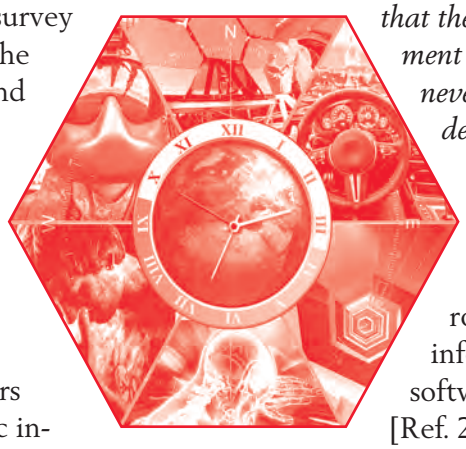
Harnessing Uncertainty in Autonomous Vehicle Safety

by Stephen L. Thomas and Dirk J. Vandenberg
Pittsburgh, Pennsylvania

Safely developing self-driving vehicles presents technical challenges. Among the key technical challenges are how to confidently demonstrate the safety of a self-driving vehicle when the number of permutations of operating conditions, scenarios, system inputs, etc. are complex, uncertain and potentially limitless. This paper provides a broad survey of the various types of uncertainty in the development of self-driving vehicles and outlines several possible strategies for handling uncertainty. Advantages and challenges of different approaches, including qualitative and quantitative methods, are also discussed.



Introduction

During the past decade, self-driving cars have rapidly moved from the academic incubator into commercial enterprise. Currently, dozens of companies, from major corporations to start-ups, have self-driving vehicles (SDVs) on the road, albeit with some form of human supervision. Many challenges — including situations where the technology didn't exist or the solution was computationally infeasible — have now been solved. However, the ultimate challenge of producing a fully autonomous car that is demonstrably safer than a car with a human driver has yet to be achieved.

The automotive industry, and other industries, have embraced the safety case approach to producing a coherent and convincing argument for the safety of an item based on evidence. A safety case is required in instances where the automotive ISO 26262 functional safety standard and the new ISO 21448 safety of the intended functionality (SOTIF) standard are applicable. In this paper, we provide a survey of the role of uncertainty in safety assurance, including the critical role of the safety case in identifying and reducing uncertainty.

Background and Literature

In their influential paper, Bloomfield made the case that uncertainty management lies at the heart of safety case arguments:

> *[A] dependability case is taken to be some reasoning, based upon assumptions and evidence, allowing certain confidence to be placed in a dependability claim. For a given claim (e.g., Probability of Failure on Demand (pfd) is smaller than 10^{-3}), the confidence — and its complement, doubt — will depend upon confidence/doubt in the truth of assumptions, in correctness of reasoning, and in "strength" of evidence. A key notion here is the recognition that there is uncertainty involved in the assessment of system dependability: It is (almost) never possible to claim with certainty that a dependability claim is true* [Ref. 1].

The above argument comes from a line of thought that has developed over several decades and was rooted in earlier research showing the infeasibility of validating safety-critical software systems via direct evidence alone [Ref. 2 & 3].

The need for practical validation methods led to the further development of formal theories of multi-legged arguments in safety cases. Analogous to the benefit of diverse redundancy in hardware, diversity in argument and evidence allows each leg of a safety case argument to compensate for weaknesses in the other leg. Bishop identified three different types of arguments [Ref. 4], while Habli cited several sources to list potential types of evidence [Ref. 5]. These results are summarized in a simple taxonomy in Table 1.

A review of Table 1 reveals that essentially all of the items are subject to uncertainty of one form or another. Some are clearly subject to aleatory uncertainty (e.g., probabilistic arguments, direct evidence, historic data, test data), while even non-stochastic arguments are subject to epistemic uncertainty (e.g., deterministic arguments, proof evidence). Additionally, all of these items are subject to the uncertainties of human performance and human judgment. It is important to emphasize that there may be uncertainty in the claims, the inference and/or the evidence of the safety case.

One mechanism to address uncertainty in the safety case is to identify and mitigate the so-called "defeaters" of the argument. Goodenough et al identify three types of defeaters [Ref. 6]:

- **Rebutting** — Providing information that contradicts the claim

Table 1 — Types of Arguments and Evidence.

Levels of Arguments	Types of Arguments	Types of Evidence	Types of Evidence
Product Process	Deterministic Probabilistic Qualitative	Direct Process Qualitative Historic Test Proof	Relevance Independence Trustworthiness

- **Undercutting** — Specifying conditions under which the claim is not necessarily true, even if the premises are true
- **Undermining** — Invalidating one or more of the premises

We will use the categories in Table 1 as a guide for identifying the types of uncertainty and show how potential defeaters can be enumerated and mitigated. The purpose of this paper is not to outline or discuss an autonomous vehicle safety case in its entirety. Rather, we give an overview of the various sources of uncertainty that may present themselves and discuss how they may be identified, assessed and/or reduced during the development of the safety case.

Types of Uncertainty

As a framework for discussion, we present an example of a highly simplified pattern for an autonomous vehicle safety case. The example safety case shows two legs — one a scenario-based argument and the other a hazards-based argument. This breakdown is generally inspired by the ISO 21448 SOTIF approach and the ISO 26262 functional safety approach, respectively.

Our intent is not to argue the merits of different approaches to safety case structure, but rather to focus on the lower levels of the diagram, i.e., the sub-goals and the evidence.

As with Table 1, there is clearly uncertainty associated with the various sub-goals and their associated

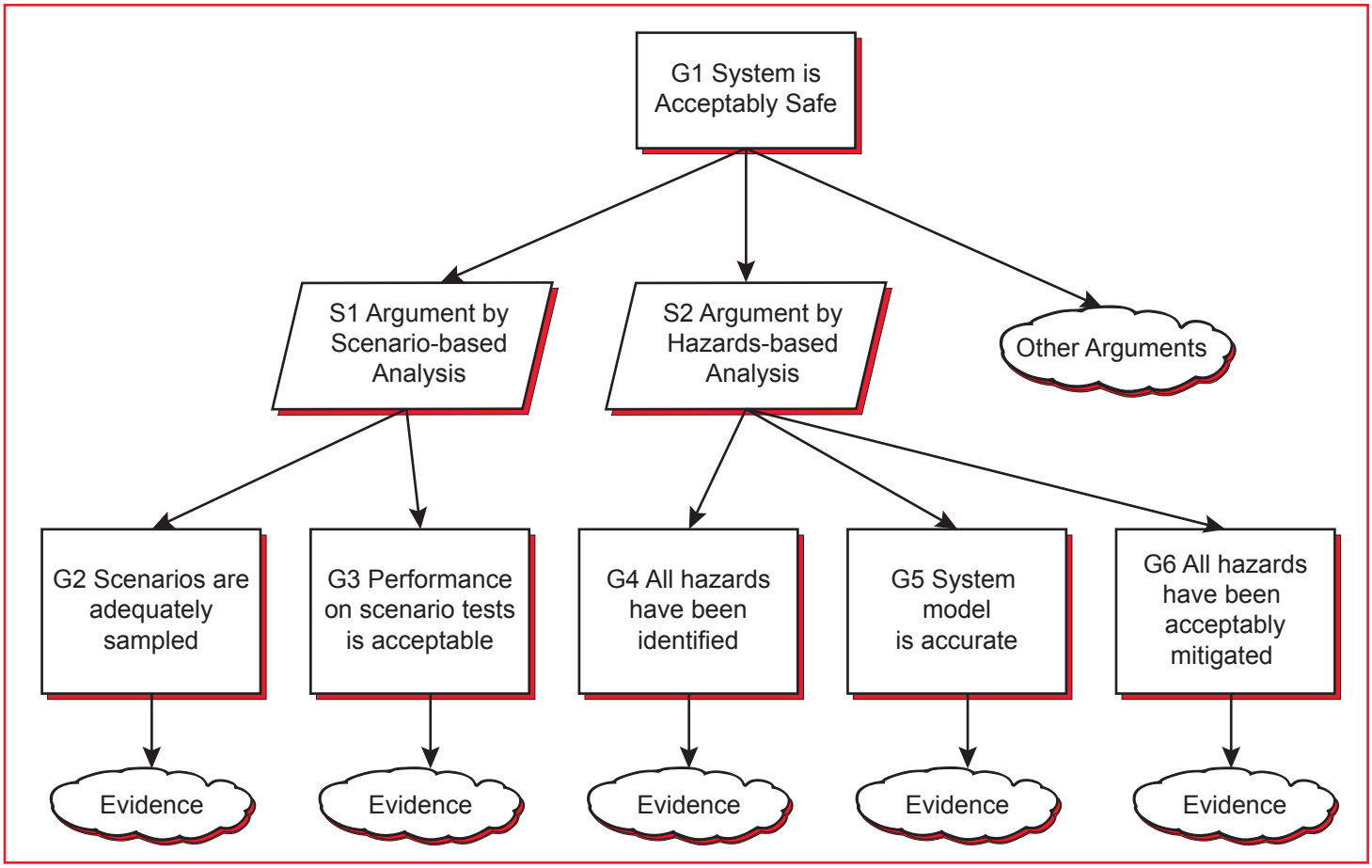


Figure 1 — A Simplified Safety Case.

Table 2 — Potential Features of Operational Design Domain.

Category	Instance Examples
Vehicle Actors	Cars, Pickups, SUVs, Mini-vans, Vans, Motorcycles, Motor scooters, Trucks, Buses, School buses, Trailers, RVs, Farm equipment, Trains, Trolleys, Police Vehicles, Emergency Vehicles
Non-vehicle Actors	Adult pedestrian, Child pedestrian, Cyclists, Strollers, Wheelchairs, Equestrians, Skateboarders, Scooters, Domestic animals, Wild animals
Non-static Objects	Parked vehicles, Trash cans, Trash bags, Road debris, Plastic bags, Dust, Steam, Exhaust, Smoke, Potholes, Standing water, Signs, Umbrellas, Billboards, Banners, Graffiti
Static Objects	Buildings, Lamp posts, Utility poles, Trees, Bushes, Mailboxes, Fences, Bus stops, Overpasses, Guard rails, Walls, Statues, Fountains, Bike racks, Parking meters, Fire hydrants, Bollards
Road Properties	Markings, Paving, Roughness, Wetness, Iciness, Color, Width, Curvature, Grade
Road Configurations	Lanes, Turn lanes, HOA lanes, Bike lanes, Reversible lanes, Speed Limit, Minimum Speed, Variable Speed Limits, Traffic Signs, Traffic Signals, Intersections, Roundabouts, Crosswalks, Driveways, Merges, Speed bumps, Construction zones, Temporary Traffic Signals, Toll booths, Accident scenes, Human traffic directors, Parking lots, Parallel Parking, Angled Parking, Tunnels, Bridges, Railroad crossings, Dividers, Islands, Shoulders, Parades
Environment Conditions	Day, Night, Sunrise, Sunset, Sunny, Rain, Snow, Hail, Sleet, Fog, Extreme Cold, Extreme Heat, Flooding, High Wind, Dusty, Insects
Observable Behaviors	Turning, Stopping, Yielding, Merging, Parking, Passing, Nudging, Pulling Over, Turn Signals, Hazard Lights, Brake Lights, Post-accident Behaviors, Speeding, Erratic Driving, Cutting Off, Pick-up, Drop-off

evidence. To conveniently frame the discussion, we will categorize the types of uncertainty as follows:

- **Domain Uncertainty** (G2 and G5) — Confidence that our model of the environment, road configurations, actor behaviors, actor characteristics, associated hazards, etc. is adequate
- **Scenario Uncertainty** (G2, G4, G5, G6) — Confidence that all models, simulations, samples, etc. adequately represent actual scenarios arising from the parameter space
- **Performance Uncertainty** (G3, G5, G6) — Confidence that the performance of the system under nominal and abnormal conditions is adequate in relevant scenarios in the domain

In the following sections, we will survey each of these categories of uncertainty, citing specific examples and discussing alternatives for assessing and mitigating these uncertainties.

Domain Uncertainty

Before we can consider individual operational scenarios, it is necessary to look at the potential features of the Operational Design Domain (ODD). Table 2 shows a sampling of some of the features and properties that might be found in a typical ODD. This list is neither comprehensive nor particularly scientific. It is simply meant to illustrate the wide array of variation found when driving in the real world. Note that this table does not even delve into the variability of each instance type, e.g., the infinite variability of pedestrian appearance.

The amazing complexity and variability of real-world driving is one reason why machine learning, and specifically deep learning algorithms, can provide valuable contributions to the development of self-driving cars. However, real-world complexity strains even the cleverest neural networks and today's fastest self-driving computers. This so-called state-space explosion cannot be solved entirely by management techniques, but these techniques can significantly reduce complex-

> "A convolutional neural network may demonstrate excellent performance in pedestrian detection, but will it successfully detect a pedestrian partially occluded by a mailbox on a snowy day at sunset? What if the pedestrian is wearing a hat? A winter coat? What if a pedestrian is wearing a full-body Wookie suit for a cosplay convention? This last one was an actual scenario encountered during SDV testing!"

ity. In this section, we briefly outline some of the ways ODD uncertainty can be corralled.

Operational Domain Constraints — The general public may be familiar with the practice of geofencing, or electronically restricting the area where an SDV may operate. However, the purpose of ODD constraints is not to limit driving areas, but to help limit uncertainty. For example, an ODD may be chosen to limit high-risk areas (e.g., school zones, high pedestrian density), high-risk maneuvers (unprotected left turns, high speeds) or conditions (e.g., night, snow). Some ODD constraints are simple rules (e.g., "don't drive at night"), but others require detailed characterization of the ODD, including real-time updates as conditions change. The challenge of continually updating and revalidating the ODD characterization may make it a non-optimal approach for deploying SDVs at scale over the long term. In the near term, characterizing and limiting the geographic ODD is critical, but it is just one means to the end of narrowing the distribution of potential driving scenarios.

High-definition Maps — Even with the ODD limited, the wide variety of potential scenes and actors is a computing challenge, including the difficulty of separating static objects from potential actors in the scene. High-definition maps can a priori identify road features (e.g., lanes, dividers, speed limits), static road objects (e.g., signals, signs), and static background objects (e.g., trees, buildings, mailboxes). Once again, this approach requires detailed characterization of the ODD, and the characterization must be evergreen since even so-called static objects change with time (e.g., seasonal growth in vegetation). One of the other challenges of the characterization is that it requires labelled data, either by labor-intensive human labeling, error-prone automation or a combination of both. Despite the challenges, even imperfect a priori segmentation improves operational uncertainty and reduces the real-time computing load on the SDV.

Vehicle Operators — No discussion of ODD uncertainty would be complete without highlighting the role of the vehicle operator. Much has already been written about the challenges and limitations of human performance in the development of SDVs [Ref. 7]. However, as of the writing of this paper, every SDV company (of which we are aware) that is operating a vehicle on public roads has relied on a vehicle operator during aspects of the development cycle. This caution is with good reason. While the methods above narrow the distribution during development and testing, the human operator can help provide important protection for the low-probability tails of the ODD distribution, at least during initial stages of SDV testing. They may help the SDV navigate unusual scenarios, such as ODD characterization errors, traffic accidents, etc.

These interventions also need not be purely reactive. Vehicle operators may be trained to pre-emptively disengage the SDS at specific locations or under certain high-risk or complex scenarios. These planned disengagements may allow expansion of the ODD while controlling the risk profile. This approach is one reason why many have argued that SDS disengagement metrics are a poor proxy for SDS safety.

Scenario Uncertainty

The characterization and constraining of the ODD are primarily focused on observable features, such as road configurations, speed limits, object types, etc. Even if the feature list were completely bounded, the potential behaviors and interactions of the features cause another form of unbounded uncertainty.

A convolutional neural network may demonstrate excellent performance in pedestrian detection, but will it successfully detect a pedestrian partially occluded by a mailbox on a snowy day at sunset? What if the pedestrian is wearing a hat? A winter coat? What if a pedestrian is wearing a full-body Wookie suit for a cosplay convention? This last one was an actual scenario encountered during SDV testing!

Clearly, a brute force sampling of the features in Table 2, along with behaviors and interactions, could quickly lead to a combinatorial explosion. Strategies are needed for harnessing this type of uncertainty as well.

Scenario Taxonomy and Completeness — Scenarios and interactions largely depend on the behavior of other actors and are therefore extremely difficult to control in practice. A potentially more tractable approach is to identify a taxonomy of potential scenario families with allowable variations within families. Much data is available to support this approach from government reports and academic naturalistic driving studies. A key challenge in this approach is understanding and avoiding overconfidence in the completeness of the taxonomy. Although road testing can certainly help identify missing scenarios, it is likely that only rigorous simulation will be able to unearth more infrequent scenarios.

Ranking and Importance-based Sampling — While it is desirable to have a taxonomy that accurately represents a wide array of real-world driving scenarios, it does lead to a paradox. A theoretically complete taxonomy of every possible driving scenario is not only impossible to build, but also impossible to test. The compromise is that the taxonomy must be *adequately* complete and the variations *adequately* detailed. Scenario families and level of variation may be prioritized based on scenario-relative frequency, relative risk, difficulty for human drivers, or difficulty for the SDV. Judicious use of these strategies may improve efficiency of testing efforts, but it is critical to ensure the scenario space is not under-sampled. Unfortunately, determination of the thresholds for sampling adequacy is largely a trial-and-error process.

“A theoretically complete taxonomy of every possible driving scenario is not only impossible to build, but also impossible to test. The compromise is that the taxonomy must be adequately complete and the variations adequately detailed. Scenario families and level of variation may be prioritized based on scenario-relative frequency, relative risk, difficulty for human drivers, or difficulty for the SDV. Judicious use of these strategies may improve efficiency of testing efforts, but it is critical to ensure the scenario space is not under-sampled.”

Scenario-independent Safeguards — The methods above are intended to limit and/or adequately sample the scenario parameter space. With any new technology, it is important to recognize that there are both “known unknowns” and “unknown unknowns.” Hypothetically, one way to address the latter would be to build simple safeguards that need no deep understanding of complex scenarios. Such a system could put a lower bound on system behavior for edge-case scenarios. One example of such a system provides simplified rules for trajectory validation to ensure the SDV does not route through known obstacles [Ref. 8]. The same principle has been applied to other portions of the SDS stack.

Unfortunately, these approaches are not a panacea, and they often suffer from a performance trade-off between false negatives and false positives. The performance trade-off may be even more difficult for these simplified safeguards, since by their nature, they lack the sophistication of the full SDS. Performance uncertainty is discussed in the next section.

Performance Uncertainty

The discussion so far has centered on understanding and controlling the mission of the SDV, but we now turn to the SDV system itself. A theoretically complete and performant SDV would not need ODD constraints or scenario-based simulation. It is necessary to acknowledge the current limitations of the technology in order to understand performance uncertainty and ensure safety.

SOTIF and the Confusion Matrix — Much has been made of the stochastic nature of machine-learned (ML) deep neural networks. Indeed, there is often a clear

trade-off at design time between false positive rate and false negative rate. However, this tradeoff is by no means unique to ML systems. In traditional functional safety applications, there is often a conscious trade-off between safety and operational reliability. For example, a 1 out of 2 (1oo2) voted system is far safer than a 2oo2 voted system. A 1oo3 system is theoretically safer than a 2oo3 system, but the latter is widely used in critical applications such as nuclear power plants.

The problem is not that ML applications are stochastic by nature, but that their uncertainty is confounded by the uncertain world of ODD scenarios. The neural network may achieve a certain selectivity or recall *on average*, but it is necessary to understand the performance in the context of ODD scenario sampling discussed previously. An algorithm that performs well on a wide variety of scenario families may fall short in unusual circumstances. Detecting and handling these anomalous situations is an active area of research, including so-called "safety cages" for neural networks that may detect when the ML application is overconfident and trigger a more deterministic safe response [Ref. 9].

Component Failures and Degraded Operation — Rigorous analysis of component failures is a staple of safety engineering and remains valuable for SDVs. However, the traditional analysis is complicated by the more sophisticated sensors in the SDV, including high-definition cameras, LIDAR and RADAR. Analysis of basic failures remains the same (e.g., no output), but the devil is in the details of partial failures and performance degradation.

A complete sensor failure analysis would include the degradation effects of ODD scenario variations, including lighting, precipitation, fog, cleanliness, etc. It would likely also be necessary to evaluate all credible permutations of the different conditions. Perhaps the most challenging aspects of the evaluation would be 1) ensuring that the SDV can detect when it is in a degraded state and 2) selecting appropriate degradation thresholds that responsibly balance safety and operational reliability goals.

Generational Performance Uncertainty — Building a safe and performant SDV, selecting and characterizing an ODD, and validating performance across a spectrum of scenarios are each a challenging undertaking. Unfortunately, none of them is a one-time task. As SDV technologies continually advance, there is a challenge to manage all forms of uncertainty across generations of technologies, including sensors, computing, middleware, applications and vehicle platforms. It is important that the investment of time and effort not be lost with each new generation.

One way to minimize generational loss is to use a modular safety case that is incrementally updated as components and modules change. This approach by itself is not entirely satisfactory, since the safety-case argument structure is largely qualitative and provides no inherent mechanism to combine inter-generational information. Advanced Bayesian techniques provide an intriguing alternative that allows quantitative knowledge, including uncertain and partially applicable knowledge, to be propagated across multiple generations [Refs. 10 & 11].

Conclusion

There is much existing literature making the case that safety assurance can be viewed as the process of systematically removing different forms of uncertainty in system and mission analysis. The safety-case structure is well established as a preferred method for systematically explicating and documenting mitigation of these uncertainties. A simplified safety-case structure was presented, focusing on aspects of uncertainty

This paper provided a broad survey of the various types of uncertainty encountered in the novel domain of self-driving vehicle development. The general uncertainty categories in the example safety-case structure were mapped to three main categories: 1) Domain Uncertainty, 2) Scenario Uncertainty and 3) Performance Uncertainty. Examples of practical uncertainties were given, along with potential mitigation strategies, including a discussion of the challenges and limitations of each approach.

The field of self-driving vehicle safety assurance contains many novel challenges, as well as novel twists on conventional safety engineering problems. This paper has attempted to provide a broad overview of the state of the art in this field using the concept of uncertainty management as a unifying mechanism to provide continuity to a variety of concerns and techniques that might otherwise seem disparate.

Acknowledgements

This work touches on engineering work conceived and developed by many contributors at the Uber Advanced Technologies Group (Uber ATG). The authors wish to acknowledge and thank these engineers for their important contributions, especially our colleagues in the system safety group.

References

1. Bloomfield, R. and P. Bishop. "Safety and Assurance Cases: Past, Present and Possible Future — An Adelard Perspective," *Making Systems Safer: Proceedings of the Eighteenth Safety-Critical Systems Symposium*, 51-67, Springer, London, 2010.
2. Butler, R. W. and G. B. Finelli. "The Infeasibility of Quantifying the Reliability of Life-Critical Real-Time Software," *IEEE Transactions on Software Engineering*, Vol.19, Issue 1, 3-12, 1993.
3. Littlewood, B. and L. Strigini. "Validation of Ultra-High Dependability for Software-Based Systems," *Predictably Dependable Computing Systems*, 473-493, Springer, Berlin, Heidelberg, 1995.
4. Bishop, P. G. and R. E. Bloomfield. "The SHIP Safety Case Approach," *Safe Comp 95: The 14th International Conference on Computer Safety, Reliability and Security*, 437-451, Springer, London, 1995.
5. Habli, I. and T. Kelly. "Achieving Integrated Process and Product Safety Arguments," *The Safety of Systems: Proceedings of the Fifteenth Safety-critical Systems Symposium*, 55-68, Springer, London, 2007.
6. Goodenough, J. B. and C. B. Weinstock. "Toward a Theory of Assurance Case Confidence," Report No. CMU/SEI-2012-TR-002, Carnegie-Mellon University Software Engineering Inst., Pittsburgh Pennsylvania, 2012.
7. Koopman, P. and B. Osyk. "Safety Argument Considerations for Public Road Testing of Autonomous Vehicles," No. 2019-01-0123, SAE Technical Paper, 2019.
8. Shalev-Shwartz, S., S. Shammah and A. Shashua. "On a Formal Model of Safe and Scalable Self-Driving Cars," arXiv:1708.06374, Cornell University, Ithaca, New York, 2017.
9. Henriksson, J., C. Berger, M. Borg, L. Tornberg, C. Englund, S. R. Sathyamoorthy and S. Ursing. "Towards Structured Evaluation of Deep Neural Network Supervisors," arXiv:1903.01263, Cornell University, Ithaca, New York, 2019.
10. Littlewood, B. and D. Wright. "A Bayesian Model that Combines Disparate Evidence for the Quantitative Assessment of System Dependability," *Safe Comp 95: The 14th International Conference on Computer Safety, Reliability and Security*, 173-188, Springer, London, 1995.
11. Droguett, E. L., F. J. Groen and A. Mosleh. "Bayesian Assessment of the Variability of Reliability Measures," *Pesquisa Operacional*, Vol. 26, Issue 1, 109-127, 2006.