The Missing Link: Developing a Safety Case for Perception Components in Automated Driving

Hiroshi KUWAJIMA Hirotoshi YASUOKA Toshihiro NAKAE
Rick SALAY Krzysztof CZARNECKI Vahdat ABDELZAD
Chengjie HUANG Maximilian KAHN Van Duong NGUYEN

Assuring safety is a central concern in the development and societal acceptance of automated driving (AD) systems. Perception is a key aspect of AD that relies heavily on Machine Learning (ML). Despite the known challenges of assuring safety using ML-based components, proposals have recently emerged for unit-level safety cases addressing these components. Unfortunately, AD safety cases express safety requirements at the system level. These efforts are missing the critical linking argument needed to integrate safety requirements at the system level with component performance requirements at the unit level. In this paper, we propose the Integration Safety Case for Perception (**ISCaP**), This generic template for such a linking safety argument specifically tailored for perception components. This template takes a deductive and formal approach to define strong traceability between levels. We demonstrate the applicability of ISCaP with a detailed case study and discuss its use as a tool to support the incremental development of perception components.

Key words :
Automated driving, perception, machine learning, safety case, safety argument, assurance case

特集

INTRODUCTION

Safety assurance is a central concern for the development and societal acceptance of Automated Driving Systems (ADS). It is no coincidence that major players in this field have made their safety strategies publicly available (e.g., [1, 2]). An ADS relies heavily on complex perception tasks to accurately determine the state of the world. These include image classification, object detection, and image segmentation using camera, LiDAR and Radar sensor data. Because these tasks are difficult to specify, machine learning (ML) is a preferred method of implementation; however, ML poses significant obstacles to safety assurance [3]. Despite this, the safety critical nature of perception requires reliable approaches to assuring their safety.

An ADS safety case aims to provide a hierarchical evidence-based argument for the claim that the ADS is acceptably safe. An important quality of a safety

argument is “rigor”, but when the steps of the argument are based on informal or non-deductive reasoning, this may be difficult to ensure. To address this, Rushby [4] has suggested that the internal decomposition steps of a safety case should be deductive, while inductive reasoning (e.g., generalizing from test results) are limited to the leaf claims that are supported directly by evidence.

This is the approach we take in this paper to define a generic argument template for a perception component within an ADS.

A limited amount of related work on safety cases in the ADS domain exists both at the whole system ADS level and at the unit-level for individual ML components. Kurd et al. [5] give a safety case for neural networks, and more recently, Burton et al. [6] give one for ML components in automated driving. Both use Goal Structuring Notation (GSN) [7], but remain high-level. Wozniak et al. [8] define a GSN argument pattern for ML components to produce an ISO 26262 style safety case. The pattern covers the refinement of unit-level safety requirements, data appropriateness, adequacy of the component design, and component training. The latter three areas are given a more detailed treatment with subclaims proposed.

Picardi et al. [9] also focus on the unit level and sketch a safety case pattern in which each ML assurance claim is supported by a series of *confidence* arguments that show why the claim is supported by context artifacts such as the test dataset, learned model, ML safety requirements, etc. These confidence arguments, in turn, draw on the ML development lifecycle [10] that specifies high-level requirements regarding these artifacts. For example, it specifies that test data should be “relevant, complete, accurate and balanced”. The authors note that the ML assurance claims are drawn from ML safety requirements, which in turn are refined from system-level safety requirements based on methods such as hazard analysis; however, they have given no details on how this refinement is done. As an illustration of how this could work, Gauerhof et al. [11] study the elicitation of safety requirements for an ML-based pedestrian detector. For example, an analysis of the system-level safety requirement “Ego shall stop at the crossing if a pedestrian is crossing” yields several object detection component safety requirements such as “Position of pedestrians shall be determined within 50cm of actual position”. Then the ML development lifecycle is used to guide the identification of detailed requirements for the confidence arguments—e.g., “The data samples shall include sufficient range of environmental factors within the scope of the ODD” is a requirement to address test data completeness. Although this illustration is compelling, it offers no general approach for connecting system-level to component unit-level safety requirements.

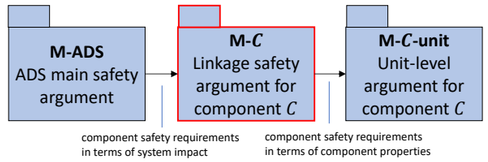

Fig. 1 Role of the linking argument M-*C* instantiated from ISCaP presented in this paper, where *C* represents a perception component. Arrows represent the “is supported by” relation where a higher-level argument is supported by a lower-level one

The recent work by Bloomfield et al. [12] is an extensive effort to define a safety case template using the Claims-Argument-Evidence (CAE) notation, for autonomous systems that include ML components. It assumes that ML components will be coupled with monitors that guard the component against bad behaviour. The template focuses on the adequacy of hazard analysis at the system level and gives high-level templates for arguments at the monitor+ML subsystem level as well as the unit level for the ML component. The reasoning approach advocated is informal with an emphasis on identifying potential *defeaters* that challenge claims. As with the work of Picardi et al. [9], there is a brief discussion about connecting the system-level requirements to the unit-level but details are missing. For example, “The number of crashes involving the AV averages at most 89 crashes per million miles driven with confidence 95%.” is given as a sample system-level claim and “YOLOv3 correctly identifies traffic lights in 87% of images containing traffic lights” is a claim at the unit-level but the parts of the argument that show how the performance of object detector YOLOv3 impacts the crash rate of the AV are not addressed by the template.

The Aurora safety case framework [13] is another recent effort defining a safety case structure addressing the entire ADS development process. It consists of a GSN decomposition tree of generic claims from the top claim that the ADS is acceptably safe. Although its scope is broad, its claims remain at high-level, are non-formal, and do not refine to component-level claims. For example, the leaf claim “The systems engineering process is appropriate for safety critical design” is typical of the most refined level of detail in the framework.

In contrast to the work discussed above, in this paper we focus specifically on addressing the connection between the system-level and unit-level arguments. The recent work of Vaicenavicius et al. [14] studies this connection formally for a specific simple driving scenario and sketches the safety case structure for this scenario. While this work is in the same spirit as ours, it is limited to a simple idealized illustration and does not consider the special characteristics of perception components such as vulnerability to environmental conditions affecting perception. In this paper, we propose the Integration Safety Case for Perception (**ISCaP**, pronounced Ice-Cap), a template that provides a systematic method of generating a formal “linkage argument” between the system-level and unit-level arguments, specifically for perception components. This is illustrated in Fig. 1 using GSN safety case modules. The safety case module **M**-*C* instantiated from ISCaP for a perception component *C*, sits between the main ADS argument module **M-ADS** that relies on a claim about component contribution to system safety and the unit-level argument module **M**-*C*-**unit** that provides evidence for specific component properties required for safety.

Our contributions are:

- ISCaP, a formal safety case template linking system and unit arguments using a safety claim decomposition method based on *hazardous misperception patterns* that separates analysis of ADS dynamics from perception.
- The use of ISCaP to identify a set of *risk-aware* performance metrics that are tailored to the perception component and the ADS in which it operates.
- A safety case structure that has desirable properties including stability with respect to component changes and support for assessing risk trade-offs in different operational situations.

The remainder of the paper is structured as follows. In the next section we give the formalization preliminaries. Then ISCaP is presented in detail using GSN notation. This is followed by a discussion on the application of ISCaP in component development. Then the use of ISCaP is illustrated with a detailed example. Finally, we give conclusions and discuss future work.

Preliminaries

In this section, we define the terms and notation used in ISCaP.

Perception Tasks

We distinguish between the perception task *T* in the ADS and a component *C* that implements the task.

For example, T may represent the task "detect and localize vehicles surrounding the ego vehicle" while C_1 and C_2 may be alternate components that implement this task with different performance characteristics.

Definition 1 (**Perception task**) *A **perception task** T is represented as function $T : X \rightarrow Y$ from input domain X to output codomain Y defining the "ground truth" of how the task should be performed. If component C implements T, then it defines function $C : X \rightarrow Y$ intended to approximate $T : X \rightarrow Y$. The unit of perception for T, called a **frame**, is denoted $(x,y) \in X \times Y$. The **frame rate** is the number of frames per unit time. The distribution $PC,_{fr}(x,y)$ denotes the probability of frame (x,y) occurring while the ADS operates a vehicle in the ODD using component C that implements T.*

For example, task `OD` for camera-based object detection defines function `OD : CImage → BBSet` from camera images to bounding box sets. Object detector `Yolo : CImage → BBSet` implements `OD`. A frame is a single camera image and corresponding output set of bounding boxes.

Drives

We define notation for describing driving scenarios.

Definition 2 (States) *A **state** s is a snapshot of all relevant ADS and environment state variables at a point in time. S is the set of possible states.*

The rate at which snapshots are taken determines how fine-grained in time driving is represented. For the argument **M**-*C*, it is convenient to take a task-centric view and assume that this is the frame rate for T.
ADS operation can be viewed as an infinite stochastic process producing a sequence of states. We can classify the states that occur during ADS operation using temporal properties.

Definition 3 (State classification) *Given state s_t, $s \in S$, $t \in Z$, $P_C(s_t \in \phi)$ (or $P_C(\phi)$) denotes the probability that a randomly chosen state during ADS operation in the ODD using component C satisfies temporal property ϕ.* `MCrash` *denotes the temporal property identifying crash states caused by some preceding sequence of misperceptions in performing task T.*

Thus, P_C(`MCrash`) denotes the probability that a misperception-caused crash state occurs when using component C.

Definition 4 (Drives) *A **drive** is a finite sequence $d \in S^*$ of states. $d' \subseteq d$ denotes that drive d' is a subsequence of d.* `Matches`*(d, s_t) is a predicate that holds iff $\forall k \in \{0, ..., n-1\} \cdot s_{t-k} = d[n-k]$, where, n is the length of d and $d[i]$ is the i^{th} state in d.*

Thus, `Matches`(d, s_t) means that s_t and its $n-1$ preceding states match d.

Misperceptions

The safety of perception tasks is impacted by the presence of misperceptions.

Definition 5 (Misperception) *Given task T, a **misperception** is a frame (x,y) such that $y \neq T(x)$. A misperception* by component *C is a frame $(x,C(x))$ that is a misperception.*

The safety impact of a misperception varies depending on the context in which it occurs. For example, in a vehicle detection task, a false negative (FN) (i.e., not detecting a vehicle) or false positive (FP) (i.e., falsely detecting a vehicle) close to the ego vehicle may be hazardous, but when these occur far away, they may be benign. Although having too many benign misperceptions can negatively impact ADS performance, in the safety argument for component C we consider only the hazardous misperceptions it can produce. In order to characterize misperceptions, we generalize from individual misperceptions to *patterns* of misperceptions.

Definition 6 (Misperception Pattern) *A **Misperception Pattern (MP)** identifies a subset of drives in which every drive contains some states with misperceptions. A **Frame Misperception Pattern*** applies to a single state and is defined as function `fMP` $: Y \rightarrow pow(Y)$ *such that* $\forall y \in Y \cdot y \notin$ `fMP`*(y), where $pow(Y)$ is the power set of Y. State s containing frame (x,y)* satisfies *frame misperception pattern* `fMP` *iff* $y \in$ `fMP`*$(T(x))$.*

Misperception patterns can be used to categorize misperceptions based on conditions of interest. For example, the frame misperception pattern $\texttt{FN}_{20}$ can denote all object detection misperceptions that include false negatives within $20m$ of the ego vehicle. Note that for some values of $T(x) \in Y$ there may be no instances of the frame misperception pattern $\texttt{FN}_{20}(T(x))$. For example, if $T(x)$ contains no detections within $20m$ of the ego vehicle then $\texttt{FN}_{20}(T(x)) = \phi$ since there is no way to get such an FN. The misperception pattern $\texttt{30FN}_{20}$ can denote the set of all drives satisfying $\texttt{FN}_{20}$ in at least 30% of its states. Thus, misperception patterns are naturally defined in terms of frame misperception patterns and we exploit this in the safety argument.

Hazardous Misperceptions

Following ISO 26262 (Functional safety) [15] and ISO 21448 (SOTIF) [16], system-level hazard analysis identifies cases where a driving scenario combined with a hazardous behaviour by the ego vehicle and particular reactions by scenario participants will result in harm (i.e., a crash). For example, the ego vehicle waiting to turn left at an intersection is a scenario in which, if the ego vehicle begins turning with an on-coming vehicle too close (hazardous behaviour), and the on-coming car cannot stop (participant reaction), a collision will occur. We define a term to represent these cases.

Definition 7 (Hazardous Behaviour Sensitive Scenario) *A **Hazardous Behaviour Sensitive Scenario (HBSS)** is a subset of drives exhibiting a particular combination of operational scenarios with participant reactions in which there are possible hazardous ego vehicle behaviours.*

In some cases, the hazardous behaviour in an HBSS can be *caused by* a sequence of frame misperceptions in performing perception task T. For example, in the HBSS with the ego vehicle turning left, the hazardous behaviour could be caused by a sequence of misperceptions in the `OD` task over several frames that leads the ADS to believe there is no vehicle coming and allows a hazardous left turn to be performed. We refer to such sequences as *hazardous misperceptions* and formalize their occurrence as a type of misperception pattern.

Definition 8 (Hazardous Misperception Pattern) *A **hazardous misperception pattern (HMP)** is a pair* ⟨`HBSS`, `MP`⟩ *of predicates over set S^* of drives where,*

- `HBSS` *is the **Hazardous Behaviour Sensitive Scenario condition** that specifies properties of the drive required for the scenario to occur and no other drive factors. At most one state in the drive can be a crash state and it must occur at the end of the drive.*
- `MP` *is the **Misperception Pattern condition** that specifies **all** sequences of misperceptions that will cause a hazardous behaviour in drives that satisfy the* `HBSS` *condition. The* `MP` *condition constrains only drive factors essential for describing the misperceptions.*

A drive $d \in \mathcal{S}^*$ satisfies the HMP iff

$$\texttt{HBSS}(d) \wedge (\exists d' \in \mathcal{S}^* \cdot \texttt{MP}(d') \wedge d' \subseteq d)$$

The condition HBSS says that the HBSS for the system hazard associated with HMP occurs over the length of the drive. Thus, a drive satisfying HBSS represents a complete scenario that ends in a crash if any hazardous behaviour by the ego vehicle, as identified by the HBSS, occurs.

Condition MP identifies the sequences of misperceptions performing task *T* that, when they occur within an HBSS drive, will cause a hazardous behaviour to occur leading to a crash. This condition would naturally be expressed in terms of various frame misperception patterns.

For example, assume HMP IL is based on the $\texttt{HBSS}_{\texttt{IL}}$ condition that identifies a drive in which the ego vehicle is waiting to turn left at an intersection. We focus on the hazardous behaviour in which the ego vehicle begins turning with an on-coming vehicle too close, causing a collision. This hazardous behaviour could be caused by misjudging the position and/or speed of the on-coming vehicle (mis-localization) or not detecting the on-coming vehicle (FN). With analysis and experimentation, it is possible to determine the exact sequences of these misperceptions that would cause the hazardous behaviour and these are used to define $\texttt{MP}_{\texttt{IL}}$.

Although drive predicates such as HBSS and MP play an important role in the safety argument, we do not specify a formal language for expressing drive predicates and instead remain at the semantic level, thinking of predicates in terms of the sets they define. For example, drive predicates can be defined using a general language such as First Order Logic (FOL), or more specialized logics such as temporal logic (e.g., Linear Temporal Logic). Remaining "syntax agnostic" gives ISCaP the flexibility to be used in different analytical contexts.

The ISCaP Safety Case Template

In this section, we describe in detail, the claims, decomposition strategies, and sources of evidence for the argument in **M**-*C* as an instantiation of the ISCaP safety case template. Fig. 2 shows the rendering of ISCaP in GSN. Claims are expressed using *Goal* nodes, decompositions of claims into sub-claims as *Strategy* nodes and evidence as *Solution* nodes. In addition, *Context* and *Assumption* nodes denote supporting information.

We exploit the modularity and templating features in GSN to encapsulate the argument instantiated from ISCaP in a separate module **M**-*C* and treat the top-level goal **G**-*C* as an *away-goal*—i.e., referenced from within the main ADS argument **M-ADS**. While details of **M-ADS** and **M**-*C*-**unit** are left unspecified and are out of scope for this paper, we highlight aspects of them that **M**-*C* depends on.

Overall, the argument takes a formal deductive approach in which claims are expressed as bounded probabilities and strategies mathematically relate the bounds between child and parent claims. This restricts all inductive steps of the argument to the solution nodes that provide evidence for leaf claims. The bound in the top claim for component *C* is first decomposed using HMPs corresponding to system-level hazards that have misperceptions as causes. The bound in each claim for a specific HMP is then factorized into claims that bound the crash rate, misperception pattern occurrence rate, and HBSS (i.e., hazardous ADS dynamics) occurrence rate, plus a guarantee that all hazardous misperceptions in the HBSS are covered. Among these claims, the bound on the misperception pattern rate is further decomposed using perception-only (PO) conditions, separating the misperception rate for each PO condition and rate of occurrence for the PO condition. Finally, the bound on the misperception pattern rate for each PO condition is decomposed into bounds on the occurrence rates of the constituent frame misperception patterns. These define risk-aware performance metrics that can be estimated via testing in the machine learning context.

Goal **G**-*C*

ISCaP assumes a single type of high severity event, which we term "crash" throughout this paper. In the section "Applying ISCaP in Practice", we discuss how to adapt ISCaP to allow for multiple severity levels.

The top-level claim, that component *C* is adequately safe, is formalized by bounding the probability $P_C(\texttt{MCrash})$ that the ADS reaches a crash state due to a misperception by component *C* in performing task *T*. Note that we take an assume-guarantee view of **M**-*C*, where claim **G**-*C* is guaranteed (up to the evidence) only if we assume all other ADS components are operating as designed (**A-Sys**). This assumption can be adjusted to accommodate different system considerations. For example, if *C* is a camera-based object detector, we could assume a particular failure rate of the camera. This would have to then be incorporated into the claims since then sometimes MCrash may occur due to camera failure rather than a misperception of the object detector.

Fig. 3 helps explain the meaning of **G**-*C*. Subset $D_{\texttt{MCrash}}$ are drives that contain hazardous misperceptions in carrying out task *T*. Subset D_C are drives that are possible when component *C* implements *T*. Component *C* may not produce some hazardous misperceptions, and other misperceptions it produces may be benign; thus, it is the occurrence of drives in the intersection $D_{\texttt{MCrash}} \cap D_C$ (bright red) that the claim **G**-*C* seeks to bound to an acceptable level. Safer components have a smaller value for $P_C(\texttt{MCrash})$, and if the current component *C* does not meet the required target level, different mitigation steps could be taken. For example, *C* could be replaced with a better component *C'* that makes fewer hazardous misperceptions, an ensemble of components can be used jointly to reduce hazardous misperceptions using redundancy, etc.

Strategy **S-HMP**

It is unrealistic to expect to provide evidence directly for **G**-*C* since this covers all possible misperception-caused crashes. The similar complexity issue in the main ADS argument is addressed by decomposing the safety claim based on system hazards that may occur. For a system hazard that can be caused by hazardous misperceptions, we can define a corresponding HMP.

The strategy **S-HMP** exploits this use of system hazards to decompose $D_{\texttt{MCrash}}$ and, correspondingly, **G**-*C* into sub-claims using HMPs. In Fig. 2, **S-HMP** is linked to context information about the ADS hazard analysis and creates sub-claims $\textbf{G-}\texttt{HMP}_i$ for the corresponding identified hazardous misperception patterns. In addition, it includes the sub-claim **G-Res**, which corresponds to the residual hazardous misperception-caused drives not covered by any HMP.

To simplify the relationship between bounds, we assume that the HMPs are chosen so that the subsets of drives they define are disjoint. We show below that it is always possible to satisfy this restriction. With this, the simple additive relationship $\gamma_C \geq \gamma_{res} + \Sigma_i \gamma_i$ holds among the sub-claims. This allows the bound on P_C(MCrash) to be conveniently seen as a "risk budget" that is allocated to different HMPs and helps to study risk trade-offs. Mathematically, this strategy can be understood as an application of the law of total probability: $P_C(\texttt{MCrash}) = P_C(\texttt{MCrash} \wedge \neg \bigvee_i \texttt{HMP}_i) + \bigvee_i P_C(\texttt{MCrash} \wedge \texttt{HMP}_i)$.

Goal $\textbf{G-}\texttt{HMP}_i$

Def. 8 defines an HMP as a drive specification in terms of HBSS and MP conditions. Fig. 3 shows an example subset of drives $D_{\texttt{HMP}_i}$ that satisfy the specification for one $\texttt{HMP}_i$. The corresponding sub-

特集

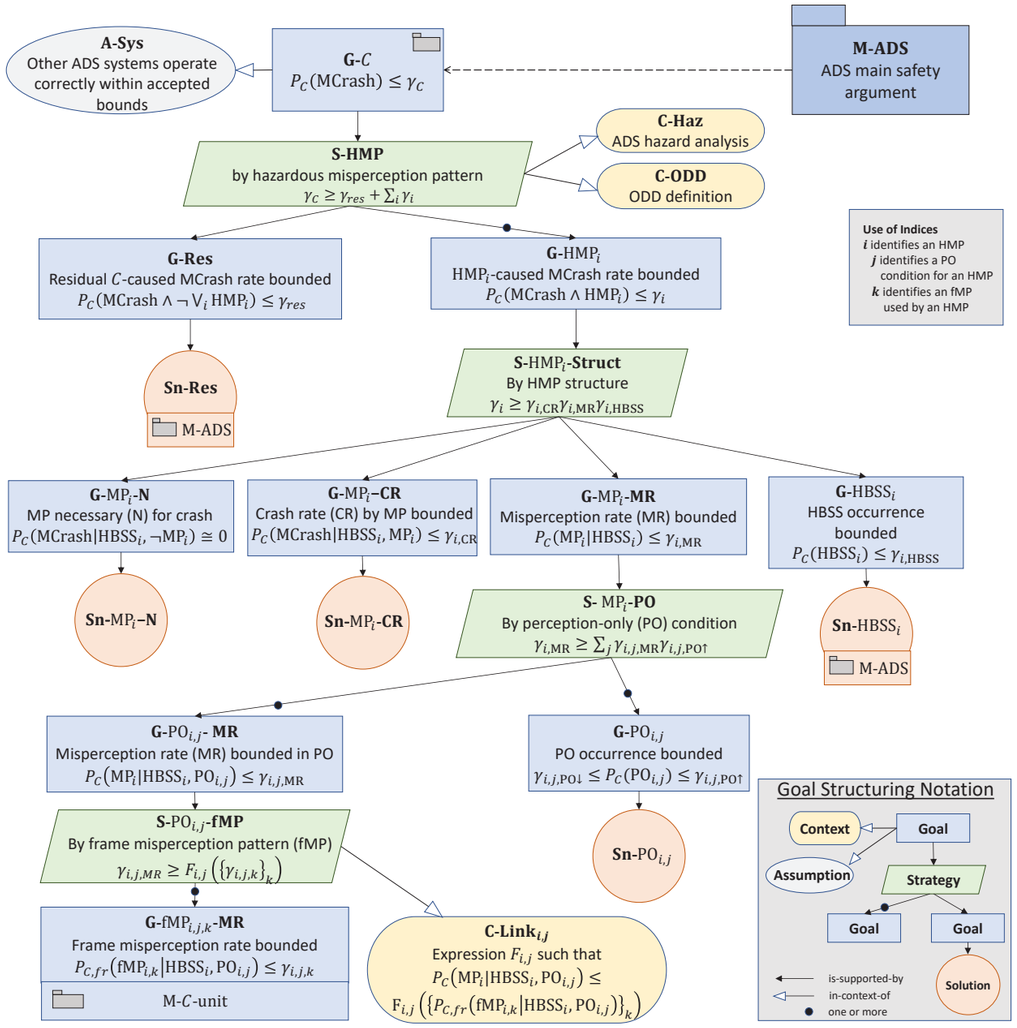

Fig. 2 The ISCaP safety case template for the argument in **M**-*C* using GSN

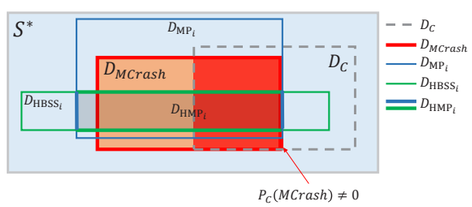

Fig. 3 Decomposition of *G-C* into sub-claims by decomposing D_{MCrash} into subsets corresponding to hazardous misperception patterns

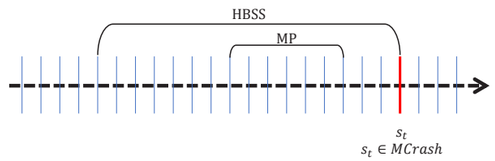

Fig. 4 Illustration of how an HMP can be used to classify cases of misperception-caused crashes

claim **G**-HMP_i bounds the occurrence of drives in the subset $D_{\text{HMP}_i} \cap D_C$ — i.e. the drives in HMP_i that could be produced when component C is used. As discussed above, we assume that HMPs identify disjoint sets of drives. This is always possible to achieve by constructing additional HMPs corresponding to overlaps. For example, if HMP_i and HMP_j overlap, we define $\text{HMP}_{i,j} \equiv \langle \text{HBSS}_i \wedge \text{HBSS}_j, \text{MP}_i \vee \text{MP}_j \rangle$ and redefine $\text{HMP}_i \equiv \langle \text{HBSS}_i \wedge \neg\text{HBSS}_j, \text{MP}_i \rangle$ and $\text{HMP}_j \equiv \langle \neg\text{HBSS}_i \wedge \text{HBSS}_j, \text{MP}_j \rangle$.

Goal **G-Res** *and Solution* **Sn-Res**

Claim **G-Res** bounds the occurrence of hazardous misperceptions in drives not covered by any HMP—those in the residual subset $(D_{\text{MCrash}} \cap D_C) \setminus \cup_i D_{\text{HMP}i}$ in Fig. 3. Since, the HBSSs are based on system-level hazards, the evidence for claim **G-Res** should be drawn from evidence in argument **M-ADS** regarding completeness of hazard analysis. We identify **Sn-Res** as an away-solution to indicate this fact.

For example, **M-ADS** may use a reference such as the NHTSA pre-crash scenario typology [17] to provide evidence for sufficient coverage of hazard scenarios. These scenarios account for 99.4% of all reported light-vehicle crashes and the HMPs could be based on the subset of these that can be caused by misperceptions. Thus, subject to the assumptions that data about human-caused crashes can be applied to ADS-caused crashes and that we are limited to reported light-vehicle crashes, the residual of 0.6% could be used as a basis for residual bound γ_{res}.

Strategy **S-**HMP_i**-Struct**

Fig. 4 shows how an HMP is related to a misperception-caused crash state. The crash state s_t occurs because the drive preceding it was an HBSS drive, and within this, an MP misperception sequence caused a hazardous behaviour. Given this fact and Def. 8, we can express $P_C(\text{MCrash} \wedge \text{HMP}_i)$ as

$$P_C(s_t \in \text{MCrash} \wedge \exists d \in \mathcal{S}^* \cdot \text{HBSS}_i(d) \wedge \text{Matches}(d, s_t) \wedge \exists d' \in \mathcal{S}^* \cdot \text{MP}_i(d') \wedge d' \subseteq d)$$

This can be decomposed into the product of three terms using the chain rule for joint probabilities:

1. Misperception-caused crash rate

$$P_C(s_t \in \text{MCrash} | \exists d \in \mathcal{S}^* \cdot \text{HBSS}_i(d) \wedge \text{Matches}(d, s_t) \wedge \exists d' \in \mathcal{S}^* \cdot \text{MP}_i(d') \wedge d' \subseteq d)$$

- Meaning: Probability that a state s_t is a misperception-caused crash given it ends an HBSS_i drive containing an MP_i sequence.
- Shorthand: $P_C(\text{MCrash}|\text{HBSS}_i, \text{MP}_i)$

2. Misperception rate

$$P_C(\exists d \in \mathcal{S}^* \cdot \text{HBSS}_i(d) \wedge \text{Matches}(d, s_t) \wedge \exists d' \in \mathcal{S}^* \cdot \text{MP}_i(d') \wedge d' \subseteq d \,|\exists d \in \mathcal{S}^* \cdot \text{HBSS}_i(d) \wedge \text{Matches}(d, s_t))$$

- Meaning: Given a state s_t ends an HBSS_i drive, the probability the state is preceded by an MP_i sequence contained within the drive.
- Shorthand: $P_C(\text{MP}_i|\,\text{HBSS}_i)$

3. Exposure of HBSS in ODD

$$P_C(\exists d \in \mathcal{S}^* \cdot \text{HBSS}_i(d) \wedge \text{Matches}(d, s_t))$$

- Meaning: The probability that a state s_t ends an HBSS_i drive.
- Shorthand: $P_C(\text{HBSS}_i)$

These three terms produce the sub-claims **G-**MP_i**-CR**, **G-**MP_i**-MR** and **G-**HBSS_i, respectively. Thus, we have that $\gamma_i \ge \gamma_{i,\text{CR}}\gamma_{i,\text{MR}}\gamma_{i,\text{HBSS}}$. In Fig. 2 the shorthand forms of the probabilities are used. The fourth sub-claim, **G-**MP_i**-N**, is needed to show that MP_i covers all hazardous misperceptions in HBSS_i.

Goal **G-**MP_i**-N** *and Solution* **Sn-**MP_i**-N**

By Def. 8, MP_i must identify all hazardous misperceptions that component C could produce in an HBSS_i drive—i.e., that MP_i is *necessary* for a

特集

misperception caused crash in this HBSS. In Fig. 3, this is equivalent to the condition:

$$(D_{\mathtt{MCrash}} \cap D_C \cap D_{\mathtt{HBSS}_i}) \subseteq (D_{\mathtt{MP}_i} \cap D_C \cap D_{\mathtt{HBSS}_i})$$

Claim **G**-$\mathtt{MP}_i$-N states this by saying that an $\mathtt{HBSS}_i$ drive not containing an $\mathtt{MP}_i$ misperception sequence has probability zero of ending in a crash.

Strong evidence for this claim may require safety analysis methods such as FTA and FMEA (or its specializations, e.g., CFMEA [18]). Note that this analysis would establish that the $\mathtt{MP}_i$ is necessary for a *hazardous behaviour to occur* under the $\mathtt{HBSS}_i$ conditions. The analysis identifying the hazardous behaviours (i.e., behaviours that lead to a crash) under $\mathtt{HBSS}_i$ is a separate analysis that should be done at the vehicle level as part of **M-ADS**. Empirical evidence could be obtained by simulating randomly sampled HBSS_i drives and checking that those that end in a crash also satisfy $\mathtt{MP}_i$.

Goal **G**-$\mathtt{MP}_i$-**CR** *and Solution* **Sn**-$\mathtt{MP}_i$-**CR**

The crash rate is the probability that a drive in $\mathtt{HMP}_i$ ends in a crash. The $\mathtt{HBSS}_i$ condition alone only guarantees that a misperception-caused hazardous behaviour is *possible* in the drive, but doesn't guarantee one occurs. The condition $\mathtt{MP}_i$ in $\mathtt{HMP}_i$ constrains $\mathtt{HBSS}_i$ to drives that contain misperceptions. When $\mathtt{MP}_i$ is sufficiently constraining so that the only misperceptions that satisfy it are hazardous (i.e., no benign misperceptions), then a crash is guaranteed, and thus $P_C(\mathtt{MCrash}|\mathtt{HBSS}_i, \mathtt{MP}_i) = 1$. This is desirable because allowing benign misperceptions into $\mathtt{HMP}_i$ forces the bound $\gamma_{i,\mathrm{CR}}$ to be less tight than necessary. However, allowing benign misperceptions does not pose a safety risk and may simplify analysis. One could also conservatively set $\gamma_{i,\mathrm{CR}} = 1$ as long as the resulting bound in the top claim is sufficiently low (this is done in the case study), otherwise a tighter bound may be needed.

Empirical evidence to check or estimate $\gamma_{i,\mathrm{CR}}$ could be obtained by simulating randomly sampled HBSS_i drives with randomly injected MP_i misperceptions and counting the number that end in a crash.

Goal **G**-$\mathtt{HBSS}_i$ *and Solution* **Sn**-$\mathtt{HBSS}_i$

The claim in this goal bounds the occurrence of $\mathtt{HBSS}_i$ drives in the ODD. In the context of ISO 26262 and ISO 21448, this is related to the level of *exposure*[1] to the $\mathtt{HBSS}_i$ scenario by the ADS.

Since the exposure to $\mathtt{HBSS}_i$ is the same as the exposure to the corresponding system-level hazard it is based on, the exposure level and evidence should already be found in argument **M-ADS**. Thus, the **Sn-HBSS**$_i$ is identified as an away-solution node.

Although this claim uses distribution P_C specific to component C, in practice, changing the component should not affect the exposure. The exposure level would only be dependent on a specific choice of C if the ADS explicitly accounted for the weaknesses of C in its driving policy. For example, if the object detector being used was known to perform poorly on busy roads, the ADS policy could be designed to avoid them, thus affecting exposure levels.

Goal **G**-$\mathtt{MP}_i$-**MR**

The misperception rate is the key subclaim of the decomposition **S**-$\mathtt{HMP}_i$-**Struct** that is affected by the behaviour of component C ; thus, it is the focal point of any development effort to improve safety by changing C. However, estimating the bound $\gamma_{i,\mathrm{MR}}$ directly may be difficult. Since the development goal for component C is to make it perform well in the ODD, $\mathtt{MP}_i$ misperceptions may not be produced by C in the majority of $\mathtt{HBSS}_i$ drives, making this a rare event and challenging for collecting empirical evidence. However, it is well known that perception performance can be dramatically affected by external conditions such as weather, lighting, object properties such as shape, color and size, spatial configurations of objects, etc. We refer to these as *Perception-Only (PO)* conditions since they identify sub-cases within $\mathtt{HBSS}_i$ with higher probability of $\mathtt{MP}_i$ misperceptions but do not constrain ADS behaviour. Distinguishing HBSS from MP and PO conditions allows a separation of analysis of ADS dynamics (i.e., planning and actuation) from perception. In a similar spirit to hazard analysis at the higher levels of the safety case, we decompose the claim **G**-$\mathtt{MP}_i$-**MR** according to PO conditions.

Strategy **S**-MP_i-**PO**

Let $\{\mathtt{PO}_{i,j}\}_j$ be a set of PO conditions for $\mathtt{HMP}_i$. PO conditions are drive predicates that further "condition" the drives within $D_{\mathtt{HBSS}i}$ in Fig. 3 by constraining drive factors that affect the occurrence of misperceptions. We assume that PO conditions define disjoint subsets of drives within $D_{\mathtt{HBSS}i}$. Finally, we define $\mathtt{PO}_{i,\mathrm{Nom}} = \neg \bigvee_{j \neq \mathrm{Nom}} \mathtt{PO}_{i,j}$ as the distinguished PO condition representing the nominal case of drives where no other PO condition holds. With this, the subsets $\{\mathtt{PO}_{i,j}\}_j$ partition the set $D_{\mathtt{HBSS}i}$. Based on these assumptions and by the law of total probability, we have:

$$P_C(\mathtt{MP}_i|\mathtt{HBSS}_i) = \sum_j P_C(\mathtt{MP}_i|\mathtt{HBSS}_i, \mathtt{PO}_{i,j}) P_C(\mathtt{PO}_{i,j}|\mathtt{HBSS}_i) \quad (1)$$

Note that if $\mathtt{PO}_{i,j}$ and $\mathtt{HBSS}_i$ are independent, we also have $P_C(\mathtt{PO}_{i,j}|\mathtt{HBSS}_i) = P_C(\mathtt{PO}_{i,j}) = P_C(\mathtt{PO}_j)$. Although the template does not require this independence, $P_C(\mathtt{PO}_j)$ can be reused for any $\mathtt{HBSS}_i$, where the independence holds, simplifying the analysis. The independence often holds in practice, but not always. For example, it is reasonable to assume that the PO condition representing lighting conditions is independent from the HBSS of crossing an intersection as in the left-turn scenario defined by $\mathtt{HBSS}_{\mathrm{IL}}$. On the other hand, the probability of the PO condition of poor visibility due to precipitation is likely higher under the HBSS condition of reduced road friction.

The sub-claims **G**-$\mathtt{PO}_{i,j}$-**MR** and **G**-$\mathtt{PO}_{i,j}$ correspond to the terms in the summation in Eqn 1 and denote the conditioned misperception rate and occurrence of the PO condition, respectively. Thus, the relationship between bounds is: $\gamma_{i,\mathrm{MR}} \geq \Sigma_j \gamma_{i,j,\mathrm{MR}} \gamma_{i,j,\mathrm{PO}\uparrow}$.

The effectiveness of this decomposition strategy depends on whether we can select a set $\{\mathtt{PO}_{i,j}\}_j$ of conditions that well cover the categories of external triggers for hazardous misperceptions defined by $\mathtt{MP}_i$. One way to achieve this is to base the set of PO conditions on a safety analysis method such as CV-HAZOP [19, 20]. CV-HAZOP defines a systematic method of exhaustively identifying modes of interference with a computer vision (CV) process by first modeling the CV process, and then using guide words in the spirit of HAZOP (Hazard and Operability Study) to identify how the process can be disturbed. The resultant list of interference modes can then be filtered and aggregated based on the specifics of the CV task [21], ODD, HBSS and MP conditions to produce a specifically applicable PO set. A similar strategy could be used to analyze any perceptual task.

Goal **G**-$\mathtt{PO}_{i,j}$ *and Solution* **Sn**-$\mathtt{PO}_{i,j}$

The claim in this goal bounds the occurrence of ADS drives in the ODD that exhibit the condition $\mathtt{PO}_{i,j}$. Here we require both upper and lower bounds to allow us to use the following for the nominal case:

$$P_C(\mathtt{PO}_{i,\mathrm{Nom}}) = 1 - \sum_{j \neq \mathrm{Nom}} P_C(\mathtt{PO}_{i,j}) \leq 1 - \sum_{j \neq \mathrm{Nom}} \gamma_{i,j,\mathrm{PO}\downarrow}$$

As with **G**-$\mathtt{HBSS}_i$, it may be reasonable to assume that the occurrence of these conditions are independent of the specifics of the ADS design and

[1] Note that in the current version of ISCaP we assume that exposure is frequency-based. Additionally supporting duration-based exposure is left as future work.

choice of component C; thus, the bounds $\gamma_{i,j,\texttt{PO}\downarrow}$ and $\gamma_{i,j,\texttt{PO}\uparrow}$ could be based on generic empirical or analytical sources of data. For example, information on precipitation frequency, visibility and sunlight variation may be sourced from weather bureaus, the frequency of different vehicle colors and shapes could come from vehicle sales statistics, etc. An important additional consideration is that the occurrence of these conditions may vary based on the geography in which the ADS will operate.

Goal **G-$\texttt{PO}_{i,j}$-MR**

This claim bounds the rate of $\texttt{MP}_i$ misperceptions within the context of a particular PO condition. Estimating the bound $\gamma_{i,j,\texttt{MR}}$ directly is difficult since $\texttt{MP}_i$ is a condition on *sequences* of misperceptions, but component C only operates at the single frame level producing individual misperceptions. Thus, there is a need to link these two levels of representation of misperceptions. The solution is to decompose the probability of $\texttt{MP}_i$ with respect to the probabilities of its constituent frame misperception patterns.

Strategy **S-$\texttt{PO}_{i,j}$-fMP**

The condition $\texttt{MP}_i$ is expressed in terms of a set $\{\texttt{fMP}_{i,k}\}_k$ of frame misperception patterns. For each $\texttt{fMP}_{i,k}$, a sub-claim **G-$\texttt{fMP}_{i,j,k}$-MR** is created. In addition, we specify a *linking expression* $F_{i,j}$, given in context node **C-Link**$_{i,j}$, that bounds the probability of $\texttt{MP}_i$ in terms of the probabilities of $\{\texttt{fMP}_{i,k}\}_k$.

Although the set $\{\texttt{fMP}_{i,k}\}_k$ is fixed for a given $\texttt{MP}_i$, the linking expression can vary depending on condition $\texttt{PO}_j$ because this can affect relationship between probabilities. For this reason, $F_{i,j}$ depends on both indices $_i$ and $_j$. For example, in the HMP `IL` the misperception pattern of repeated frame mis-localization events can cause an incorrect speed estimate of the on-coming vehicle, leading to a hazardous left turn. If a sequence of mis-localization events are caused by interference from precipitation, then each event can be considered to be independent in the context of precipitation PO condition. However, if the mis-localizations are caused by the reflectance characteristics of the on-coming car's surface (i.e., reflectance PO condition), then the events are not independent because the reflectance problem will continue to occur in all events.

Goal **G-$\texttt{fMP}_{i,j,k}$-MR**

Since each $\texttt{fMP}_{i,k}$ is defined over individual frames, the claim for **G-$\texttt{fMP}_{i,j,k}$-MR** uses the distribution $P_{C,fr}$ of frames rather than the distribution P_C of states in ADS drives. However, the probability is still conditional, restricted to frames occurring during drives satisfying $\texttt{HBSS}_i$ and $\texttt{PO}_j$. Since this goal involves testing or analysis at the unit level, it is marked as an away goal for **M-*C*-unit**.

Note that different HMPs may share reliance on the same frame misperception patterns. For example, the fMP $\texttt{FN}_{20}$, representing FNs within 20 meters of the ego vehicle, may be used in definitions of different MP conditions. However, even if in $\texttt{HMP}_i$ and $\texttt{HMP}_{i'}$, $\texttt{fMP}_{i,k} = \texttt{fMP}_{i',k'} = \texttt{FN}_{20}$, the goals **G-$\texttt{fMP}_{i,j,k}$-MR** and **G-$\texttt{fMP}_{i',j',k'}$-MR** that bound its occurrence remain distinct because they depend on different HBSS and PO conditions.

A key contribution of ISCaP is that the set of claims $\{$**G-$\texttt{fMP}_{i,j,k}$-MR**$\}_{i,j,k}$ can be seen to define a set of *risk-aware performance metrics* for evaluating a component C implementing task T:

Definition 9 (Risk-aware Performance Metric) *Given test dataset* $\text{TDS}_{i,j} = \{(x_l, y_l)\}_l$ *drawn from conditional distribution* $P_{C,fr}((x,y)|\texttt{HBSS}_i,\texttt{PO}_j)$, *a* ***risk-aware performance metric*** $m_{i,j,k}$ *for component C is defined as:*

$$m_{i,j,k} = \frac{1}{|\texttt{TDS}_{i,j}|} \sum_{(x,y)\in \texttt{TDS}_{i,j}} \mathbf{1}[C(x) \notin \texttt{fMP}_{i,k}(y)]$$

Metric $m_{i,j,k}$ is a *performance* metric because it computes a measure of the misperceptions produced by C. It is *risk-aware* because it only counts hazardous misperceptions and ignores benign ones. Unlike "generic" performance metrics typically used for evaluating perception components (e.g., recall, mAP, AuPR) that count any deviation from ground truth as bad, here only deviations that satisfy $\texttt{fMP}_{i,k}$ are considered bad. Furthermore, $\texttt{fMP}_{i,k}$ is derived from, and is traceable to, safety claims about C via the argument in ISCaP. Finally, another benefit is that each metric $m_{i,j,k}$ focuses on a different system hazard and perception context (via HBSS i and PO j) and a different aspect of the performance of C (via fMP k), allowing a more fine-grained tuning of how C impacts system safety.

Goal **G-$\texttt{fMP}_{i,j,k}$-MR** is a unit-level claim on the performance of component C; thus, it is expressed as an away-goal that links to module **M-*C*-unit**. While the full details of this argument are beyond the scope of this paper[2], the metric $m_{i,j,k}$ can be used as part of a black-box testing argument since, from a statistically standpoint, it is a sample estimate of $P_{C,fr}(\texttt{fMP}_{i,k}|\texttt{HBSS}_i,\texttt{PO}_{i,j})$.

To use $m_{i,j,k}$ to compute $\gamma_{i,j,k}$ we need to take into account the sampling error of this estimate with sample size $N = |\texttt{TDS}_{i,j}|$. The sampling distribution is binomial but can be approximated with a Normal distribution when $N > 30$. Then the upper bound σ_q of the q% confidence interval on the error of $m_{i,j,k}$ is given by

$$\sigma_q = z(q)\sqrt{\frac{m_{i,j,k}(1-m_{i,j,k})}{N}}$$

Where, $z(q)$ is the $\frac{100+q}{200}$ quantile of the standard normal distribution. For example, if $q = 99$ then $z(q) = 2.58$. Then we can define $\gamma_{i,j,k} = m_{i,j,k} + \sigma_q$. This bound assumes that $\texttt{TDS}_{i,j}$ is a representative sample and data adequacy arguments in **M-*C*-unit** are applicable here.

[2] See the recent related work discussed in the introduction for approaches to the argument in **M-*C*-unit**.

Applying ISCaP in Practice

In this section, we discuss various topics related to the use of the ISCaP during the development of C and for constructing argument **M-*C***.

Accommodating Multiple Severity Levels

ISCaP is designed for a single high severity event (i.e., crash); however, safety cases typically address multiple severity levels (e.g., four severity levels in ISO 26262). A simple and flexible way to do this is to create multiple *parallel* arguments by adding an implicit parameter L to the template representing N severity levels. Then each GSN node can be interpreted as an array of N nodes corresponding to the severity levels. For example, goal **G-$\texttt{PO}_{i,j}$-MR** is interpreted as $P_C(\texttt{MP}_i[L]|\texttt{HBSS}_i[L],\texttt{PO}_{i,j}[L]) \leq \gamma_{i,j,\texttt{MR}}[L]$ allowing the definitions of conditions $\texttt{MP}_i$, $\texttt{HBSS}_i$, $\texttt{PO}_{i,j}$ and bound $\gamma_{i,j,\texttt{MR}}$ to be qualified by severity level. When a condition or bound is not dependent on severity, the severity parameter need not be considered in the definition.

For example, in the HBSS for HMP `IL` (discussed in the preliminaries section), the severity of the hazardous behaviour of turning left when the on-coming vehicle is too close varies depending on the speed of the on-coming vehicle. Thus, $\texttt{HBSS}_{\texttt{IL}}[L]$ represents a version of the HBSS condition with a different speed depending on L. Correspondingly, in $\texttt{MP}_{\texttt{IL}}[L]$ the number of required $\texttt{FN}_{20}$ misperceptions may reduce with increasing speed. However, if $\texttt{MP}_{\texttt{IL}}$ represents misperceptions of a direct speed measurement (e.g., via Radar) then it would not be dependent on L.

Top-down vs. Bottom-up Development

ISCaP takes a formal deductive approach by using the decompositional structure of the argument to express

特集

bound γ_C in the top-level claim in terms of similar bounds in lower level claims. The mathematical relationship between the bounds given in each strategy defines formal traceability from every lower-level claim to the top-level claim.

The formal traceability allows γ_C to be interpreted in either of two ways: i) as a *safety target* that component C must achieve, or, ii) as a conservative (i.e, upper bounding) estimate of the probability P_C (MCrash) based on its sub-claims. Interpretation (i) supports a top-down development strategy in which a safety target from the main ADS argument **M-ADS** is systematically allocated to different sub-claims to define component level requirements for **M**-*C*-**unit**. Interpretation (ii) supports a bottom-up development strategy in which confidence about leaf claims (as expressed by the bounds on these) are correctly propagated upward to **M-ADS**. This can be used to guide development effort on the component by identifying which HMPs, fMPs, and PO conditions contribute the most to risk.

Iterative and Continuous Development

Developing a safety case is costly, thus any iteration regarding a component that impacts its safety case must incur a cost. However, ML-based perception components may be frequently re-trained as new useful training samples become identified or when domain shift occurs. In addition, some development methods, such as active learning, require multiple retraining.

ISCaP has the beneficial property that it largely independent of the particular choice of component C. The structure (i.e., choice of claims) is determined by perception task T and is independent of C. The only claims that are directly affected by C are **G**-$\texttt{fMP}_{i,j,k}$-**MR** because the bounds $\gamma_{i,j,k}$ must be recomputed or rechecked when C changes (although, the test datasets $\texttt{TDS}_{i,j}$ are independent of C). The bounds for higher claims can be automatically recomputed from these. Thus, the impact of iterating C is well localized to limit the safety case change cost.

Case Study

In this section, we demonstrate the instantiation of ISCaP for an object detection task OD and define the claims associated with one HMP in detail—HMP `SCA` (“stopped car ahead”). The component implementing the task `OD` is PoPi, the Point Pillars object detector for LiDAR point clouds [22]. Knowing the internal details of `PoPi` is not needed for understanding this case study.

The ODD we assume consists of driving conditions represented in the KITTI object detection dataset [23], which is naturalistic, summer, clear weather, day-time driving in a small city (based on Karlsruhe, Germany). The ADS we used in which `PoPi` operates has the following details relevant to the case study:

- In the perception pipeline, the output of `PoPi` feeds a *Tracker* component that maintains a model of past and predicted (near future) trajectories of all relevant road users. It takes $n_{trk} = 9$ consecutively missed frames by `PoPi` for tracker to lose the track of an object.
- The comfortable and maximum (i.e., emergency) braking rates capable by the ego vehicle are $a_{b,comf} = 2.01 m/s^2$ and $a_{b,emerg} = 2.86 m/s^2$, respectively. The maximum acceleration is $a_{max} = 3.02 m/s^2$. Maximum speed is $11.11 m/s$ ($40 km/h$).
- The frame rate for task OD is $10\,f/s$

We use FOL to sketch formal definitions throughout this section.

Goal **G**-`PoPi`

The task `OD` is defined by ground truth function $\texttt{OD} : X \rightarrow Y$ where X are LiDAR point clouds and Y are sets of bounding boxes classified as car, pedestrian or cyclist. Component `PoPi` implementing `OD` defines function PoPi : $X \rightarrow Y$ and may deviate from `OD` resulting in FNs and FPs. Fig. 5 shows two examples in the context of a hazardous frame misperception pattern `FNA` defined below. The objective of the claim in this goal is to show that the occurrence of hazardous misperceptions by `PoPi` is bounded to an acceptable level $\gamma_{\texttt{PoPi}}$ as determined by the main ADS argument **M-ADS**.

Strategy **S-HMP**

In this example, we are only considering the HMP `SCA`.

Goal **G**-$\texttt{HMP}_{\text{SCA}}$

We assume that, in the hazard analysis used by the main ADS argument, the following hazardous operational situation is identified: the ego vehicle, gets close enough to a stopped car ahead, such that, unless braking is applied by the ego vehicle a crash (i.e., collision with stopped vehicle) will result. Thus, any *braking interruption* of sufficient time length to cause an accident is a hazardous behaviour of the ADS in this situation.

This system-level hazard could have different causes including slippery roads and malfunctioning brakes but since a misperception in performing OD can also be a cause, we create the HMP $\texttt{SCA}= (\texttt{HBSS}_{\text{SCA}}, \texttt{MP}_{\text{SCA}})$ to represent this case within argument **M**-`PoPi`.

To define $\texttt{HBSS}_{\text{SCA}}$, we assume the following safe driving policy: a stopped vehicle should be detected sufficiently far ahead to allow the ego vehicle, braking comfortably, to stop at a stand-still distance of 4m from it. In addition, based on simple physics-based modeling we observe that if the ego vehicle is travelling at speed v, it must brake with at least a_b to avoid a collision with a stopped vehicle distance $x = \frac{v^2}{2a_b}$ ahead.

Definition 10 ($\texttt{HBSS}_{\text{SCA}}(d)$) *For all drives* $d \in D$, $\texttt{HBSS}_{\text{SCA}}(d)$ *iff there is a stopped vehicle* $x_{SC} = \frac{v_{init}^2}{2a_{b,comf}} + 4$ *meters ahead of the ego vehicle and the ego speed in state* $d[1]$ *is* $v_{init} > 0$.

Thus, $\texttt{HBSS}_{\text{SCA}}(d)$ then, under normal operation, the ego vehicle stops safely. However, if there is a braking interruption, it will need to compensate by braking more intensely (up to a maximum of $a_{b,emerg}$) once braking resumes. The braking interruption becomes a hazardous behaviour if at any point, $x < \frac{v^2}{2a_{b,emerg}}$, where v is the ego vehicle speed and x is the remaining distance to the stopped vehicle, since a collision will result.

(y_a) (y_b)

Fig. 5 Illustration of frame misperception pattern `FNA` on two point cloud outputs of task `OD` with same input x (predictions are dashed boxes, ground truth are solid boxes). In output y_a, `FNA` occurs since the vehicle ahead of ego vehicle is FN (shown larger), thus $y_a \in \texttt{FNA}(\texttt{OD}(x))$. In output y_b, the vehicle ahead is TP, thus FNA does not occur and $y_b \notin \texttt{FNA}(\texttt{OD}(x))$. Accuracy on other objects is irrelevant for `FNA`. Classes: Car (red), Pedestrian (orange), Cyclist (purple), Other (gray)

Different misperceptions could lead the ADS to interrupt braking, including: not detecting the stopped car, misjudging the position of the stopped car as not obstructing the ego vehicle, and misjudging the speed of the car ahead as not stopped. In this study, we focus on the first of these—missed detections (i.e.,

特集

FNs)—and assume that if the car is detected, other information about it will be correctly perceived.

To define $\texttt{MP}_{\texttt{SCA}}$, we first define the frame misperception pattern FNA : $Y \rightarrow pow(Y)$ identifying an FN for the vehicle ahead.

Definition 11 (Frame misperception pattern FNA)

$$\forall y, y_{\texttt{gt}} \in Y \cdot y \in \texttt{FNA}(y_{\texttt{gt}}) \textit{ iff } \langle \exists bb_{\texttt{gt}} \in y_{\texttt{gt}} \cdot \texttt{AheadOf}(y_{\texttt{gt}}, bb_{\texttt{gt}}, \texttt{Ego}) \rangle \wedge (\neg \exists bb \in y \cdot \texttt{TPMatch}(bb_{\texttt{gt}}, bb))$$

Where, $\texttt{AheadOf}(y,bb,bb')$ *iff bounding box bb* $\in y$ *is positioned ahead of bb' with no other in between,* Ego *is the bounding box for the ego vehicle and* $\texttt{TPMatch}(bb,bb')$ *iff bb' matches bb well enough to be considered a true positive prediction for bb.*

Fig. 5 illustrates FNA. To define $\texttt{MP}_{\texttt{SCA}}$ we must determine what sequences of FNA occurrences yield enough a braking interruption to be hazardous. To do this, we conducted a physics-based analysis combined with ADS simulation and determined that the worst-case (minimum) hazardous braking interruption is a single interruption of $t_{\text{crash}} = 0.48s$ beginning $19.65m$ from the stopped vehicle with $v_{init} = 11.11m/s$ (i.e., speed limit) and an ego vehicle acceleration of $a_{max} = 3.02m/s^2$ during the interruption. Given the frame rate of $10f/s$, this means that braking must be interrupted for 5 frames. However, since the tracker component requires $n_{trk} = 9$ consecutive FNs before the track is lost, PoPi must exhibit at least $n_{crash} = 5 + 9 = 14$ FNA occurrences to cause a hazardous braking interruption.

Definition 12 ($\texttt{MP}_{\texttt{SCA}}$) *For all drives* $d \in D$, $\texttt{MP}_{\texttt{SCA}}(d)$ *iff it contains at least* $n_{crash} = 14$, *not necessarily consecutive, occurrences of* FNA

Note that consecutiveness is not specified here because there are nonconsecutive sequences of FNAs that can be hazardous, but based on the above analysis, more than 14 FNAs are required.

Table 1 Computation of bounds $\gamma_{\texttt{SCA},j,\texttt{PO}\downarrow}$ and $\gamma_{\texttt{SCA},j,\texttt{PO}\uparrow}$.

j	$\frac{\lvert\texttt{TDS}_j\rvert}{\lvert\texttt{TDS}\rvert}$	$\sigma_{0.99}$	$\gamma_{\texttt{SCA},j,\texttt{PO}\downarrow}$	$\gamma_{\texttt{SCA},j,\texttt{PO}\uparrow}$
Nom	0.965	0.008	0.957	0.973
Crowd	0.035	0.008	0.027	0.043

Goal **G-Res** *and Solution* **Sn-Res**

Since we are only considering a single HMP in this case study, this goal is not applicable.

Goal **G-**$\texttt{MP}_{\texttt{SCA}}$**-N** *and Solution* **Sn-**$\texttt{MP}_{\texttt{SCA}}$**-N**

Evidence for this claim is based on the argument that since we are limiting the focus to braking interruptions due to OD misperceptions, this could only have been caused by occurrences of FNA misleading the ADS into believing there is no car ahead. Furthermore, by the physics/simulation analysis discussed in the section for goal **G-**$\texttt{HMP}_{\texttt{SCA}}$, producing a hazardous braking interruption requires at least 14 FNAs. Finally, by Def. 5, $\texttt{MP}_{\texttt{SCA}}$ represents all possible sequences containing at least 14 FNAs.

Therefore, we conclude $P_{\texttt{PoPi}}(\texttt{MCrash} \mid \texttt{HBSS}_{\texttt{SCA}}, \neg\texttt{MP}_{\texttt{SCA}}) \cong 0$ (approximation to account for a potential small error in analysis).

Goal **G-**$\texttt{MP}_{\texttt{SCA}}$**-CR** *and Solution* **Sn-**$\texttt{MP}_{\texttt{SCA}}$**-CR**

The physics/simulation analysis says that fewer than 14 FNAs cannot produce the hazardous behaviour, but it is still possible that 14 or more nonconsecutive FNAs may not be hazardous. For example, having two nonconsecutive groups of nine FNAs is nonhazardous due to the tracker compensation. Although a simulation study could yield a more accurate estimate of this crash rate, for this claim we take a conservative stance and set $\gamma_{\texttt{SCA,CR}} = 1$.

Goal **G-**$\texttt{HBSS}_{\texttt{SCA}}$ *and Solution* **Sn-**$\texttt{HBSS}_{\texttt{SCA}}$

We assume that the bound in this claim is based on information from the corresponding system-level hazard in **M-ADS**. For this case study, we conservatively assume that $\texttt{HBSS}_{\texttt{SCA}}$ occurs at every intersection. If the ego vehicle is travelling at the speed limit ($11.11m/s$) and the average distance between intersections is $500m$, then we let $\gamma_{\texttt{SCA,HBSS}} = 1.11/500 = 0.0022$ since $1.11m$ is travelled per state and one state ends an $\texttt{HBSS}_{\texttt{SCA}}$ drive per $500m$.

Goal **G-**$\texttt{MP}_{\texttt{SCA}}$**-MR** *and Strategy* **S-**$\texttt{MP}_{\texttt{SCA}}$**-PO**

The claim is $P_{\texttt{PoPi}}(\texttt{MP}_{\texttt{SCA}} \mid \texttt{HBSS}_{\texttt{SCA}}) \leq \gamma_{\texttt{SCA,MR}}$. To decompose HMP SCA we consider a single non-nominal PO condition: $\texttt{PO}_{\texttt{Crowd}}$ identifying crowded scenes defined as those containing more than 10 objects within 40m of the ego vehicle. This is based on the assumption that crowded scenes will contain more FNs than uncrowded.

Goals **G-**$\texttt{PO}_{\texttt{SCA},j}$ *and Solutions* **Sn-**$\texttt{PO}_{\texttt{SCA},j}$

We estimate the probability of crowded scenes by the proportion of such scenes in the KITTI test dataset and use the 99% confidence interval to account for sampling error. The bound estimates are given in Table 1.

Table 2 Computation of metric $m_{\texttt{SCA},j,\texttt{FNA}}$.

j	$\lvert\texttt{TDS}_{\texttt{SCA},j}\rvert$	$m_{\texttt{SCA},j,\texttt{FNA}}$
Nom	1479	0.052
Crowd	67	0.045

Table 3 Computation of bounds $\gamma_{\texttt{SCA},j,\texttt{FNA}}$ and $\gamma_{\texttt{SCA},j,\texttt{MR}}$.

j	$\sigma_{0.99}$	$\gamma_{\texttt{SCA},j,\texttt{FNA}}$	$\gamma_{\texttt{SCA},j,\texttt{MR}}$
Nom	0.015	0.067	1.15×10^{-5}
Crowd	0.065	0.110	2.06×10^{-3}

Goals **G-**$\texttt{PO}_{\texttt{SCA},j}$**-MR** *and Strategies* **S-**$\texttt{PO}_{\texttt{SCA},j}$**-fMP**

We decompose each goal **G-**$\texttt{PO}_{\texttt{SCA},j}$**-MR** using the frame misperception pattern FNA. If we assume each FNA occurrence is independent, then given Def. 12, we have the following expression as expected value of the cumulative binomial over drives of varying length n,

$$P_{\texttt{PoPi}}(\texttt{MP}_{\texttt{SCA}} \mid \texttt{HBSS}_{\texttt{SCA}}, \texttt{PO}_{\texttt{SCA},j}) = \mathbb{E}_n\left[\sum_{l=14}^{n} \binom{n}{l} p_j^l (1-p_j)^{(n-l)}\right]$$

Where, $p_j = P_{\texttt{PoPi},fr}(\texttt{FNA} \mid \texttt{HBSS}_{\texttt{SCA}}, \texttt{PO}_{\texttt{SCA},j})$. Note that the cumulative binomial is monotonic in n and max $n = 55$ occurs when starting with max $v_{init} = 11.11m/s$. Thus, for the linking expression we use:

$$P_{\texttt{PoPi}}(\texttt{MP}_{\texttt{SCA}} \mid \texttt{HBSS}_{\texttt{SCA}}, \texttt{PO}_{\texttt{SCA},j}) \leq \sum_{l=14}^{55} \binom{55}{l} p_j^l (1-p_j)^{(55-l)}$$

Goals **G-**$\texttt{fMP}_{\texttt{SCA},j,\texttt{FNA}}$**-MR**

The objective here is to estimate $P_{\texttt{PoPi}}(\texttt{FNA} \mid \texttt{HBSS}_{\texttt{SCA}}, \texttt{PO}_{\texttt{SCA},j})$. To do this, we extracted frames from the KITTI dataset conforming to $\texttt{HBSS}_{\texttt{SCA}}$ and $\texttt{PO}_{\texttt{SCA},j}$ to form datasets $\texttt{TDS}_{\texttt{SCA},j}$ and used these to test PoPi and compute the risk-aware metrics $m_{\texttt{SCA},j,\texttt{FNA}}$. Specifically, $\texttt{TDS}_{\texttt{SCA},j}$ consisted of frames in which there was a car ahead of the ego vehicle. These frames over-approximate the set of frames from $\texttt{HBSS}_{\texttt{SCA}}$ drives because it doesn't consider the speed of the ego vehicle (v_{init}) or whether the car ahead is stopped or moving; however, the single-frame misperception performance is unaffected by this. Tables 2 and Tables 3 give the results assuming a 99% confidence bound.

Combining the Case Study Results

By propagating the values determined for the bounds in the leaf claims upward using the expressions in the strategies (See Fig. 2), we obtain bound values for higher level claims. Based on the values in Tables 1 and Table 3 for PO conditions Nom and Crowd, we have:

$$\begin{aligned}\gamma_{\texttt{SCA,MR}} &= \gamma_{\texttt{SCA,Nom,PO}\uparrow}\gamma_{\texttt{SCA,Nom,MR}} + \gamma_{\texttt{SCA,Crowd,PO}\uparrow}\gamma_{\texttt{SCA,Crowd,MR}} \\ &= (0.973)(1.15 \times 10^{-5}) + (0.043)(2.06 \times 10^{-3}) \\ &= 1.12 \times 10^{-5} + 8.86 \times 10^{-5} \\ &= 9.98 \times 10^{-5}\end{aligned}$$

特集

We can see that the dominant contribution is from PO condition Crowd. The bound for **G-HMP**$_{SCA}$ can then be computed as:

$$\gamma_{SCA} = \gamma_{SCA,CR}\gamma_{SCA,MR}\gamma_{SCA,HBSS} = (1)(9.98 \times 10^{-5})(2.2 \times 10^{-3}) = 2.20 \times 10^{-7}$$

Finally, the bound for the top claim is:

$$\gamma_{PoPi} = \gamma_{res} + \gamma_{SCA} = \gamma_{res} + 2.20 \times 10^{-7}$$

Since this analysis is based on only one HMP, it is partial, and we leave the residual bound γ_{res} as an unspecified variable term.

Summary/Conclusions

In this paper, we address a gap in the research on safety cases for automated driving and ML by proposing the ISCaP template safety argument linking the system-level arguments and unit-level arguments. The template provides a formally deductive claim decomposition approach tailored to perception and identifies a set of risk-aware safety metrics that can be used to evaluate perception components. We demonstrate the applicability of ISCaP through a detailed case study.

As part of future work, we explore several directions. First, the issue of *confidence* needs special consideration. It is well known that the strength of an argument depends on the level of confidence in claims generated by the evidence and there is much research on defining and propagating confidence within an argument. We have addressed this using confidence intervals in some claims, but it requires a more systematic treatment throughout the argument in ISCaP. Second, while an underlying methodology for identifying HMPs and PO conditions is suggested in various claims, this needs comprehensive elaboration.

Third, we intend to extend ISCaP to address some common variations, including the following:

- Currently, only a single linking expression is allowed for connecting MP conditions to their constituent frame misperception patterns but it is clear that the expression could also depend on other factors. For example, in the case study, we assume independence between occurrences of `FNA`s. This is appropriate for "typical" cars, but for cars that are unusual it is likely that if there is one `FNA`, then all subsequent detections will be `FNA`, so the independence assumption is not valid. Thus, having different linking expressions for typical and atypical cars is needed here.
- Currently, HBSS exposure is assumed to be frequency-based, which is appropriate for cases such as when stopping for a stopped vehicle. All uses of HBSS exposure in ISCaP are based on this assumption. However, ISO 26262 also allows for duration-based HBSS exposure, such as when following a vehicle at a steady pace. The template should allow either approach for generality.
- Currently, the frame rates of input and output of perception task T are assumed to be the same, but in general, they do not have to match. For example, a multi-rate tracker might use 30Hz camera input, a 10Hz LiDAR input and output at 20Hz. The template should be generalized to accommodate such cases.
- Currently, PO conditions are assumed to represent special (non-nominal) cases and the distinguished PO condition `Nom` captures all remaining (assumed to be nominal) cases. For greater generality, the `Nom` PO condition should be split to further distinguish nominal cases from any residual non-nominal cases that may exist.

Finally, we intend to do a detailed feasibility study of the approach and identify places it can be improved with the hope that eventually ISCaP can be adopted to support industrial practice.

References

1. M. Wood, P. Robbel, M. Maass, R. D. Tebbens, M. Meijs, M. Harb, and P. Schlicht, "Safety first for automated driving," *Aptiv, Audi, BMW, Baidu, Continental Teves, Daimler, FCA, HERE, Infineon Technologies, Intel, Volkswagen*, 2019.
2. N. Webb, D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov, and T. Daniel, "Waymo's safety methodologies and safety readiness determinations," *arXiv preprint arXiv:2011.00054*, 2020.
3. R. Salay, R. Queiroz, and K. Czarnecki, "An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software," *SAE Technical Paper*, 2018, doi:10.4271/2018-01-1075.
4. J. Rushby, "The interpretation and evaluation of assurance cases," *Comp. Science Laboratory, SRI International, Tech. Rep. SRI-CSL-15-01*, 2015.
5. Z. Kurd, T. Kelly, and J. Austin, "Developing artificial neural networks for safety critical systems," *Neural Computing and Applications*, vol. 16, no. 1, pp. 11–19, 2007, doi:10.1007/s00521-006-0039-9.
6. S. Burton, L. Gauerhof, and C. Heinzemann, "Making the case for safety of machine learning in highly automated driving," in *International Conference on Computer Safety, Reliability, and Security*, pp. 5–16, Springer, 2017.
7. A. C. W. Group *et al.*, "Goal structuring notation community standard (version 3)," 2021.
8. E. Wozniak, C. Cârlan, E. Acar-Celik, and H. J. Putzer, "A safety case pattern for systems with machine learning components," in *International Conference on Computer Safety, Reliability, and Security*, pp. 370–382, Springer, 2020.
9. C. Picardi, C. Paterson, R. D. Hawkins, R. Calinescu, and I. Habli, "Assurance argument patterns and processes for machine learning in safety-related systems," in *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020)*, pp. 23–30, CEUR Workshop Proceedings, 2020.
10. R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *arXiv preprint arXiv:1905.04223*, 2019.
11. L. Gauerhof, R. Hawkins, C. Picardi, C. Paterson, Y. Hagiwara, and I. Habli, "Assuring the safety of machine learning for pedestrian detection at crossings," in *International Conference on Computer Safety, Reliability, and Security*, pp. 197–212, Springer, 2020.
12. R. Bloomfield, G. Fletcher, H. Khlaaf, L. Hinde, and P. Ryan, "Safety case templates for autonomous systems," *arXiv preprint arXiv:2102.02625*, 2021.
13. Aurora, "Aurora Safety Case Framework, Version 1," 2021.
14. J. Vaicenavicius, T. Wiklund, A. Grigaite, A. Kalkauskas, I. Vysniauskas, and S. Keen, "Self-driving car safety quantification via component-level analysis," *arXiv preprint arXiv:2009.01119*, 2020.
15. International Organization for Standardization, *ISO 26262: Road Vehicles – Functional Safety*, 2018. 2nd edition.
16. International Organization for Standardization, *ISO/AWI PAS 21448: Road Vehicles – Safety of the Intended Functionality*, 2019.
17. W. G. Najm, J. D. Smith, M. Yanagisawa, *et al.*, "Pre-crash scenario typology for crash avoidance research," tech. rep., United States. National Highway Traffic Safety Administration, 2007.
18. R. Salay, M. Angus, and K. Czarnecki, "A safety analysis method for perceptual components in automated driving," in *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 24–34, IEEE, 2019.
19. H. Kuwajima, H. Yasuoka, and T. Nakae, "Engineering problems in machine learning systems," *Machine Learning*, vol. 109, no. 5, pp. 1103–1126, 2020.
20. O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner, "Cv-hazop: Introducing test data validation for computer vision," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2066–2074, 2015.
21. O. Zendel, K. Honauer, M. Murschitz, M. Humenberger, and G. Fernandez Dominguez, "Analyzing computer vision data-the good, the bad and the ugly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1980–1990, 2017.
22. A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.
23. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, IEEE, 2012.

Acknowledgments

The University of Waterloo authors were supported, in part, by DENSO CORPORATION.

Definitions/Abbreviations

ISCaP	Integration Safety Case for Perception
ADS	automated driving system
ML	machine learning
OD	object detection
ODD	operational design domain
GSN	Goal Structuring Notation
MP	misperception pattern
FOL	First Order Logic
HMP	hazardous misperception pattern
fMP	frame misperception pattern
HBSS	hazardous behaviour sensitive scenario
PO	perception-only
LiDAR	Light Detection and Ranging

著者

桑島 洋
くわじま ひろし

ソフト生産革新部
AI 品質関連の社内プロセス構築と渉外活動に従事

安岡 宏俊
やすおか ひろとし

博士（情報科学）
ソフト生産革新部
AI 品質関連の要素技術開発に従事

中江 俊博
なかえ としひろ

ソフト生産革新部
AI とソフトウェア工学関連のプロジェクトマネジメントに従事

Rick Salay, Ph.D.

ウォータールー大学 Waterloo Intelligent Systems Engineering Lab 研究員
システム工学の研究に従事

Krzysztof Czarnecki, Ph.D.

ウォータールー大学 Electrical and Computer Engineering 教授
自動運転の研究に従事

Vahdat Abdelzad, Ph.D.

ウォータールー大学 Waterloo Intelligent Systems Engineering Lab 研究員
機械学習モデルの安全の研究に従事

Chengjie Huang

ウォータールー大学 Waterloo Intelligent Systems Engineering Lab
自動運転の研究に従事

Maximilian Kahn

ウォータールー大学 Waterloo Intelligent Systems Engineering Lab
自動運転の研究に従事

Van Duong Nguyen

ウォータールー大学 Waterloo Intelligent Systems Engineering Lab
自動運転の研究に従事