles included in OC_i the probability of catastrophic failure/accident per mile can plausibly be assumed the same, pfm_i. The pfm_i can, however, vary across OC_i, ($i = 1, 2, \ldots, n$)[6].

The recently developed model [12] (PPR-model), which builds on the work by Adams [16], lands itself well to dealing with the problem at hand under the assumptions we have made so far. Appendix 1 provides the essence of the PPR-model.

The model of Bayesian assessment with an ODD, thus, leads to a *double-stochastic multivariate model* in which:

- we capture the likelihood of selecting a mile at random from partition OC_i using a probability distribution defined on the set of partitions OC, $P(OC_i)$[7], and
- the joint distribution $f_{\theta_1,\theta_2,\ldots,\theta_n}(\theta_1, \theta_2, \ldots, \theta_n)$, which characterizes the uncertainty in the value of pfm_i in different OC_i and possibly the stochastic dependencies between the variates, $\Theta_1, \Theta_2, \ldots, \Theta_n$, of the multivariate distribution, $f_{\theta_1,\theta_2,\ldots,\theta_n}(\theta_1, \theta_2, \ldots, \theta_n)$.

To simplify the analysis, we make a couple of *additional* assumptions:

- We ignore the details on how AV moves within OC_i and assume that each mile driven by an AV is *chosen at random* from the respective OC_i and model the selection as a Bernoulli trial. This assumption is clearly simplistic. **Fig. 1** shows an alternative model – a vehicle moving along a trajectory through different *OC*s. Later in the paper (Section 5) we discuss further how stochastic state-based models can replace the Bernoulli trial model.
- pfm_i are assumed *independently distributed* random variables and we use the notation Θ_i and $f_{\theta_i}(\cdot)$ for the random variables and the probability density functions of Θ_i, respectively, for $i = 1,\ldots, n$. In other words, we assume that changes in $f_{\theta_i}(\cdot)$ do

[6] We assume that each mile can be attributed reliably to a particular operating condition. This assumption, the implications of incorrect attribution of miles/accidents on the prediction outputs is further discussed later in the paper together with ways of relaxing it.

[7] A more refined definition of the partition model may require further details, e.g., a state model of AV moving between the *OC*s in which the *OC*s are states and the transitions between the *OC*s are defined stochastically.

not affect $f_{\theta_j}(\cdot), i \neq j$. We discuss the implications of this assumption and ways of relaxing it in Section 5 of the paper, too. Appendix 2 provides further details on the implications of the assumption that pfms are independently distributed random variables by constructing $f_{\theta_1,\theta_2,\ldots,\theta_n}(\theta_1, \theta_2, \ldots, \theta_n)$ using Copula to capture the dependencies between the variates $\Theta_1, \Theta_2, \ldots, \Theta_n$ and their impact on a weighted sum of the variates.

$P(OC_i)$ may vary over time or be subject to epistemic uncertainty, which we capture by using a random variable, Ψ_i with a probability density function $f_{\psi_i}(\cdot)$. Since the operating conditions form a partition of the space of miles, the constraint $\sum_{i=1}^{n} \Psi_i = 1$ applies: a mile with certainty will be selected from one of the partitions[8].
We now express the joint distribution $\mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n})$, which captures the epistemic uncertainty associated with the selection of a mile from the space of all miles. This distribution is known in software reliability engineering and safety as *operational profile*. A suitable analytic *multivariate distribution* which can be used here to capture the uncertainty in the operational profile and its variation over time is the *Dirichlet* distribution, which for *n* variates, $\Psi_i \ldots, \Psi_n$ is defined as [17]:

$$\begin{aligned}\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\boldsymbol{\alpha}) &\equiv \mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n}) \\ &= \frac{\Gamma(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}_\mathrm{i})}{\prod_{\mathrm{i}=1}^{\mathrm{n}} \Gamma(\mathrm{a}_\mathrm{i})} \left[\prod_{\mathrm{i}=1}^{\mathrm{n-1}} \psi_\mathrm{i}^{\mathrm{a_i}-1}\right]\left[1-\sum_{\mathrm{i}=1}^{\mathrm{n-1}} \psi_\mathrm{i}\right]^{\mathrm{a_n}-1}\end{aligned} \quad (1)$$

Using the joint distribution $\mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n})$ we can now express the marginal distribution of the *system pfm* $f_\theta^{WB}(x)$ as follows (see Appendix 1 for further details):

$$\begin{aligned} & f_\theta^{WB}(x) \\ &= \int f_{\theta|\psi_1,\psi_2,\ldots\psi_n}(x|\psi_1, \psi_2, \ldots, \psi_n)\mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n})d\psi_1\psi x_2 \ldots d\psi_n \\ &= \int \left[f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \ldots * f_{\theta\psi_n}(x) \times \mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n})\right] d\psi_1 d\psi_2 \ldots d\psi_n \end{aligned} \quad (2)$$

Let us now consider how *new operational evidence* from driving an AV would affect the distribution $f_\theta^{WB}(x)$. Let us consider that we have received operational evidence in the form $\{(N_1, r_1), (N_2, r_2), \ldots, (N_n, r_n)\}$ of the miles driven, N_i, and failures/accidents observed, r_i, $0 \leq r_i \leq N_i$, in each of the operating conditions, OC_i. We can account for the new operational data by conducting a Bayesian inference in the following steps:

- ***Step 1***: Update the uncertainty related to the operational profile, $\mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n}|N_1, N_2, \ldots, N_n)$. Note that the updated operational profile is not affected by the number of failures/accidents that have been observed. The posterior distribution only depends on the number of miles driven in different operating conditions. For instance, if we capture the operational profile uncertainty using a Dirichlet distribution, $\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\boldsymbol{\alpha})$, then the new evidence (i.e. $\{(N_1, r_1), (N_2, r_2), \ldots, (N_n, r_n)\}$)

[8] Although dealing with "out of ODD" is outside the scope of the paper we note that accounting for "out of ODD" would simply add an additional partition, $OC_{out\text{-}of\text{-}ODD}$.

will lead to a new Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}^{\boldsymbol{post}})$, which is derived from $\mathrm{Dir}(\boldsymbol{\alpha})$ by a simple modification of the parameters of the Dirichlet distribution:

$$\mathrm{Dir}(\boldsymbol{\alpha}^{\boldsymbol{post}})=\mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_n;\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_n + N_n) \quad (3)$$

- ***Step 2***: The conditional distributions, $f_{\theta_i}(x|N_i, r_i)$ of failure/accident per mile in OC_i will be updated to reflect the new evidence by conducting Bayesian inferences on conditional distributions $f_{\theta_i}(x|N_i, r_i)$ in each of OC_i as follows:

$$f_{\theta_i}(x|N_i, r_i) = \frac{f_{\theta_i}(x)\times L(N_i, r_i|x)}{\int_{x=0}^{1} f_{\theta_i}(x)\times L(N_i, r_i|x)dx} \quad (4)$$

where $L(N_i, r_i|x)$ is the likelihood of observing r_i accidents in N_i miles. For Bernoulli trial a binomial likelihood is used, $L(N_i, r_i|x) = \binom{N_i}{r_i} x^{r_i}(1-x)^{N_i - r_i}$.

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 updating the conditional distributions in each of the operating conditions is affected by both the number of miles, N_i, and the number of failures/accidents, 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 the conditional probability of failure, Θ_i, but do not affect Θ_j for the other operating conditions.

- **Step 3**: derive $f_{\theta\psi_i}(x|N_i, r_i)$ from $f_{\theta_i}(x|N_i, r_i)$ using (A3 of Appendix 1).
- ***Step 4***: Using the distributions updated in Step 1 and Step 2 above we apply (2) and derive the marginal distribution of the probability of system failure, $f_{\theta}^{WB_{post}}(x|N_1, r_1, N_2, r_2, \ldots, N_n, r_n)$ as follows:

$$f_{\theta}^{WB_{post}}(x|N_1, r_1, N_2, r_2, \ldots, N_n, r_n) = \int\left[f_{\theta\psi_1}(x|N_1, r_1) * f_{\theta\psi_2}(x|N_2, r_2) * \ldots * f_{\theta\psi_n}(x|N_n, r_n)\right]\mathrm{Dir}(\boldsymbol{\alpha}^{\boldsymbol{post}})dx_1 dx_2 \ldots dx_n \quad (5)$$

The symbol "*" is the convolution operator of independently distributed random variables $\theta\psi_i$ with probability density functions $f_{\theta\psi_i}(x|N_i, r_i)$, respectively.

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

A more detailed discussion of the derivation can be found in Appendix 1.

So far, we have not specified explicitly how the data needed in the Bayesian inferences would be collected. The counts $\{(N_1^1, r_1^1), (N_2^1, r_2^1), \ldots (N_n^1, r_n^1)\}$ could come from an individual AV or from a fleet of AVs. The inference procedures will be the same irrespective of whether the counts are collected for a single AV instance or for a fleet of AVs.

Let us now look at the differences between using the data collected from an AV instance or from a fleet of AVs of the same brand.

3.2 Safety assessment for the AV vendor

Let us consider the case of a fleet of L AV instances from a given AV brand being deployed in operation. Each AV instance will be used within the specified ODD, but the operational profile, which applies to instance AV_m, i.e., $f^{(m)}{}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_n)$, may differ from the profile $f^{(l)}{}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_n)$ which applies to instance AV_l. For each OC_i the instance AV_m will collect:

- The number of miles driven, $N_i^{(m)}$, in OC_i: some miles will be without failure/accident, some – with failures/accidents.
- The number of miles with failures/accidents, $r_i^{(m)}$ for each OC_i.

These numbers will be periodically passed to the vendor[9] (e.g., by installing on each AV a device dedicated to collecting $N_i^{(m)}$ and $r_i^{(m)}$ and sending these over to the vendor). The vendor will then be able to aggregate the observations received from all AV instances and compute the following sums:

$$M_i = \sum_{m=1}^{L} N_i^{(m)}$$
$$r_i = \sum_{m=1}^{L} r_i^{(m)} \quad (6)$$

M_i and r_i will be used by the vendor to conduct an inference to assess the *pfm* for the AV brand. Note that under this scenario, the inference will account for the evidence (miles driven and failures observed) collected from the entire fleet of deployed AV instances and with the level of details required by the multivariate model using the ODD. The posterior distributions $f_{\theta_i}(\cdot\,|M_i, r_i)$ will account for every piece of operational data that has been seen in OC_i by the fleet of AVs deployed by the vendor. The operational profile (which we will refer to as an "operational profile on average"), derived with the M_i counts (the number of miles seen in operation condition OC_i calculated as shown in (3)), however, may well differ from the operational profiles of all deployed AV instances. If this is the case, the "operational profile on average" derived by the vendor, may not be useful to assess the safety of *any* of the deployed instances.

With the aggregated M_i and r_i counts, the vendor could apply an inference on $f_{\theta_i}(\cdot\,|M_i, r_i)$ and using (3), (4) and (5) derive the posterior distributions of the AV brand, i.e., "on average" over the entire fleet of deployed AV instances. Given the fact that the inference for the "operational profile on average" using (3) may not be immediately useful to any of the deployed AV, one wonders what benefits the inference

[9] Clearly, some synchronization is needed between the vendor and the deployed AV instances so that the counts collected by the instances for a particular epoch of observation are accurately passed to the vendor. We acknowledge that deploying a robust synchronization procedure is an important implementation detail, but one which is outside the scope of this paper

conducted by the vendor can offer to the deployed instances. Here are a couple of considerations.

- $f_{\theta_i}(\cdot\,|M_i, r_i)$ would capture the current knowledge about *pfm*$_i$ in *OC*$_i$ arrived at by using the data from the entire fleet. This data is *valuable* and can be used not only to derive $f_{\theta}^{WB_{post}}(x|M_1, r_1, M_2, r_2, \ldots, M_n, r_n)$ for "the operational profile on average", but also for *any other operational profile*, judged by the vendor as important. For instance, the vendor may be interested in conducting a safety assessment "on average" for a geographical region in which the vehicles of the particular AV brand have seen *no or very little operational exposure* to date. The vendor may be able to define an anticipated operational profile for that region (e.g., using any data that may be available for the man-driven vehicles in the region). Once the anticipated operational profile is defined, the posterior distributions $f_{\theta_i}(x|M_i, r_i)$ derived for observations collected under a *different* operational profile, can be used to construct the distribution of the *system pfm* "on average" for the anticipated new (regional) operational environment.
  In extreme cases it is possible, of course, that the ODD for the region of interest may only *partially overlap* with the ODD, for which the operational data has been collected, i.e., the anticipated ODD may contain *OC*s for which *no data* has been collected from the deployed fleet. For instance, in regions of extreme weather (e.g., polar circle) there may be *OC*s for which data collected in moderate climates will offer no observations. In this case, of course, the benefits from reusing the available M_i and r_i counts may be limited.
  Note that reusing the data from the AV fleet is only possible with an inference model where the conditional *pfm*$_i$ are explicitly accounted for as separate random variables, $\Theta_1, \Theta_2, \ldots, \Theta_n$. A similar "extrapolation" from one operational profile to another would require *pfm*s being assessed under the first operational profile.
- Knowledge about the operational exposure accumulated by the fleet of AVs to date may be *useful for the vendor* in the limited sense of finding out how the deployed AV fleet is used "on average". Comparing the "observed" profile "on average" with the "target" profile (i.e., the one for which the safety has been claimed) may itself provide the vendor with either an assurance that the assumptions made about the *operational profile* in the safety assessment prior to AV deployment are (broadly) correct, or that the assumptions are "biased". The latter, in turn, may trigger a safety review to check if the AV for the observed "operational profile on average" is sufficiently safe even if the AV safety for the observed environment differs significantly from the one used before the AV fleet deployment. Again, such an analysis is only possible if the inference is based on a model, in which the operational profile and the conditional *pfm*s are derived from data separately.
- Finally, in addition to the analysis "on average", the vendor may conduct a run-time safety assessment of the individual vehicles following the procedure explained in section 3.3 using the data provided by the individual AVs with the vendor. Should the safety of some AVs become inadequate, the vendor may issue them with a

warning. Such a "targeted warning" campaign is quite different from what is done by the car manufacturers at the moment[10].

3.3 *DyAVSA* for safety assessment of individual AVs

The Bayesian inference discussed above using the observations from the entire fleet of data, especially the inference of $f_{\theta_i}(x|fleet\ data)$, can be useful for the deployed instances. A newly deployed AV instance could be bootstrapped by the vendor at the time of deployment with prior distributions $f_{\theta_i}(x)$, each of which may be the posterior distribution derived by the vendor using the observations accumulated from the entire fleet of previously deployed AV instances. In other words, the newly deployed AV instance will be provided with up-to-date priors $f_{\theta_i}(x)$, which account for the evidence from the entire AV fleet of deployed AVs. Regarding the operational profile of the newly deployed AV instance, it can be the profile "on average" computed by the vendor to date, or another profile, when there are reasons to trust the alternative profile more than the profile "on average".

From this point in time on, a newly deployed AV_m, will rely on the data it collects, $N_i^{(m)}$ and $r_i^{(m)}$. Quite clearly, the operational exposure of AV_m will be limited in comparison with the exposure of the entire AV fleet (one assumes that the fleet will be large, of course). Accounting for own observations only, however, will allow AV_m to learn (e.g., to improve the confidence in the value of the chosen measure of safety) slowly in comparison with learning from the experience of the entire AV fleet. Consider the following scenario, in which the data collected by the *vendor* from all AV instances is organized in *epochs* of observations: E_1, E_2, …, E_n ,… and $\left(M_1^{(E)}, r_1^{(E)}\right), \left(M_2^{(E)}, r_2^{(E)}\right), \dots, \left(M_n^{(E)}, r_n^{(E)}\right)$ represents the data accumulated by the vendor during epoch *E*. $M_i^{(E)}, r_i^{(E)}$ are the counts we defined in (6) for epoch *E*. The vendor at the end of epoch *E* could broadcast $\left(M_1^{(E)}, r_1^{(E)}\right), \left(M_2^{(E)}, r_2^{(E)}\right), \dots, \left(M_n^{(E)}, r_n^{(E)}\right)$ to all deployed AV instances. With these aggregated counts each AV instance will be able to update the own distributions $f_{\theta_i}(x|fleet\ data)$[11].

AV_m will update its own operational profile using its *own observations only*, i.e., if the operational profile, if expressed as a Dirichlet distribution for AV_m the profile will become:

$$\mathrm{Dir}_m(\boldsymbol{\alpha^{post}}) = \mathrm{Dir}\left(\psi_1, \psi_2, \dots, \psi_n; \alpha_1 + N_1^{(m)}, \alpha_2 + N_2^{(m)}, \dots, \alpha_n + N_n^{(m)}\right)$$

Thus, the posterior distribution of $f_\theta^{WB_{post}}(x|fleet\ data, own\ data)$ becomes:

$$f_{\theta_{m_{fleet}}}^{WB_{post}}\left(x|M_1, r_1, M_2, r_2, \dots, M_n, r_n, N_1^{(m)}, N_2^{(m)}, \dots, N_n^{(m)}\right) =$$

---

[10] To recall the entire fleet of vehicles following a serious incident.

[11] $fleet\ data$ is a shortcut for $\left(M_1^{(E)}, r_1^{(E)}\right), \left(M_2^{(E)}, r_2^{(E)}\right), \dots, \left(M_n^{(E)}, r_n^{(E)}\right)$. It is a technical (implementation) detail whether broadcasting $\left(M_1^{(E)}, r_1^{(E)}\right), \left(M_2^{(E)}, r_2^{(E)}\right), \dots, \left(M_n^{(E)}, r_n^{(E)}\right)$ will be more efficient (computationally and in terms of communication bandwidth) than broadcasting the conditional distributions $f_{\theta_i}(x|fleet\ data)$.

$$\int\left[f_{\theta\psi_1}(x|M_1,r_1) * f_{\theta\psi_2}(x|M_2,r_2) * \ldots * f_{\theta\psi_n}(x|M_n,r_n)\right] \times \mathrm{Dir}_m(\boldsymbol{\alpha}^{\boldsymbol{post}})^{12} \quad (7)$$

As indicated earlier, we call the assessment leading to (7) "Dynamic AV Safety Assessment (DyAVSA)".

If instead of using the fleet data $\mathrm{AV_m}$ only relied on its own observations in updating the operational profile and the distribution of the conditional *pfm*s, then the posterior marginal distribution would be:

$$f_{\theta m_{own}}^{WB_{post}}\left(x|\, N_1^{(m)}, r_1^{(m)}, N_2^{(m)}, r_2^{(m)}, \ldots, N_n^{(m)}, r_n^{(m)}\right) =$$
$$\int\left[f_{\theta\psi_1}\left(x|N_1^{(m)}, r_1^{(m)}\right) * f_{\theta\psi_2}\left(x|N_2^{(m)}, r_2^{(m)}\right) * \ldots * f_{\theta\psi_n}\left(x|N_n^{(m)}, r_n^{(m)}\right)\right] \times \mathrm{Dir}_m(\boldsymbol{\alpha}^{\boldsymbol{post}}) \quad (8)$$

In summary, conducting a continuous safety assessment of AV instances either using *DyAVSA* or counting on own observations only will allow for monitoring how safety will vary over the lifetime of individual AV instances. Continuous safety assessment can be conducted in a *decentralized fashion*, which brings advantages. For instance, if the vendor's server is down, then DyAVSA may temporarily be disabled, too. In this case, the AVs can continue the assessment switching to using their own data only. As soon as the vendor's server is back up, DyAVSA can be enabled.

DyAVSA brings a clear advantage for the AV instances: they can learn *faster* about the values of the conditional *pfm*s in the different operating conditions than if they had counted on the own observations only, which in turn will reduce the uncertainties in the conditional *pfm*s and of the marginal *system pfm*.

Fig. 2 below illustrates the data flow used in DyAVSA and how the data exchanged between the AV instances, and the vendor affects the predictions by an AV instance and by the AV vendor.

The key elements in DyAVSA are:

- AVs-to-Vendor communication. The AV instances send the observations they have collected, $\left(N_1^{(E)}, r_1^{(E)}\right), \left(N_2^{(E)}, r_2^{(E)}\right), \ldots, \left(N_n^{(E)}, r_n^{(E)}\right)$, for each "epoch" E of observation respectively, to the Vendor Data Centre (VDC), where the data is aggregated, anonymized as necessary, and used by the vendor to derive the multivariate posterior "operational profile on average" and the conditional distributions $f_{\theta_i}^{E}(\cdot\,|M_i, r_i)$, in each of the operating conditions. This communication is sufficient for the vendor to monitor the safety "on average", but also the safety of the individual AVs.
- Vendor-to-AVs communication. The vendor shares with the AV instances the distributions of the conditional *pfm*s, $f_{\theta_i}^{E}(\cdot\,|M_i, r_i)$ (denoted as $f_{\theta_i}(\cdot\,|fleet\ data)$ in the figure) in all operating conditions. These distributions can be used by the individual AVs to compute the distribution of the marginal *system pfm* for the particular AV. Note, that the vendor does not share with the AV instances the operational profile "on average", as this profile is of no use to the individual AVs. In the absence of Vendor – to - AVs communication, the AV instances will not be receiving

[12] We omitted the indexes referring to epochs from (7) to simplify the expression. All counts M_i, r_i and $N_1^{(m)}$ are aggregated during the respective epoch.

$f_{\theta_i}^{E}(\cdot\,|fleet\ data)$, hence will be unable to benefit from the observations collected by other AVs. In this case they still can monitor their own safety relying only on the data they have collected themselves. Under the "Individual AV perspective" the diagram shows *two multivariate inferences* labelled "Own data only" and DyAVSA ("All fleet data"), respectively, thus illustrating the differences between the inferences based on locally collected data only or on data shared by the vendor, respectively.

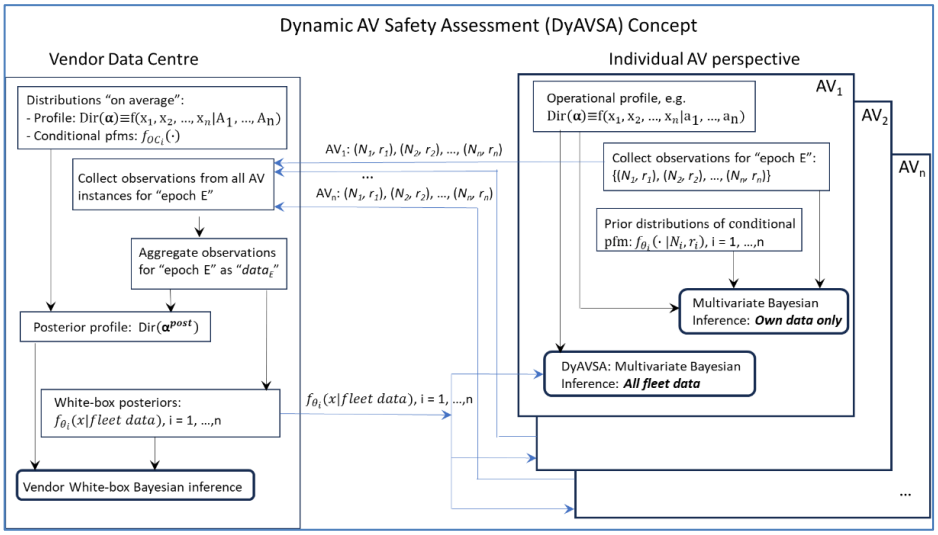

Fig. 2. Dynamic AV Safety Management (DyAVSA) concept in comparison with dynamic safety management by the AV vendor. The indexes "E" referring to data collection epochs have been omitted to simplify the figure. For the same reason we use $f_{\theta_i}(\cdot\,|fleet\ data)$ as a shortcut for the set of conditional distributions $\{f_{\theta_i}(\cdot\,|M_i, r_i), i = 1, \ldots, n\}$.

4 Contrived examples

We use several *contrived examples* to illustrate how the multivariate Bayesian predictions are affected by whether DyAVSA is used or not.

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 and the prior distributions of the conditional pfm_i are defined as *Beta* distributions with the following parameters[13]:

$f_{\theta_1}(x) \equiv Beta(\alpha = 2, \beta = 299)$,

$f_{\theta_2}(x) \equiv Beta(\alpha = 2, \beta = 800)$,

[13] Using Beta distributions is not essential for the method. A different type of distributions can be used for the conditional pfm_i. In the latter case 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.

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

The parameters of the Beta distributions are chosen to illustrate the possibility that *OC*s may differ both in terms of expected pfm_i and in terms of the *uncertainty* in the values of the conditional pfm_i in the respective *OC*s.

We assume that the vendor assessed the operational profile and expressed it as a Dirichlet distribution $\mathrm{Dir}(\psi_1, \psi_2, ..., \psi_n; \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10)$. In the examples used in this section this prior profile is assigned to all AVs and to the vendor.

With the defined prior operational profile and distributions $f_{\theta_i}(x)$ of the conditional *pfm*s the marginal prior distribution of the system *pfm* can be derived using (2).

Now let us consider a fleet of 5 AVs (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***

Table 1. Observations by AV_1 … AV_5.

Table 1 shows two observations. *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.

Each AV is assumed to have driven 100 additional miles with both *Observation 1* and *Observation 2*. Thus, both observations consist of 500 miles (the sum of the miles driven by all AVs).

4.1 AV instance inference: own data only vs. data from the fleet of AVs

In this sub-section we compare the Bayesian predictions by AV *instances* for the following cases:

- An AV instance uses *its own data* only.
- An AV instance uses DyAVSA, defined above.

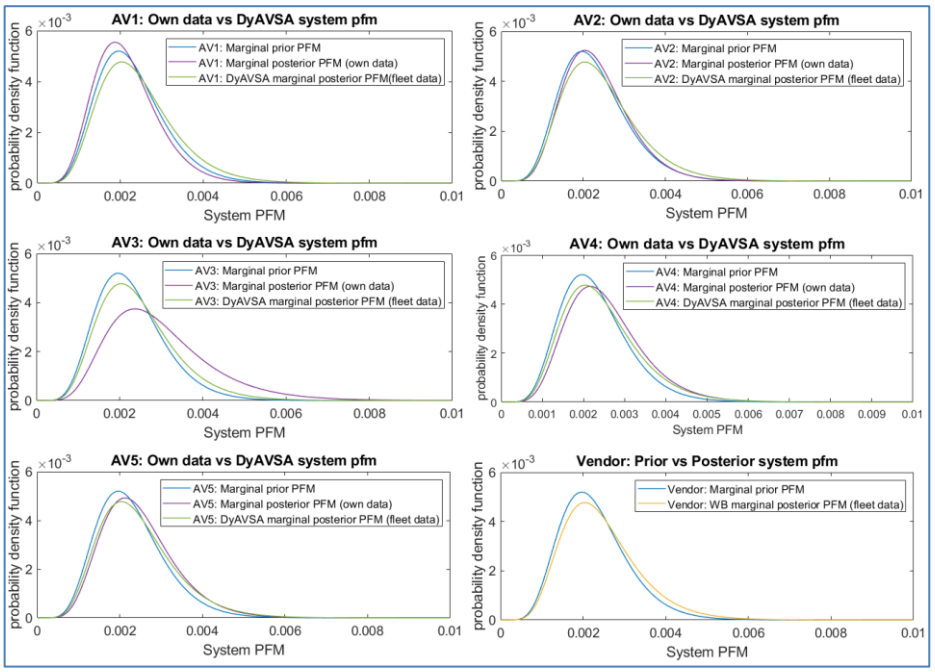

Fig. 3. Illustration of the marginal *system pfm* of individual AVs using either own data only or the aggregated data for the entire fleet (as proposed with DyAVSA) and the vendor for **Observation 1** (no failures).

The predictions are derived from the observations described above: Observation 1 – with no failures observed, and Observation 2 – with some failures observed by AV_3. The results are captured in **Fig. 3** (Observation 1) and **Fig. 4** (Observation 2), respectively.

The plots show the distributions of the marginal *system pfm* for the AV instances computed with and without DyAVSA, and for the vendor. The marginal prior distribution of *system pfm*, the same for all AVs and the vendor, is also shown.

We can make a few observations from **Fig. 3**:

- The impact of the fleet data is clearly visible – the posteriors by the AV instances based on fleet data differ more significantly from the priors than the AV posteriors based on own data only.
- The case of AV_1 provides an interesting insight. The predictions for AV_1 based on own observations only are *more optimistic than the prior*. Looking at the number of miles driven by AV_1, we note that it spent only 7 miles in OC_1, less than the prior operational profile would suggest ("on average"). Hence, AV_1 benefits from the own observations in two ways: it observes no failures, hence the conditional probabilities of failure in OC_1 will be predicted to get "stochastically smaller", and the new AV_1 operational profile (after driving additional 100 miles) will make OC_1 even *less likely*

than it was in the prior operational profile. Since OC_1 is the worst OC (with the biggest expected prior *pfm* of all *OC*s) the reduction of its weight in the operational profile stochastically reduces the posterior system *pfm* for AV_1.

- The results for AV_3 are quite interesting, too. Its white-box posterior distribution of *system pfm* based on own data only is significantly more pessimistic than the predictions of *system pfm* based on the fleet data. This is a consequence of AV_3 spending 45% of the driving in OC_1, the worst operating condition. As a result, the weight of OC_1 in the operational profile *increases* significantly (the posterior probability of selecting a mile from OC_1 will become 55/200, e.g., more than 27%). The number of additional miles driven in OC_1 will only marginally reduce the conditional pfm_1. In comparison, with the fleet data the posterior distribution of the pfm_1 will be more optimistic than with own data only. Hence, with own data only the overall effect of the additional driving is that the posterior pfm_3 is now worse (i.e., more pessimistic) than with the fleet data. Both predictions (with own data only and with the fleet data) are more pessimistic than the prior.
- AV_2, AV_4 and AV_5 offer further interesting insight. The posterior *system pfm* for AV_4 and AV_5 using own data only are slightly more pessimistic than the predictions based on the fleet data. The two posteriors for AV_2 are *not stochastically ordered* – the posterior probability density functions have a *cross-over point*: the tail of the predictions of the *system pfm* with the fleet data is "thicker" than the predictions with own data only.

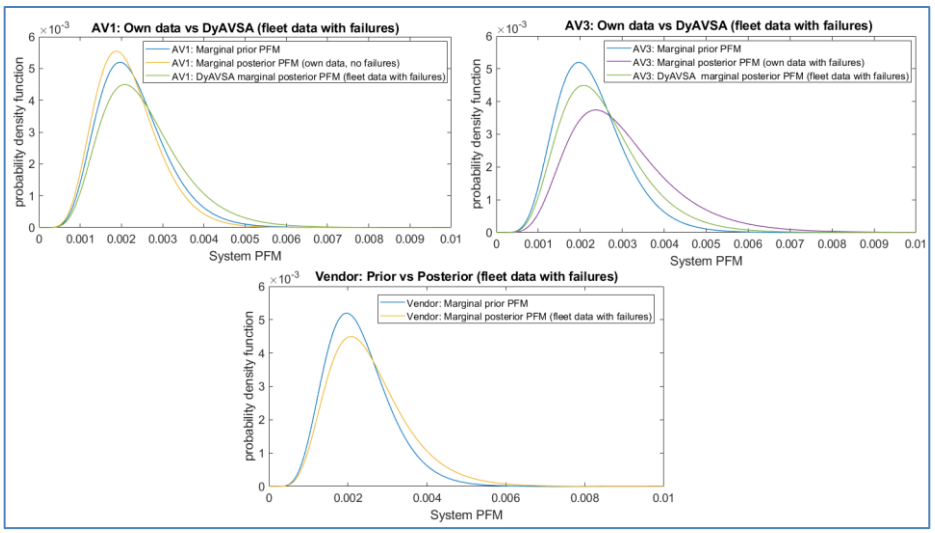

Fig. 4. Black-box vs. White-box inference of the marginal probability of catastrophic failure of **individual AVs** using either own data or fleet data (DyAVSA) for **Observation 2** (with failures of AV_3).

Fig. 4 only shows the posteriors for two of the AVs: AV_1, which did not experience any failures of its own, and AV_3 which experienced two failures in OC_1 and OC_2, respectively.

The plots indicate patterns similar to those recorded with Observation 1 (shown in **Fig. 3**):

- The posterior distributions based on *own data* of AV_1 (and of AV_2, AV_4 and AV_5) are identical to those recorded in **Fig. 3**, which is to be expected since these AVs did not observe any failures of their own under Observation 2. The posterior distributions recorded for AV_3 based on own observations are similar but *slightly worse* than the posteriors derived for AV_3 with Observation 1.
- The predictions based on fleet data and own operational profile are quite similar to the ones that we recorded with Observation 1. The posterior distribution of *system pfm* for AV_3 based on own data is again worse than the predictions based on the fleet data and is subject to *much greater uncertainty* as evident from the spread of the probability mass of the posterior distribution of *pfm* of AV_3.
- Finally, the patterns that we observe for the vendor are also similar to those recorded for Observation 1: the posterior is worse than the prior due to the observed failures of AV_3 and the complex interplay between the operational profile (changed due to the additional miles driven) and how the numbers of miles and accidents affect the predicted *system pfm*.

We can conclude from **Fig. 3** and **Fig. 4** that the predictions of the marginal *pfm* are quite *sensitive to the data* used in the inference and indicate that DyAVSA can bring significant advantages to the individual AVs. The results clearly indicate that using the data collected from the entire fleet affects the predicted distributions of the marginal *system pfm* of the AVs.

4.2 Conditional probability of failure in $OC_1 - OC_5$: prior vs posterior, AVs own data only vs. fleet data

The next two figures, **Fig. 5** and **Fig. 6**, provide further details on how the distributions of the conditional pfm_i in $OC_1 - OC_5$ are affected by the data used in the inference: own data by AV_1, …, AV_5 only or data collected by the entire fleet.

Fig. 5 shows that using own data or fleet data makes a considerable difference in all *OC*s. The magnitude is most significant in OC_1. **Fig. 5** also plots together (the bottom right of the figure) the posterior distributions of the marginal system *pfm*, computed by the AVs based on own data and by the vendor using in full the fleet data.

Comparing the plots of the conditional *pfm*s on OC_1, …, OC_5 does not indicate visible differences between the predicted $f_{\theta_i}(x|own\ data)$, computed by the AVs and $f_{\theta_i}(x|fleet\ data)$, computed by the vendor. We would expect that this would imply similarity to the predicted system *pfm*. Surprisingly, however, the posterior system *pfm* (shown in the bottom right plot of the figure) computed by AV_3 (marked with the arrow labelled "1" in the right-bottom plot in the figure) stands out and is visibly more pessimistic than the predicted system *pfm* of the other AVs and of the vendor, which are close to one another. This observation suggests that counting on the vendor predictions "on average" only (labelled with "2" in the right- bottom plot of the figure) may be misleading, as in this case the spread of safety predictions by the individual AV

instances may be significant (e.g. as is with AV_3) but will remain unseen.

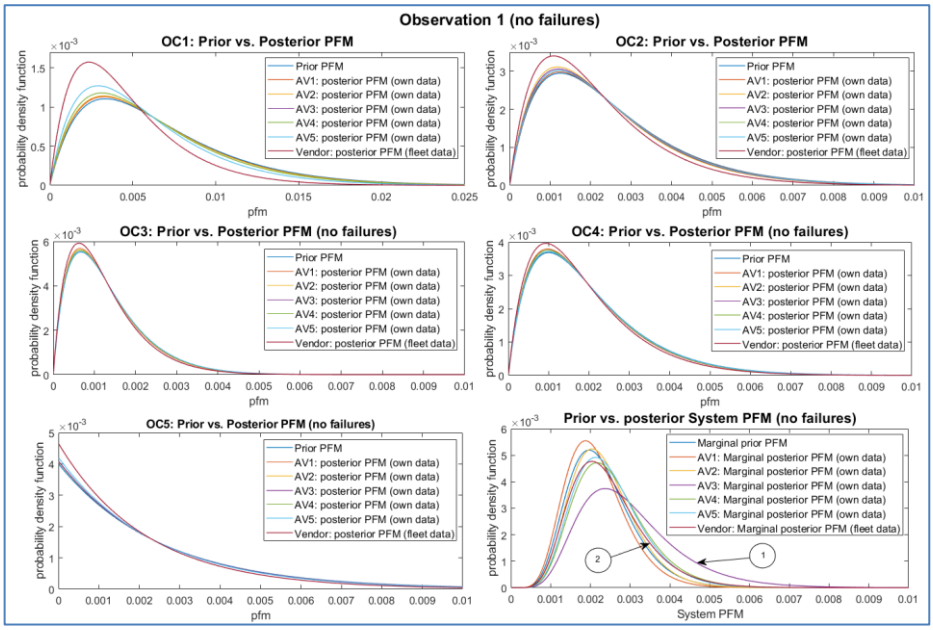

Fig. 5. Bayesian predictions on the conditional pfm_i in $OC_1 - OC_5$ for Observation 1 (no failures): own data vs. fleet data.

Another noteworthy observation from **Fig. 5** is that the AVs predictions vary in terms of how they compare with the prior distribution of system *pfm*: some of the posterior distributions are *more optimistic* than the prior (i.e., the tails of the respective distributions are "thinner" than the tail of the prior), while other posteriors – are more pessimistic than the prior (i.e. their tails are "thicker than the tail of the prior"). These differences are due to the complex dependence of the system *pfm* distribution on the operational profile of the AVs (or the vendor) and how the additional miles of AV driving have changed the distributions of the conditional *pfm*s in the operating conditions. Counting on the vendor's predictions alone will not allow one to see that individual AVs' predicted system *pfm* may be close and even violate a safety claim.

Under Observation 2 (**Fig. 6**) the posterior distributions $f_{\theta_1}^{(3)}\left(x|N_1^{(3)}, r_1^{(3)}\right)$ and $f_{\theta_2}^{(3)}\left(x|N_2^{(3)}, r_2^{(3)}\right)$ by AV_3 of the conditional *pfm*s in OC_1 and OC_2, respectively, are visibly different from the predictions of the other AVs and of the vendor, which is expected as these are the two *OC*s in which AV_3 has observed accidents. Interestingly, while the posteriors for AV_3 are visibly worse than the prior, the posterior distributions of pfm_1 and pfm_2 by the vendor are more optimistic than the prior. Clearly, a single failure by AV_3 in both OC_1 and OC_2 is insufficient to make the two posteriors calculated by the vendor for OC_1 and OC_2, respectively, to become more pessimistic than the priors (assumed in the example the same for all AVs and the vendor) for these *OC*s. This example clearly shows the possibility for the "predictions on average" to "smooth over"

the impact of sudden "reverse of fortune" when some AVs observe accidents in some of the *OCs* and reiterates the point that we have already made above that the predictions "on average" may be biased and hide important information related to safety of individual AVs.

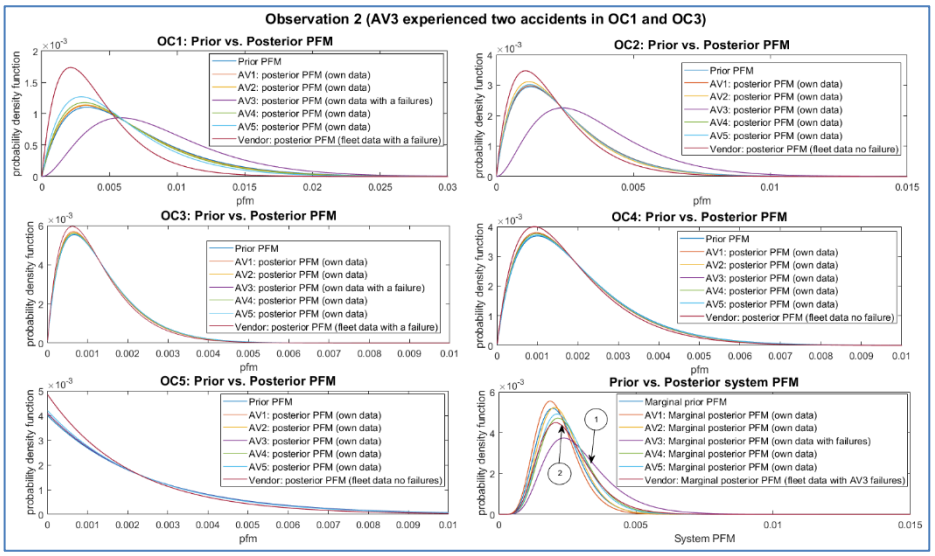

Fig. 6. Bayesian predictions on the conditional pfm_i in OC_1 – OC_5 for Observation 2 (failures observed): own data vs. fleet data.

5 Discussion and threats to validity

The results from the contrived examples demonstrate that the effect of the data used in the Bayesian inference may be *quite significant*. We observed that multivariate probabilistic models, which account for a variable operational profile, bring the following advantages:

- The predictions are in tune with the needs of AV safety assessment and account for fluctuations in operational profiles by individual AVs. The advantages of the multivariate inference over a univariate inference are extensively discussed in our recent work [13].
- Forces assessors to collect operational data, which is suitable to *porting a safety claim to a new operational profile* which may differ significantly from the profile for which the data has been collected,
- Serves the needs of the AV vendors and of the individual AVs, which are quite different: the individual AVs, via DyAVSA, may benefit considerably from the data collected from a fleet of AVs, and accounts for the own operational profile.

Clearly making use of DyAVSA would depend on how reliably the AV can

discriminate between different operating conditions. Unreliable discrimination will affect the accuracy of counting the miles driven in different operating conditions, the accuracy of the updates sent by AV instances to the vendor servers, and the accuracy of the aggregation of the data by the vendor servers shared with all AV instances. Likewise, failures to detect "out-of-ODD" may lead to incorrect attributions of miles and accidents which are outside ODD to some of the operating conditions[14]. Such failures may affect the predictions but are outside the scope of this paper. We intend to address this concern in our future work.

DyAVSA is clearly dependent on the necessary communication infrastructure such as a suitable mechanism for sending observations from each AV instance to a centralized collector (e.g., a server operated by the AV vendor) where the data is aggregated as necessary (including addressing the privacy concerns) and suitable mechanisms for sharing the aggregated data among the deployed AV instances. We acknowledge the importance of the implementation details, e.g., the aggregated data should account for the needs of each AV instance. Some instances will require frequent updates of aggregated data, while other AV instances will be used less intensively and therefore may require less frequent access to the aggregated data. Clearly, the data sharing mechanism should account for the needs of all AVs, especially if the aggregated counts (of miles and accidents) only are shared. Sharing the distributions of the conditional pfm_i seems easier to implement, as it will only require the aggregator to send the current snapshot of the distributions of the conditional pfm_i.

A related concern is whether the current communication technologies are good enough to allow a scalable DyAVSA deployment with large fleets of AVs. Our preliminary calculations indicate that the current communication technologies used in modern vehicles (e.g., 4G and 5G mobile networks) provide plenty of bandwidth to allow scalable deployment of DyAVSA with millions of AVs. We envisage that each AV will require infrequent updates, say no more than once a minute, and will likely require an exchange of a few kilobytes of data with the vendors' servers in each direction, as illustrated in **Fig. 2**. The current cloud computing technologies have a typical communication bandwidth of terabytes per second, which should easily meet DyAVSA requirements. Via load balancing the vendor servers can easily handle millions of connections with individual AV instances. Technologies which require continuous communications between modern vehicles and a centralized service exist, e.g., the Mobileye REM (Road Experience Management)[15], which offers "crowdsourced, continuously updated map of the world". REM uses a two-way communication between Mobileye terminals installed on individual AVs and the company's cloud service. It seems that REM's requirements for communication bandwidth exceed significantly the requirements by DyAVSA.

Among the threats to the validity of our results we would like to acknowledge the following:

- The proposed method of multivariate Bayesian inference relies on several

[14] A related problem, that may affect the inference, is reliability of data communication, including failures due to malicious activities. We assume that sufficiently reliable and secure data communication will be used making negligible the adverse effect of data miscommunication.

[15] https://www.mobileye.com/technology/rem/

assumptions:

- the conditional *pfm*s of driving in different operating conditions are assumed *independently distributed random variables*. This assumption seems plausible but may in fact turn out to be difficult to justify. The problem is not new and has been discussed in the past, e.g., in [18]. Conceivably, a failure may be traced to a fault, which can be triggered in more than one operating condition, thus promoting the idea that beliefs about conditional *pfm*s in operating conditions should be dependent. Technically, the independence assumption can be relaxed, e.g., by using suitably chosen *Copulas*[16] *[19]* to capture the dependencies between the random variables (in this case - the conditional *pfm*s in different operating conditions). In Appendix 2 a detailed discussion is provided about applying a Gaussian Copula to capture possible dependencies among the distributions of the conditional *pfm*s. We also illustrate the implications of dependency among the distributions of the conditional *pfm* for the distribution of a weighted sum of the dependent variates aligned with an operational profile captured by a Dirichlet distribution. 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 envisage two important aspects of this future work: i) is it plausible to assume that the dependencies captured by a Copula will remain unaffected over time. Such a view would be consistent with the spirit of Copulas – a Copula functional can be applied to any marginal distributions (in our case – to prior and posterior distribution of *pfm*s.), and ii) how can one elicit the parameters of a Copula applied to marginal distributions which represent *epistemic uncertainty*, and more importantly, for which the dependence may be difficult to "measure". Contrary to typical applications of Copulas, e.g., in finance to represent dependencies between risks of different stocks, which are directly observable and measurable, in our case Copula will capture dependences among epistemic uncertainties, which are difficult to capture. On the one hand, there is a clear intuition behind dependence, e.g., changes of driving policy may affect the true *pfm*s in several operating conditions (i.e., there is a "common factor" affecting several *pfm*s). Whether this implies that one should opt for modelling the dependence between the respective marginal distributions (i.e., the epistemic uncertainties) or just let the Bayesian inference eventually update the marginal uncertainties is unclear. Further detailed analysis is needed to understand the phenomenon (of "common factors") and what the best way of modelling it is. We also envisage that adding dependence among the distributions of the *pfm*s may be a way of introducing a degree of *conservatism* in Bayesian predictions. As our illustrations in Appendix 2 show assuming significant degree of positive correlation between the distributions of the *pfm*s leads to an increase of the tail of the measure of interest – the distribution of system probability of accident per mile (i.e., the weighted sum of *pfm*s). If adding dependence will be a way of introducing conservatism in predictions, it may be

[16] Copulas are a specific way of modelling the dependence between random variables. The interested reader may check https://en.wikipedia.org/wiki/Copula_(probability_theory) for further details

useful to leave the decision about the *degree of conservatism* to the stakeholders – the AV vendors or individual AV owners/users. Again, it seems that further extensive research and analysis is needed before DyAVAS adopts a model of dependence among the marginal distributions of *pfm*s. Without such research adopting dependence among the marginal distributions of *pfm*s will in our opinion be premature.

- Bayesian inference is undertaken under the assumption of reliable recording of the counts of miles driven, and accidents observed in different operation conditions. Clearly, there may be errors due to various factors, e.g. misclassification of OCs, or due to failures to detect "out-of-ODD", which in turn may lead to attributing accidents (and miles without accidents) that occur "out-of-ODD" to some of the operational OCs defined in an ODD. Conceptually, accounting for these possibilities is straightforward – one needs to allow for misclassifications of OCs. This concern is conceptually similar to the following two concerns: i) "oracle perfection" in software testing and its impact on software reliability assessed via software testing [20], and ii) the checker coverage in asymmetric architectures such as "primary - checker" (e.g. an AV safety monitor) and its impact of reliability assessment of the asymmetric architecture, a concern which has been studied in the past, including in own work [21]. We intend to address these concerns in our future work, too.
- We assume that the AV operational profile is adequately captured by a Dirichlet distribution. Although this type of multivariate distribution has been used by many[17] in the past and, more importantly, seems quite plausible for the task, 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. A similar approach has been taken in our recent work [22].
- In this work we relied on the prior work by others [2], whereby the key parameter of interest is the *pfm* of driving and on the critical assumption that success/failure of driving a set of randomly chosen miles can be modelled as a Bernoulli trial. Clearly the successive miles of driving may not be quite like a Bernoulli trial, although the recent work [23] provides a rationale suggesting that the implications of the assumption for the mathematical rigor are insignificant. An alternative approach to modelling AV driving would be to consider the *duration* (in miles) in the same operating condition and see the AV driving as a trajectory via different *OC*s defined by an ODD. We took this approach in a recent study [22, 24]. Such a model may reveal a different insight. Developing this alternative approach in detail is also an area for future research.
- Finally, a separate strand of research deals with the observations from "microscopic traffic simulation" tools, e.g. [25], which differs conceptually from our work as it seems to rely on a different safety measure. Adopting a different safety measure may reveal a different insight, too, an area for future research.

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

6 Related research

An idea somewhat related to DyAVSA are the Safety Performance Indicators (SPI) [9], the authors of which argue that SPIs must be quantitative, and their assessment should be done by collecting "operational data", i.e., data from the deployed AVs. The main advantage of SPIs would be that they can provide "early warnings" of possibly insufficient safety and thus would allow the AV vendor to improve over time the safety of the AV brand. For instance, if an SPI is related to failures to detect a pedestrian on the road, the AV vendors may act upon such data without having to wait for an actual accident. The key difference between SPIs and DyAVSA, apart from the purpose and the scope of data collection, is that DyAVSA allows different stakeholders, including the owners/drivers of individual AVs to benefit from the operational data collected by a fleet of AVs and the fact that DyAVSA provides a *complete computational procedure* to make use of the collected data in targeted run-time safety assessment while with the SPIs the AV vendor is left to decide how to make use of them.

An alternative approach to Bayesian assessment is offered in [26]. This work offers a hierarchical Bayesian inference and builds on a previous publication relying on the use of *extreme values theory* for safety assessment of AV [27].

An important work is [28], which provides a theoretical Bayesian hierarchical extreme value model integrating several conflict indicators such as the modified time to collision (MTTC), the post encroachment time (PET), and the deceleration rate to avoid a crash (DRAC) and demonstrates that the multivariate model outperforms the respective univariate and bi-variate models which use fewer measures of interest. The paper builds on a multivariate Bayesian inference and in this sense is similar to the approach developed in [13] and used in the presented paper. There are, however, significant differences, too. The variates used in [28] are "conflict indicators", while in own work the conditional *pfm*s of driving in different operating conditions are used. While in own work (following the development in [13]) we assume that the variates are independently distributed random variables, the model developed in [28] treats the indicators as stochastically dependent random variables. There is also similarity between the conclusions reported in [28] and [13] – in both cases the multivariate models are said to be superior in terms of the prediction accuracy to their univariate counterparts (and a bivariate prediction model referred to in [28]).

A separate strand of related work deals with the "surrogate safety measures" (SSM). As the name suggests SSM are looking for useful ways of evaluating road safety in the *absence of accident data*, which are typically rare. SSM are used to address a variety of use cases, among them the impact of AV and connected AV (CAV) [25] on road safety. A recent survey summarizing the advances with SSM is provided in [29]. SSM focus is very different from the focus of the work presented here. The main premise in SSM related studies is that the scarcity of accident data makes safety assessment difficult. The approach taken in this paper is radically different. The multivariate Bayesian inference, on which DyAVSA is based, can take any observations, including *no accidents at all*, to derive predictions about the chosen measure of AV safety. The scope of SSM is also different from the scope of the current work: SSM are used to assess how a large fleet of AV/CAV on the public roads in the future can affect the road safety of

all participants in the road traffic. The scope of our work is entirely focused on the benefits that sharing observations (miles driven and accidents observed or lack thereof) among the AVs of a fleet of AVs can bring to the owners/users of the individuals AVs.

There is a conceptual similarity between aforementioned Mobileye Road Experience Management (REM) and DyAVSA although the focus of data sharing is different. REM is focused on constructing an up-to-date *road map* using the data coming from AV instances and sent to a cloud-service where the data is aggregated, and maps are constructed. Part of the constructed road map is then shared with the AVs based on their current location. DyAVSA instead collects and shares data useful for predicting the conditional *pfm*s and foreseeing violations of a safety claim.

Finally, DyAVSA is based on a multivariate Bayesian assessment, a topic extensively developed by many over the years, other methods based on Bayesian reasoning have been widely used in risk assessment of various systems. An authoritative text on Bayesian risk assessment is [30], which provides a foundational introduction to Bayesian Belief Networks (BBN) and contrast them with alternative formalisms of dealing with uncertainty in risk assessment such as statistics and causal reasoning. A very extensive literature exists on the use of BBN for risk assessment in safety critical systems. The seminal work by Littlewood and Strigini [31] on demonstrating infeasibility of demonstrating ultra-high software reliability via testing. A similar conclusion was reached by Butler and Finelli [30] at approximately the same time.

Among the examples of using BBN in other related domains, e.g. in maritime operations, we would like to acknowledge the contributions of Zaili Yang: [32] on BBN based risk assessment of seaports, and [33] on BBN - based risk assessment of the operations of maritime autonomous surface ships. In [34] Jingbo Yin applies a BBN risk assessment to cargo operations at seaports.

We already acknowledged the value of Copula in capturing the dependencies between the distributions of random variables. A few examples are included in Appendix 2. We would like also to acknowledge other examples relevant to the multivariate Bayesian inference DyAVSA is based. [35] offers a generic Bayesian hierarchical Copula model. In comparison with the alternative models used for Bayesian inference using layers of a hierarchy, the authors claim that their approach provides increased flexibility and allows Copulas (e.g., Archimedean and Gaussian) to be adapted as required by the specific context. Although [35] is focused on dealing with clusters of data sources, there is conceptual *similarity* with the multivariate inference used in DyAVSA: the model used in DyAVSA can be seen as a 2-layered model: the "upper" layer deals with the operational profile uncertainty, while the "lower" layer deals with the uncertainty of the *pfm*s.

A recent reprint [36] deals with a problem which is highly relevant to a possible extension of DyAVSA and the multivariate inference it relies upon. The author of the reprint reports on the effect of using Copula on the distribution of a sum of non-independent random variables. Similarly to our own observations presented in Appendix 2, the author concludes that the impact of dependence among the margins of a Copula may impact significantly (in excess of 10%) the high quantiles (i.e., the tail) of the distribution of the sum.

7 Conclusions and future work

This paper proposes the Dynamic AV Safety Assessment (DyAVSA), an approach to run-time AV safety assessment of a particular AV product, whereby the data on miles driven and failures/accidents observed, are collected by individual AV instances, and passed to a centralized AV Vendor Server for safety monitoring. Under DyAVSA the individual AV instances can monitor their own safety using the operational data collected from the entire fleet of AVs of the same brand collected by the AV vendor.
We demonstrate that a two-stage Bayesian inference procedure, we developed recently [12] and adapted to the needs of AV safety assessment [13], can serve the run-time safety assessment needs of different stakeholders: i) the AV vendors, can collect data from the entire deployed AV fleet of AVs and assess the safety of the fleet "on average"; ii) should the vendor periodically share the aggregated measurements with all deployed AVs instances, then the users/owners of individual AV will be able to monitor the safety of their own AVs themselves relying on the data records of accidents/successes for a given ODD aggregated for the entire fleet of AVs and using their unique operational profile. We illustrated the advantages of the proposed method over the alternatives – relying on the predictions "on average" made by the vendor, or on the predictions by the AV instances relying on their own operational data only. To the best of our knowledge a concept similar to DyAVSA *has not* been used before.

We already identified in the previous section a few areas for future development to address some of the recognized threats to validity of our work and findings, among them relaxing the assumptions on which the multivariate inference is based upon – that the miles of AV driving can be modelled as a Bernoulli trial and that the conditional probabilities of accident per mile of driving in different driving conditions are independently distributed random variable.

Developing a highly efficient computation procedure which would allow for fast multivariate inference aligned with the needs of run-time safety assessment is another area of research which we intend to address in the future.

In passing we mentioned that a safety claim, linked implicitly to a given ODD should address the following concerns affecting the AV safety: i) detecting reliably "out-of-ODD (OoODD)", and ii) accounting the impact on AV safety of unreliability of responding to OoODD. Both detecting OoODD itself and the implemented response to a detected OoODD, may be subject to failure. The impact of both failures should be accounted for in a complete safety analysis [6], and we intend to address this problem in our future research.

Acknowledgement

This work has been partially supported by the Intel Collaborative Research Institute on safety of autonomous vehicles (ICRI-SAVe).

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. Wachenfeld, W. and H. Winner, *Die Freigabe des autonomen Fahrens*, in *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, M. Maurer, et al., Editors. 2015, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 439-464.
4. 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.
5. 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.
6. Standards&Engagement, *ANSI/UL 4600: Evaluation of Autonomous Products*. 2023, ANSI.
7. IAEA, *Periodic Safety Review for Nuclear Power Plants*, in *Specific Safety Guide No. SSG-25*. 2013, International Atomic Energy Agency: Viena. p. 128.
8. EASA, *Annual Safety Review 2023*. 2023: European Union Aviation Safety Agency.
9. Johansson, R. and P. Koopman, *Continuous Learning Approach to Safety Engineering.* , in *CARS - Critical Automotive applications: Robustness & Safety*. 2022, HAL: Zaragoza, Spain. p. 5.
10. NHTSA *SGO Incident Reports ADS January 202*. Standing General Order on Crash Reporting, 2021.
11. British Standards Institute, B.S.I., *PAS 1883:2020 Operational Design Domain (ODD) taxonomy for an automated driving system (ADS) - Specification*. 2020, BSI Standards Limited: London. UK. p. 26.
12. 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.
13. Popov, P., *Why Black-Box Bayesian Safety Assessment of Autonomous Vehicles is Problematic and What Can be Done About it?* IEEE Transactions on Intelligent Vehicles, 2024. **(under review)**: p. 13.
14. NAIIC, *The Fukushima Nuclear Accident Independent Investigation Commission*, in *The National Diet of Japan* 2012. p. 82.
15. Zhao, X., et al., *Assessing safety-critical systems from operational testing: A study on autonomous vehicles.* Information and Software Technology, 2020. **128**: p. 106393.
16. Adams, T., *Total Variance Approach to Software Reliability Estimation.* IEEE Transactions on Software Engineering, 1996. **22**(9): p. 687-688.
17. Albert, I. and J.-B. Denis *Dirichlet and multinomial distributions: properties and uses in Jags*. Unité Mathématiques et Informatique Appliquées, 2012. 28.
18. Klotz, J., *Statistical Inference in Bernoulli Trials with Dependence* The Annals of Statistics, 1973. **1**(2): p. 373–379.
19. Nelsen, R.B., *An Introduction to Copulas*. Springer Series in Statistics. 2006: Springer New York, NY. 272.
20. Littlewood, B. and D. Wright, *A Bayesian Model that combines disparate evidence for the quantitative assessment of system dependability", In - Mathematics of Dependable Systems, II, (V Stavridou, Eds.), pp. 243-258, Clarendon Press, Oxford, 1997.* 1997.

21. Popov, P. and L. Strigini, *Assessing Asymmetric Fault-Tolerant Software*, in *IEEE 21st International Symposium on Software Reliability Engineering*. 2010, IEEE: San Jose, CA, USA, . p. 41-50.
22. Buerkle, C., et al., *Road Hazards on Road Intersections and Stochastic Modelling of their Effect on Safety of Autonomous Vehicles*, U.o.L. City, Editor. 2023 (under review). p. 18.
23. 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.
24. 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.
25. Papadoulis, A., M. Quddus, and M. Imprialou, *Evaluating the safety impact of connected and autonomous vehicles on motorways.* Accident Analysis & Prevention, 2019. **124**: p. 12-22.
26. Kamel, A., T. Sayed, and C. Fu, *Real-time safety analysis using autonomous vehicle data: a Bayesian hierarchical extreme value model.* Transportmetrica B: Transport Dynamics, 2022. **11**(1): p. 826-846.
27. Reyad, P., et al., *Real-Time Crash-Risk Optimization at Signalized Intersections.* Transportation Research Record, 2022. **2676**(12): p. 32-50.
28. Fu, C., T. Sayed, and L. Zheng, *Multivariate Bayesian hierarchical modeling of the non-stationary traffic conflict extremes for crash estimation.* Analytic Methods in Accident Research, 2020. **28**: p. 100135.
29. Wang, C., et al., *A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling.* Accident Analysis & Prevention, 2021. **157**: p. 106157.
30. Fenton, N. and M. Neil, *Risk Assessment and Decision Analysis with Bayesian Networks (2nd ed.)* 2018, NY: Chapman and Hall/CRC.
31. Littlewood, B. and L. Strigini, *Validating Ultra-High Dependability for Software-Based Systems*, in *PDCS 2nd year Report*. 1991, PDCS.
32. Yang, Z., S. Bonsall, and J. Wang, *Fuzzy Rule-Based Bayesian Reasoning Approach for Prioritization of Failures in FMEA.* IEEE Transactions on Reliability, 2008. **57**(3): p. 517-528.
33. Chang, C.-H., et al., *Risk assessment of the operations of maritime autonomous surface ships.* Reliability Engineering & System Safety, 2021. **207**: p. 107324.
34. Khan, R.U., et al., *Seaport infrastructure risk assessment for hazardous cargo operations using Bayesian networks.* Marine Pollution Bulletin, 2024. **208**: p. 116966.
35. Zhuang, H., L. Diao, and G.Y. Yi, *A Bayesian hierarchical copula model.* Electronic Journal of Statistics, 2020. **14**(2): p. 4457-4488.
36. Schneider, W. *On the distribution of the sum of dependent standard normally distributed random variables using copulas*. 2021. 10 DOI: https://doi.org/10.48550/arXiv.2107.00007.
37. Sklar, M.J. *Fonctions de repartition a n dimensions et leurs marges*. 1959.
38. Gijbels, I. and K. Herrmann, *On the distribution of sums of random variables with copula-induced dependence.* Insurance: Mathematics and Economics, 2014. **59**: p. 27-44.
39. Elfadaly, F.G. and P.H. Garthwaite, *Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models.* Statistics and Computing, 2017. **27**(2): p. 449-467.

9 Appendix 1: Multivariate Bayesian inference

A multivariate Bayesian assessment applicable to systems with multiple operating conditions, e.g., demand space partitions, partition testing, autonomous vehicles used in different operating conditions OC_1, OC_2, …, OC_n, as defined by an operational design domain (ODD), etc., can use a *double-stochastic multivariate model*, developed

recently [12] and adapted to the needs of a AV used with a defined ODD [13]. The model and the Bayesian inference procedure included in this appendix are derived from [13].
The model captures:

- A partition of operating conditions OC = {OC_1, OC_2, …, OC_m} such that iff i ≠ j then $OC_i \cap OC_j = \emptyset$ defined with a probabilistic measure on OC, $P(OC_i)$, and
- In each OC_i the AV drives a sequence of miles. Each mile may be either successfully completed or lead to an accident. We model the driving in OC_i as *Bernoulli* process with a parameter pfm_i, which is treated as a random variable, Θ_i. A joint distribution $f_{\theta_1,\theta_2,\ldots,\theta_n}(\theta_1, \theta_2, \ldots, \theta_n)$, characterizes the uncertainty in the values of pfm_i in different OC_i and the stochastic dependencies between the variates, $\Theta_1, \Theta_2, \ldots, \Theta_n$, of the multivariate distribution, $f_{\theta_1,\theta_2,\ldots,\theta_n}(\theta_1, \theta_2, \ldots, \theta_n)$.

  To simplify the analysis, we make an *additional simplifying* assumption that $\Theta_1, \Theta_2, \ldots, \Theta_n$ are *independently distributed* random variables. $f_{\theta_i}(\cdot)$ denotes the probability density function of Θ_i, for i = 1,…, n. In other words, we assume that changes in $f_{\theta_i}(\cdot)$ do not affect $f_{\theta_j}(\cdot), i \neq j$.

$P(OC_i)$ may vary over time or be subject to epistemic uncertainty, which we capture by using a random variable, Ψ_i with a probability density function $f_{\psi_i}(\cdot)$. Since the operating conditions form a *partition* of the space of miles, the constraint $\sum_{i=1}^{n} \Psi_i = 1$ applies: a mile with certainty will be selected from one of the partitions.

The joint distribution $f_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_n)$ captures the epistemic uncertainty associated with the selection of a mile from the space of all miles. A suitable analytic multivariate distribution which can be adopted here is the Dirichlet distribution, which for n variates, $\Psi_i \ldots, \Psi_n$ is defined as [17]:

$$\begin{aligned} \mathrm{Dir}(\psi_1, \psi_2,\ldots, \psi_n;\boldsymbol{\alpha}) &\equiv f_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_n;a_1, \ldots, a_n) \\ &= \frac{\Gamma(\sum_{i=1}^{n} a_i)}{\prod_{i=1}^{n} \Gamma(a_i)} \left[\prod_{i=1}^{n-1} \psi_i^{a_i-1}\right]\left[1-\sum_{i=1}^{n-1} \psi_i\right]^{a_n-1} \end{aligned} \quad \text{(A1)}$$

where $\boldsymbol{\alpha}$ is a vector $a_1, \ldots, a_n$ and defines the parameters of the Dirichlet distribution. The sum of the variates $\sum_1^n \Psi_i = 1$.

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

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

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

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

The marginal distribution of each variate, Ψ_i, is a Beta distribution, Beta(ψ;a_i, A-a_i), [17].

Now, let us consider the case of an ODD known with certainty, i.e., $P(OC_1) = \psi_1, P(OC_2) = \psi_2, \ldots P(OC_n) = \psi_n$, where ψ_i $(i = 1, \ldots n)$ are known constants. The random variable Θ, which represents *pfm*, is then the weighted sum of the random variables Θi, weights being the probabilities $\psi_1, \psi_2, \ldots, \psi_n$, respectively.

$$\Theta_{\psi_1,\psi_2,\ldots,\psi_n} = \sum_{i=1}^{n} \Theta_i \psi_i \quad \text{(A2)}$$

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 themselves independently distributed random variable. Let us denote the probability density function 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:

$$f_{\theta\psi_i}(x) = \frac{1}{|\psi_i|} f_{\theta_i}\left(\frac{x}{\psi_i}\right) \tag{A3}$$

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

$$f_{\theta|\psi_1,\psi_2,\ldots\psi_n}(x|\Psi_1 = \psi_1, \Psi_2 = \psi_2, \ldots, \Psi_n = \psi_n) = f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \ldots * f_{\theta\psi_n}(x) \tag{A4}$$

where the "*" sign indicates a convolution of the respective probability density functions.

Finally, we can now remove the condition that the operational profile is known with certainty (captured by $\Psi_1 = \psi_1, \Psi_2 = \psi_2, \ldots, \Psi_n = \psi_n$) using the joint distribution defined by (A1):

$$f_\theta^{WB}(x) \int f_{\theta|\psi_1,\psi_2,\ldots\psi_n}(x|\psi_1, \psi_2, \ldots, \psi_n) \mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n};\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n}) d\psi_1 \psi x_2 \ldots d\psi_n$$
$$= \int \left[f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \ldots * f_{\theta\psi_n}(x) \times \mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n})\right] d\psi_1 d\psi_2 \ldots d\psi_n \tag{A5}$$

The integration in the last expression (A5) is done with respect to all dimensions $\psi_1, \psi_2 \ldots \psi_n$ of the ODD. One can see that (A5) provides us with the marginal distribution of system *pfm* and accounts for the epistemic uncertainty of both the operational profile – this is captured by the joint distribution $\mathrm{f}_{\psi_1,\psi_2,\ldots\psi_n}(\psi_1, \psi_2,\ldots, \psi_\mathrm{n})$ – and the conditional probabilities of catastrophic failure in partitions $f_{\theta_i}(x)$. Clearly, the latter will affect the convolution, $f_{\theta\psi_1}(x) * f_{\theta\psi_2}(x) * \ldots * f_{\theta\psi_n}(x)$, representing the distribution of the sum $\Theta_{\psi_1,\psi_2,\ldots,\psi_n}$ expressed by (A4).

We labelled (A5) with "WB" to signify the fact that this distribution is derived using a "white box" model of both the ODD and how likely the AV is to fail in each of the operating conditions.

The marginal distribution of *system pfm*, $f_\theta^{WB}(x)$, can be used in different ways. 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), by looking at the tail of the distribution of system *pfm*:

$$P(\Theta \geq T) = \int_T^1 f_\theta^{WB}(x) dx \tag{A6}$$

10 Appendix 2: Using Copula to model dependencies between *pfm*s

In this section we illustrate the impact of relaxing the assumption that the conditional probabilities of accident per mile of driving (*pfm*s) in different operating conditions are independently distributed random variables. This is done by adopting a Copula functional to define a structure of dependence between the *pfm*s (the marginals). We use the resulting Copula functional to assess the impact of dependence on the distributions of a weighted sum of the uncertain *pfm*s by comparing the distribution of the sum

assuming the *pfm*s independently and non-independently distributed random variables, respectively.

Definition: A function $C: [0,1]^d \rightarrow [0,1]$ is called d-copula (or short copula), if C is the distribution function of a d-dimensional random vector $\boldsymbol{U} = (U_1, \dots, U_d)$ with standard uniform marginals, i.e. $\mathbb{P}[U_k \leq u_k] = u_k$ for all $k \in \{1, \dots, d\}$ and $u_k \in [0,1]$.

According to the Sklar's theorem [37] if H is a d-dimensional distribution function with "margins", $F_1, \dots, F_d$, then there exists a d-copula C such that for all $x \in \overline{\mathbb{R}}^d$ $H(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big)$.

In our formulation of AV safety assessment problem, $F_1(x_1), \dots, F_d(x_d)$ would be the marginal probability distributions ("margins") of the uncertain conditional *pfm*s, $\Theta_1, \Theta_2, \dots, \Theta_n$, of AV driving in operating conditions OC_1, …, OC_d, and $H(x_1, \dots, x_d)$ – would be the joint distribution of the margins $\Theta_1, \Theta_2, \dots, \Theta_n$, in the general case of their being non-independently distributed. The Sklar theorem further asserts that if the margins $F_1(x_1), \dots, F_d(x_d)$ are all continuous, then C is unique.

Copulas can take different form (https://en.wikipedia.org/wiki/Copula_(statistics)). In this appendix we use a *Gaussian*[18] Copula to illustrate the impact of the strength of dependence among the margins $\Theta_1, \Theta_2, \dots, \Theta_n$ on their weighted sum. In other words, we look at the distribution of the sum $S = w_1\Theta_1 + w_2\Theta_2 + \cdots + w_d\Theta_d$, where $w_i, i \in [1, \dots d]$, represent the weights of the random variables $\Theta_i, i \in [1, \dots d]$, respectively.

We also look at the impact of the uncertainty in the weighting coefficients, W_i by assuming them random variables; following the assumption made in Appendix 1 that a Dirichlet distribution is used to capture their joint distribution.

10.1 Effect of dependence among margins on distribution of their sum

In this appendix we take the contrived examples developed in the paper with 5 operating conditions, OC_1 – OC_5, with Beta distributed $\Theta_1, \Theta_2, \dots, \Theta_5$ and use a Gaussian Copula with an increasing coefficient of correlation, ρ_{ij} between all pairs of margins, $(\Theta_1, \Theta_2), \dots, (\Theta_4, \Theta_5)$. The illustration is done under the assumption that the *same level of correlation* applies to all pairs of margins, i.e., if the correlation is set to 0.1, then the following covariance matrix is used to define ρ_{ij}, used in generating a copula[19]:

ρ_{ij} = [0.1,0.0,0.0,0.0,0.0;
0.0, 0.1,0.0,0.0,0.0;
0.0,0.0, 0.1,0.0,0.0;
0.0,0.0,0.0, 0.1,0.0;
0.0,0.0,0.0,0.0, 0.1];

The joint distribution of dependent $\Theta_1, \Theta_2, \dots, \Theta_5$is defined for values of the correlation ρ_{ij} from the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. Although hypothetically

---

[18] The choice of Gaussian copula is motivated by its wide use in other studies, e.g. [38], [39], and the fact that the sole purpose of this appendix is an illustration of the magnitude of the difference between the distribution of the sum of margins computed under the assumption of independence and the presence of dependence.

[19] The specific format of the covariance matrix is as required in the MATLAB function mvnrnd() used in the calculations conducted in this appendix. The full MATLA script(s) used in the calculations can be found at: https://openaccess.city.ac.uk/id/eprint/35206/.

the correlation may be negative, there is no reason to expect negative correlation between the distributions of the conditional *pfm*s, hence negative correlation has been excluded from the analysis presented here. $\Theta_1, \Theta_2, \dots, \Theta_5$ were defined as Beta distributed with parameters as follows:

Θ_1: Beta(2, 299)
Θ_2: Beta(2, 800)
Θ_3: Beta(2, 1500)
Θ_4: Beta(2, 1000)
Θ_5: Beta(1, 400)

These are the parameter values of the prior distributions used in the contrived examples of the paper.

Once the joint distribution of $\Theta_1, \Theta_2, \dots, \Theta_5$ was defined for a given degree of correlation (0.1 – 0.9), as described above, a Monte Carlo simulation was used to generate samples of 200,000 random vectors $(x_1, x_2, \dots, x_5)$ from the respective joint distributions. Each vector was used to compute the weighted sum $w_1x_1 + w_2x_2 + \dots + w_5x_5$. The weights used in this illustration are set equal to the *expected values* of the variates used in the Dirichlet distribution of the contrived examples, computed using the formula:

$$w_i = \frac{a_i}{\sum_{i=1}^{5} a_i}$$

With the parameters of the prior Dirichlet $a_1 = 10, a_2 = 10, a_3 = 40, a_4 = 30, a_5 = 10$, the values of the respective weights are as follows: $w_1 = w_2 = w_5 = 10/100 = 0.1$, $w_3 = 40/100 = 0.4$ and $w_4 = 30/100 = 0.3$.

The parameters of the posterior Dirichlet distribution are different: $a_1 = 137, a_2 = 133, a_3 = 149, a_4 = 106, a_5 = 75$, which leads to $w_1 = 137/(137 + 133 + 149 + 106 + 75) = 137/600 = 0.228333333$. Similarly, $w_2 = 133/600 = 0.221666667$, $w_3 = 149/600 = 0.248333333$, $w_4 = 106/600 = 0.176666667$, and $w_5 = 75/600 = 0.125$.

The 200,000 instances of the weighted sum, derived for the sample of random vectors, were used to estimate the experimental distribution of the sum, under the increasing degrees of correlation (0.1 – 0.9). Similar calculations are conducted with both the prior and posterior distributions as used in the contrived examples in the paper.

With the increase of the correlation coefficient the variance of the sum $(\Theta_1 + \Theta_2 + \dots + \Theta_5)$ increases, too, as is shown in **Table 2** below, which is to be expected.

Table 2. Effect of correlation of the margins on the distribution of the sum (without weights) of the *prior* marginal distributions: Beta(2, 299), Beta(2, 800), Beta(2, 1500), Beta(2, 1000), Beta(1, 400)

Correlation	Variance
No correlation	6.343472e-07
0.1	8.338104e-07
0.2	1.037454e-06
0.3	1.252600e-06
0.4	1.454592e-06
0.5	1.674819e-06

0.6	1.909173e-06
0.7	2.128853e-06
0.8	2.373866e-06
0.9	2.591027e-06
Posterior marginal distribution Beta(2, 306), Beta(2, 809), Beta(2, 1545), Beta(2, 1030), Beta(1, 409)	
Correlation	Variance
No correlation	1.440540e-06
0.1	1.740738e-06
0.2	2.038486e-06
0.3	2.357392e-06
0.4	2.709229e-06
0.5	3.015632e-06
0.6	3.403947e-06
0.7	3.736954e-06
0.8	4.086776e-06
0.9	4.468203e-06

The variance of the weighted sum computed for the priors is slightly greater than the weighted sum computed for the posterior distributions. This is due to the impact of the change in the uncertainty about the variables $\Theta_1, \Theta_2, \ldots, \Theta_5$ represented by the marginals of the joint distribution and due to the changed operational profile.

The cumulative distribution functions (*cdf*s) computed with the parameters described above for prior and posterior distributions, respectively, are shown in **Fig. 7** below. The plots indicate a visible increase of the spread of the distributions moving from the prior to the posterior, which was already captured by the variances of the respective distributions reported in Table 2, above.

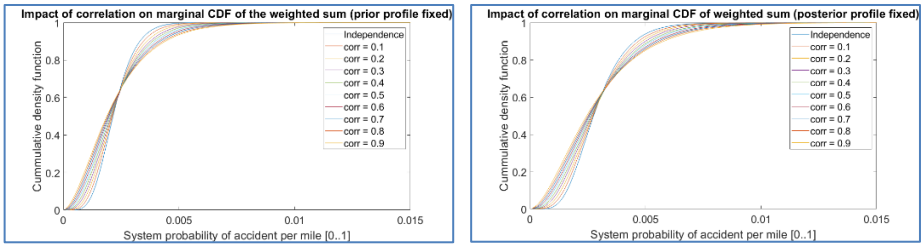

Fig. 7. Cdfs of the distributions of the weighted sum of the conditional probabilities of accident per mile assuming a fixed operational profile (weights used in the weighted sum computation are set to the expected values of the variates of the prior and posterior Dirichlet, respectively).

10.2 Combined effect of Copula and of the uncertain operational profile

In the final part of this Appendix, we compute the distribution of the weighted sum

of the margins assuming that the weights are now assumed *uncertain* and modelled as random variables with joint distribution represented by the prior and posterior Dirichlet distributions, respectively. The dependency between the conditional *pfm*s is captured by a Copula functional and we apply the same Copulas as in 10.1 to: i) the prior distributions of the conditional probabilities of accident per mile of driving, and ii) to the posterior distributions. The operational profiles, too, are the ones assumed in the prior Dirichlet(10, 10, 40, 10, 30) and Dirichlet(137, 133, 149, 106, 75) used in the contrived examples, respectively.

The calculations are done using the generated 200, 000 samples from the respective joint distributions (assuming independence of the margins and a correlation coefficient of between 0.1 and 0.9, as described above). Dirichlet distribution is applied using discretization with 20 values in each of its 5 dimensions, which in turn leads to 3876 distinct points of the Dirichlet distribution (which represent a vector of 5 values – each representing a specific value of the probability of the operating conditions OC_1 – OC_5). Each of these points would account (as a result of Dirichlet discretization) for a slice of the probability mass of the Dirichlet distribution. These probabilities will be used as weights of the margins in the sum.

The distribution of the weighted sum for a given set of weights is computed directly using the Monte Carlo simulated 200,000 instances sampled from the joint distribution with dependencies between the margins. This distribution is weighted with the mass associated with the vector of weights and added to the marginal distribution of the weighted sum. Once all 3876 vectors of weights (i.e., points of discretization of the Dirichlet distribution) are accounted for, the marginal distribution of the weighted sum of the margins consistent with the Dirichlet distribution will be derived.

For each of the snapshots – prior and posterior distributions – we show cdfs of the marginal distribution of the weighted sum of the Copula margins. We also computed the deviation of sums derived with Copulas from the case assuming "independence" between the margins (see Appendix 1).

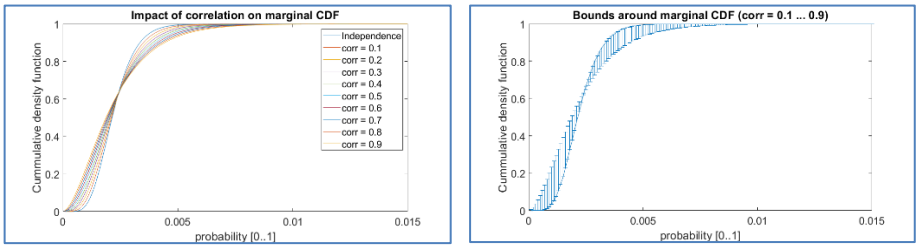

Fig. 8. Impact of dependency among the margins on the marginal probability of accidents per mile (a Gaussian Copula is applied to *prior* distributions). The figure on the left plots the cdfs, derived from the Monte Carlo generated sample of 200, 000 random vectors. of the system probability of accident per mile of driving for different degrees of dependence among the margins: starting with "independence" and increasing the correlation between 0.1 and 0.9. The figure on the right plots the magnitude of the "error" of assuming "independence" among the margins in comparison with the cases of positive correlation using a Gaussian Copula.

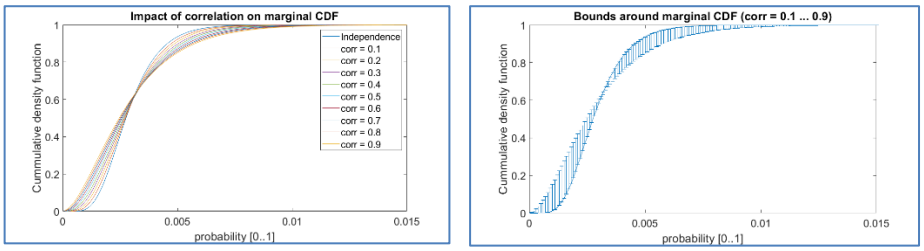

Fig. 9. Impact of dependency among margins on the marginal probability of accidents per mile (a Gaussian Copula applied to *posterior* distributions). The arrangement of the plots on the left and on the right is as in Fig. 7 above, but the Copula is applied to the posterior marginal *pfm*s.

Clearly, introducing dependency between the margins via a Gaussian Copula with the particular dependencies structure does not lead to stochastic ordering between the *cdfs* of the system probability of accident per mile of driving computed under "independence" and with any degree of correlation between the margins. **Fig. 8** and **Fig. 9** show a clear trend that the *cdfs* with correlation tend to be stochastically smaller at the beginning of the distributions (for values closer to 0 end of the distribution support) while their tails tend to become stochastically greater than the distribution under "independence".

It is also clear from **Fig. 8** and **Fig. 9** that with the increase of the correlation parameter p_{ij}, between the margins, the distribution of the weighted sum of the dependent margins deviates more strongly from the sum under "independence", and becomes quite noticeable for the highest value of p_{ij} = 0.9. For the much smallest values of p_{ij} (0.1 or 0.2), however, the deviation of the sums with "dependence" from the *sum* under "independence" is quite small.