

CITY UNIVERSITY LONDON

“Safety Assessment of Autonomous Vehicles: Approaches and Challenges

Peter Popov

City, St George’s, University of London, United Kingdom

EDCC’26: CARS Workshop

Keynote Talk

7th April 2026

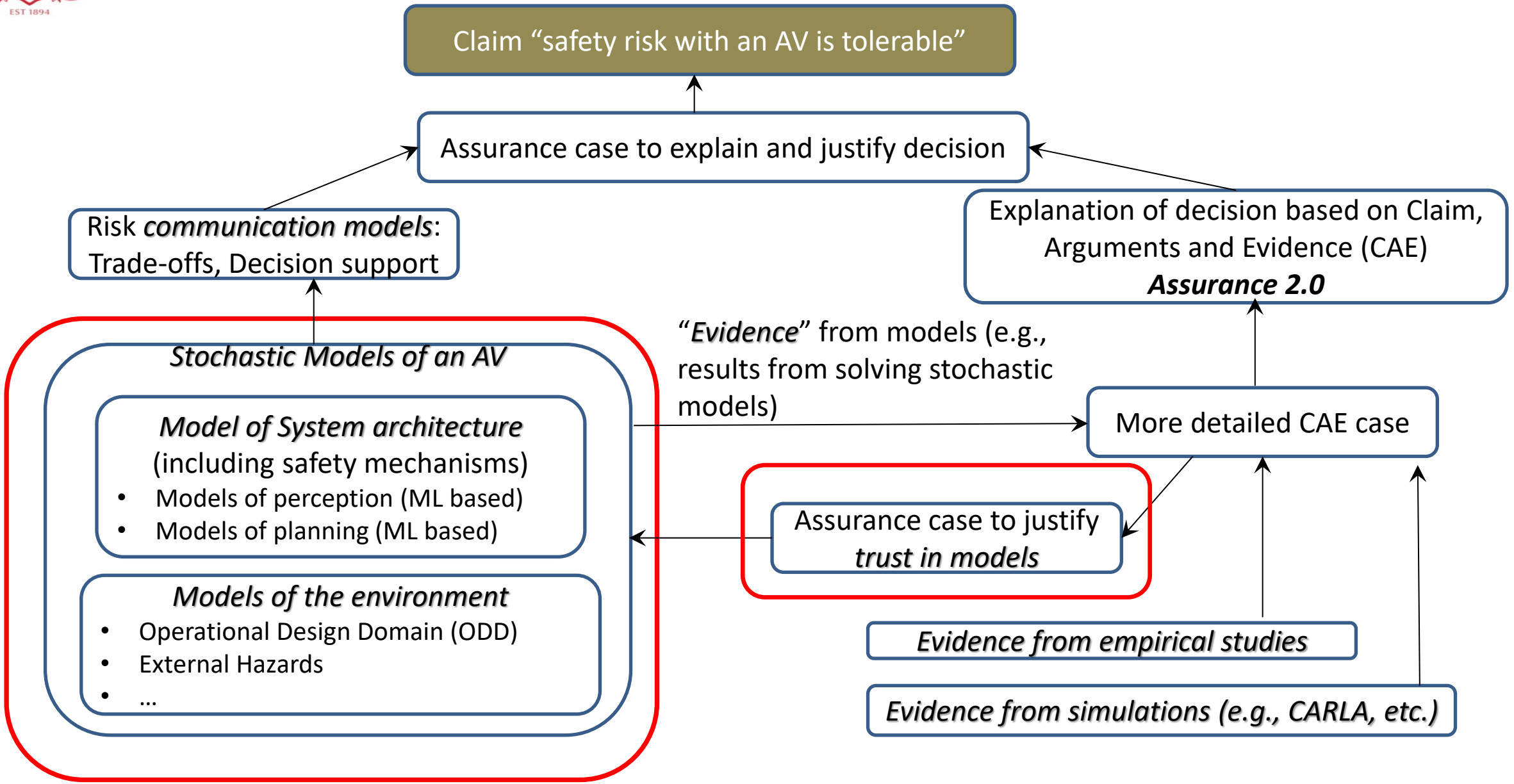
Canterbury, United Kingdom

CITY UNIVERSITY LONDON

Talk outline

- Probabilistic Safety Assessment/Assurance of Autonomous Vehicles (AV)
  - Driving to safety accounting for ***Operational Design Domain*** (ODD)
    - Univariate Bayesian Safety Assessment model
      - The RAND corporation model
      - "Conservative Bayesian inference (CBI)"
    - A multivariate Bayesian inference accounting for ODD
      - "on average) (i.e. fleet of AVs) vs "in particular" (i.e. a specific AV instance)
- Probabilistic AV safety assessment based on road hazards
  - Imperfection of AV ***perception*** and ***safety monitors*** and their effects on AV safety (work with Intel Labs)
  - A ***basic*** model of road AV safety in the presence of road hazards
  - A model with road ***Intersection –*** an example of safety assessment with ODD
- AV Safety Assessment Challenges (work in progress, collaboration is invited)
  - Stochastic framework for safety assessment of AV based for ***Scenario-based testing***
    - Assumptions ("difficult scenarios", ODD, trajectories, etc.)
    - Imperfection of simulators and their effect on AV safety assessment
    - Littlewood ***asymptotic result*** for semi-Markov models (developed for software, but relevant to AV/ODD)
    - Inference with a non-homogeneous Poisson process in the presence of ***epistemic*** uncertainty
    - Inference when safety is assessed by combining real driving, testing on "proving grounds" and with simulators
- Asymmetric system and their relevance to AV safety monitors
- Conclusions and future work

General approach to AV Safety assurance



CITY UNIVERSITY LONDON

General approach to Safety Assurance

Assurance case based on Claims-Arguments-Evidence framework:

- Top claim: "An AV in its Operational Design Domain (ODD) is safe enough"
- ***Assurance 2.0*** (Bloomfield and Rushby) gives more prominence in Safety Assurance to (probabilistic) models (see R. Bloomfield and J. Rushby *Assurance 2.0: A Manifesto. 2020, arXiv:2004.10474v2 [cs.SE] 9 June 2020.*).
  - Probabilistic models allow us to deal with various *sources of uncertainty* which affect AV safety.
  - *Aleatory uncertainty*
    - The nature of phenomena affecting safety in the real world is driven by ***random processes***. The events of interest can only be described ***probabilistically***.
    - ODD (difficult to specify how ODD would change over time), component failures, external hazards occur randomly.
  - *Epistemic uncertainty*
    - Our knowledge and understanding of the phenomena we are dealing with are *limited*.
    - We typically know only ***imprecisely*** (i.e., with a degree of doubt) the parameters of the random processes of interest. Can gradually ***learn them better with observations***, but knowing all model parameters with certainty is unlikely.
    - ***Known unknown*** (comprehensively describing all ODDs – seems infeasible). Even describing a single ODD is challenging:
      - » PAS 1883 and other international attempts to define ODD ***struggle with complexity***.
    - The "***unknown unknown***" (malicious actors, legislation, etc.) ***exacerbate difficulties***.

CITY UNIVERSITY LONDON

Motivation for the work

- Safety Assurance of AV is a challenging task
- Standards address some of the concerns:
  - ISO 21448 “Safety of the Intended Functionality (SOTIF)”
  - ISO 26262 Series of Standards (“Road Vehicles – Functional Safety”)
    - Does not deal with an AV as a system and only look at “E/E components” (i.e., its parts).
  - National Guidelines, e.g., in Germany, in the US, in the UK, etc.
  - ANSI UL4600 “**Evaluation of Autonomous Products**” (https://webstore.ansi.org/standards/ul/ul4600ed2023)
- Alternative approaches exist, too,
  - E.g., “Driving to safety”, proposed by researchers from RAND Corporation and further extended by colleagues at City
    - Fundamentally: makes Bayesian predictions from operational data that AV is “safe enough”
  - Mobileye’s concept of “Responsibility-Sensitive Safety (RSS)” (https://www.mobileye.com/technology/responsibility-sensitive-safety/) has attracted serious interest in industry and in academia.
    - Essentially, RSS defines a set of ***deterministic*** rules (using physics laws and geometry), which if followed by all participants in the road traffic should eliminate crashes (or at least allow for the “blame” for accidents to be decided more easily than is possible currently).

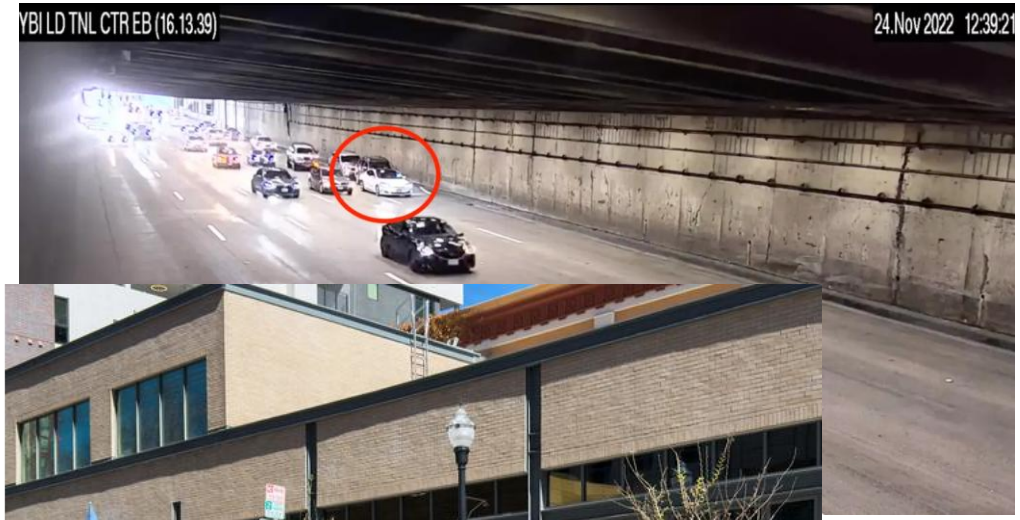
- The problem that is addressed in this talk is:

How can one demonstrate that an AV is safe enough to participate in regular traffic (i.e., to be certified as “safe” for use on the public roads)?

Is the problem of AV safety assurance real?

A few examples of AV failures

- Tesla vehicle crashes in a tunnel:
  - https://uk.video.search.yahoo.com/yhs/search?fr=yhs-tro-freshy&ei=UTF-8&hsimp=yhs-freshy&hspart=tro&p=tesla+collision+in+a+tunnel&type=Y219_F
    0e6eba0c9
- In the news:







AARIAN MARSHALL GEAR MAR 31, 2026 1:00 PM

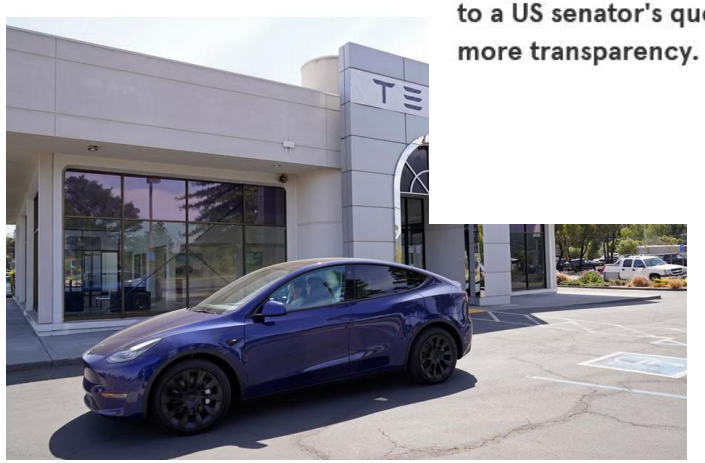
Tesla Admits Its Robotaxis Are Sometimes Driven by Remote Humans

The electric-car maker says it happens rarely and at speeds below 10 mph. But the disclosure—in response to a US senator's questions—occasioned a call for more transparency.

California places Tesla's Driving' under review

California's Department of Motor Vehicles is reviewing whether Tes advertising its vehicles as being fully autonomous without meeting

Via AP news wire • Tuesday 18 May 2021 01:02 BST





A Cruise car on the street in San Francisco

About 600 self-driving cars made by a unit of General Motors are being investigated by regulators after reports of injuries to pedestrians.



Statistics on crashes with AVs (source: NHTSA in US)

U.S. Department of Transportation
National Highway Traffic Safety Administration

NHTSA

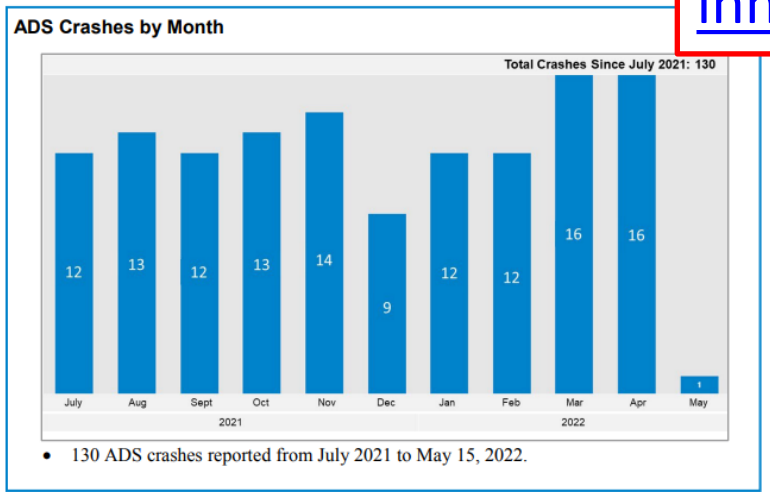
DOT HS 813 324 June 2022

- NHTSA (the National Highway Traffic Safety Administration) of the United States is responsible for collecting data and reporting on crashes of different vehicles (man driven and vehicles with different ***degrees of autonomy***).
  - SAE J3016 "Taxonomy and Definitions for Terms Related to Driving Automation S… 2021 defines the diff…
- Standing General Order…
  - The report from Jun… vehicles equipped w…
    - 130 accidents (c…

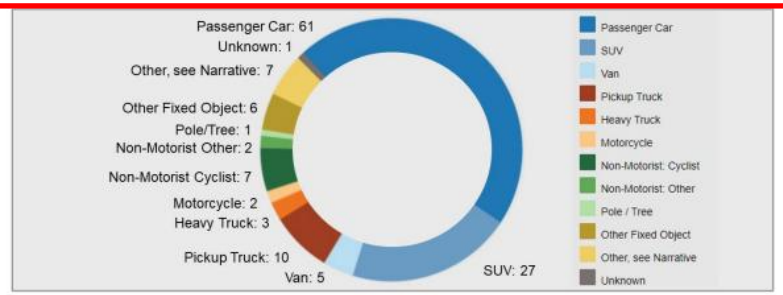
Summary Report: Standing General Order on Crash … mated

Vehicles with an automated driving system, which some refer to as "self-driving" cars, are a ***future technology*** – not a technology you're able to purchase and use today.

Source NHTSA: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

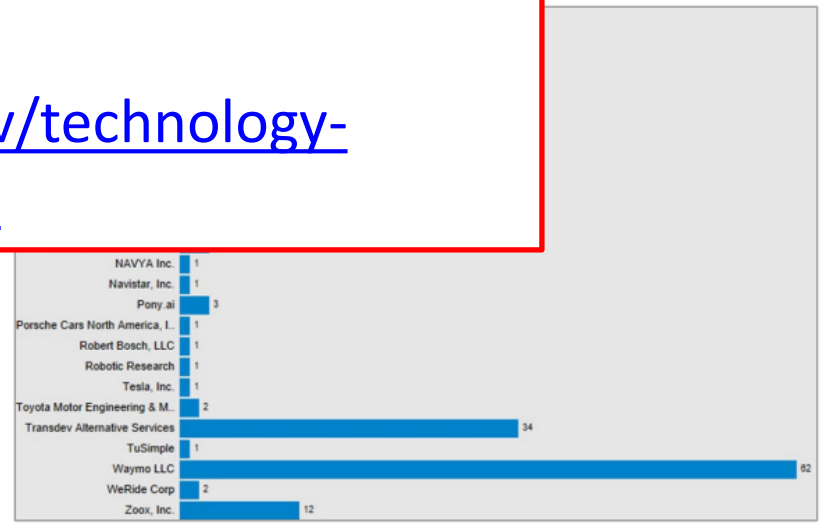


- 130 ADS crashes reported from July 2021 to May 15, 2022.



- 108 of ADS-involved crashes reported collisions with another vehicle.
- 11 crashes involved a vulnerable road user[2] (7 with cyclists, 2 with motorcycles, 2 *Non-Motorist: Other* crashes were with electric scooters).

Note: The sum of the results shown in this figure may be greater than the total number of crashes because multiple entities may report different information for the same crash.



- Waymo, Transdev, and Cruise reported the most ADS crashes.

Note: The sum of the results shown in this figure are greater than the total number of crashes because multiple entities can report the same crash.

Operational Design Domain (ODD) definitions

- PAS 1883 (a UK standard)
  - operating conditions under which a given driving automation system or feature thereof is specifically designed to function
  - *This includes, but is not limited to:*
    - *Scenery*
    - *Environmental conditions,*
    - *Dynamic elements (presence or absence of certain traffic or roadway characteristics).*

PAS defines a set of attributes which can/should be used to define ODD.

- ISO 34503:2023 Road Vehicles – Test Scenarios for automated driving systems – Specification for operational design domain
- PAS 1883:2025) Operational design domain – implementation of BSI ISO 34503:2023 - Guide

The ODD definitions vary by the region.

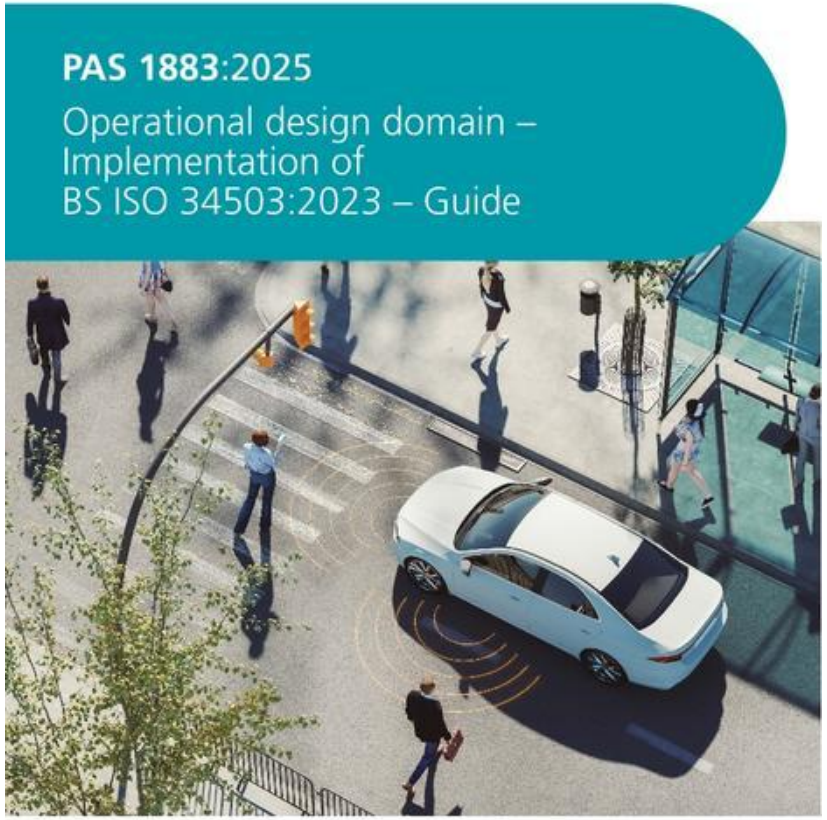
ODD is also defined in SAE J3016 (2018).

Similar definitions given in a number of other standards:

- ISO/TR 4804 (Safety and cyber-security for ADS), etc.

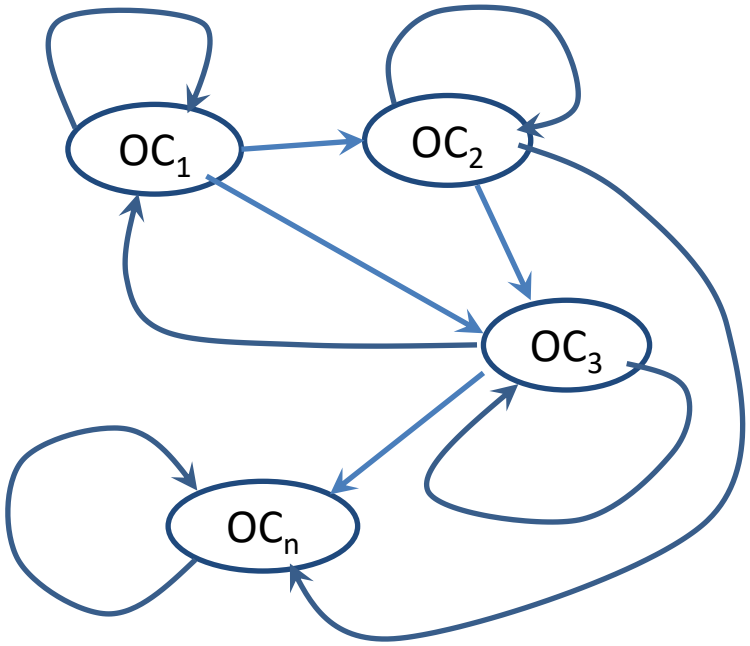
ODD fundamentally constrains the ***scope of a claim*** that an AV is safe enough.

A claim of safe operation in ALL conditions is ***impossible to justify*** (except for Level 5 ADS) Justifying a safety claim for a given ODD, however, is feasible.



Modelling ODD as a stochastic state-machine

- A particular ODD can be characterised by a *set of states of the operational environment*, which characterise the operating conditions (OC) of an AV (ADS).
- An AV trip then can be characterised by a ***trajectory*** via the states:
  - Sojourn time in an OC would be a random variable
  - Moving from a state (i.e., a particular OC) to another is a *random* process.
  - Different trips will be *instances* of this *random* process.
- The dynamic process of an AV traversing through OC can be characterised differently:
  - a Continuous Time Markov Chain
  - Semi-Markov process (sojourn time in states may have an arbitrary probability distribution),
  - Constructed from collected empirical data (datasets exist, which allow for such model construction, e.g., the ***highD*** dataset, referred to later in talk)
  - Etc.
- Some of the trips will fall within the ODD, some will violate the ODD (i.e. the Current Operational Domain may become a superset of ODD).
  - Detection of "out of ODD" reliably is itself a ***serious technical problem***.



CITY UNIVERSITY LONDON

EST 1894

Driving to Safety

CITY UNIVERSITY LONDON

Overview of the "Driving to Safety" Approach

- "Driving to safety" has been developed by researchers from the RAND corporation in the US:
  - N. Kalra 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.
    - A method, based on Bayesian inference, which allows one to ***compute the amount of AV driving (e.g., in miles)*** needed to reach levels of AV safety comparable with the safety of man-driven vehicles.
    - The required amount of driving is (prohibitively) high (10s of millions of miles without an accident).
      - This is seen as infeasible, hence motivated the need for alternative ***innovative methods*** for AV safety assessment.
  - The initial publication (above) triggered interest among many, including several colleagues at City, University of London, who developed a "***Conservative Bayesian Inference (CBI)***" method:
    - CBI spells out formally how one can compute conservatively, given the observations, the number of miles required for "driving to safety". Conservatism is stated mathematically precisely: CBI selected the "prior" used in the inference which guarantees conservative predictions of the confidence in a particular safety claim.

CITY UNIVERSITY LONDON

Driving to Safety Technical Weaknesses

- Most of the known models of "driving to safety" have a number of ***conceptual weaknesses***:
  - They rely on a ***univariate*** Bayesian inference model:
    - the measure of interest is the "***probability of accident per mile of driving*** (PAMD)", which allows for easy comparison of the AV safety with the safety of the man-driven vehicles.
      - statistics on man-driven cars exist, e.g., from NHTSA reports.
    - This univariate model ***ignores ODD*** (and indeed the "common sense") and implicitly assumes that PAMD is the ***same for all driving conditions***, which is clearly ***implausible***: PAMD varies (possibly very significantly) over different operating conditions (e.g., sunny day vs. heavy rain, etc.).
  - The known models typically consider the special case of "***driving with no accidents***". Given the statistics on vehicles with ADS (shown earlier), driving millions of miles with no accident is of ***limited/no use***.
  - Different stakeholders' perspectives on AV safety are not addressed (if addressed, the degree of conservatism is often excessive):
    - "Driving to safety" addresses the ***AV vendor*** needs only: it uses the evidence from the ***entire deployed fleet*** of AV vehicles of a particular make and model.
    - The needs of the ***owners/users of individual AV vehicles*** require different approach and the data from the fleet of AV should be used differently.
      - Different AVs are used differently (intensity of use will be different, the mix of operating conditions will be different, etc.). These differences must be accounted for in the Bayesian inference.

CITY UNIVERSITY LONDON

Driving to Safety Advances based on own work

- Own work addressed most of the weaknesses of the previous work on “driving to safety” (details in [1]).
  - Two papers have been published recently [1] and [2] which develop a methodology for driving to safety, which accounts for ODD and the need of the individual users.
- The essence of own work is replacing the ***univariate*** model with a ***multi-variate*** one (a multivariate Bayesian inference) to account for a given ODD defining a ***set of operating conditions***. The methods is based on another prior publication:
  - R. Pietrantuono, ***P. Popov***, and S. Russo, *Reliability assessment of service-based software under operational profile uncertainty.* Reliability Engineering & System Safety, 2020. **204**: article 107193.
- Advantages of the multivariate Bayesian inference:
  - Black – box predictions may be ***overoptimistic*** (in comparison with the “white-box” multivariate prediction), contrary to the claims by the proponents of the “conservative Bayesian Inference (CBI)”.
    - CBI seems to optimise for conservatism starting with a “wrong” model used in the Bayesian inference.
  - Not only the vendor, but the owners of individual AV vehicles, too, can potentially benefit from ***data sharing by the AVs*** in the entire deployed fleet of AV vehicles. This is formalised under additional ***plausible assumptions***.
    - Operational profile “on average” (for the fleet) and “in particular” (i.e. of the individual vehicles) may differ considerably;
    - Observations from the fleet (miles with and without accident) can be ***shared among all AVs***, thus allowing them to ***learn more quickly*** about the reliability of the particular AV model (as a design artifact) than if only own observations have been used.

Blackbox vs. White-box inference models

Black – box inference takes the view:

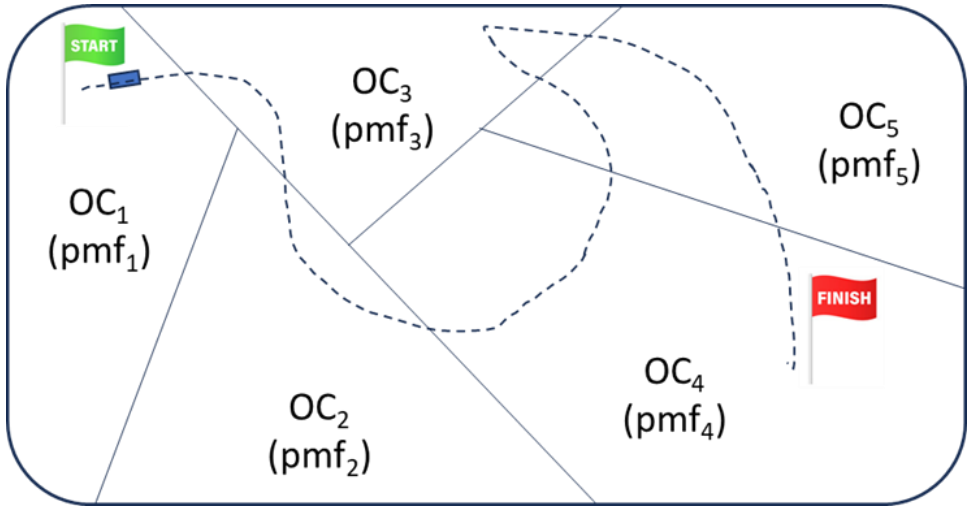
- A single ***pfm*** (probability of accident per mile of driving) is of interest
- No need to consider explicitly ODD and be concerned with its ***variation over time***.
  - No way to discriminate between different operational profiles (e.g. between different AVs , or between a profile "on average" (for a fleet of AVs) and a profile "in particular" (i.e. of an individual AV)
  - Inference is only concerned with the update of the marginal distribution of ***pfm***.
  - Vendors undertake extensive testing (e.g. Waymo publish AV safety related reports claiming multi-million miles driven by their fleets (56.700 million miles through January 2025), which address the problem of "infeasibility" of driving to safety.
    - *Kusano, K. D., et al. (2025). "Comparison of Waymo Rider-Only crash rates by crash type to human benchmarks at 56.7 million miles." Traffic Injury Prevention 26(sup1): S8-S20.*

The impact of operational profile in any reliability/safety/risk assessment is fundamental. Ignoring it is simply a poor practice.

- ***The most reliable and safest artifact is the one no one uses.***

Blackbox vs. White-box inferences

- ***White box model*** is defined in terms of:
  - ODD as a partition with different ***operating conditions***.
  - A key measure of interest in the model is the "operational profile (OP)" defined on an ODD.
    - OP is a probabilistic measure defined on ODD (some partitions are more likely than others, and the likelihood may vary over time), P($\boldsymbol{OC_i}$). OP is a ***fundamental concept*** in reliability and safety assessment.
  - Each partition (i.e. $\boldsymbol{OC_i}$) is characterised by the probability of accident per mile of driving ($\boldsymbol{pfm_i}$) conditional on the mile being drawn from partition $\boldsymbol{OC_i}$.
    - This is radically different from the standard "driving to safety" where a single parameter is of interest – the (marginal) "probability of accident per mile (***pfm***)"
  - ***pfm*** (used with a black-box inference model) can be derived from the "operational profile" P($\boldsymbol{OC_i}$) and $\boldsymbol{pfm_i}$ and used to compare AV safety against the safety of a human driver.
- Both, P($\boldsymbol{OC_i}$) and $\boldsymbol{pfm_i}$ are subject to ***uncertainty***, which is dealt with using the Bayesian inference, which updates both:
  - The operational profile, P($\boldsymbol{OC_i}$) and
  - the distributions of conditional $\boldsymbol{pfm_i}$.

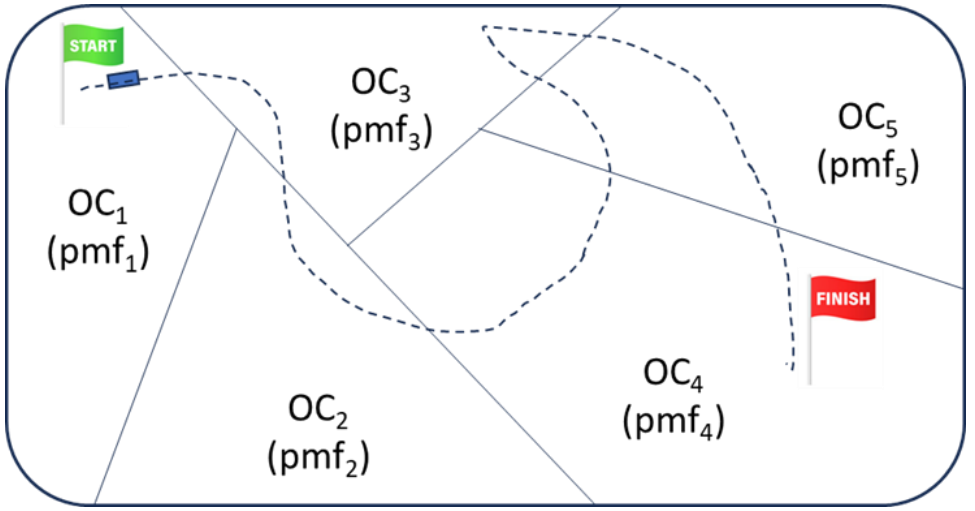


- The inference starts with an explicit definition of prior distributions:
  - of the conditional probabilities, $\boldsymbol{pfm_i}$. These can be assumed ***independently distributed random variables***.
  - Of the operational profile P($\boldsymbol{OC_i}$). These are ***dependent variables*** as their sum must be 1 (a randomly chosen mile must belong to one of the $\boldsymbol{OC_i}$).

Blackbox vs. White-box inferences (2)

Both Bayesian inferences are similar:

Prior Distribution(s) —(Observations)→ Posterior Distribution(s)

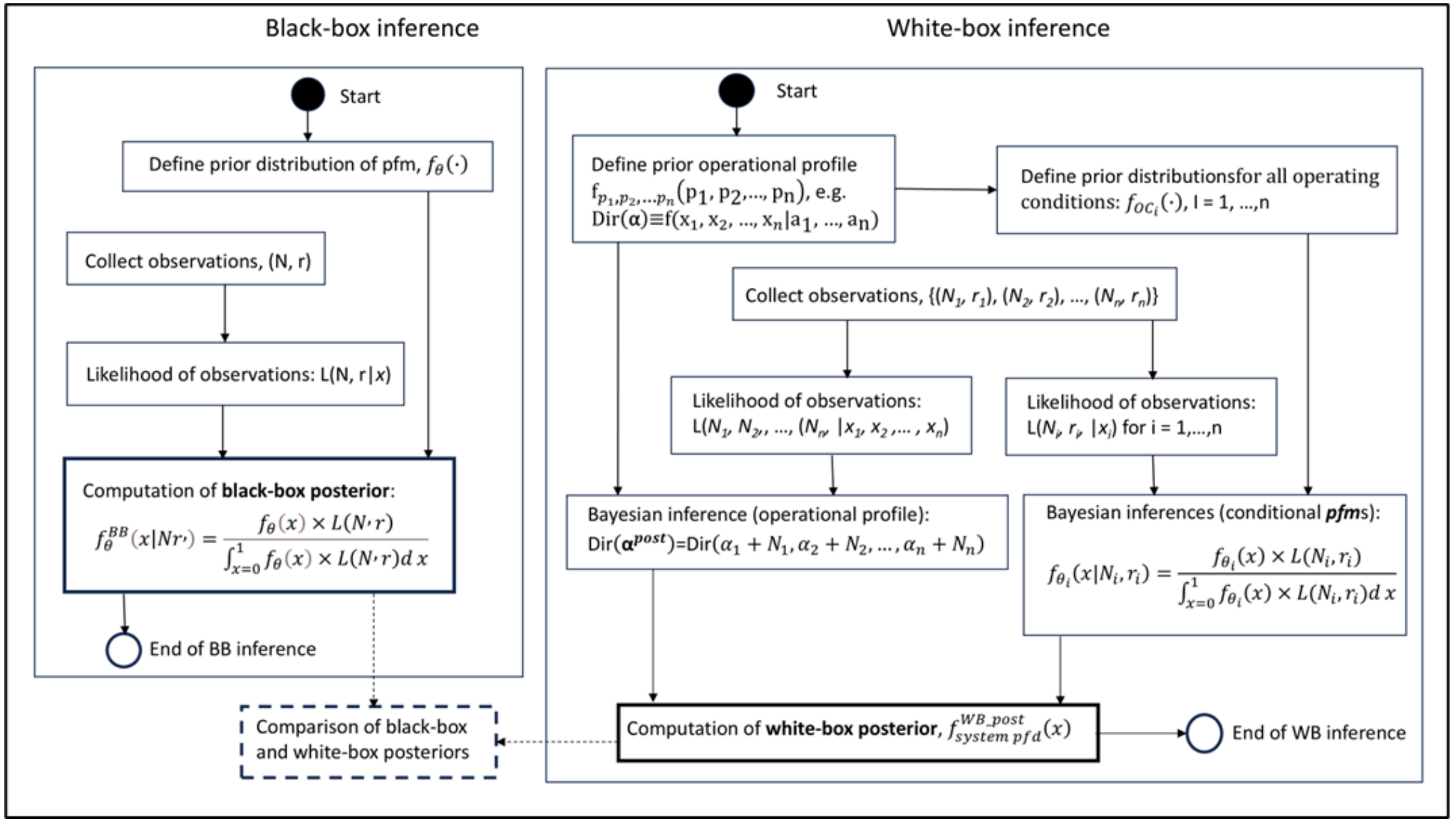


The differences are:

- The priors are different:
  - Black box:
    - univariate (of ***pfm***)
  - White – box:
    - multivariate (of $\boldsymbol{pfm}_i$) of conditional probabilities of accidents in different operating conditions $\boldsymbol{OC}_i$.
    - multivariate operational profile (P($\boldsymbol{OC}_i$))
- Observations
  - Black box: (***n***, ***r***), where ***n*** – number of miles driven, ***r*** - number of accidents observed.
  - White box: {($\boldsymbol{n}_1$, $\boldsymbol{r}_1$), …, ($\boldsymbol{n}_i$, $\boldsymbol{r}_i$), …, } – tuples ($\boldsymbol{n}_i$, $\boldsymbol{r}_i$) for all operating conditions.
- The predictions from both inferences can be compared, provided:
  - The priors are "comparable"
  - The prediction of the marginal distribution of ***pfm*** is used in comparison
    - White-box predictions need an "aggregation": ***pfm*** is a sum of the conditional $\boldsymbol{pfm}_i$ weighted by P($\boldsymbol{OC}_i$).

CITY UNIVERSITY LONDON

Black-box vs White-box inference



Does the inference really matter?

- It turns out that it does matter.
- Some "surprising" counterintuitive results were recorded
  - ***No accident*** is observed, but the ***"operational profile" is "evolving"***: partitions judged to be rare become more frequent.
    - If the uncertainty about $\boldsymbol{pfm_i}$ of rare partitions was significant (e.g. due to less testing), then a profile change may make the predicted marginal ***pfm*** worse than the prior despite the good news of "no accidents observed".
      - The $\boldsymbol{pfm_i}$ will all get better with no accidents
      - But the operational profile is evolving (cannot get "better" in all partitions, just becomes different)
    - Such an observation will be invisible with a black – box inference.

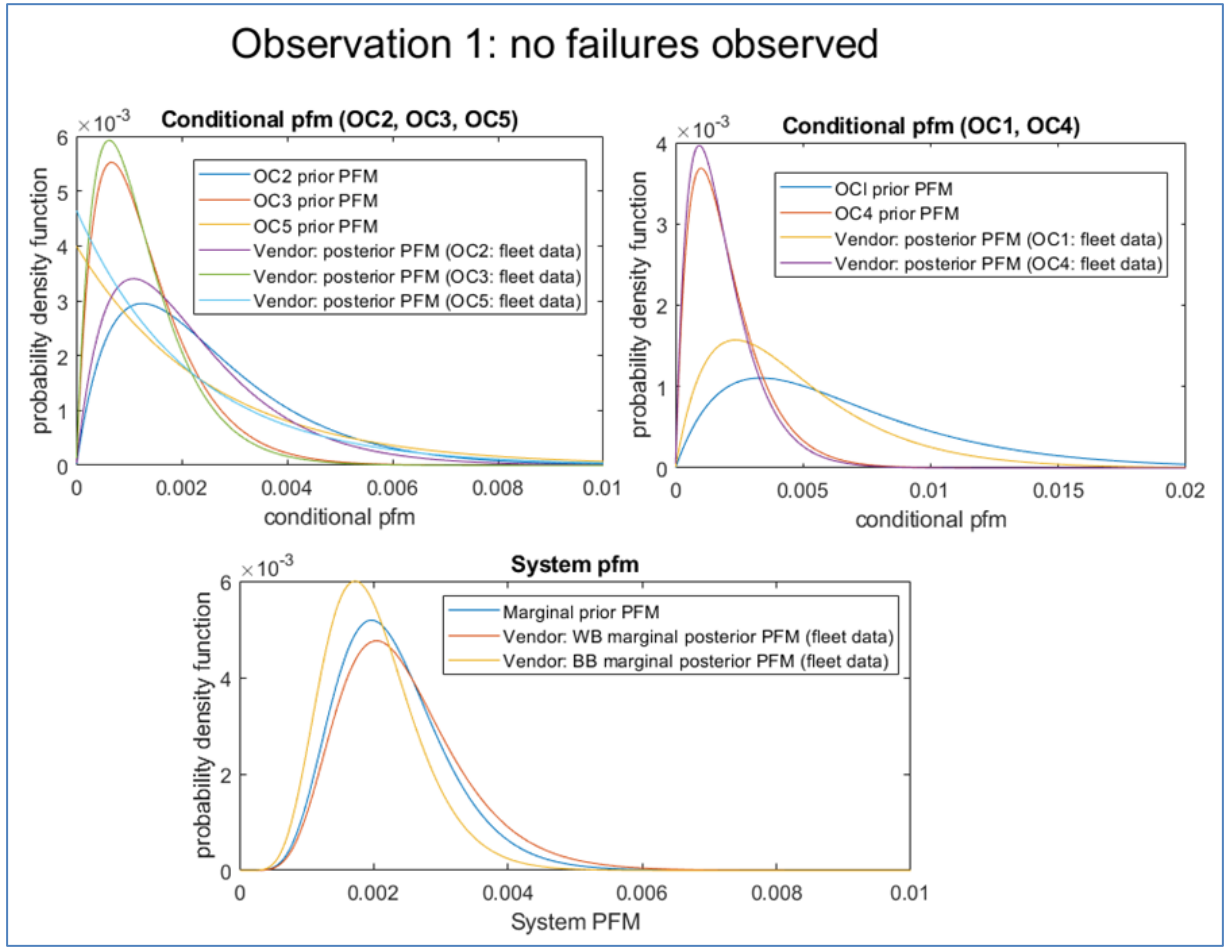
$$\mathrm{Dir}(\psi_1, \psi_2, \ldots, \psi_n; \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10)$$

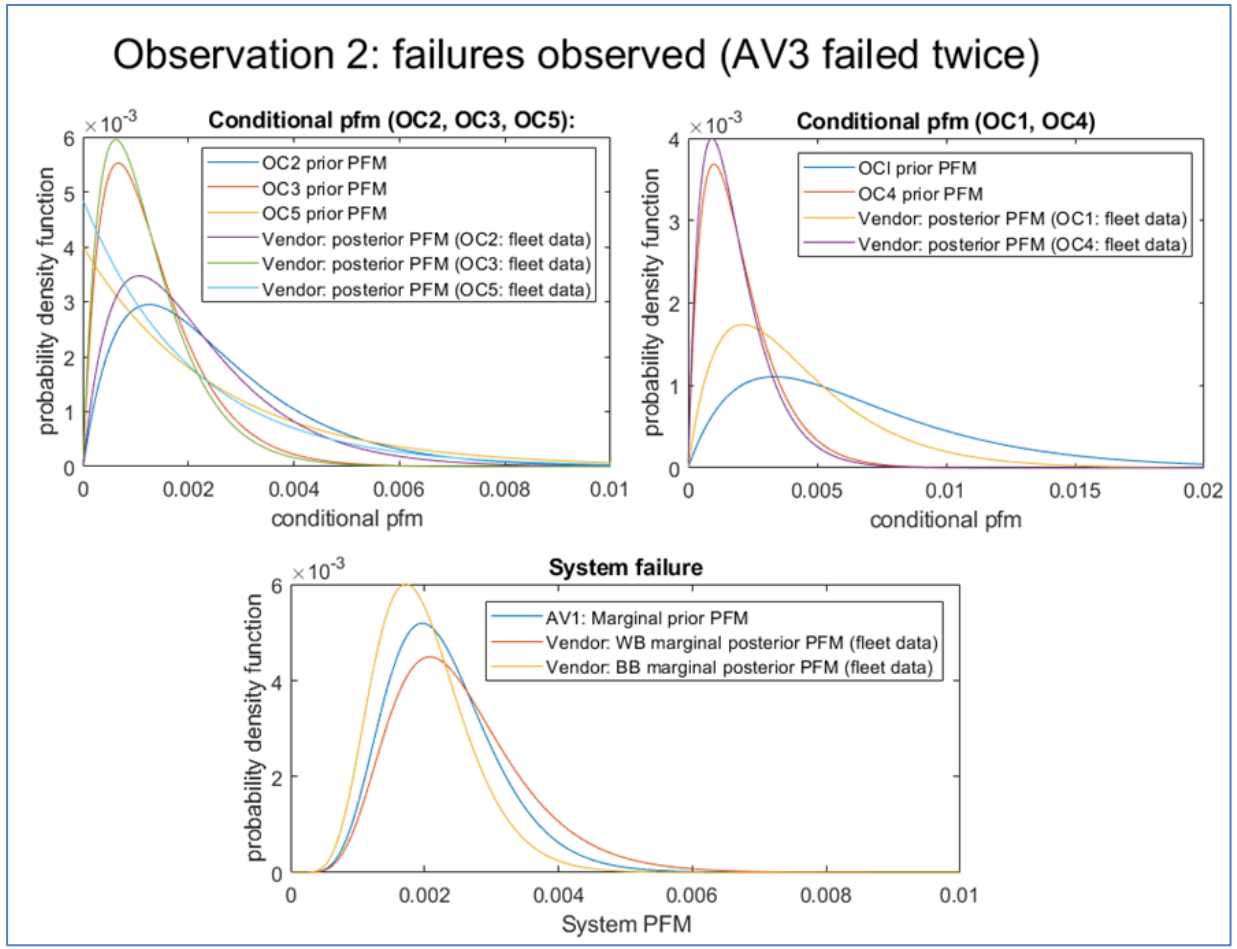
$$f_{\theta_1}(x) \equiv Beta(\alpha = 2, \beta = 299) \qquad f_{\theta_4}(x) \equiv Beta(\alpha = 2, \beta = 1000)$$
$$f_{\theta_2}(x) \equiv Beta(\alpha = 2, \beta = 800) \qquad f_{\theta_5}(x) \equiv Beta(\alpha = 1, \beta = 400)$$
$$f_{\theta_3}(x) \equiv Beta(\alpha = 2, \beta = 1500)$$

	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**

CITY UNIVERSITY LONDON

Black-box vs white-box inference illustrations





CITY UNIVERSITY LONDON

Conservative Bayesian Inference

It turned out that the Conservative Bayesian Inference (CBI) is ***not guaranteed*** to provide conservative predictions or the predictions are too conservative:

- ***Observation 1*** (No accidents): CBI predictions are more optimistic than the predictions derived with the "white-box" model
  - The difference in predictions is small.

- ***Observation 2***: (some accidents): CBI predictions are ***massively conservative*** (to the point of become not really useful, at all) even with only 2 accidents in 500 miles.
  - For any number of accidents (even 1-2) CBI provides very conservative predictions, i.e. it is very "brittle" to be useful for practical safety assessment.
  - This observation is not new. Have observed it with other "conservative models".
    - It seems a common pattern: conservative models lead to "side effects", which are counterintuitive, often simply unjustifiable.

Full details on the studies are in reference [1].

CITY UNIVERSITY LONDON

EST 1894

Dynamic AV Safety Assessment (DyAVSA)

CITY UNIVERSITY LONDON

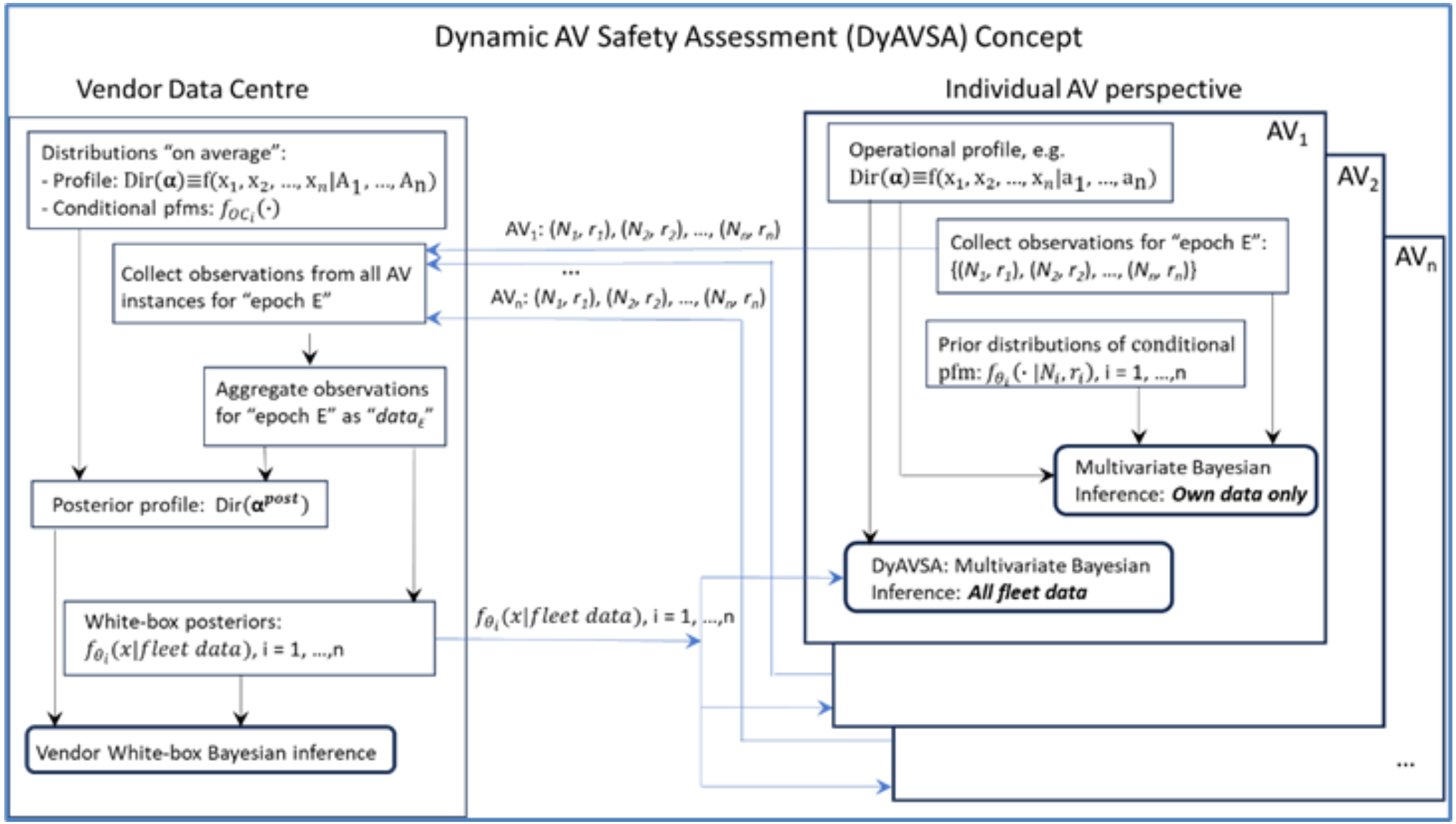
Dynamic AV Safety Assessment (DyAVSA)

Multivariate Bayesian inference does not care what observations will be used to derive the posterior predictions. The math works in exactly the same manor using observations from:

- A ***single AV***
  - The observations will be limited, even for the entire lifetime of an AV
- Data from a ***fleet of AVs***
  - e.g. aggregated somehow by the vendor: computing the sum of miles driven and accidents observed in different operating conditions
  - The inference will be useful for the vendor – data from all deployed AVs will be counted
  - The operational profile will be one “on average” (over the entire fleet), which may not apply to any particular AV.
  - The conditional $\boldsymbol{pfm_i}$, however, can be established with high accuracy since the number of observations will grow very rapidly.
- Hybrid inference (DyAVSA):
  - AV will use own data to predict the current ***own operational profile***
  - AV will use the aggregated data from the AV fleet to get estimates of the distributions of the $\boldsymbol{pfm_i}$.

CITY UNIVERSITY LONDON

Safety “on average” vs “in particular”



Does it really matter?

$$\text{Dir}(\psi_1, \psi_2, \ldots, \psi_n, \alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 40, \alpha_4 = 30, \alpha_5 = 10)$$

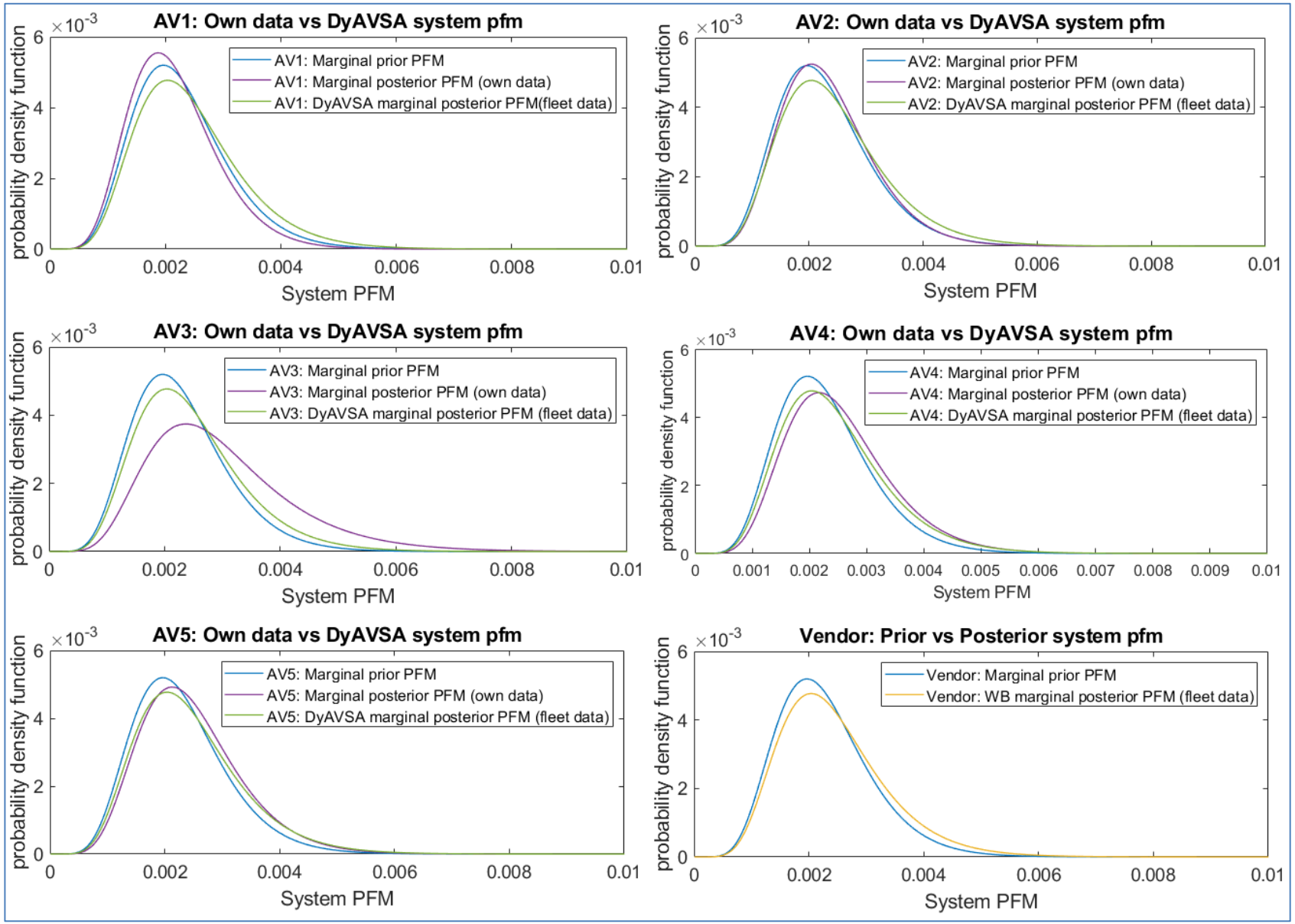
- It turns out that it does.
- Consider the same parameterisation of the prior distributions and the observations with two hypothetical observations:
  - Observation 1: ***No accidents***
  - Observation 2: Accidents observed with one of the AVs (AV_3) in two of the operating conditions (OC_1 and OC_3)
- We then conduct three inferences:
  - Each AV uses own data only
  - Inference for the fleet (of 5 AVs)
  - DyAVSA (hybrid inference as define previously)

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

	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**

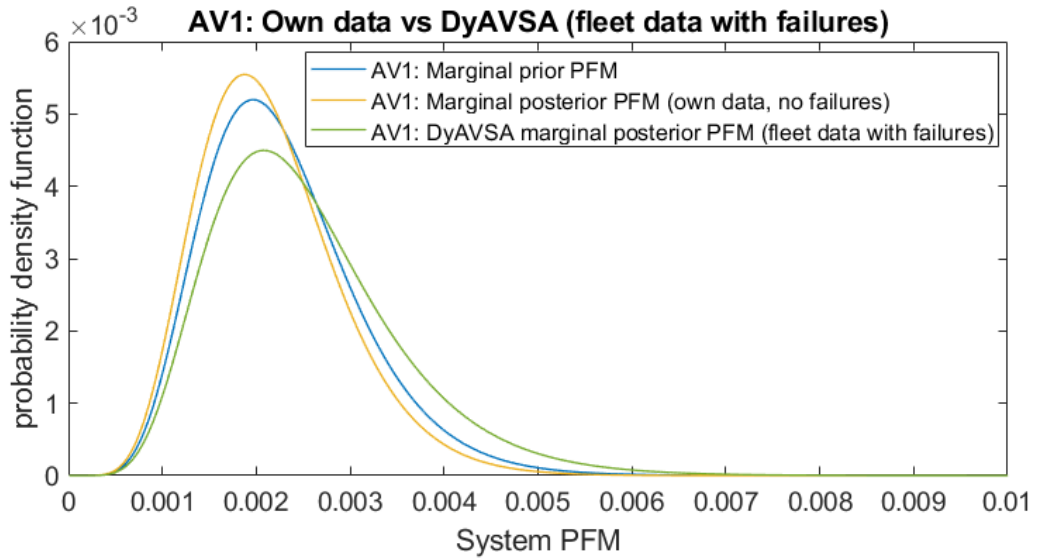
CITY UNIVERSITY LONDON

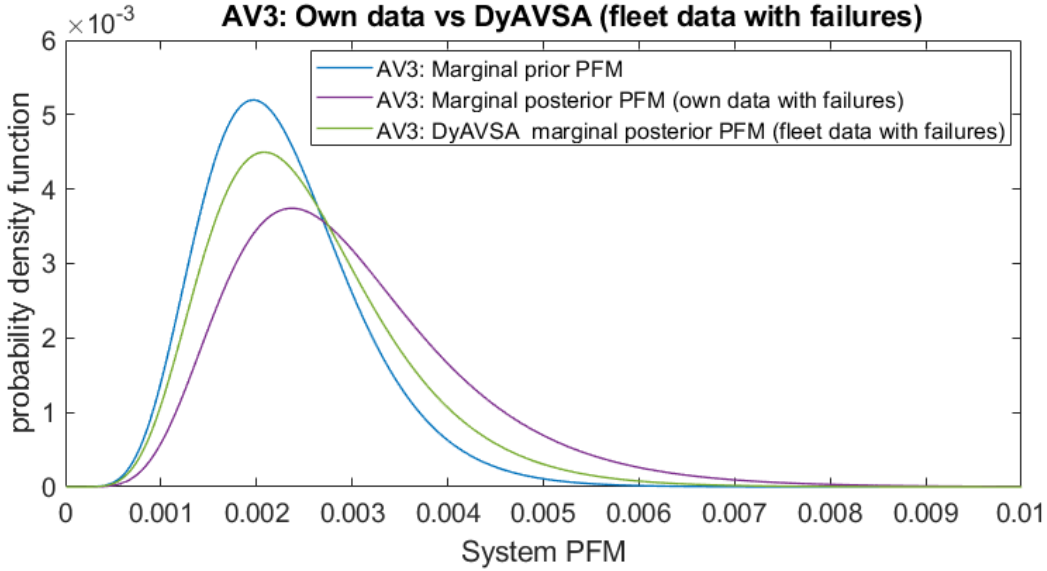
Own data vs fleet data vs. DyAVSA (no accidents)

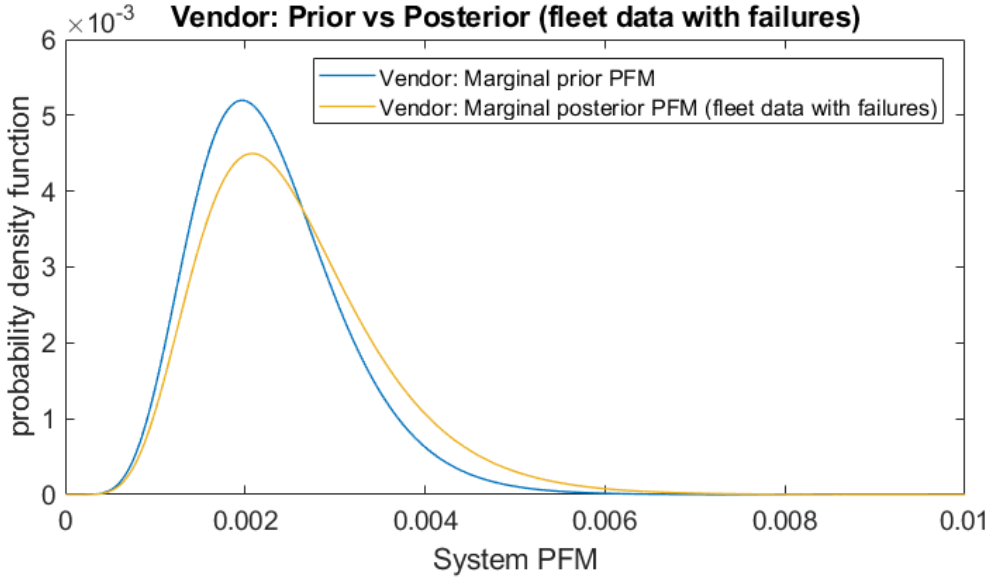


CITY UNIVERSITY LONDON

Own data vs fleet data vs. DyAVSA (with accidents)







CITY UNIVERSITY LONDON

DyAVSA Summary of observations

- Clearly, the predictions derived with own data only, will differ from the data derived with DyAVSA (details available in [2]):
  - A ***single AV*** and DyAVSA are suitable for safety assessment of a specific AV.
    - Should the risk (i.e. confidence) of accident cross a given target (e.g. agreed in a safety case) the vendor could raise a ***warning to the particular AV only***
    - The safety assessment will account for AV ***own operational profile*** and the observations on accidents and miles driven in different OCs collected from the entire AV fleet.
  - Data from a ***fleet of AVs***
    - The vendor can apply safety assessment using AV data and, which is significant, when necessary "translate" the assessment to operational profiles different from the profile based on fleet observations
      - E.g. if the vendor considers engaging with a ***new geographic areas*** (with anticipated different operational profile), DyAVSA will allow the vendor to use the current estimates of $\boldsymbol{pfm_i}$ and check if the AV type is likely to operate sufficiently safely in the new target area using the anticipated operational profile.
  - Feasibility of DyAVSA
    - Mobileye REM (Road Experience Management, https://www.mobileye.com/technology/rem/) seems to offer the communication infrastructure sufficient for the needs of DyAVSA.
    - Attributes used for ODD specification are sufficient to discriminate between different OCs.

CITY UNIVERSITY LONDON

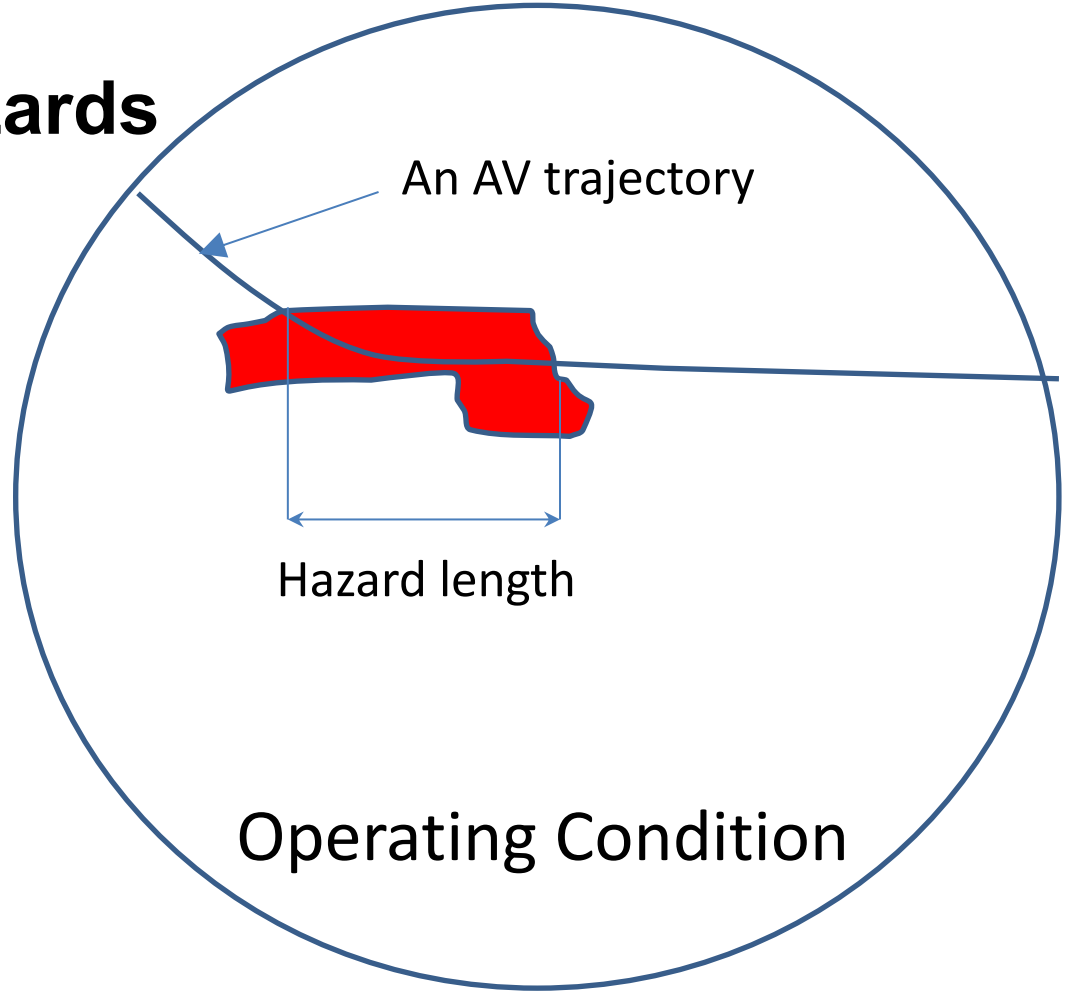
EST 1894

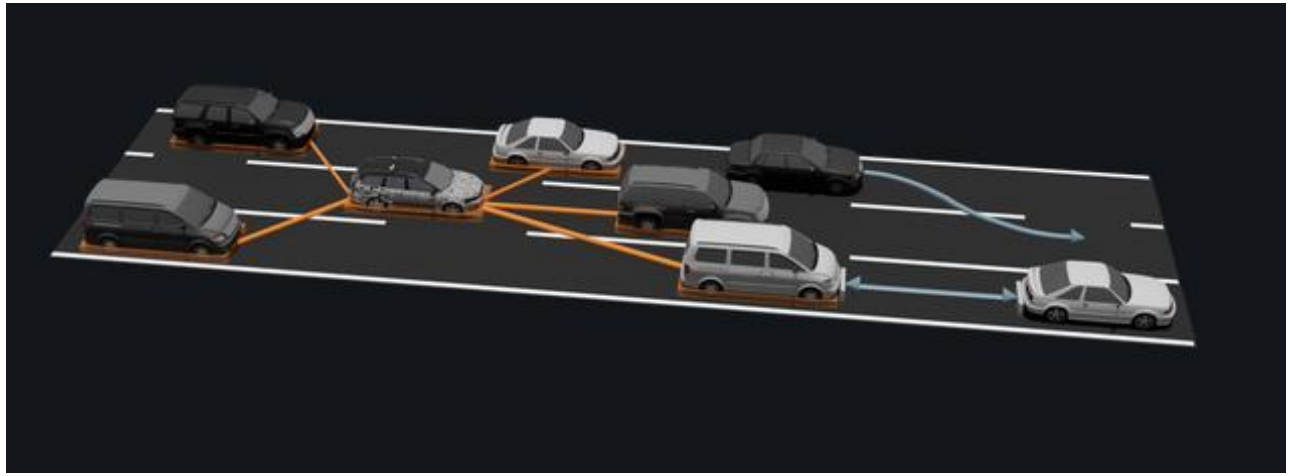
State-based probabilistic models in AV safety assessment

(Collaboration with Intel Labs Germany in ICRI SAVe)

Modelling road Hazards

- External (exogenous) hazards can occur on the road
  - ISO 26262 suggests that these be modelled by their:
    - *Rate* of hazard occurrence
    - *Duration* of hazard being present on the road (i.e., hazard's length in time)
- Road hazards can be captured using a *stochastic state machine*.
  - Then one can consider the implications of AV components failures (e.g. misclassification of an object on the road or failure of a safety monitor) which occur simultaneously with hazards.

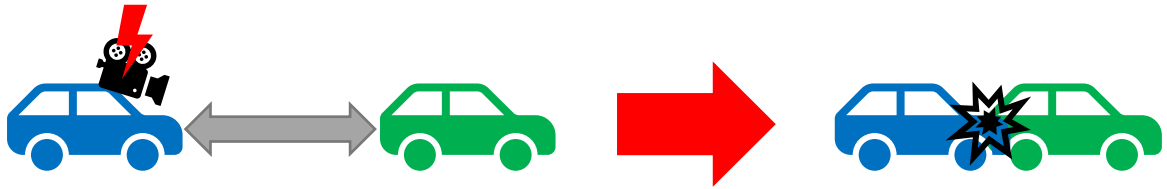




In part of this work probabilistic models were used to study the implications of AV perception system failures and failures of safety monitors on AV safety in the presence of road hazards.

Approach to safety: Road Hazards

- The view taken in the studies, presented in this talk, is that accidents are a consequence of road ***hazards***.



System Failure &
Road Hazard

Catastrophic Failure
(Accident)

  - Accidents on the road are ***much less likely without road hazards than with road hazards.***
  - The AV perception system may ***detect hazards*** correctly or may fail to detect them. The presented studies look at several distinct ***failure modes***:
    - Overlooking a road hazard when it occurs. Hazard may be detected ***with a delay***.
    - "Safety monitors", meant to guarantee AV safety, may themselves fail.
    - "False alarms" – flagging a situation on the road as hazardous while there is no hazard. These create hazards for the other AVs and vehicles with drivers.
- We study how these failure modes impact AV safety.

This view emphasises the importance of road hazards. Traditional "functional safety" focusses on developing mechanisms to counter hazards and keep the process (chemical plan, nuclear plant, etc.) within acceptable boundaries. Making the safety mechanisms fault-tolerant (e.g., using defence in depth) is an essential part of functional safety.

CITY UNIVERSITY LONDON

EST 1894

Basic Model

CITY UNIVERSITY LONDON

EST 1894

Study 1: A Basic model of Road Hazards

- Operates at a relatively high level of abstraction (details are in [4])
- Does not discriminate between different types of road hazards (e.g., due to scenery, bad weather, or extreme congestion, etc.)
- Accounts for:
  - Failures of perception system such as overlooking manoeuvres by other vehicles on the road
    - This failure may be temporal
    - Rate of accident when a road hazard is overlooked is likely ***larger*** than if the hazard is detected on time.
  - Failures of "safety monitors".
  - False alarms, which may also cause accidents by "surprising" other participants in the road traffic.

https://ieeexplore.ieee.org/document/9922283



Modelling road hazards and the effect on AV safety of hazardous failures

Cornelius Buerkle[1], Fabian Oboril[1], Michael Paulitsch[1], Peter Popov[2] and Lorenzo Strigini[2]

***Abstract*— Autonomous vehicles (AV) are about to appear on our roads within the next few years. However, to achieve the final breakthrough, not only functional progress is required, but also fundamental safety questions must be solved. Among those, a question demanding special attention is the need to assess the overall safety of an AV and quantify that it is safe enough to take part in normal traffic despite its inherent imperfections. Therefore, this paper describes a probabilistic model, which allows to study how imperfections of an AV perception system and of mechanisms responsible for AV safety (e.g., Safety Monitors), can impact AV safety in the presence of road hazards. We also demonstrate how the model can be used to validate if the AV is safe enough, to understand the criticality of (perception) errors, and to identify areas/parameters that have more influence on safety than others.**

I. INTRODUCTION

Assessing safety of autonomous vehicles (AV) poses new challenges. The known approaches to AV safety assessment, e.g., functional safety [1], have been tried in the past, but dealing with components based on machine learning remains problematic. Consequently, new alternatives are evolving, among them is the standard on "Safety of the Intended Functionality (SOTIF)" [2] (ISO 21448), or the upcoming standard on safety of AV decision making components [3]. In addition, governmental authorities are currently setting up the required legislative boundaries for the public use of AVs. For example, the German government announced that an AV has to operate at least as safely as a human driver to receive certification [4]. However, while these standards define guardrails in which manufacturers must operate, an important question remains: *How to prove that an AV is safe enough to receive certification to participate in regular traffic?*

To demonstrate that an AV is safe enough, manufacturers need to validate that the rate of catastrophic failures (i.e., accident) is below the required threshold. Targeting the performance of a human driver, this means that the accident rate must be below 10^{-5} (1 severe accident every 10^5 hours of driving). One solution to this problem is to derive safety arguments from the large number of miles driven by AVs [5]. This, however, seems impractical for future mass deployment with frequent software updates. Instead, formal approaches are required, that can be either mathematically verified or validated on a smaller amount of data.

As illustrated in Figure 1, a catastrophic failure (i.e., an accident caused by the AV) is the consequence of two events that happen "simultaneously". First, a road hazard is required (dangerous driving situation), and second, a system failure (e.g., a failure of the AV's perception system to detect a

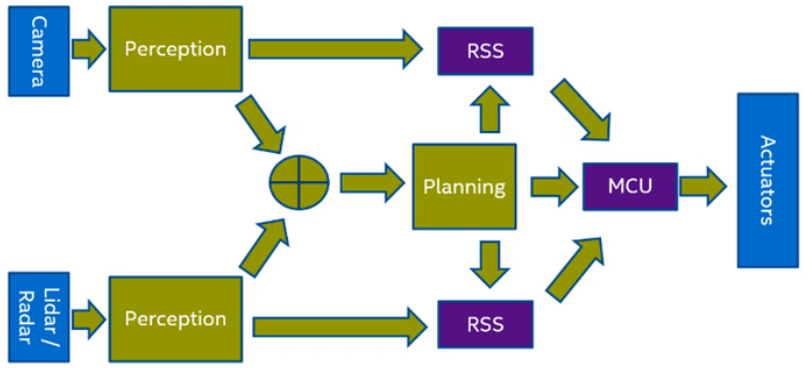
Figure 1 Illustration of road hazards.

In this regard, it is worth noting that for example even the best state-of-the-art perception systems have only a modest accuracy. As indicated by some, e.g., [6], detection accuracy of objects (lights, pedestrian, etc.) by a good machine-learning (ML)-based perception device is in the range of 80% - 99% per object. Although these numbers can be improved by combining several perception approaches and using multi-modal sensor technologies, one question remains [1]: *What is the required failure rate of a perception system, to achieve the necessary level of vehicle safety* (given the situational-awareness mentioned before)?

In this work we address the key questions raised before by a formal approach to safety assessment. We propose a comprehensive model to estimate the probability of a catastrophic failure of an AV over a period of time (thousands of hours of driving), that connects road hazards to the aforementioned failures in the AV processing pipeline (e.g., perception or planning). For this purpose, we model the failures and their occurrences as non-homogeneous stochastic processes. As we demonstrate, the model can be fed with naturalistic driving data on the distribution of road hazards (their rate of occurrence and durations as recommended by ISO 26262) and yield important insight on acceptable failure rates of AV components. Our evaluations reveal that perception stack and safety monitors must meet stringent reliability requirements. Yet it is not required that their failure rates are on par or below the target catastrophic failure rate, as non-hazardous traffic situations act as masking factors. We also report that some of the parameters used in our analysis have negligible impact on system safety, thus they need not be estimated with high confidence. In summary, our proposed model captures SOTIF performance limitations with situational awareness and their impact at system level (catastrophic failures), which can be used early in the design process to make appropriate design decisions.

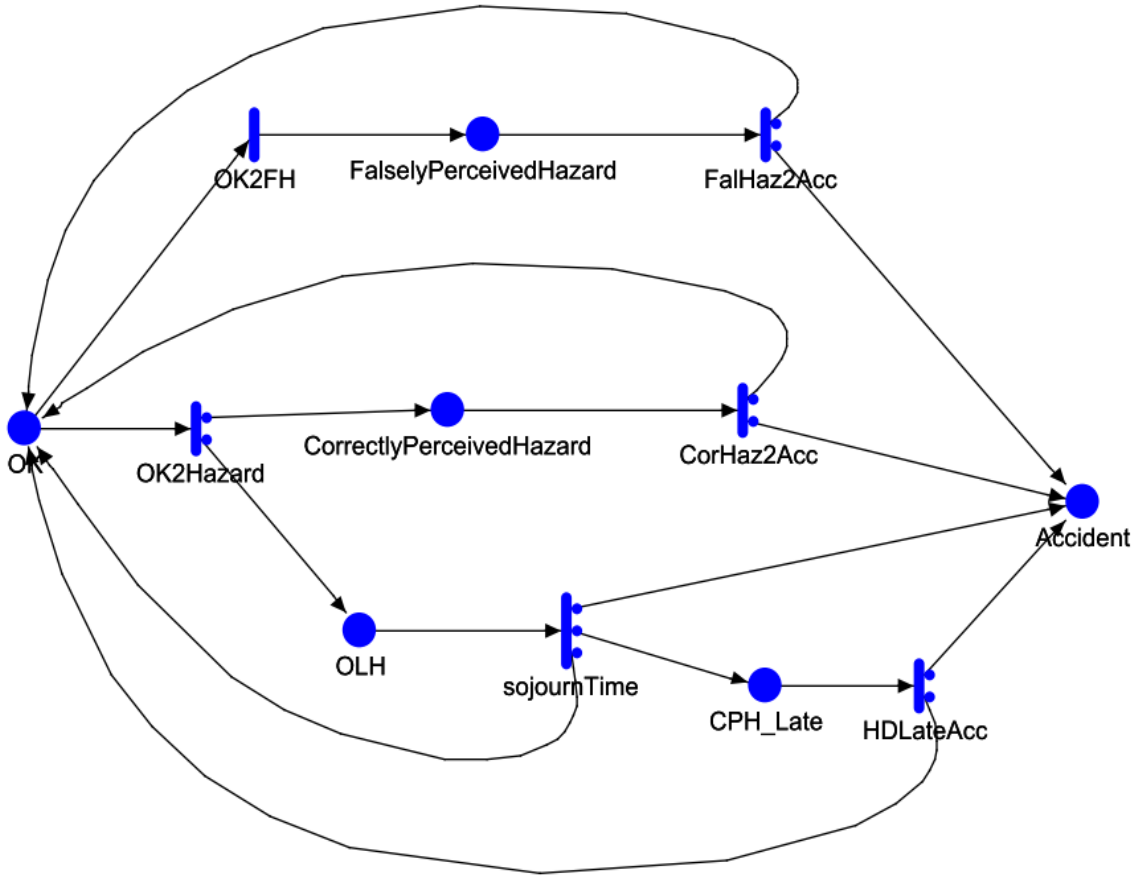
In the rest of this paper, Section II defines essential concepts used in our proposed model. The model itself is presented in Section III. Section IV covers our evaluations, followed by a discussion in Section V and conclusions in Section VII.

Architecture



CITY UNIVERSITY LONDON

The probabilistic state-based model

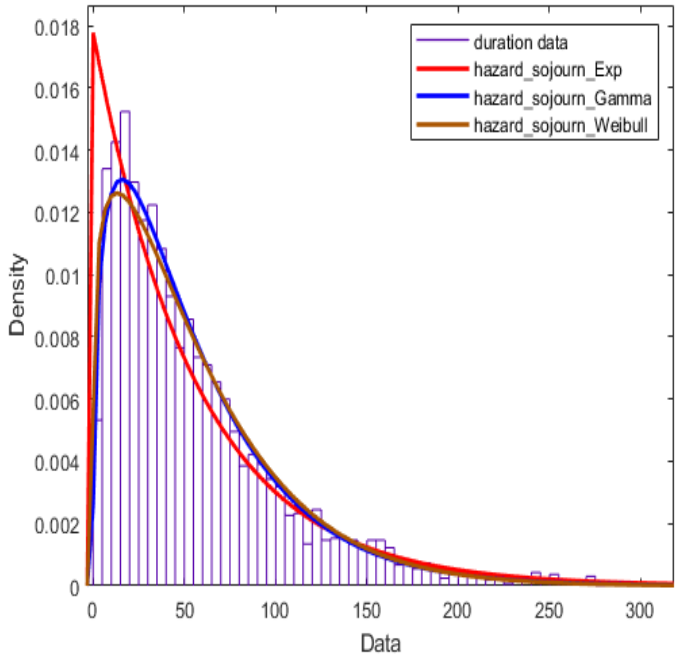


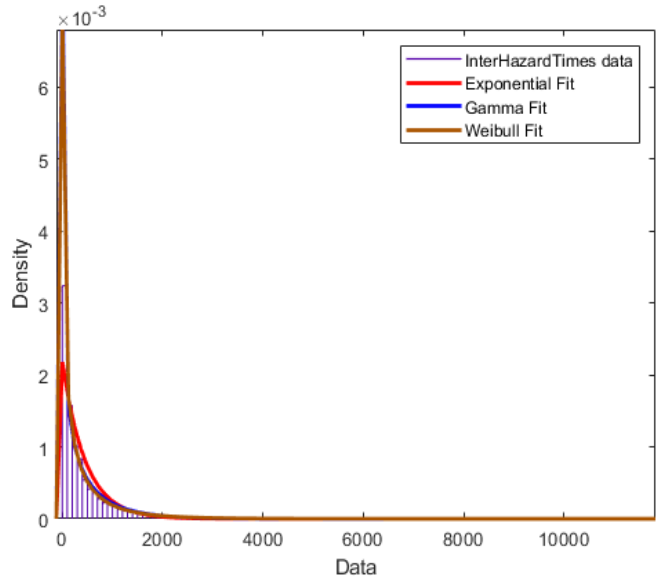
- The formalism of ***stochastic activity networks*** (SAN) is used to model AV safety in the presence of road hazards and failures of AV perception and safety monitors:
  - The states of an AV in the model are:
    - Normal (OK)
    - FalselyPercievedHazard
    - CorrectlyPercievedHazard
    - OLH (OverlookedHazard)
    - CPH_Late (CorrectlyPercievedHazard_Late)
    - Accident (an absorbing state)
  - Transitions between states are driven by ***timed activities***.
    - We assume (for convenience) ***exponential distributions*** for all timed activities.
  - The model starts in "OK" state and makes transitions to other states until AV enters the ***absorbing*** state ("Accident").
  - Measure of interest (i.e., of AV safety) is the ***probability of entering state "Accident" within predefined intervals*** of operation (100, 1100, …, 9100 hours) since the start.
    - At speed of 100 km per hour 9100 hours of operation is almost ***1 million km***.
  - ***Case probabilities*** in the model represent branches in model behaviour, e.g., either AV perception detects a hazard correctly upon its occurrence or not. These are important model parameters.

CITY UNIVERSITY LONDON

Model parameterisation: estimable from data

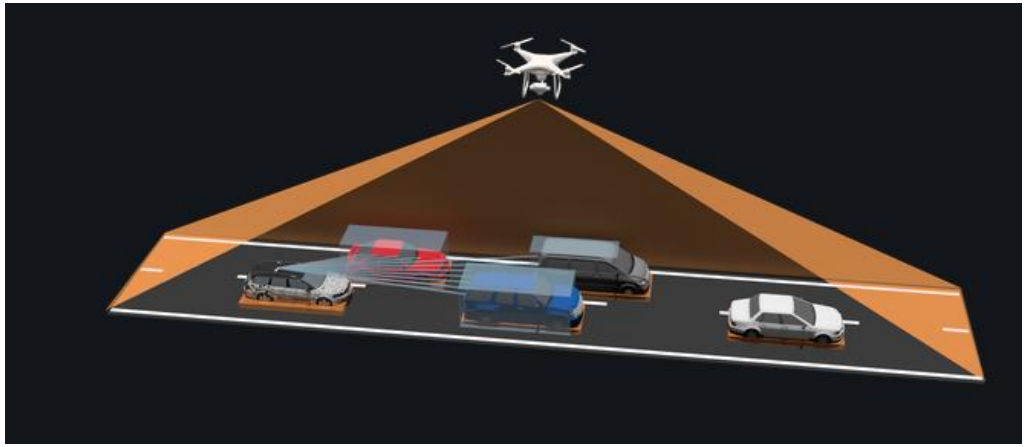
- Some of model parameters are ***directly estimable*** from empirically collected datasets:
  - Intervals between hazards (e.g., timed activity “OK2Hazard”)
  - Duration of hazards (e.g., timed activity “CorHaz2Acc”, “sojournTime”, etc.)
- For the study these model parameters are estimated using two publicly available datasets:
  - ***HighD*** dataset – a drone-recorded dataset for human driving behaviour on German highways.
  - ***Lyft*** dataset
    - was used to get a ***ball-park estimate*** of the duration of perception errors.
  - Further details on datasets and how they were used to estimate model parameter are given in the paper.





CITY UNIVERSITY LONDON

The highD dataset



The highD dataset (https://levelxdata.com/highd-dataset/) is a dataset of naturalistic vehicle trajectories recorded on ***German highways*** collected with a drone.

Traffic was ***recorded*** at six different locations and includes more than 110 500 vehicles. Each vehicle's trajectory, including vehicle type, size and manoeuvres, is automatically ***extracted***.

The positioning error is typically less than ten centimetres. The dataset was created for the safety validation of highly automated vehicles.

It is also suitable for many other tasks such as the analysis of traffic patterns or the parameterization of driver models.



CITY UNIVERSITY LONDON

Model parameters: sensitivity analysis

- For other model parameters, not estimable from data, we conducted "sensitivity analysis" within a "plausible" range of values.
  - These parameters are ***unknown with certainty*** (e.g., they are difficult to estimate empirically. No suitable datasets seem available).
- "Sensitivity analysis" serves ***two purposes***:
  - Tells us which parameters matter for the analysis (i.e., if they must be estimated accurately):
    - little (or negligible) effect of a parameter value on AV safety indicates that the parameter value ***does not matter***. In this case a "ball-park" figure can be used and the effort on accurate estimation can be saved.
  - For those parameters which ***matter***, the sensitivity analysis shows the range of parameter values, which would make the AV unsafe.
    - The analysis, thus, allows one to establish the ***requirements*** for the reliability of the sub-systems modelled by the respective model parameters (e.g., how good the perception system should be) so that the AV is safe.

Parameter	Values	Description
CH2Acc_prob	0	Case probability 2 of FalHaz2Acc timed activity.
CorHaz_sojournTime FH_sojournTime	856.3	Rate of timed activities CorHaz2Acc and FalHaz2Acc
FH2Acc_prob	0	Case probability of CorHaz2Acc timed activity.
HDLateAcc_prob	$[10^{-4}, 2x10^{-4}]$	Case probability of timed activity HDLateAcc leading to Accident.
HDLate_sojournTime	1000.0	Rate of timed activity HDLateAcc.
HazardRate	197.4	Rate of timed activity OK2Hazard (intervals between hazards).
OH_sojournTime	[1125, 2250, 4500]	Rate of timed activity sojournTime (duration of overlooked hazard)
OLH2Acc_prob	$[10^{-5}, 2x10^{-5}, 5x10^{-5}]$	Case probability of timed activity sojournTime leading to Accident
OLH2CHLate_prob	[0.99, 0.991, 0.995]	Case probability of timed activity sojournTime leading to CPH_Late.
falseHazardProb	0.0	Probability of false hazard detection.
missHazardProb	$[0, 10^{-4}, 5x10^{-4}]$	Probability of omitting a hazard given the hazard occurred

CITY UNIVERSITY LONDON

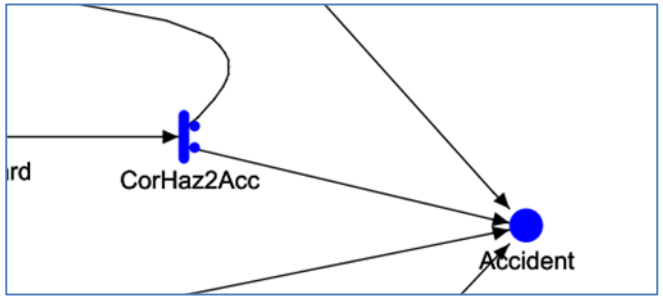
Results

The paper shows the results from 6 different studies related to AV safety.

All results are obtained by "solving" the SAN models with the chosen parameters using the ***numeric solvers*** built-in the tool (Möbius) for continuous time Markov chains (CTMC).

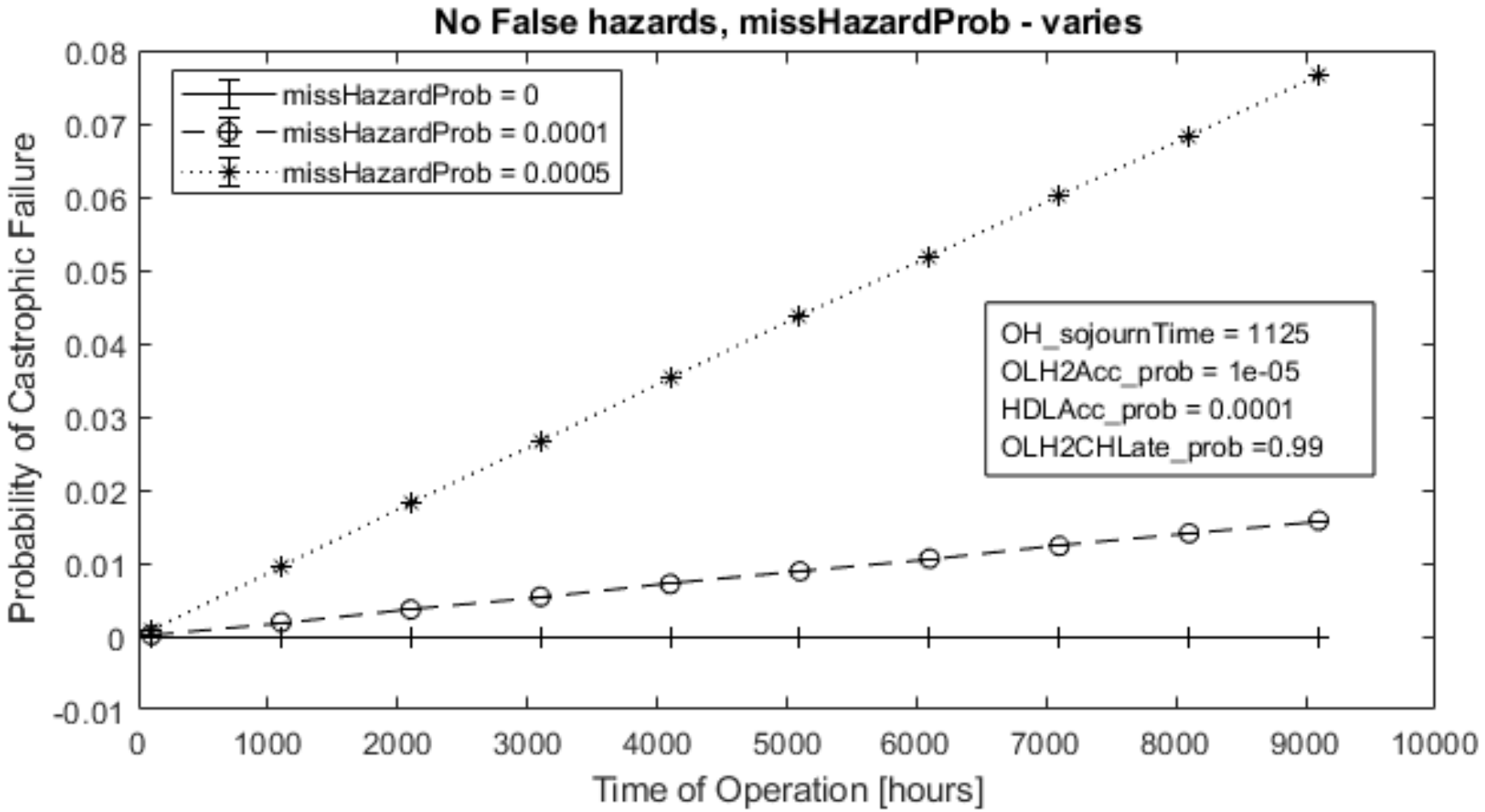
The studies included in the publication were arranged in two groups:

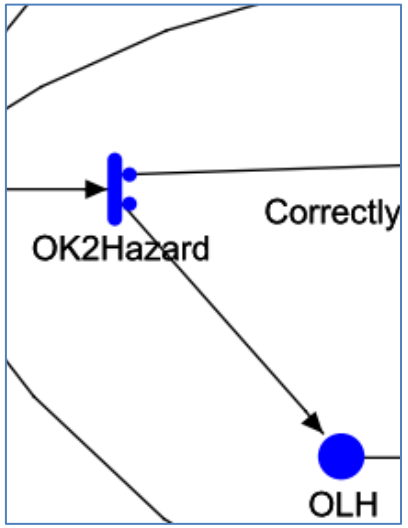
- ***Group 1***: Conducted under the assumption that if a hazard is detected ***immediately and correctly*** then an accident cannot occur:
  - Case 1: Effect of the ***probability of overlooking a hazard*** on system safety
  - Case 2: Effect of ***duration*** of overlooking a hazard on system safety
  - Case 3: Effect of the rate of accident from the state of ***late hazard detection*** on system safety
- **Group 2:** Assuming that even when a hazard is immediately detected correctly upon its occurrence, an accident still may occur (with a small probability)
  - Case 4: Effect of the rate of accident on system safety
  - Case 5: Effect of ***false alarms*** on system safety
  - Case 6: Combination of case 4 and 5 above



CITY UNIVERSITY LONDON

Results: Case 1



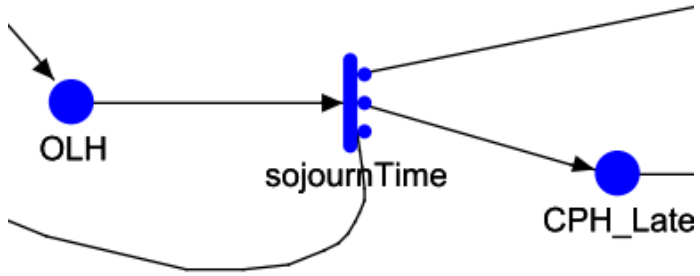


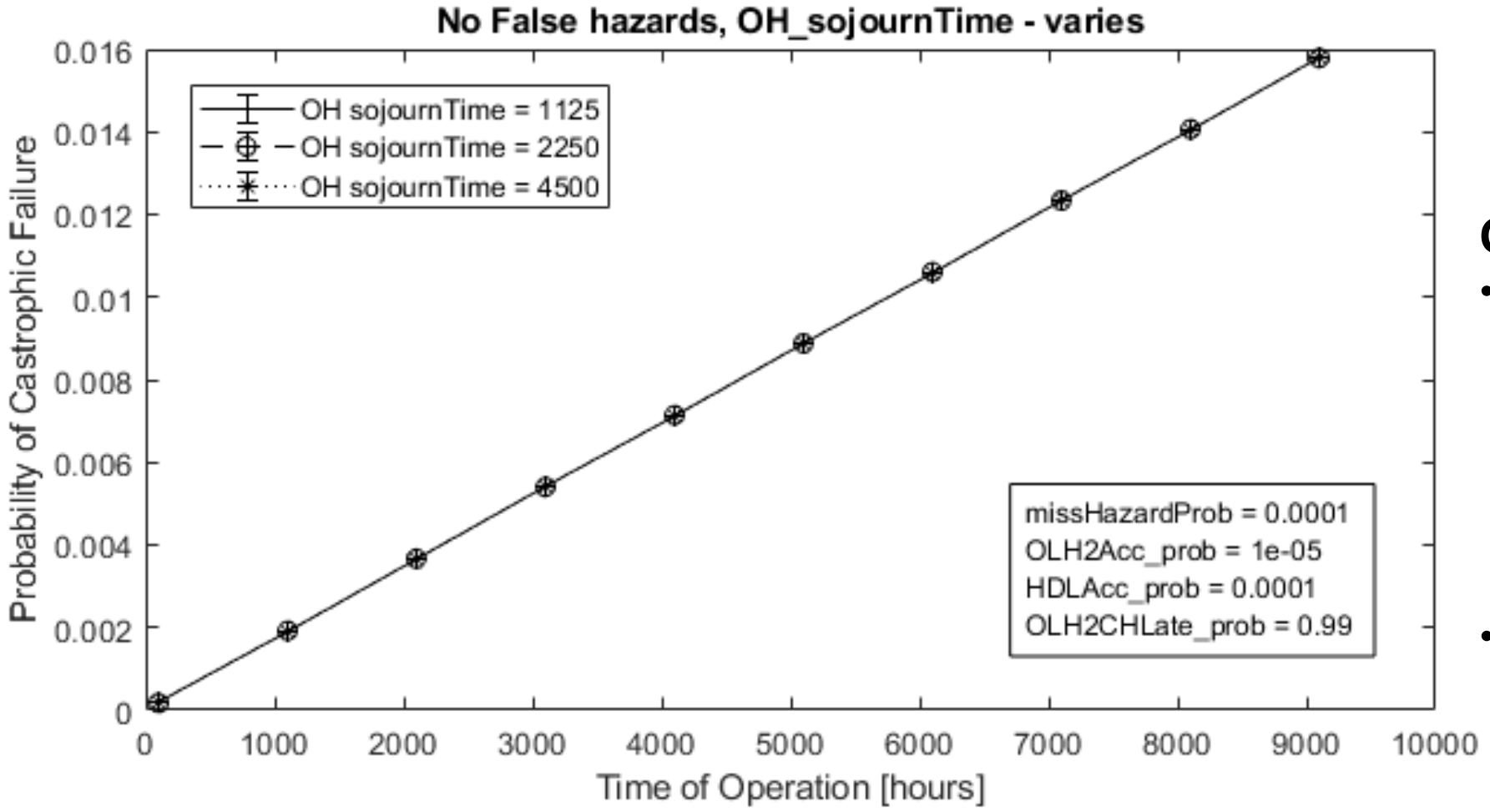
Observations

- Intensity of overlooked hazards affects system safety ***visibly***
- Probability of system failure remains low 9000+ hours of operation (i.e., over a year of continuous driving).

CITY UNIVERSITY LONDON

Results: Case 2



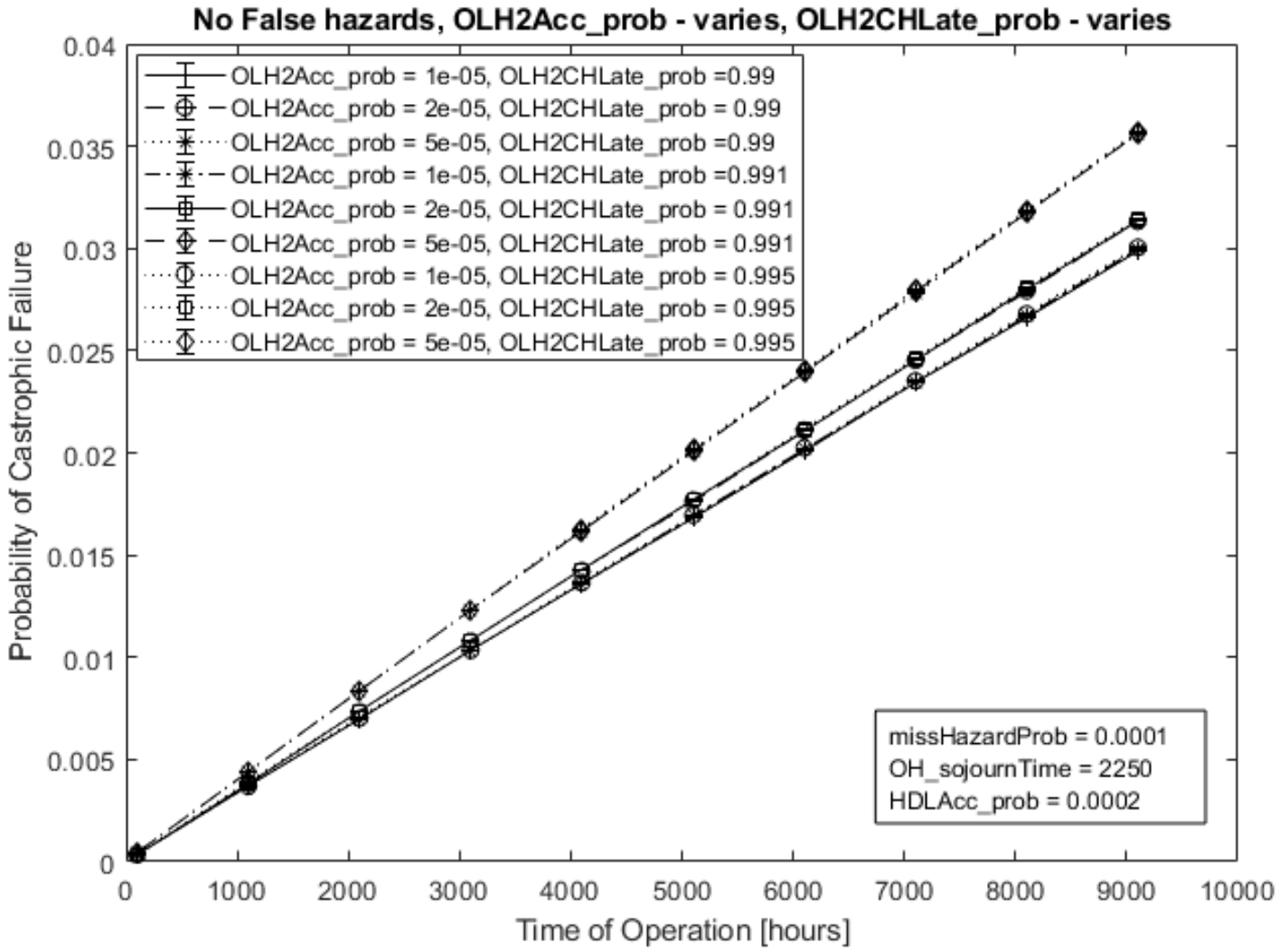


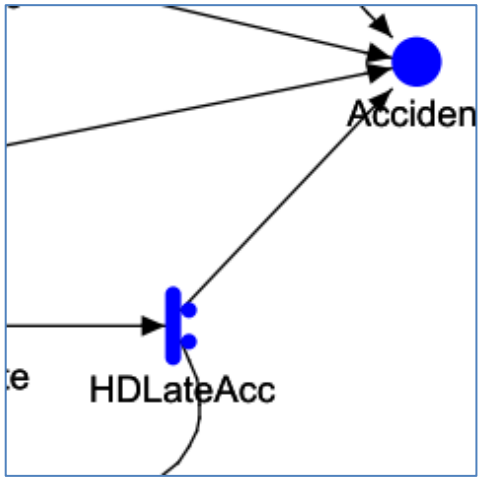
Observations

- Duration of overlooked hazards does not matter
  - For the chosen range of variation, of course (derived from the Lyft dataset)
- Probability of system failure remains low (~1.5%) after 9000+ hours of operation

CITY UNIVERSITY LONDON

Results: Case 3



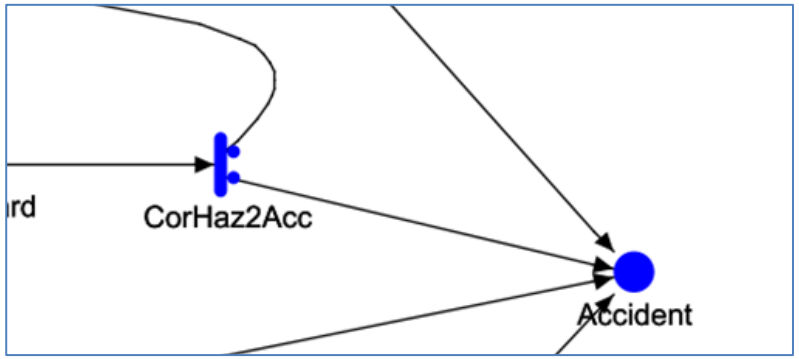


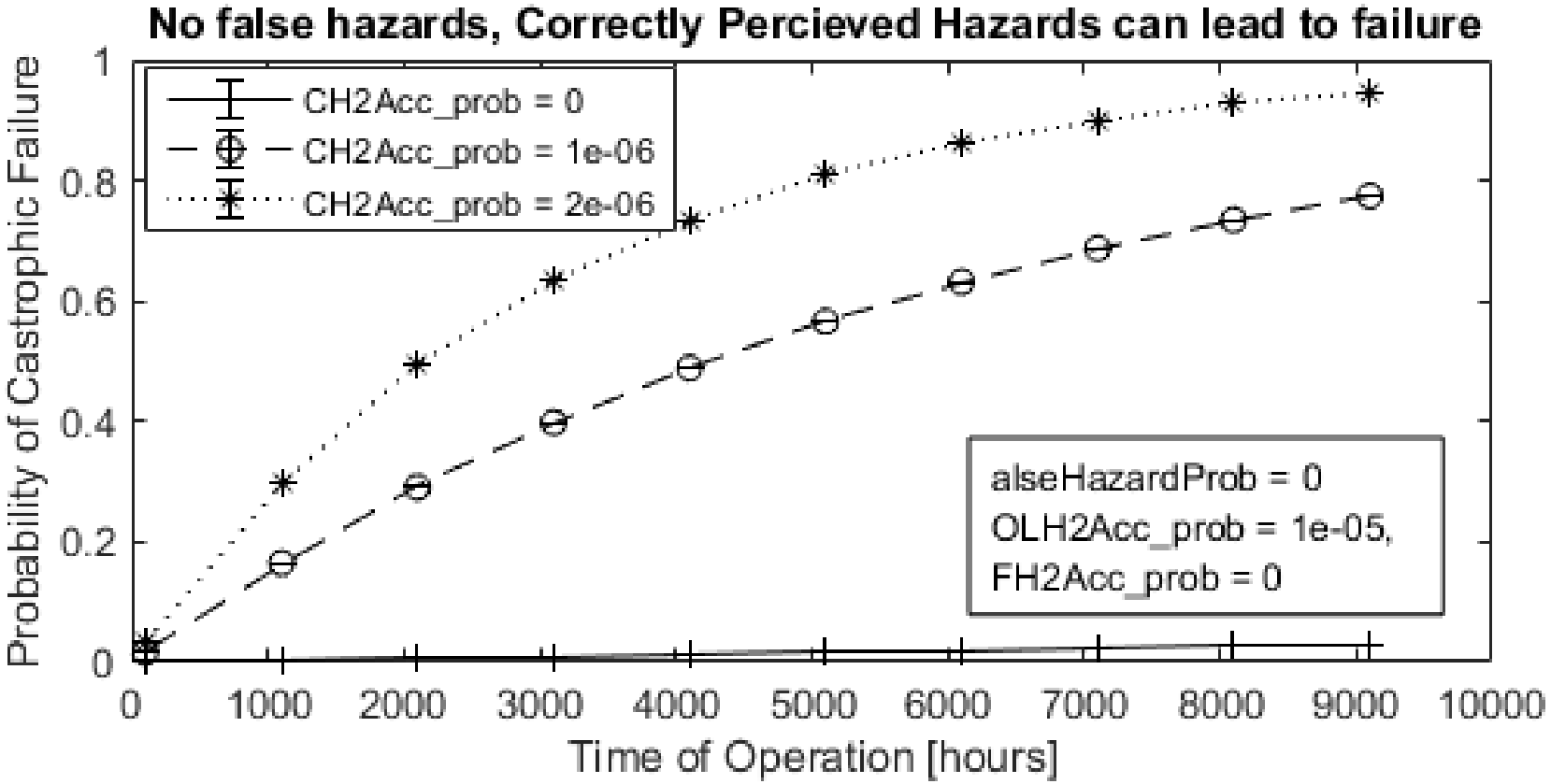
Observations

- The rate of accident from late detection of hazards affects visibly the probability of accident for a mission of given size.
- System safety remains good (< 4%) after 9000+ hours of operation.

CITY UNIVERSITY LONDON

Results: Case 4





Observations

- Even a tiny rate (10^{-6}) of accident from the state of correctly perceived hazard has a dramatic detrimental effect on system safety.
- AV safety becomes ***grossly inadequate***: probability of accident reaches very significant values even for modest number of hours on the road
- After 9000+ hours of operation the probability of accident reaches 0.8 and approaches 1.0 for rate of accident of 10^{-5} and 10^{-6}, respectively.
- This case models the impact of “safety monitors” on AV safety. Monitors must be very reliable indeed.

CITY UNIVERSITY LONDON

EST 1894

Intersection Model

CITY UNIVERSITY LONDON

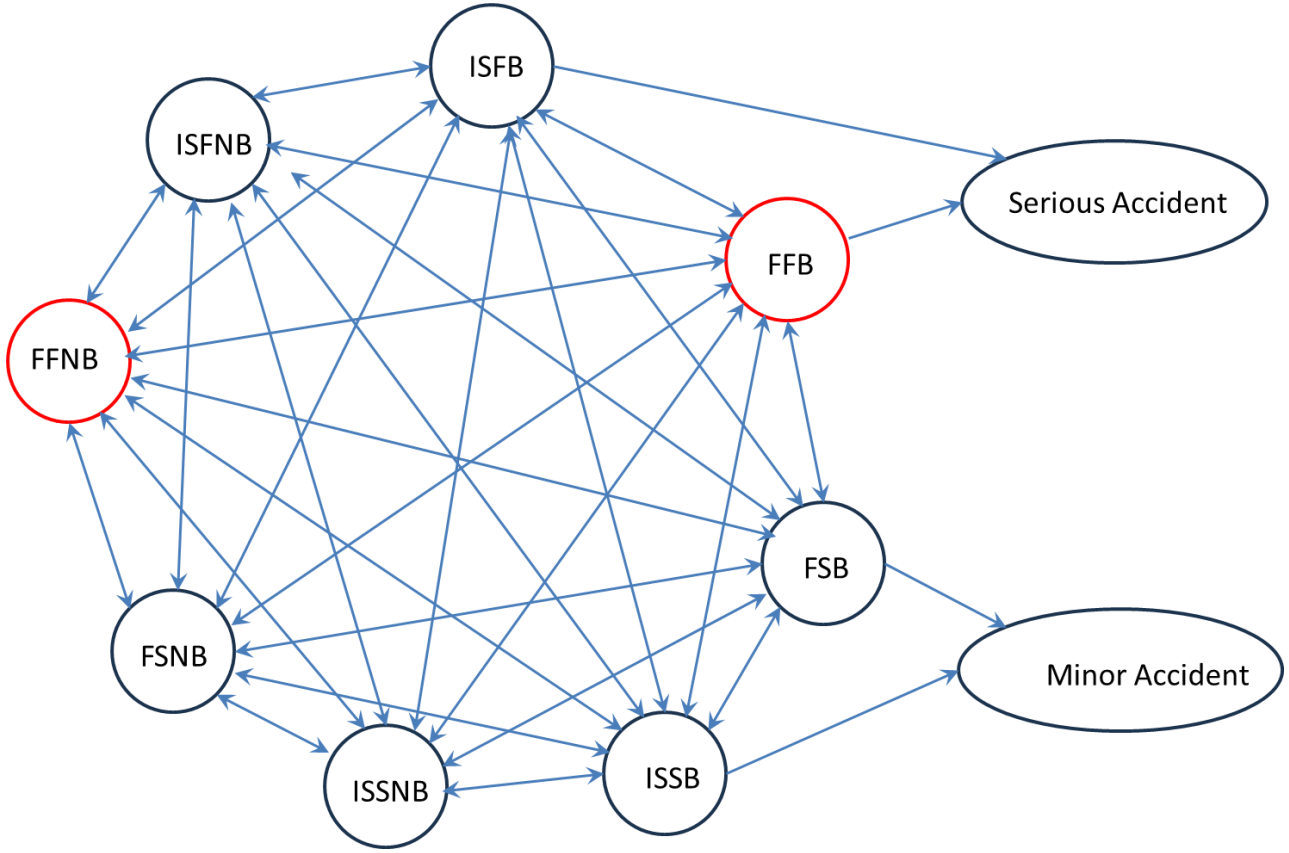
Study 2: A model of intersection (and "free" driving between intersections)

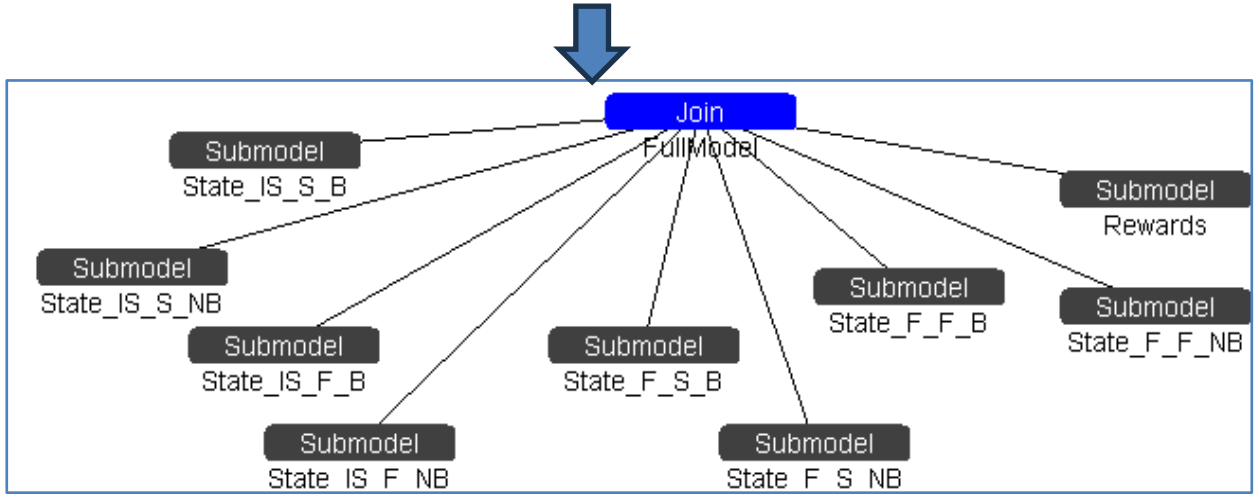
- This extension is important as it ***accounts for ODD***, mentioned earlier (details available in [5]). Intersections are explicitly referred to in ISO 34503:2023
  - The work shows that a complex model with ODD can be constructed, parameterised using publicly available datasets, and solved.
- Operates at a ***more detailed*** level of abstraction than the basic model
  - Discriminates between different types of road hazards
    - Intersection/"Free driving"
    - Driving "fast" vs. driving "slow"
  - Discriminate between vehicle "braking" and "not braking"
    - "braking" is used as a proxy for ***road hazards*** occurrence
- Accounts for:
  - ***Failure*** of perception system:
    - Omission failure (overlooking road hazards) may be temporal (hazards may eventually be detected correctly, but with a delay)
    - Rate of failure when road hazard is overlooked is likely larger than if the hazard is detected (the same assumption as in Study 1)
  - ***Failures of safety monitors*** are taken into account as in Study 1.
  - False alarms are NOT used in the model.



The Model

- The ***states*** of the model (of the AV environment) are labelled using three ***binary categories***:
  - ***Location***: {“free”, “intersection”}
  - ***Speed***: {“fast”, “slow”}
  - ***Road hazard***: {“braking”, “non-braking”}
- With these parameters the model has ***8 states*** as shown in the graph on the right:
  - F-F-B (FFB)
  - F-F-NB (FFNB)
  - …
  - IS –S-B (ISSB)
  - IS-S-NB (ISSNB)
- The vehicle spends in each state a certain period of time and then moves to one of the remaining 7 states:
  - A ***fully connected graph*** of transitions between states is a ***plausible*** description
    - Some transitions, however, would be ***illogical*** (and may not occur in a given data set).





The Model (3)

Models of states with “braking” :

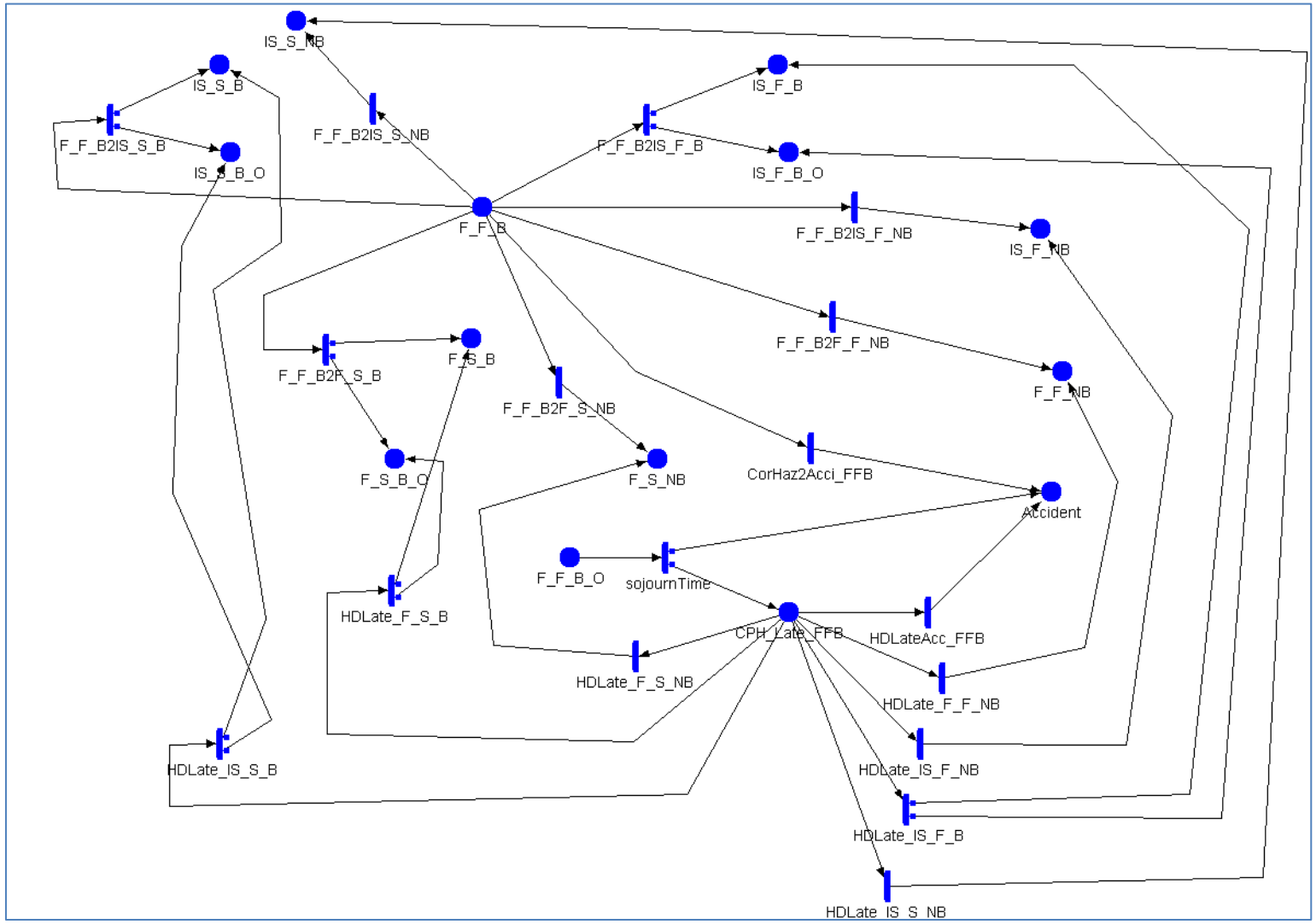
- F-F-B
- F-S-B
- IS-B-B
- IS-S-B

} Similar to the model used in Study 1.

There are transitions:

- to the other 7 states including branching to “sub-states” “-B” or “-B_O”, which indicate if hazards in “braking” states are detected by the perception system or not.
- Transitions to “internal sub-states” corresponding to failures of the perception system.
- Transitions to absorbing state “Accident” are possible.

FFB



The Model (4)

FFNB

Model of the “non-braking” states

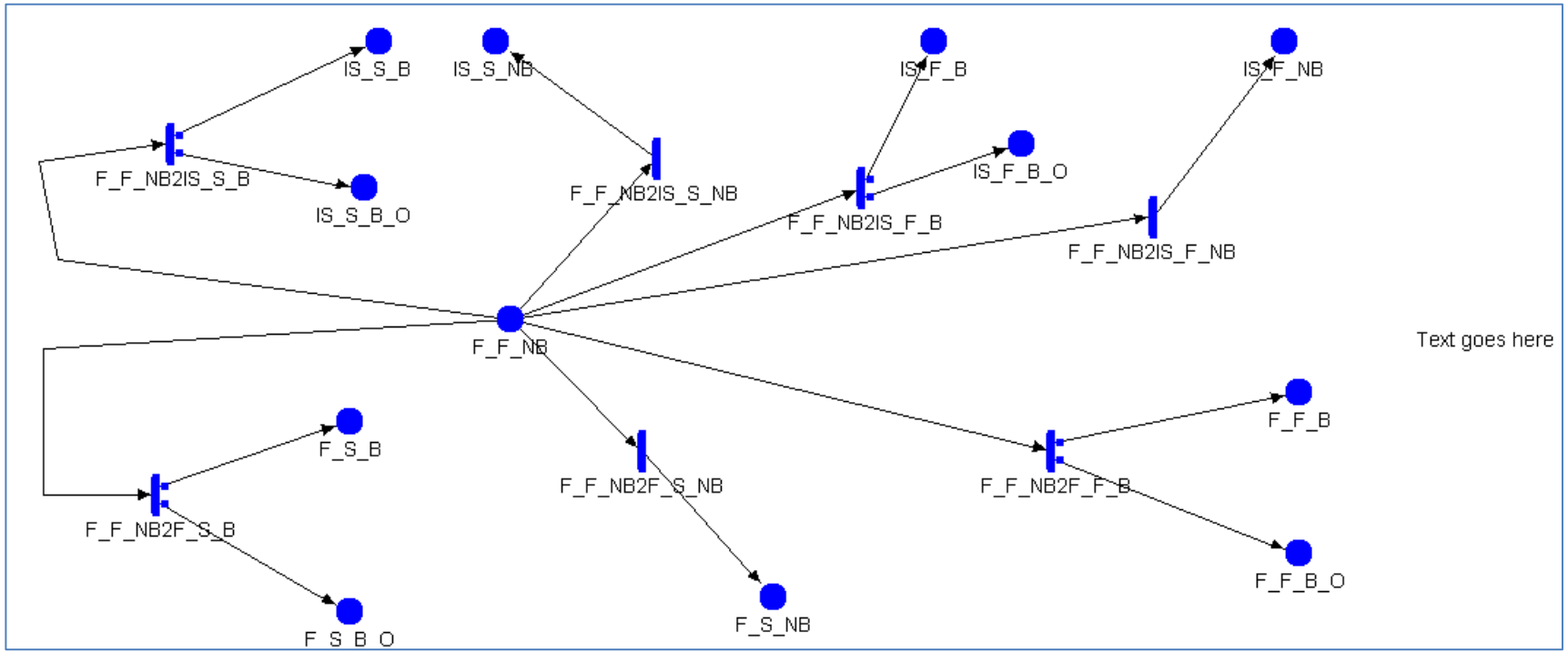
- F – F - NB
- F – S - NB
- IS – B - NB
- IS – S - NB

There are transitions:

- to the other 7 states including branching to “sub-states” “-B” or “-B_O”, which indicate if hazards in “braking” states are detected by the perception system or not.

BUT

- **No transitions to the absorbing state “Accident”** (this is an assumption made in this analysis).



CITY UNIVERSITY LONDON

Safety measure of interest (Reward)

Measure of AV safety is ***the same as with the basic model***: the probability that the system will enter the state “Accident” within predefined intervals of operation:

- 100 hours of operation
- 1100 hours of operation
- …,
- 9100 hours of operation

We discriminate between 2 accident types:

- ***Serious Accident***: an accident when the vehicle drives “fast” (these can result while the model is in states *FFB* and ISFB)
- ***Minor Accident***: an accident when the vehicle drives “slowly” (these can result while the model is in states FSB and ISSB).

CITY UNIVERSITY LONDON

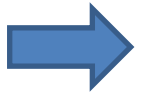
Model Parameters

- Transitions between the 8 states characterise the 8 driving ("operating") conditions the AV can be in.
  - The rates of transition are ***estimated from a proprietary dataset*** (provided by Intel Labs for the study) and are assumed ***known constants*** in all studies conducted with the model.
- Parameters characterising the imperfections of the perception system and of the safety monitors in the 4 "- B" states on the road are difficult to estimate. We opted for ***sensitivity analysis*** on them:
  - Rates of catastrophic failure of the AV when road hazards are correctly and immediately recognised by the perception system. These are parameters (for all "**-B**" states) and characterise the reliability of the "safety monitors" in respective states (operating conditions).
  - Rate of failure of the perception system to recognise a road hazard when it occurs
    - In the model this type of failure occurs upon entering a "braking" state (captured by the ***case probabilities*** leading to the "- B_O" states")
  - Rate of occurrence of an Accident (Serious or Minor) while an existing road ***hazard is overlooked.***
  - Rate of occurrence of an Accident (Serious or Minor) once the road hazard is ***eventually correctly detected*** but with some delay.

Dataset used for model parameterisation

- Parameters of the timed activities which capture the ***transitions between the 8 states*** of the environment are elicited by processing a proprietary dataset, which includes ***statistical data of natural driving by human drivers***.

#start_frame	driving_type	speed_bucket	braking_event	duration
0	free	fast	nobrake	48
48	free	fast	brake	19
67	free	slow	brake	27
94	free	slow	nobrake	21
115	intersection	fast	nobrake	18
133	free	fast	nobrake	18



rate_F_F_B2F_F_NB	609.87
rate_F_F_B2F_S_B	676.71
rate_F_F_B2F_S_NB	319.02
rate_F_F_B2IS_F_B	616.29
rate_F_F_B2IS_F_NB	551.1
rate_F_F_B2IS_S_B	549.09
rate_F_F_B2IS_S_NB	310.79
rate_F_F_NB2F_F_B	170.61
rate_F_F_NB2F_S_B	252.79
rate_F_F_NB2F_S_NB	361.58
rate_F_F_NB2IS_F_B	209.77
rate_F_F_NB2IS_F_NB	202.345
rate_F_F_NB2IS_S_B	361.87
rate_F_F_NB2IS_S_NB	273.315
rate_F_S_B2F_F_B	1080.0
rate_F_S_B2F_F_NB	604.64
rate_F_S_B2F_S_NB	465.98
rate_F_S_B2IS_F_B	0.0
rate_F_S_B2IS_F_NB	439.22
rate_F_S_B2IS_S_B	940.01

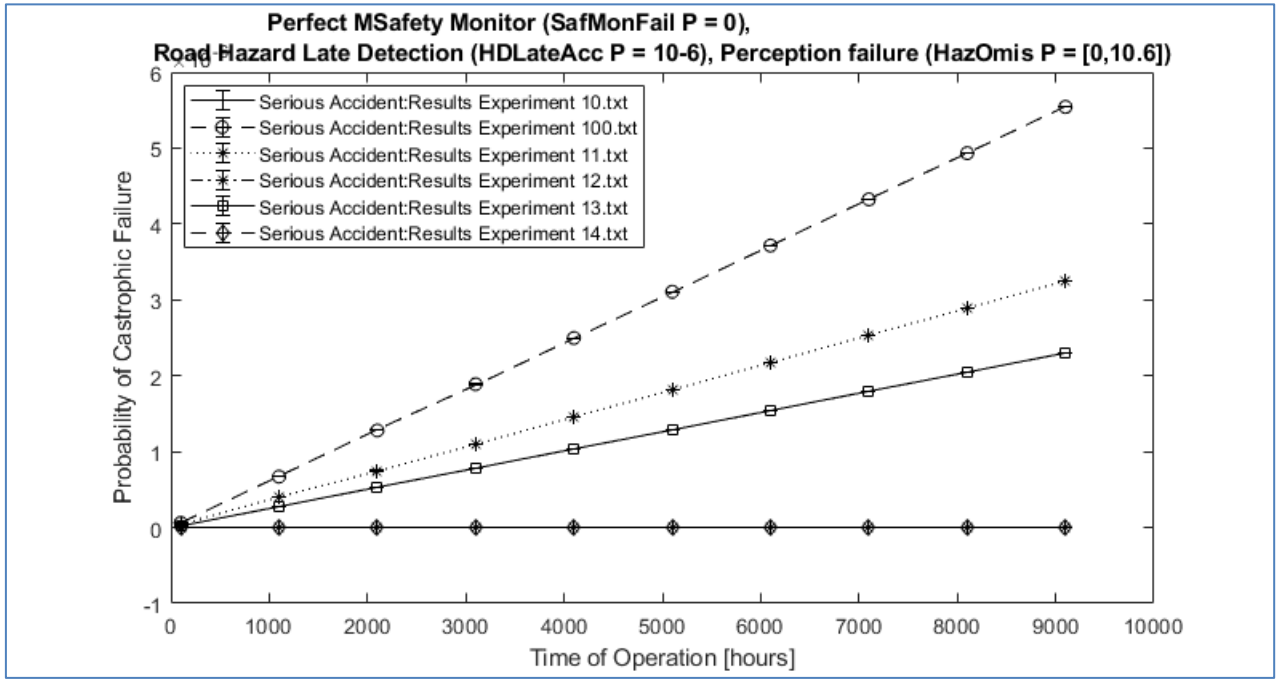
- All timed activities used in the model are assumed ***exponentially distributed***.
  - A ***convenience*** assumption, which allows one to use highly efficient ***numeric solvers*** for the SAN models (part of the Mobius tool)
  - The assumption can be dropped (i.e., use the empirical distributions derived from the dataset). In this case the models will be solved via Monte Carlo simulation.
  - ***Asymptotic solutions*** have also been derived (based on Littlewood's work for semi-Markov systems with an absorbing state) under additional assumptions, which seem plausible for the study.

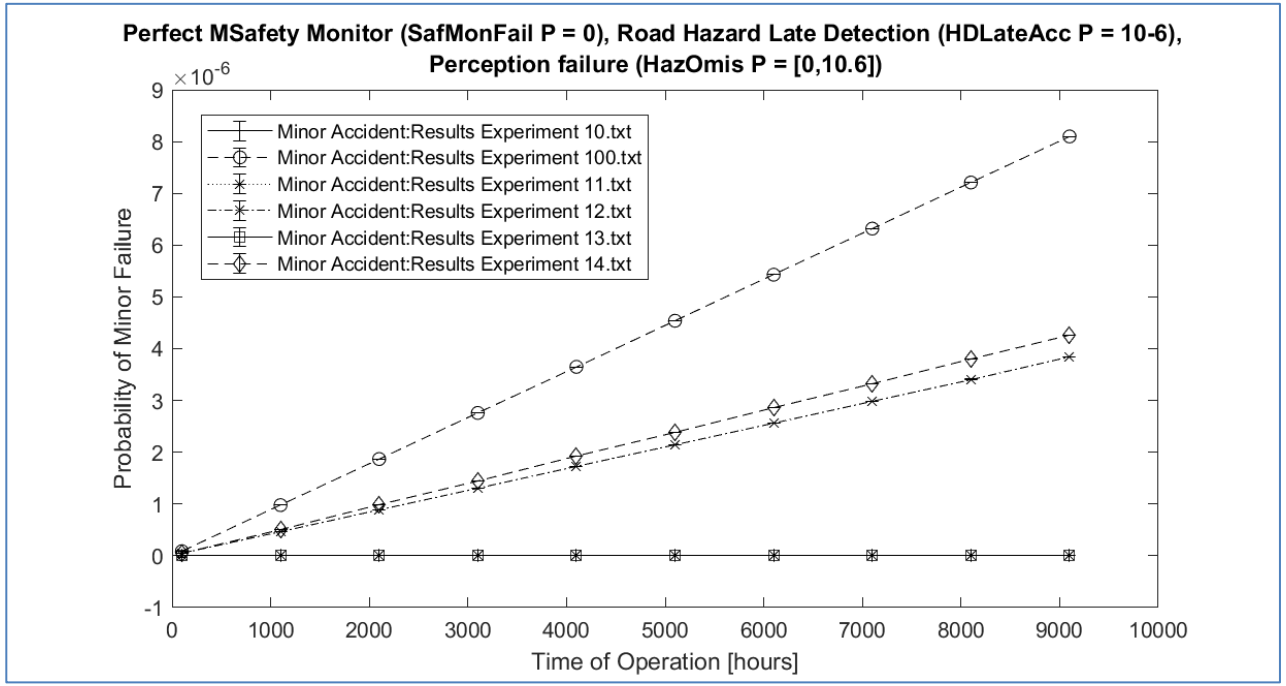
CITY UNIVERSITY LONDON

Results:

Observation 1: Perception failures in different "braking" states affect system safety differently.

	Experiment					
	10	**100**	**11**	**12**	**13**	**14**
F_F_B_O_prob	0.0	1.0E-6	1.0E-6	0.0	0.0	0.0
F_S_B_O_prob	0.0	1.0E-6	0.0	1.0E-6	0.0	0.0
HDLateAcc_FFB_prob	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6
HDLateAcc_FSB_prob	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6
HDLateAcc_ISFB_prob	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6
HDLateAcc_ISSB_prob	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6	1.0E-6
IS_F_B_O_prob	0.0	1.0E-6	0.0	0.0	1.0E-6	0.0
IS_S_B_O_prob	0.0	1.0E-6	0.0	0.0	0.0	1.0E-6
MisHazFailFFB_prob	0.0	0.0	0.0	0.0	0.0	0.0
MisHazFailFSB_prob	0.0	0.0	0.0	0.0	0.0	0.0
MisHazFailISFB_prob	0.0	0.0	0.0	0.0	0.0	0.0
MisHazFailISSB_prob	0.0	0.0	0.0	0.0	0.0	0.0
SafMonFailFSB_prob	0.0	0.0	0.0	0.0	0.0	0.0
SafMonFailISFB_prob	0.0	0.0	0.0	0.0	0.0	0.0
SafMonFailISSB_prob	0.0	0.0	0.0	0.0	0.0	0.0
SafMonFail_FFB_prob	0.0	0.0	0.0	0.0	0.0	0.0



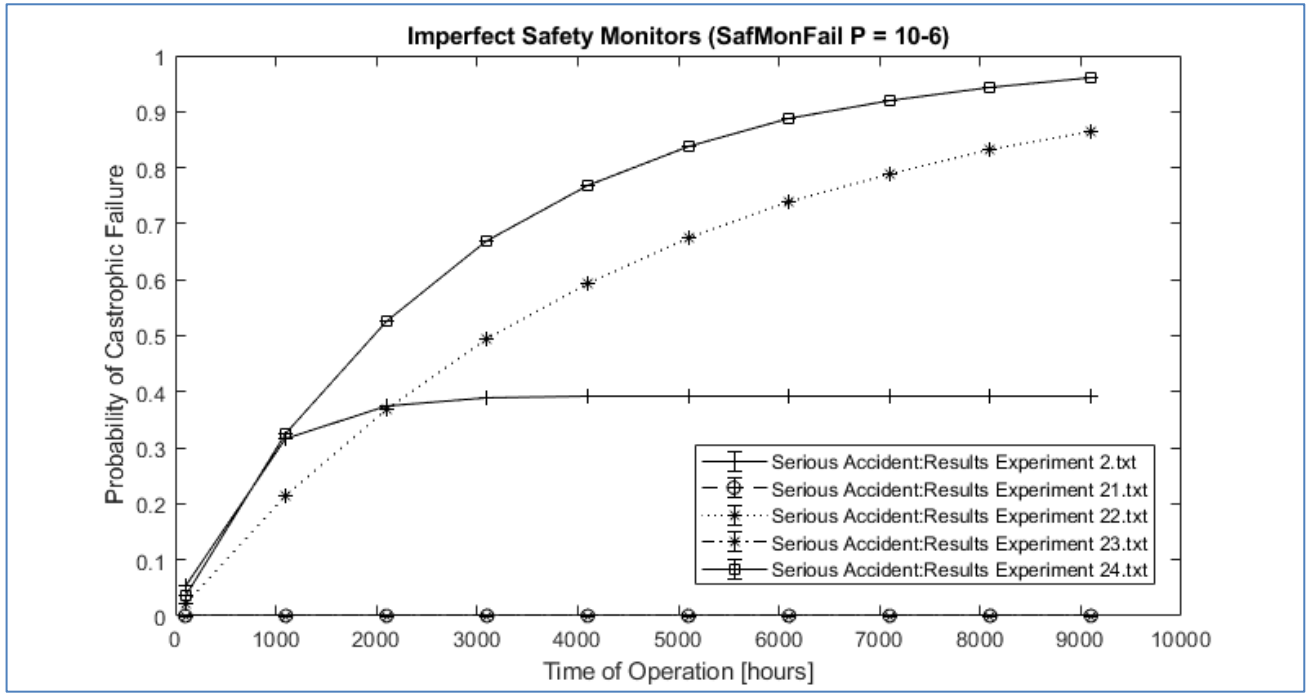


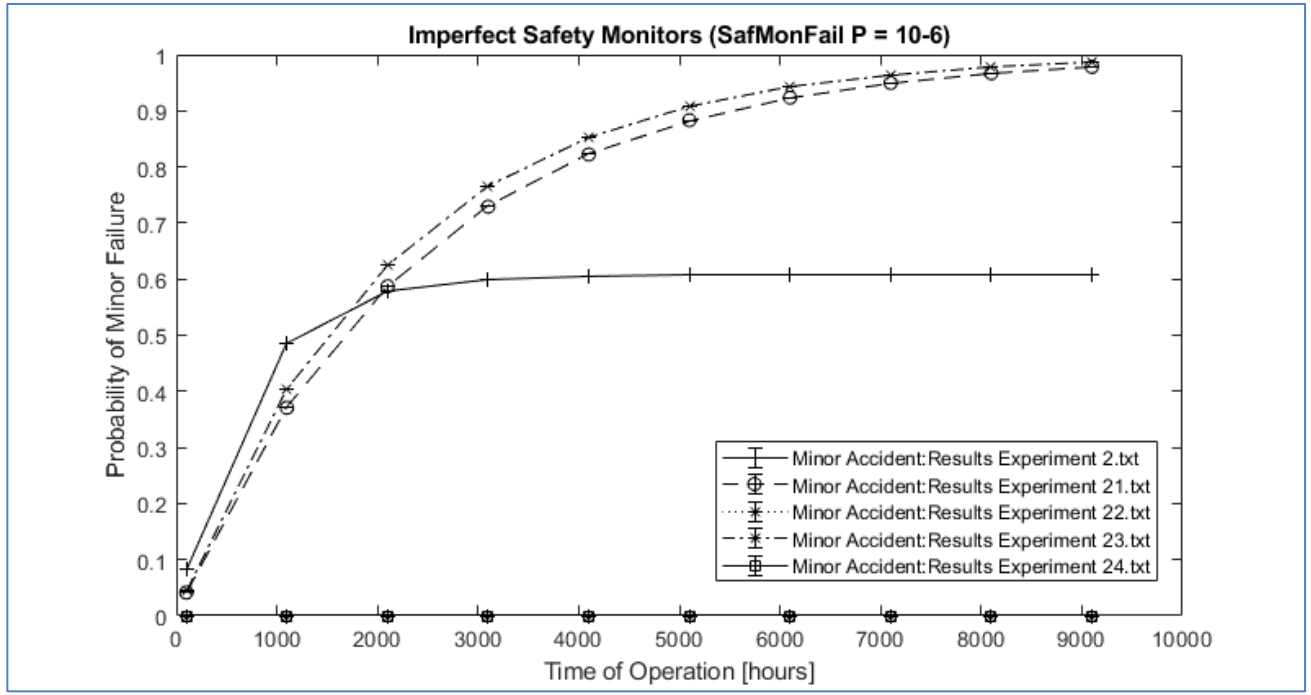
CITY UNIVERSITY LONDON

Results (2)

Observation 2: AV safety is very sensitive to intensity of failure of Safety Monitors .
Confirmation of a well-known fact. Magnitude of the impact on AV safety can be quantified.

	Experiment				
	2	**21**	**22**	**23**	**23**
F_F_B_O_prob	0.0	0.0	0.0	0.0	0.0
F_S_B_O_prob	0.0	0.0	0.0	0.0	0.0
HDLateAcc_FFB_prob	0.0	0.0	0.0	0.0	0.0
HDLateAcc_FSB_prob	0.0	0.0	0.0	0.0	0.0
HDLateAcc_ISFB_prob	0.0	0.0	0.0	0.0	0.0
HDLateAcc_ISSB_prob	0.0	0.0	0.0	0.0	0.0
IS_F_B_O_prob	0.0	0.0	0.0	0.0	0.0
IS_S_B_O_prob	0.0	0.0	0.0	0.0	0.0
MisHazFailFFB_prob	0.0	0.0	0.0	0.0	0.0
MisHazFailFSB_prob	0.0	0.0	0.0	0.0	0.0
MisHazFailISFB_prob	0.0	0.0	0.0	0.0	0.0
MisHazFailISSB_prob	0.0	0.0	0.0	0.0	1.0E-6
SafMonFailFSB_prob	1.0E-6	1.0E-6	0.0	0.0	0.0
SafMonFailISFB_prob	1.0E-6	0.0	1.0E-6	0.0	0.0
SafMonFailISSB_prob	1.0E-6	0.0	0.0	1.0E-6	0.0
SafMonFail_FFB_prob	1.0E-6	0.0	0.0	0.0	1.0E-6





CITY UNIVERSITY LONDON

Summary of the findings from state-based models

- The proposed model-based approach to safety assessment of AV is ***feasible*** and provides useful insight:
  - Despite complexity, ***many model parameters*** can be derived using ***publicly available datasets***.
  - "Sensitivity analysis" can be deployed for those parameters, which cannot be estimated
    - the impact of reliability of essential components such as AV perception and safety monitors on AV safety can be quantified.
    - The findings from the sensitivity analysis can be used to establish ***reliability requirements*** on the AV perception and safety monitors.
- Feasibility of the method is demonstrated on two studies which operate at different level of abstraction. Some of the ***observations are non-obvious:***
  - Duration of overlooking a hazard (e.g., initially omitting an obstacle on the road) for a short time (1 – 3 sec) does not seem to affect visibly the AV safety;
  - False alarms (studied in the basic model), do not worsen the system safety visibly, either.
- Perception system's reliability affects system safety ***significantly***.
- Unreliability of "safety monitors" has the ***most dramatic impact*** on AV safety.
  - The results demonstrate a constructive way of establishing reliability targets on "safety monitors" so that AV safety can be assured.

CITY UNIVERSITY LONDON

EST 1894

Work in progress on AV scenario – based testing for safety assessment

CITY UNIVERSITY LONDON

Scenario – based testing for AV safety assessment

- "Driving to safety" criticised as ***very expensive*** hence "infeasible" (10s to 100 million km/miles needed to reach a claim of safety comparable with the safety of a vehicle driven by a ***typical driver***).
  - Waymo are reporting 56+ million of miles already and make claims that their Cruiser is safer than a human driver.
- Scenario – based testing (SBT) has been adopted as a "feasible" (and much cheaper) alternative
- ***My understanding*** of SBT is that it is based on a ***number of assumptions*** without justification.
  - An international ***standard(s)*** on scenario-based testing already exist (e.g., ISO 34502)
  - ***Key assumption:*** Testing on "difficult scenarios" sufficient to demonstrate AV safety. ***No need*** to test on "easy scenarios".
    - "Difficult" scenarios derived from real scenarios by augmenting them ("synthetic scenarios")
  - Certification (e.g. in EU) will be based on a set of scenarios provided by the (EU) regulator (EU Regulation 2022/1426).
  - Each scenario is ***run on at least one platform*** (from the list: real driving, "proving grounds" or simulators)
  - Simulators' behaviour must be compared against real driving/"proving grounds"
    - Unclear what this comparison with reality should include and how the findings (e.g. faults in simulators) should be used in interpreting the simulators results. References to ***general software quality standards***.
  - It is unclear what safety claim will (or can) be made from scenario-based testing.
    - Will the claim(s) be comparable with those one can derive via "driving to safety"?
    - Will ***confidence in the claim*** be explicitly stated (as is done with "driving to safety")?
      - A few tests will only produce low confidence. How will ***high confidence in high safety*** be achieved?

CITY UNIVERSITY LONDON

Scenario – based testing: Research questions

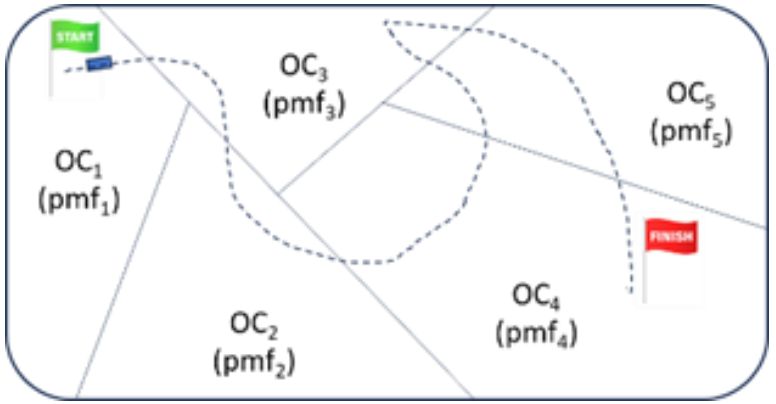
- ***RQ1:*** Is scenario-based testing a credible alternative to AV safety assessment, capable to deliver a safety claim (with high confidence) comparable to a "driving to safety" claim? (i.e. that an AV is comparable or even safer than a vehicle driven by a typical human driver).

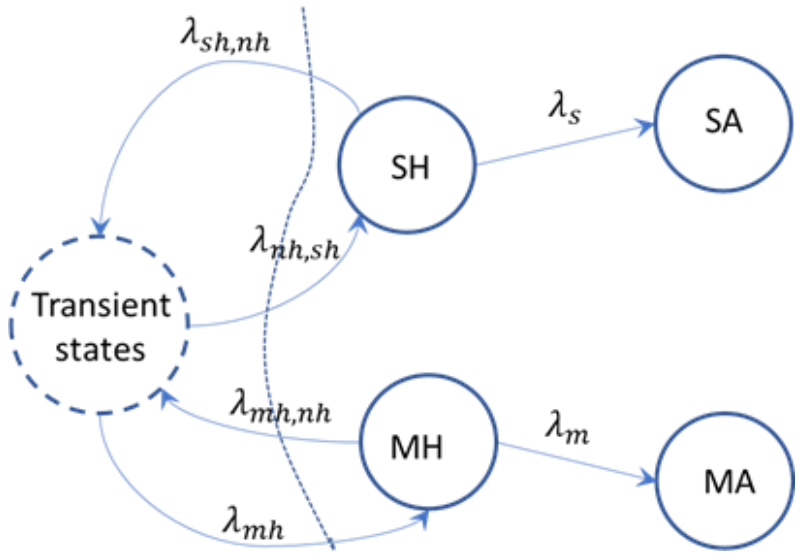
- ***RQ2:*** If the current practice of safety assessment via scenario – based testing does not allow a justifiable claim with high confidence to be made, what are the ***barriers*** that need to be addressed?

Own work on Scenario – based testing: Challenges

Early days of this work:

- C1. Stochastic model (e.g. based on ***trajectories*** via different operating conditions of a given ODD)
  - An approach similar to the work with Intel Labs is taken:
    - A large stochastic model (under development) will capture the movement of an "ego" AV via conditions in ODD.
    - The model will have ***many parameters*** (difficult to assess, the values likely to change over time)
  - Littlewood's ***asymptotic results*** for semi-Markov processes with small rates of transition to failure (accident) seems very relevant.
    - The time to accident with such models is asymptotically ***exponentially distributed*** (for arbitrary sojourn times) and only depends on a much smaller set of parameters, estimating which seems *feasible* by monitoring AVs.





$$\lambda = \sum_1^n \lambda_i P(S_i)$$

Littlewood, B., *A reliability model for systems with Markov structure.* J. Roy. Statist. Soc., Ser C (Applied Statistics), 1975. **24**(2): p. 172-177.

CITY UNIVERSITY LONDON

Scenario – based testing: Challenges (2)

- C2. An inference procedure for a non-homogeneous Poisson process is needed:
  - ***Variable operational profile***.
  - Conditional rates of transition to an ***absorbing state Accident*** likely to vary, too (inference is needed).
  - ***Similarities with the multivariate model*** developed for "driving to safety". Need to work out details.
    - The key expression of interest is the Littlewood's ***rate of accident***, which itself is a sum of random variables.
    - Under certain conditions this sum is a direct replacement of the ***total probability of accident*** per mile of driving.
- C3. Imperfections of simulators: ***fault-negatives*** (missing accident in simulators) ***particularly problematic***.
  - Assessing these imperfections credibly and, ***more importantly***, factor this imperfection in the inference seems essential for a credible safety assessment using simulators.
- C4. Defining noteworthy "difficult scenarios" is a ***best effort activity*** (i.e. some can be overlooked). What is the risk for safety assessment of overlooking important "difficult scenarios"?
  - A "blended" safety assessment (i.e. a mixture of road driving, tests on proving grounds and testing with simulators) should be used so that the risks are reduced to an acceptable level (ideally with a minimal effort).
    - I am not aware of any work trying to address this challenge. Looking for collaborators.

CITY UNIVERSITY LONDON

Conclusions

- AV safety assessment and assurance are under very *active* development.
- AV safety assurance, based on the traditional safety justification, is currently ***problematic***.
  - Reliance on "expert judgment" is problematic, especially given that critical AV components such as perception, planning, etc., rely on machine learning (ML). Such components often lack a formal specification and may be evolving over time (e.g., via retraining) , hence cannot be subjected to the scrutiny typical for safety engineering.
- The trends in safety assurance (e.g., the aforementioned Assurance 2.0) are towards ***using models*** and ***automation in safety reasoning*** more extensively than is done now.
- Probabilistic models (state-based, Bayesian inference, etc.) seem particularly ***well-suited*** to face the challenges of AV safety assurance:
  - A probabilistic model of the entire AV as a system, placed in hazardous environment, allows one to assess the impact of reliability of different sub-systems (perception system reliability, safety monitors, etc.) and establish "reliability targets" for each of these sub-systems which will allow for safety claims to be justified.
    - Often work on a particular component (e.g., perception), is undertaken to assess ***if it is safe*** (e.g., ANSI UL4600 seems to recommend this view) ***without the system context*** (of the entire AV or indeed of the entire ecosystem of vehicles on the road both AVs and man-driven). ***Safety is a system property***.
  - Assessing the achieved levels of safety and confidence by an AV from operational data is possible with Bayesian inference. The inference, however, should rely on a *credible model* aligned with the applicable ODD.
  - Scenario based testing is attractive but needs further work to allow for a confidence to be added to a safety claim.

CITY UNIVERSITY LONDON

Recently completed papers

1. 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, 2025, (under review): p. 13.
2. Popov, P., "*Dynamic safety assessment of Autonomous Vehicle based on Multivariate Bayesian Inference (DyAVSA)*", Journal of Reliable Intelligent Environments, 2025, p. 28, https://link.springer.com/article/10.1007/s40860-025-00252-4.
3. Popov, P., Strigini, L., Stankovic, V., "*On Estimating the Effectiveness of Run-Time Error Detection*", to be submitted to IEEE TDSC, 2025 (under internal review), p.28.
4. Buerkle , C., Oboril, F., Paulitsch, M., ***Popov, P.*** and Strigini, L., "Modelling road hazards and the effect on AV safety of hazardous failures". IEEE 25th International Conference on Intelligent Transportation Systems (ITSC'2022). Macau, China**:** 1886-1893.
5. ***Popov, P.*** , Strigini, L., Buerkle, C., and Oboril, F., "Stochastic Modeling of Road Hazards on Intersections and their Effect on Safety of Autonomous Vehicles", IEEE Trans on Intelligent Transportation Systems, 2025 (under review). p. 18., . arXiv:2506.02688. ArXiv. CMU arXiv, https://arxiv.org/abs/2506.02688.
6. Zhao, X., Aghazadeh-Chakherlou, R., Cheng, C.-H., ***Popov, P***., and Strigini, L. (2025) *On the need for a statistical foundation in scenario-based testing of autonomous vehicles.* In: The IEEE International Conference on Intelligent Transportation Systems (ITSC), Gold Coast, Australia, 18-21 Nov 2025. Published in: International Conference on Intelligent Transportation (ITCS'2025), Gold Coast, Australia.
7. Aghazadeh-Chakherlou R., Guo Q., Khastgir S., ***Popov P.***, Zhang X, Zhao X., "*A Hierarchical Imprecise Probability Approach to Reliability Assessment of Large Language Models*", Reliability Engineering and System Safety, vol. 272, part 2, 29 p, Article number 112615, Mar 2026.

CITY UNIVERSITY LONDON

EST 1894

Questions

Thank you!