A consistent safety case argumentation for artificial intelligence in safety related automotive systems

Alexander Rudolph, Stefan Voget, Jürgen Mottok

A consistent safety case argumentation for artificial intelligence in safety related automotive systems

An Evaluation of a New Conceptual Functional Safety Approach

Alexander Rudolph[(1)], Stefan Voget[(2)]

[(1)] Department Safety, Safety-in-Use & Cybersecurity
Continental Teves AG & Co. OHG, Frankfurt/Main
[(2)] Artificial Intelligence and Robotics Laboratory, AIR Lab
Continental Automotive GmbH, Regensburg
Germany
{alexander.rudolph, stefan.voget}@continental-corporation.com

Jürgen Mottok
Laboratory for Safe and Secure Systems, LaS³
Ostbayerische Technische Hochschule (OTH) Regensburg
Germany
juergen.mottok@oth-regensburg

***Abstract* — regarding the actual automotive safety norms the use of artificial intelligence (AI) in safety critical environments like autonomous driving is not possible. This paper introduces a new conceptual safety modelling approach and a safety argumentation to certify AI algorithms in a safety related context. Therefore, a model of an AI-system is presented first. Afterwards, methods and safety argumentation are applied to the model, whereas it is limited to a specific subset of AI-systems, i.e. off-board learning deterministic neural networks in this case. Other cases are left over for future research. The result is a consistent safety analysis approach that applies state of the art safety argumentations from other domains to the automotive domain. This will enforce the adaptation of the functional safety norm ISO26262 to enable general AI methods in safety critical systems in future.**

Keywords — Functional Safety, SIL, ASIL, Artificial Intelligence, Behavior, Goal Structure Notation.

I. INTRODUCTION

Artificial intelligence (AI) has made progress from a vision to real usage in the automotive domain. The cars of the future will be trained and will use their knowledge on the streets. But, machine learning has still its limitations, especially regarding questions related to functional safety. This already starts with the unsureness given in the definition itself. “Artificial intelligence automatizes intelligent behavior. The term is not precisely defined as there is no precise definition of intelligence. In general, artificial intelligence is related to the trial, to build machines and to program computers in such a way, that they are able to solve problems by themselves” [9].

From a computational point of view there are a couple of advantages of using artificial intelligence methods to be mentioned [10]:

- Inherent distributed representation of a function which is the base for realization in parallel programming units.
- Representation and processing of fuzziness which is usually difficult to model in classical system engineering processes. This especially includes cases when there is little understanding between the relationships of input and output patterns.
- Highly parallel and distributed computational efficiency with the ability of high performance and high failure tolerance.
- The ability to learn. This enables the application in cases whose intentionally complete algorithmic specification cannot be determined at the initial stages of development. Therefore the neural network uses learning algorithms and training sets to learn new features associated with the desired function.

For the use in safety related environments these systems have to solve the requirements in norms like IEC61508 [2] or ISO26262 [3]. The IEC61508 explicitly mentions artificial intelligence as a not to be proposed technic/measure (IEC61508 table A2-5).

The ISO26262 does not include an explicit mentioning of artificial intelligence. Nevertheless, also regarding this norm the usage is not possible, as artificial intelligence systems often (depends on the used technology) do not fulfill the basic principle of deterministic reaction of a system. This basic principle of the functional safety was so far not discussable in the past. AI seems to be nearly unrealistic to be used in safety related area.

But, AI is attractive for safety-critical fields. Despite safety regulations don’t recommend AI, the future AI usage is a field of research. However, there have been few success cases, for the AI technique is usually a lack of determinism and predictability, which is usually regarded as a disqualifier in a safety context [1].

The structure of the paper is as follows. The main concepts are based on related work which is presented in chapter II. The concept presentation is mainly split into a model definition chapter III and the safety case argumentation in chapter VI. The glue between both parts are filled by a hazard analysis in chapter IV and a limitation of scope for the further considerations in chapter V. The paper finalizes with a conclusion and outlook.

II. Related Work

In [7] Rasmussen introduces a model of human behavior based on three levels of performance of skilled human operators. These levels are not alternatives to each other but interact with each other to fulfill an action. We use the model as a base definition to transfer the logic from human behavior to the argumentation related to usage of artificial intelligence in the automotive domain.

The paper of Kurd et.al [1] outlines the safety criteria which if enforced, would contribute to justifying the safety of neural networks. The criteria are a set of safety requirements for the behavior of neural networks. A potential neural network model is also outlined and is based upon representing knowledge in symbolic form. The paper also presents a safety lifecycle for artificial neural networks. This lifecycle focuses on managing behavior represented by neural networks and contributes to providing acceptable forms of safety assurance. We will use the approach from Kurd et.al [1] to apply the safety argumentation to the automotive domain.

Safety analysis methods like functional hazard analysis [11] are standard in several industries. But so far they are not applied consequently and integrated in the literature to a systematic approach for safety argumentation for systems using artificial intelligence.

III. Definitions

To run a safety analysis of an AI-based functionality requires some initial definitions and the understanding of its nature from a system and a lifecycle perspective.

A. Structural definitions

The term "system" is used as defined by ISO26262 [3], i.e. "a set of elements that relates at least a sensor, a controller and an actuator with one another". An AI-system is a system with the capability to learn. As illustrated in Figure 1, the main elements of an AI-system are

1. the AI-system itself
2. a learning procedure
3. the learning content
4. a learning goal

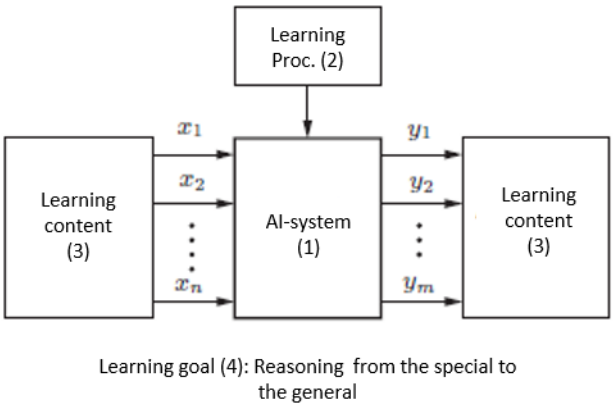

Figure 1: Elements of the AI-system [8]

For an AI-system one differentiates between the learning phase and the operation phase. During the learning procedure (2) the AI-system (1) is triggered by input vectors $\underline{x}^T=(x_1, \dots, x_n)$ that are given by a learning content (3) and the AI-system adapts itself such that the output vectors $\underline{y}^T=(y_1, \dots, y_m)$ of the learning content (3) are reached.

The learning goal (4) specifies the operational contribution of the AI-system – strictly speaking it is exactly the general behavior of the AI-system. Its determination requires awareness of the context on which the AI-system is embedded in.

Following a logical point of view, during operation time an AI-system (1) reasons for each possible input vector $\underline{x\sim}^T=(x\sim_1, \dots, x\sim_n)$ based on the internal structure that has grown up due to the learning content (3). I.e. it reasons from a special learned input vector $\underline{x}^T$ to a general one $\underline{x\sim}^T$.

B. Behavioral definitions

With the definitions given so far the AI-system is still considered as a black box system. To enable a white box view one has to have a deeper look into how the AI-system behaves. Rasmussen [7] introduced a view which is oriented on a human behavior model. It is based on the principle that "humans are not simply deterministic input-output devices but goal-oriented creatures who actively select their goals … Human activity in a familiar environment will not be goal-controlled; rather, it will be oriented towards the goal and controlled by a set of rules which has proven successful previously". In this section we adapt this model to the automotive domain such that we can base the safety arguments on this model later in the paper.

A safety-critical context is mostly determined by the control structure the AI-system is embedded in. Achieving safety in practice means to find an appropriate, hazard-minimizing control action [6]. In automotive vehicle control usually quantitative models for systems design and performance analysis are used. The considerations in this paper use extensions of these models to higher level of human decision making.

AI-systems have also to support the process model (see Figure 2) of a controller. In many expert systems, the process model represents computerized causal human knowledge [8].

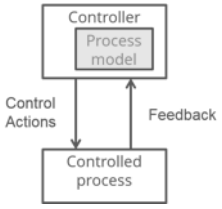

Figure 2: Process model of a controller [6]

Compared to categories of human performance, the tasks of the process model are typically categorized into skill-based behavior, rule-based behavior and knowledge-based behavior [7]. The skill-based behavior represents the sensory-motor action which is done in an automated way without conscious control. At the rule-based behavior level typically the action is controlled by a sequence of stored procedures. The knowledge-based behavior level occurs in unfamiliar situations, when a goal controlled performance is needed.

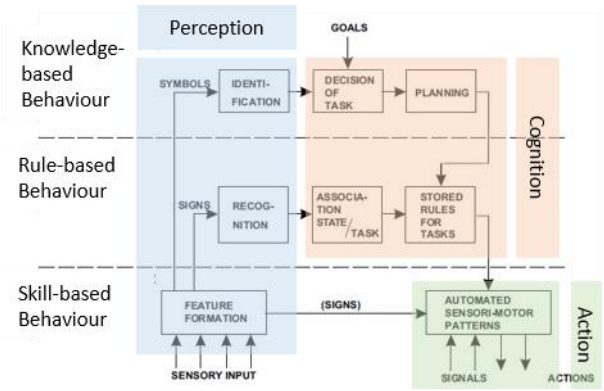

Figure 3: Behavioral Levels of AI-system (adopted from [7])

Figure 3 describes the interaction at and between each level in the steps perception, cognition and action. Following this Table 2 below.

From the criticality classification, the safety integrity level (SIL/ASIL) can be derived as guidance for activities during system design, development and operations. If protection functions outside the AI-system are available in the control context, they can be used to relax the criticality assignment.

Dependent on the FHA made in a given context, risk mitigation mechanisms introduced during system design could be for example:

- a collision avoidance function based on dedicated distance-measuring devices

concept, the principle functional behaviors of AI-systems can be summarized as follows:

Table 1: Functions of AI-system in Context

	Perception	Cognition	Action
Knowledge-based	To identify symbols	To decide next task(s)	
Rule-based	To recognize signs	To select rule for task	
Skill-based	To form features		To apply control pattern

In traditional automotive control systems the safety argumentation is based on the rule-based level only. The structure of the safety argumentation for AI-systems as worked out in this paper comprises all levels.

IV. Functional Hazard Analysis

With the above considerations on the design of a specific AI-system in its context (which may contain multiple AI-systems) the safety analyst is ready to start a functional hazard analysis (FHA) [11] on the functions defined in Table 1. The FHA identifies the relationships between functions and hazards, thereby identifying the safety-significant functions of a system as well as the hazards associated with that functionality. This identification provides a foundation for the safety program to scope additional safety analyzes and level of rigor analysis and verification of the system software.

For an exemplary illustration of a FHA, we use a simple sign recognition function. A sensor identifies a sign with a symbol on it. Based on the recognition on the rule-based behavior level and the identification on the knowledge-based behavior level, actions are taken on an actuator. The actuator is not explicitly defined here, as it is not needed for the safety argumentation in the example. An initial FHA is contained in

- traditional envelope protection functions as ABS or ESC
- an emergency stop function

These functions can mitigate situations of AI-system failure. They can incorporate own AI-systems, but should be designed independently.

Table 2: Exemplary FHA for Sign Recognition

Functional Failure Mode	Existing Sign not recognized	Non-existing sign recognized	Wrong recognition	Untimely recognition	Undetermined recognition
Worst-case Consequence	Applicable rule missed	Wrong rule applied	Wrong rule applied	Wrong rule applied	No situational awareness
Criticality	**hazardous**	**hazardous**	**hazardous**	**hazardous**	**major**
Comment	Collision avoidance, envelope protection present				Emergency operation triggered
Causal Factors during Design	Inadequate AI-system design Inadequate off-line procedure Inadequate content Potentially unrealistic goal				
Causal Factors during Runtime	System Runtime Failure Inadequate on-line procedure Reasonably foreseeable event Unforeseeable event				

V. CLARIFICATION OF SCOPE

Until now the considerations were based on a general definition of AI-systems. For the remaining paper it is necessary to structure the types of AI-systems in more detail.

1. On-operation versus Off-operation learning

   Off-operation means that the learning and the operation phase are separated from each other in the product lifecycle process. Usually the learning is done on a high performance PC or in a processing service center and the operation environment is an automotive embedded board. On-operation enables adaptations of the AI-system with the help of learning during operation, i.e. during driving within the car.

2. Supervised- versus unsupervised- versus reinforcement-learning

   This classification mainly determines which control is given during the learning phase through the process [5].

   In supervised learning an external teacher presents the network with desired input-output mappings.

   In unsupervised learning the desired outputs are not known in advance during training. The neural network is allowed to settle into suitable parameter states and during optimization the neural network develops its representation according to the inputs received.

   Reinforcement learning is used when learning examples (inputs and outputs) are not available. To determine the suitability of the neural network given an input a 'critic' element produces a 'correct' or 'incorrect' signal. This signal is generated by observing the interaction of the neural network with the environment.

3. Deterministic- versus stochastic- learning approaches

   Some learning methods include an optimization step. This step is often done using stochastic optimization methods.

For the case of deterministic off-operation learning we will introduce a safety argumentation for an AI-system in the following chapter.

In case of a deterministic on-operation learning an AI-system applies itself to the environment which makes the safety concept much more difficult to argue. Using stochastic approaches nevertheless makes the argumentation even more complex and in case of on-operation learning impossible regarding the requirement to fulfill the basic principle of deterministic reaction of a system. Both cases are topics for further future research.

As a complete consideration of all possible combinations of the classifications would be beyond the limited space of this paper, we will limit us in the remaining paper to off-operation, deterministic learning approach exemplarily analyzed using neural networks.

VI. SAFETY ARGUMENTS FOR NEURAL NETWORKS

Based on the safety integrity level classification given by the hazard analysis, the safety case responds with a structured argument intended to justify that the system is acceptably safe for the specific AI-system in a specific operating environment. Figure 4 shows the safety argumentation documented by using the goal structuring notation (GSN) [12].

The Goal Structuring Notation (GSN) is a graphical argumentation notation – explicitly represents the individual elements of any safety argument (requirements, claims, evidence and context) and (perhaps more significantly) the relationships that exist between these elements (i.e. how individual requirements are supported by specific claims, how claims are supported by evidence and the assumed context that is defined for the argument) [13].

We used a hierarchical GSN modeling approach: Figure 4 shows the goal structure for neural networks on a top level. The

GSN details for Goal 4 "(Diverse) redundancy and/or monitoring to dedicated ASIL" are shown in Figure 5. The GSN details for G6 "Neural network hazards have been eliminated" as derived in Figure 7. Figure 7 shows the GSN details of rule-based supervision to hazardous outputs of neural networks. The rule-handling approach of Table 1 is used.

Leafs of the GSN tree represent a list of methods applicable in the safety argumentation. Such methods are listed in, e.g. [14], [15], [16], [17], and [18].

As one may imagine the safety argumentation for neural networks are on a coarse level not different to other adaptive systems. Some of the solutions in the GSN tree are generic applicable also to neural networks. Some others are specific in their implementation for neural networks. Let's consider three parts in the GSN tree in more detail, as these are of specific interest for the safety argumentation regarding neural networks

A. *Plausibility checks as solution for G4.2 – monitoring*

The solution "plausibility checks" is a substitute for a set of methods that check the plausibility of an outcome with the help of more or less formal methods which go beyond visualization.

E.g. Heatmapping [22] is a visualization method that quantifies the "importance" of individual pixels with respect to the classification decision and allow a visualization in terms of a heatmap. A more formal analysis of Bayesian networks help to estimate regularization parameters, and to predict the width of the outcome distributions generated by the model.

In [23] Ribeiro e.a. state "Understanding the reasons behind predictions is, however, quite important in assessing trust in a model". They introduce the Lime procedure (local interpretable model-agnostic explanations) to create explanations that reflect the behavior of the classifier "around" the instance being predicted.

B. *Solutions regarding G10 – Restriction of outcome space*

A restriction of the outcome space provides a reduction of complexity and in some solutions the possibility to ensure stability between input and output. The solutions listed in this branch of the GSN tree have been previously mentioned in [14] and [16].

Extreme value theory is a method that has been used to promote confidence by a quantification of the probability estimates made at the tails of the used activation functions. By a formal analysis of assumptions about extreme values the possible outcome space is bounded with a probability based argumentation. With a similar goal the method novelty detection classifies test data that differs from the training data. Both methods do not restrict the values of the outcome space completely but reduce the probability of failure classification.

To ensure stability and convergence during real-time operation some kind of envelop tools help to predict and avoid regions of instability. Also real-time rage limiters on learning state space as well as on input space may ensure more stability. To ensure that the neural network actively controls only when appropriate, an engage/disengage mechanism could be the method of choice. The usage of dead bands such that learning is allowed only when useful command/response dynamics are available should also ensure that the neural network does not adapt to noise or drift away from good solutions.

C. *Solutions regarding G11 – Formal methods*

In the literature one may find an increasing set of formal methods that are at all an attempt to get the structures of neural networks understandable, reasonable and under control. Several neural network specification languages have been developed, like CONNECT [17], nn [17], NSL [19], NeuroML [20], or EpsiloNN [21]. CONNECT, nn, and NSL concentrate on the definition of a dedicated structure of the neural network. NeuroML is created with respect to the exchange of descriptions of neuronal cell and network models. EpxiloNN provides a high-level description of artificial and biology-oriented neural networks with the main objective to support the inherent parallelism of neural networks.

Specification languages directly influence the structure of neural networks. Another technique is the usage of linear models. Using explicit linear models for the activation functions of a node, neural networks provide a theoretical framework to demonstrate stability and evaluate stability margins, which is lacking in widely used heuristic approaches to non-linear control [17].

The other solutions presented in this branch of the GSN tree are attempts to transform the neural networks into structures that are easier to handle by formal analysis methods. To convert a neural network to a decision tree is one of these solutions. Examining a decision tree representing the knowledge of the neural network is more understandable than examining the neural networks actual structure. The decision tree information can be utilized, to check against requirements and to provide confidence in the neural network.

Last but not least the mapping of the empirical model onto a structural model of domain expertise provides more analysis methods, e.g. sensitivity analysis or Bayesian regularization. Bayesian regularization is a mathematical process that enables to consider the neural network in the context of a statistical problem to enable the applicability of regression analytics.

VII. ASSIGNMENT OF SAFETY METHODS TO BEHAVIOR LEVELS

All the solutions presented in the previous section make more or less use of the knowledge that is coded in the neural network. In Table 3 the solutions are mapped to the knowledge levels introduced in Figure 3.

Table 3: Relationships Safety Methods to behavior levels

Behavior Level	Available Technical Safety Methods
Knowledge-based Identification	• Fault removal • Diverse redundancy & voting • Neural network tolerates faults in inputs

	• Fault detection in weights • Dealing with novel inputs metrics • Fault detection in the activation function • Detect uncompleteness of learning data • Specification languages • NN to decision tree • Rule extraction • Bayesian regulation • Sensitivity analysis • Use linear models
Rule-based Recognition	• Build-In self test • Plausibility checks • Extreme value theory • Novelty detection • Predict and avoid regions of instability • Real-time range limiter on learning sate space • Real-time range limiter on input space • Use engage / disengage mechanisms • Dead band on adaptive system inputs
Skill-based Formation	• Program-Flow • Memory Overflow • Time-out given training samples • Time-out given input pattern • Timing guarantees

The list of technical measures is neither complete nor sufficient for the complete fulfillment of a safety argumentation in a concrete automotive application context.

However, the measures are necessary and constitute a foundation from which the safety argument can be started and expanded. Some of them are mentioned in research literature but so far not applied in the automotive industry context. Nevertheless, the assignment to the knowledge levels gives already an overview about the generality of the solutions. The more knowledge is needed for the application of a solution the more effort has to be done during the application.

VIII. Conclusion And Outlook

The ultimate goal to provide guarantees for all unexpected cases cannot be achieved by definition. So, for example, there always exists a point where the physical system is damaged/changed to such an extent that adaptation toward controllable behavior is simply not possible [4]. As this is generally accepted for human intelligence, it should also be for artificial one. However, the transition requires a careful argumentation from which the existence of adequate means of protection becomes comprehensible for the public.

This paper proposes a structural framework in which safety aspects of AI become arguable. In order to establish a defensible position for a certain use of AI, this is a relevant and a significant step.

The list of solutions presented in this paper is however not complete to argue the applicability of a neural network in an automotive safety application. So far no structured and at least sufficiently complete approach for the safety argumentation can be found in the literature. More research is needed to ensure the functional safety in the upcoming future application of neural networks in the automotive applications.

Acknowledgment

This paper is developed in a research partnership of the Laboratory for Safe and Secure Systems (LaS³) of the OTH Regensburg, the department Safety, Safety-in-Use & Cybersecurity of Continental Teves AG & Co. OHG, and the department Artificial Intelligence and Robotics (AIR) of Continental Automotive GmbH.

References

[1] Zeshan Kurd, Tim Kelly and Jim Austin: Safety Criteria and Safety Lifecycle for Artificial Neural Networks.

[2] CEI/IEC 61508-6:2000: Functional safety of electrical/electronic/programmable electronic safety-related systems.

[3] ISO 26262: Road vehicles - Functional safety, International Organization for Standardization, 2011.

[4] Hohann Schuhmann, Pramod Gupta, Stacy Nelson: On Verification & Validation of Neural Network Based Controllers.

[5] Z. Kurd, "Artificial Neural Networks in Safety-critical Applications," First Year Dissertation, Department of Computer Science, University of York, 2002.

[6] N. G. Leveson. Engineering A Safer World: Systems Thinking Applied to Safety, MIT Press. Cambridge, MA. 2011.

[7] Jens Rasmussen. Skills, Rules and Knowledge in Human Performance Model, IEEE Transactions on Systems, Men and Cybernetics, vol.1, no.3, 1983.

[8] Jürgen Adamy. Fuzzy-Logic, Neural Network, Evolutionary Algorithms, Lecture Notes, Control Methods and Robotics Group, Institute of Automatic Control, Darmstadt University of Technology.

[9] www.wikpedia.de, www.wikipeida.org; keyword "artificial intelligence.

[10] Günther Görz (editor), Handbuch der Künstlichen Intelligenz, Oldenbourg Verlag München Wien, 2003.

[11] Adam sharl, Kevin Stottlar, Rani Kady; Functional Hazard Analysis (FHA) Methodology Tutorial, International system safety training symposium, St. Lous, Missourri, 2014.

[12] GSN working group. Goal Structuring Notation, http://www.goalstructuringnotation.info/, 2017.

[13] Tim Kelly, Rob Weaver: The Goal Structuring Notation – A Safety Argument Notation, DSN 2004.

[14] P.J.G. Lisboa; Industrial use of safety-related artificial neural networks; Health & Safety Executive Books; ISBN 071761910; 2001.

[15] X. Huang, M. Kwiatkowska, S. Wang, M. Wu; Safety verification of deep neural netwoks; EPSRC Programme grant on mobile autonomy (EP/M019918/1); University of Oxford; 2017.

[16] C. Wilkinson, J. Lynch, R. Bharadwaj, K. Woodham; Verification of Adaptive systems; U.S. department of transportation; final report, April 2016.

[17] B. Taylor, M. Darrah, C. Moats; Verification and validation of neural networks: a sampling of research in progress; Proceedings of SPIE Vol. 5103; 2003.

[18] U. B. Ramachandran; Issues in verification and validation of artificial neural network based approaches for fault

diagnosis in autonomous systems; Thesis Concordia University Montreal; 2005.

[19] T. Borchert; Neural Network Specification Language; ECEN 5523; December 2005.

[20] NeuroML; https://www.neuroml.org

[21] A. Strey; EpsiloNN - A specification language for the efficient parallel simulation of neural networks.

[22] www.heatmapping.org

[23] M.T. Ribeiro, S. Singh, C. Guestrin; "Why should I trust you?" Explaining the predictions of any classifier; arXiv; 2016.

[24] Jürgen Adamy, Peter Bechtel; Sicherheit mobiler Roboter (Safety of mobile robots); Automatisierungstechnik / Methoden und Anwen-dungen der Steuerungs-, Regelungs- und Informationstechnik 51.10/2003: 435-444; 2003.

APPENDIX: GSN DIAGRAMS

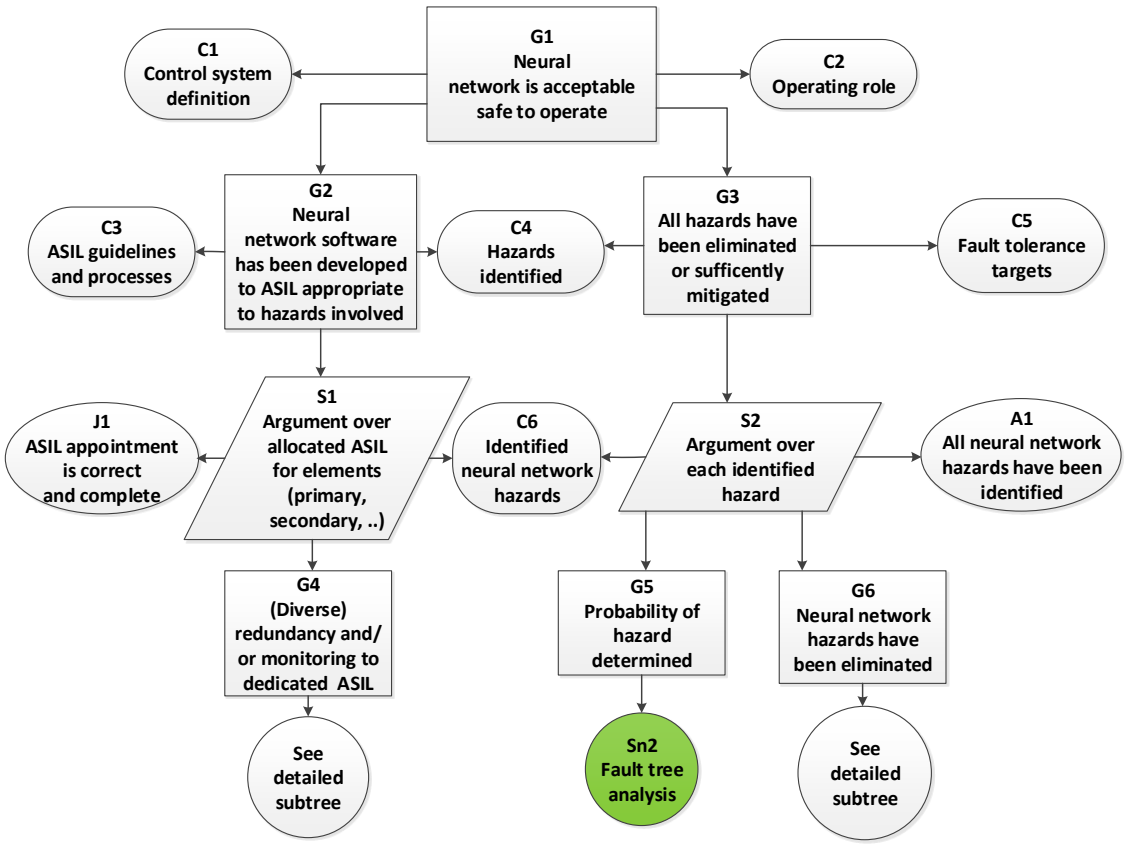

Figure 4: Goal Structuring Notation (GSN) diagram to represent the safety case argumentation regarding neural networks

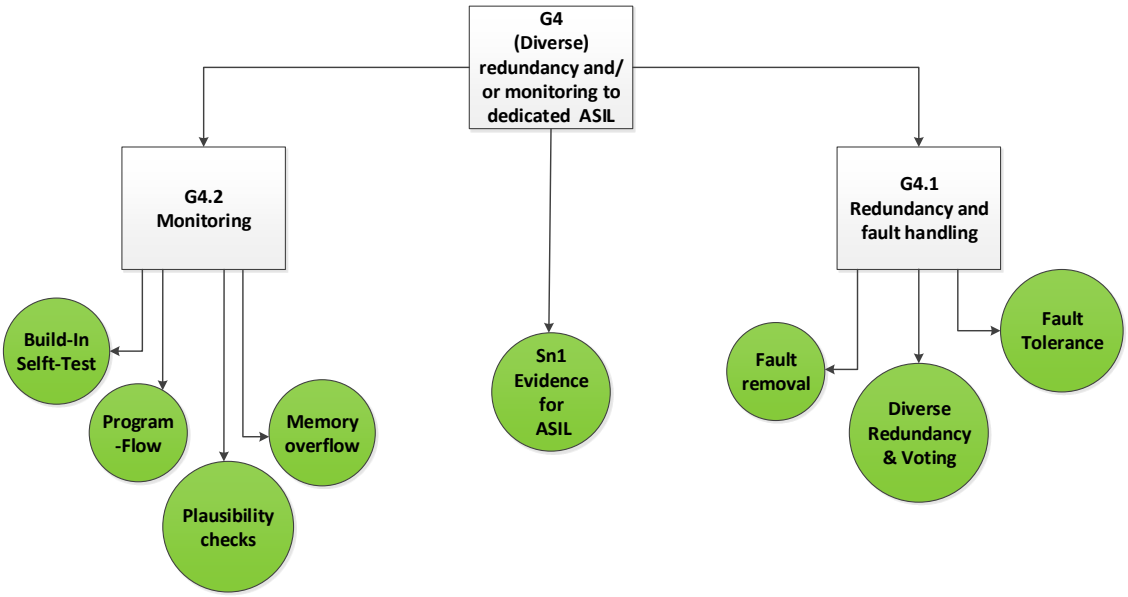

Figure 5: GSN details for G4 as derived in Figure 4

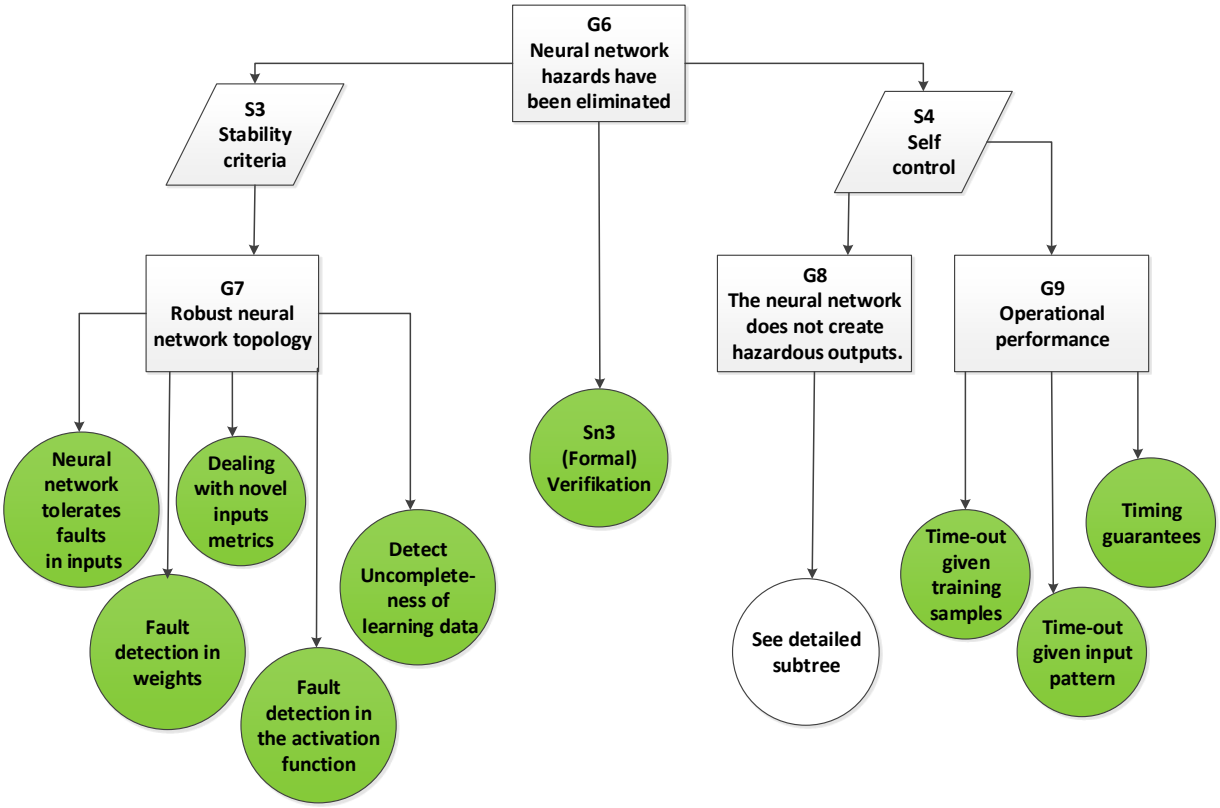

Figure 6: GSN details for G6 as derived in Figure 4

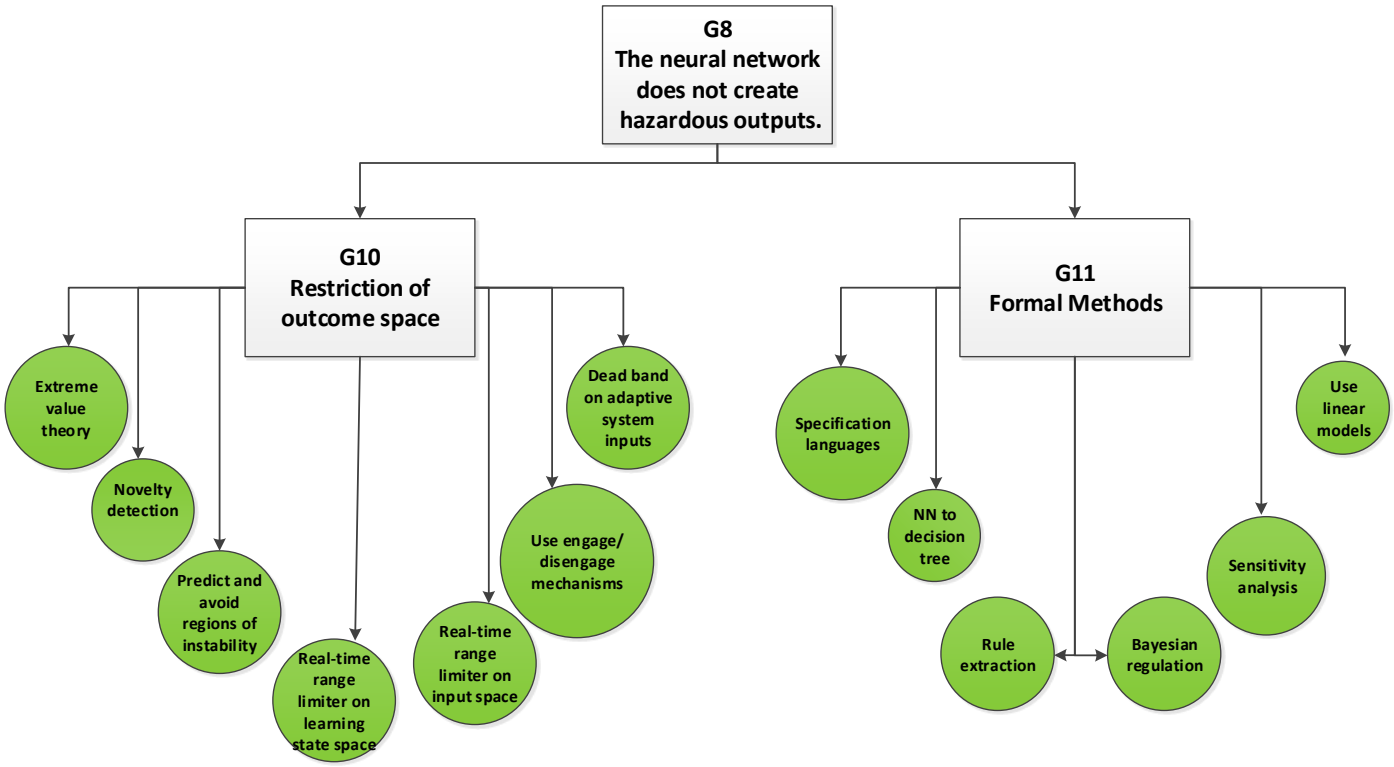

Figure 7: GSN details for G8 as derived in Figure 6