A Hybrid-AI Approach for Competence Assessment of Automated Driving Functions

Jan-Pieter Paardekooper[1,2 *], **Mauro Comi**[1 *], **Corrado Grappiolo**[3 *],
Ron Snijders[4 *], **Willeke van Vught**[5], **Rutger Beekelaar** [1]
[1]TNO - Integrated Vehicle Safety, Helmond, The Netherlands
[2]Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands.
[3]TNO - Data Science, Den Haag, The Netherlands
[4]TNO - Monitoring & Control Services, Groningen, The Netherlands
[5]TNO - Perceptual and Cognitive Systems, Soesterberg, The Netherlands
jan-pieter.paardekooper@tno.nl

Abstract

An increasing number of tasks is being taken over from the human driver as automated driving technology is developed. Accidents have been reported in situations where the automated driving technology was not able to function according to specifications. As data-driven Artificial Intelligence (AI) systems are becoming more ubiquitous in automated vehicles, it is increasingly important to make AI systems situational aware. One aspect of this is determining whether these systems are competent in the current and immediate traffic situation, or that they should hand over control to the driver or safety system.

We aim to increase the safety of automated driving functions by combining data-driven AI systems with knowledge-based AI into a hybrid-AI system that can reason about competence in the traffic state now and in the next few seconds.

We showcase our method using an intention prediction algorithm that is based on a deep neural network and trained with real-world data of traffic participants performing a cut-in maneuver in front of the vehicle. This is combined with a unified, quantitative representation of the situation on the road represented by an ontology-based knowledge graph and first-order logic inference rules, that takes as input both the observations of the sensors of the automated vehicle as well as the output of the intention prediction algorithm. The knowledge graph utilises the two features of *importance*, based on domain knowledge, and *doubt*, based on the observations and information about the dataset, to reason about the competence of the intention prediction algorithm.

We have applied the competence assessment of the intention prediction algorithm to two cut-in scenarios: a traffic situation that is well within the operational design domain described by the training data set, and a traffic situation that includes an unknown entity in the form of a motorcycle that was not part of the training set. In the latter case the knowledge graph correctly reasoned that the intention prediction algorithm was incapable of producing a reliable prediction.

This shows that hybrid AI for situational awareness holds great promise to reduce the risk of automated driving functions in an open world containing unknowns.

Automated driving is one of the most appealing applications of artificial intelligence in an open world. It holds the promise of reducing the number of casualties (1.35 million yearly (WHO 2018)), increasing the comfort of travel by taking over the driving task from humans, and bringing mobility to those unable to drive. While fleets of fully automated vehicles that can run unrestrained in an open world are still far away (Koopman and Wagner 2016), many vehicles are already equipped with Advanced Driver Assistence Systems (Okuda, Kajiwara, and Terashima 2014), like Lane Keep Assist and Adaptive Cruise Control. According to The Geneva Convention on road traffic of 1949 and the Vienna Convention on road traffic 1968, on which many countries base their national traffic laws, a human driver has to be present in the vehicle (Vellinga 2019). Artificial Intelligence (AI) opens up the possibility of automation in increasingly complex situations, but also makes it increasingly complex for human drivers to understand the limitations of the system (Thill, Hemeren, and Nilsson 2014).

The tremendous success of Deep Neural Networks (DNNs) in the recent years (LeCun, Bengio, and Hinton 2015) has lead to many applications in automated driving, ranging from perception (Cordts et al. 2016) and trajectory prediction (Deo and Trivedi 2018) to decision making (Bansal, Krizhevsky, and Ogale 2019). The strength of DNNs is the capability to deal with complex problems, but one important drawback for their application in safety-critical systems is how they deal with new situations (Hendrycks and Gimpel 2017; McAllister et al. 2017). DNNs learn a (possibly very complex) mapping from input data to output, but they lack an understanding of the deeper causes of this output. Hence, these algorithms cannot reason about whether they are competent to produce reliable output based on the input data. To safely apply DNNs (or any learning algorithm) in automated vehicles, we need to add situational awareness: the comprehension whether the system understands the current environment and is capable of producing reliable output.

In this work we describe a hybrid-AI approach (van Harmelen and ten Teije 2019; Meyer-Vitali et al. 2019) to situational awareness. In this approach, a data-driven AI is coupled to a knowledge graph with reasoning capabilities.

[*]Authors contributed equally

The current application is a DNN that predicts the intention of other road users to merge into the lane of the ego vehicle (cut-in maneuver). This is combined with a knowledge graph of the traffic state that relates the current situation to what the predictor has learned from the training data. The knowledge graph reasoner returns an estimate on the reliability of the predictor, which it forecasts into the immediate future (2 seconds ahead) to be able to warn the driver or safety system in advance that takeover of control is imminent in the near future.

Related work

In the automotive domain, situation awareness (Endsley 1995) is a term often used to describe the readiness of human drivers to make good decisions (Endsley 2020). It is based on perception of the environment, comprehension of the current situation, and projection into future. Here we extend situation awareness to an AI system in the vehicle, where we add the assessment of competence in the current situation to the comprehension of the current situation.

Machine learning methods tend to underperform when the distributions of the test dataset and training dataset differ significantly. Throughout the paper, we refer to data samples drawn from the training set distribution as in-distribution (ID), while samples drawn from a different distribution as out-of-distribution (OOD). DNNs often attribute high confidence to the classification or prediction of OOD samples (Hein, Andriushchenko, and Bitterwolf 2019); this behaviour, which is especially valid for softmax classifiers (Hendrycks and Gimpel 2016), can have dramatic consequences in applications where model reliability and safety are priorities. Various papers attempt to increase models' robustness by calibrating the predicting probability estimates (Guo et al. 2017) or by injecting small perturbations to the input data (Liang, Li, and Srikant 2017). Density estimation methods are also leveraged to detect OOD observations: the likelihood over the in-distribution sample space can be approximated (Dinh, Sohl-Dickstein, and Bengio 2016) (Ren et al. 2019) and used to compute the likelihood of new observations, thus detecting those samples that lie in low-density regions.

Another way of dealing with OOD observations is to have the DNN output more accurate certainty values. Several approaches have been described in the literature, ranging from Monte Carlo Dropout (Gal and Ghahramani 2016) to adding a Gaussian distribution over the weights in the last layer of a ReLU network (Kristiadi, Hein, and Hennig 2020). It has been shown that Bayesian deep learning is important for the safety of automated vehicles (McAllister et al. 2017).

Method

To assess the competence of data-driven-AI automated-driving capabilities we propose a pipelined framework, depicted in Figure 1. The framework receives as input the observations of the current road situation and, via a pipelined information flow, outputs the decision on whether the driving mode should remain autonomous or should be handed over to the human driver or backup safety system. The framework's internal structure is divided into three modules: Intention Predictor, Reasoner and Competence Assessment.

Raw observations related to each target vehicle, such as their speed and acceleration, are fed to the Intention Predictor. This module processes the information via two submodules. The first one is a deep neural network trained to output (predict) whether a given target vehicle will perform a cut-in maneuver (Cut-in Classifier). The second sub-module is a Feature Uncertainty Estimator. It holds univariate densities of the classifier's training set input features and provides information on the in-distribution likelihood of the network's input data.

The Intention Predictor's output, the observations related to road geometry (e.g. presence of entry lanes) and lane visibility are fed to the framework's second module, the Reasoner. The reasoner — characterised by an ontology and first-order logic rules — fuses the input observations with domain knowledge (encoded in the ontology and in the rules) into a knowledge graph. The graph realises the framework's situational awareness, as it holds a unified representation of the current situation and is aware of what entities are important and doubtful.

The graph is then fed to the last module of our framework, Competence Assessment. The module first organises its present and past situation-aware knowledge. Then it projects such knowledge into the future. Finally, it decides whether such forecast is outside the autonomous system's competence level.

In the next part, we will describe each module in more detail.

Simulation Environment

For simulation we use CARLA, an Open Source simulator which aims to support the development, training, and validation of autonomous driving systems (Dosovitskiy et al. 2017). The scenarios are defined using OpenSCENARIO[1], an open format used to describe synchronized maneuvers of vehicles.

For a given location of the Ego Vehicle (EV), we use the API of CARLA to extract a world model of the road situation. This world model includes the number of lanes, the presence of an entrance lane and all Target Vehicles (TVs). For each TV, the velocity, acceleration, angle and position relative towards EV is determined. The lane visibility v is calculated as $v = \frac{d}{s}$, where $0 \leq v \leq 1$. Here, d is the distance of the closest TV on that lane and s the scope of EV in meters ($s := 50m$ by default).

Intention predictor

Predicting the intention of a vehicle to perform a cut-in can be framed as a binary classification task; the two labels to classify are "cut-in" and "not cut-in". A data point labeled as "cut-in" refers to the collected information at timestep t for a TV that performs a cut-in between t and t + 2s. Since more than one vehicle can be present at the same time, multiple data points can be collected at t.

[1] https://www.asam.net/standards/detail/openscenario/

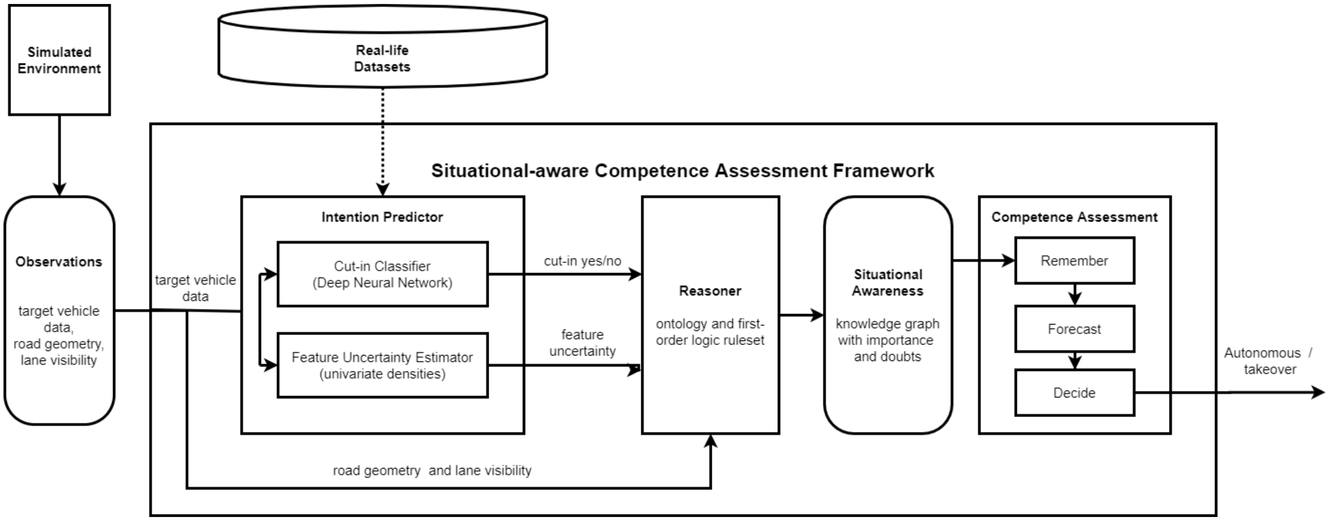

Simulated Environment
Real-life Datasets
Situational-aware Competence Assessment Framework
Observations
target vehicle data, road geometry, lane visibility
target vehicle data
Intention Predictor
Cut-in Classifier (Deep Neural Network)
Feature Uncertainty Estimator (univariate densities)
cut-in yes/no
feature uncertainty
Reasoner
ontology and first-order logic ruleset
Situational Awareness
knowledge graph with importance and doubts
Competence Assessment
Remember
Forecast
Decide
Autonomous / takeover
road geometry and lane visibility

Figure 1: The overall architecture of our situation-aware competence assessment framework. The dotted arrow from Real-life Datasets to Intention Predictor is not part of the online information flow.

The dataset consists of 24305 data points, divided into 6348 instances labeled as ”cut-in” and 17957 ones labeled as ”not cut-in”, drawn from the StreetWise database (Paardekooper et al. 2019). While completeness measures of this database have been developed (de Gelder et al. 2019), the dataset used for training the intention predictor does not cover the entire spectrum of cut-ins that are to be expected in real-life traffic. However, for the purpose of this work a complete dataset is not essential, as we are interested in the situations where the intention predictor has not been trained for.

To date, a variety of physics-based and data-driven approaches have been developed to detect spatio-temporal patterns in road users' behaviour. Specifically, DNNs have been commonly adopted for classification purposes as they can often outperform other methods for high-dimensional data (Sakr et al. 2017).

The DNN we developed for this study is a two-layer fully connected network trained with gradient-based backpropagation. The input $\mathbf{x} \in \mathbb{R}^m$ is mapped into an output $\mathbf{y} \in \mathbb{R}^n$, where $m = 30$ and $n = 1$ are respectively the input and output dimensionality. The two hidden layers contain 512 neurons each and are activated by a ReLU function (Nair and Hinton 2010). The 30 features used as input represent continuous values related to the dynamics of a TV present at a given time t. Some of these variables, such as the TV's speed, acceleration, and the relative lateral and longitudinal distance to EV, have been directly collected in real-life driving scenarios; other variables are the result of feature engineering techniques to develop expressive variables. The output, which is a single non-linear sigmoid layer defined over a domain $o \in [0, 1]$, represents the predicted confidence in the TV performing a cut-in within the next 3 seconds. The result of this two-class logistic regression is converted into a binary output by defining a maximum threshold λ on the output.

During training, the cross-entropy logarithmic loss is weighted for the two different classes to take into account their imbalance in the dataset. The threshold λ, the learning rate and the number of neurons per layer are fine-tuned using a Bayesian approach for global optimization (Brochu, Cora, and De Freitas 2010). To reduce overfitting, dropout and early stopping are used.

Feature uncertainty estimator To assess the robustness of the trained DNN predictor on unseen test scenarios, we first analyzed the univariate distribution of each feature $\mathbf{x}_i$ in the training set $X = \{\mathbf{x}_1, ..., \mathbf{x}_{30}\}$. Among these features, the most expressive ones for situational awareness were extracted for further analysis. The following features were chosen: the absolute EV and TVs' velocity and acceleration, their relative velocity and acceleration, the relative longitudinal and lateral distance between the vehicles, their relative heading, and the distance between EGO and the closest lane marker. A desired characteristic of these features concerns their distribution. We observed that the distribution of these data samples can be approximated to multimodal skewed distributions when the dynamic properties of the vehicles change incrementally over time. Such distributions can be approximated by traditional non-parametric density estimations methods, such as the Kernel Density Estimation (KDE) (Parzen 1962).

KDE is a technique used to reconstruct the probability density function of given data samples, and it can be adopted for a single feature (univariate KDE) or to multiple features (multivariate KDE). In the case of the univariate version, this technique consists of fitting a kernel function, such as a Gaussian, over each of the k samples in the chosen feature vector. The resulting k densities are then summed and normalized to return the final density estimate of the feature. The main hyperparameter of KDE, the bandwidth h, controls the variance of the kernel function; its value determines how smooth the final density estimate is. The optimization of this parameter, which is necessary to guarantee that the kernel function fits the data samples correctly, was optimized using the Maximum Likelihood Cross-Validation (MLCV) approach (Habbema et al. 1974):

$$MLCV = \frac{1}{k}\sum_{i=1}^{k} log\left[\sum_{j\neq i} K\left(\frac{x_j - x_i}{h}\right)\right] - log[(k-1)h] \tag{1}$$

where k is number of data samples to fit, $K(\cdot)$ is a Gaussian kernel, x_j is a data point over the defined domain chosen, and x_i is the i-th sample in the feature vector. Once the final density estimate is computed, it is possible to evaluate the likelihood of new samples for each feature; this computation can be performed synchronously with the observation of new data in unseen scenarios, as required in our study. For practical purposes, the log-likelihood of the samples is used instead of the likelihood.

A main assumption in our investigation is that samples with low likelihood lead to higher uncertainty on the DNN's competence. To quantify this intuition, we define the ratio r_i as:

$$r_i = \frac{\mathcal{L}(x_i \mid \mathrm{M}_i)}{\mathcal{L}_{max,i}} \tag{2}$$

where x_i is an observation that belongs to the i-th feature. The value r_i represents the ratio between the estimated log-likelihood of the new sample x_i given the fitted model M_i and the maximum log-likelihood $\mathcal{L}_{max,i}$ observed for the i-th feature. The maximum log-likelihood was pre-computed and stored during the kernel fitting phase. Finally, we define the *feature uncertainty* ϕ as:

$$\phi = 1 - \frac{1}{m}\sum_{i=0}^{m} r_i \tag{3}$$

where m is the dimensionality of the feature space $\mathbf{x}$. This quantity is intrinsically related to the frequency of the observation in the training set and reflects our previously mentioned assumption on the DNN's competence. The subtraction guarantees that the feature uncertainty tends to 1 when all the features are out-of-distribution, thus maximizing the uncertainty in the predictor's output, and to 0 when the features are in-distribution, thus following the same trend as the competence.

Reasoner The second module of our framework, the Reasoner, is in charge of aggregating all observations — namely target vehicle data, road geometry, lane visibility and the output of the Intention Predictor — to have a unified and quantitative representation of the situation on the road. This view is represented by means of a knowledge graph based on an underlying ontology and a set of first-order logic inference rules. We will hereafter refer to the ontology-rule pair as the schema. The Reasoner is implemented in Grakn [2]. The ontology specifies (part of) the automotive domain via entities, attributes and relations. Example of entities are vehicles and road lanes. An example of relations is "drive-on", linking vehicles and lanes. An example of attributes is "distance-from-ego", which both vehicles and lanes have.

Given a set of observations in input, the reasoner first initialises the related knowledge graph by creating nodes — corresponding to entities and attributes — and edges — corresponding to relations. Subsequently, the rules, defined as Horn clauses (Horn 1951), augment the graph by creating the two attributes ”importance” and ”doubt”, linked to entities and relations. The importance aims to encode domain expert knowledge of the automotive domain. Its purpose is to categorise and rank nodes and edges. The doubt, on the other hand, can be interpreted as a measure of uncertainty associated to the nodes and edges. Its purpose is to assign a unique type of weight across the whole graph elements. The two features are orthogonal to each other: the schema could specify that a fully visible entry lane is important independently on its doubt value. On the other hand, the cut-in classifier prediction of a target vehicle that drives far away from ego, yet in an erratic way (high feature uncertainty), could have a high doubt value associated to it and, concurrently, a low importance value because of its position. We consider three possible importance values, namely low, $medium$ and $high$, and 11 doubt values, bounded in the $[0, 1]$ interval, equidistant from each other $(0,\ 0.1,\ 0.2,\ \ldots 1)$.

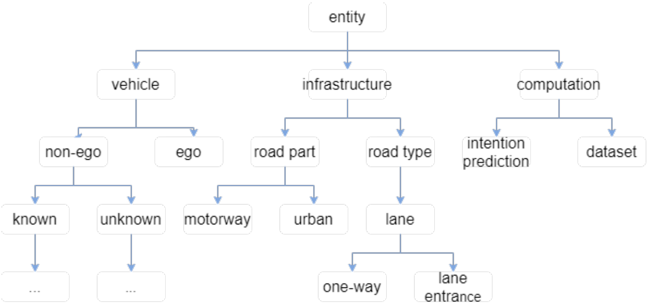


Figure 2: The main schematics of the entities of our ontology and their hierarchical organisation.

An excerpt of the ontology's entity organisation is depicted in Fig.2, whilst a schematic representation of the relations is shown in Fig. 3. The entities are organised hierarchically and along three main branches: one representing the possible vehicles, one the driving infrastructure, and one the computational models external to the reasoner. The non-ego vehicles are divided into two key-categories: known and unknown. The categorisation is done based on the types of vehicles present in the dataset the cut-in classifier was trained on. For instance, if the dataset contained only passenger cars, such entity would be placed under the known branch, whilst other vehicles such as lorries and motorcycles would be inferred as children of the unknown entity. The known/unknown information associated to observed TVs is used by the rules to assign doubt values to the classifier's output and importance values to the graph nodes. The driving infrastructure describes all non-vehicle entities present on the road, such as lanes, ramps and signs, in accordance with (Zhao et al. 2015; Czarnecki 2018a,b). Lanes have a

[2] https://grakn.ai/. Last accessed 18 December 2020.

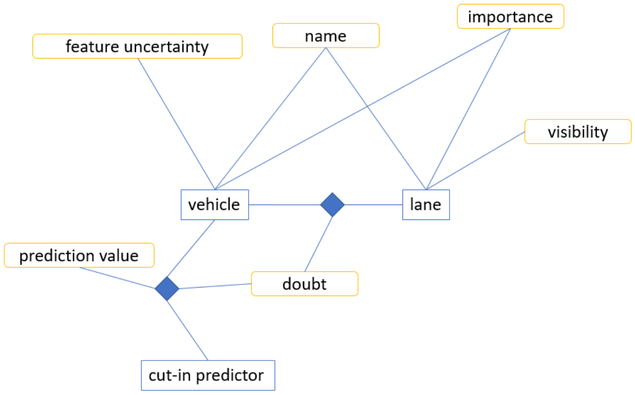


Figure 3: The observation-relation implemented in our schema. Orange items are attributes, blue items are entities, *has* entity-attribute relations are straightforward, whilst the entity-entity relation is depicted with a rhombus.

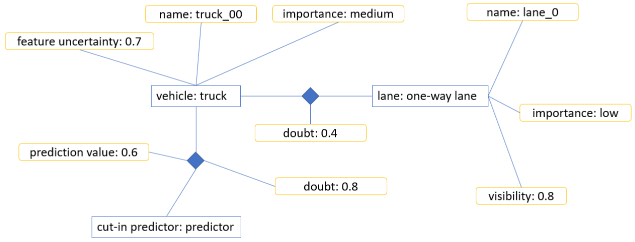


Figure 4: The instantiation of the schema of Figure 2 given fictional observations. A truck drives on a one-way lane. The intention predictor, due to the truck's high feature uncertainty, assigns a cut-in probability of 0.6. This rather uncertain value leads to a high doubt value. Nonetheless, the truck is rather far from EV, as it can be hinted by the high visitibility value of the lane[3]. Hence, the importance value for the lane is low. The truck has a higher importance value due to its non-likely behaviour. Finally, the doubt value associated to the truck-lane relation is low mainly because of the lane, though not extremly low due to the feature uncertainty related to the vehicle and the fact that it is not entirely sure whether it will perform a cut-in.

fundamental attribute: visibility. The rules implement a negative correlation: the lower the visibility, the higher the doubt associated to that lane. In this way, the framework aims to speculate about the possible existence of hidden entities in adjacent lanes. Computation entities represent framework models which process raw observations to generate new information, in our case the cut-in classifier. In case the models are machine learned, the Reasoner infers, via positive correlation, doubt values associated to the model outputs depending on the related in-distribution likelihood values: the lower the likelihood, the lower the doubt. An example of an observation-reasoned knowledge graph is shown in Fig. 4.

Competence Assessment

The last module in our framework — Competence Assessment — leverages previous and current knowledge graphs to determine whether the EV should maintain an autonomous driving modality or leave the control to the human driver or backup safety system. Competence Assessment follows a remember-forecast-decide processing flow.

Remember A time-indexed memory of η graph embeddings

$$e_1, \; \dots e_\eta$$

is kept. The embedding of a particular time corresponds to a single value encoding the graph related to that particular time's road observations. Currently, the embedding procedure corresponds to a weighted average of all doubts, where the weights are associated to the relative importance values: the higher the importance, the higher the weight.

Forecast The remembered embeddings represent, albeit in a compact way, reasoned (importance/doubt-aware) situations. Intuitively, the lower an embedding, the more competent the autonomous vehicle was in that situation, as low importance and doubt attributes would predominantly exist in the corresponding graph. We therefore define the Competence related to a graph embedding as:

$$c_i \;=\; 1 \;-\; e_i, \; \forall \; i \;\in [1, \; \dots \eta] \tag{4}$$

Intuitively, c_η corresponds to the latest (current) competence value. The c_i values are then fed to a regressor to estimate ρ future competence values

$$\hat{c}_1, \dots \; \hat{c}_\rho$$

Currently, the framework implements a linear regressor, based on the assumption that a short-term linear dependency across observations holds.

Decide The decision whether the driving should remain autonomous or handed over to a human is made based on the lowest future competence value

$$\hat{c}_{min} \;=\; \min \; \hat{c}_i, \; \forall \; i \;\in\; [1, \; \dots \rho] \tag{5}$$

and by comparing it to an *assessing* threshold τ_c

$$\text{decision} \;=\; \begin{cases} \text{takeover} & if \; \exists \, c_i \; < \; \tau_c \\ \text{AD mode} & otherwise \end{cases} \tag{6}$$

where AD stands for Autonomous Driving.

Results and Discussion

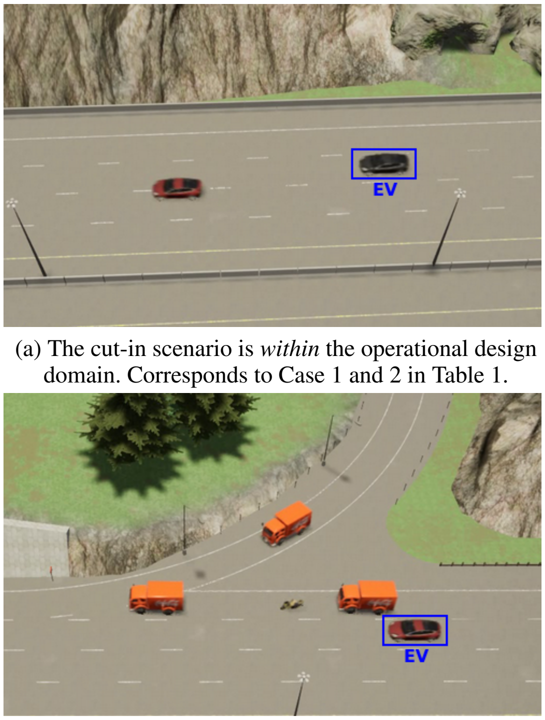
We have trained the DNN for intention prediction on 24305 instances, divided into 6348 cut-ins and 17957 non cut-ins. Since the dataset was unbalanced, we weighted the loss function to compensate for the difference in the observations per class. The algorithm was tested on 7200 instances, resulting in a $F_{score} = 0.98$ (accuracy = 0.99).

We have assessed the competence of the intention predictor in two cut-in scenarios. The first scenario describes a cut-in by a passenger car on an otherwise empty road (Fig. 5a).

			Competence			Decision	
Case	**Potential Risk**	**Reasoner**	**Current**	**Minimum Future**	**1 - ϕ**	**w/o Reasoner**	**w/ Reasoner**
			c_η	$\hat{c}_{min}$		τ_ϕ = 0.7	τ_c = 0.7
1	Low	not present	-	-	0.57	takeover	-
2	Low	present	0.71	0.84	0.57	-	AD mode
3	High	not present	-	-	0.31	takeover	-
4	High	present	0.15	0.14	0.31	-	takeover

Table 1: Results on the four different cases tested with and without the Reasoner.

The velocity, distance and driving profile of the TV was designed not to pose any risk to the EV. In addition, every vehicle present in the scenario was known to the knowledge graph.



(a) The cut-in scenario is *within* the operational design domain. Corresponds to Case 1 and 2 in Table 1.

(b) The lane entrance scenario which is *outside* the operational design domain. Corresponds to Case 3 and 4 in Table 1.

Figure 5: Snapshots from the two different scenarios as shown in the CARLA simulator.

The second scenario (Fig. 5b) describes multiple vehicles (two trucks and a motorcycle) on the first right-most lane and a truck approaching from the entrance lane. The EV is in the left lane and cannot see the approaching truck as it is occluded by the vehicles on its right. The features in this scenario are out-of-distribution, as only two features lie within the training set domain (Fig. 6). Moreover, the scenario includes an unknown entity in the form of a motorcycle that was not part of the training set. The rationale for this is that a type of vehicle not present in the training set might display a driving profile that the intention prediction does not expect. In other words, the output of the predictor might be incorrect since it relies on the detection of spatio-temporal patterns in the vehicle's driving behaviour. The visibility on the road was reduced by the traffic on the first lane; this lane was considered of high importance due to the road entrance. This scenario was designed to pose potential risk to the autonomous system, due to the out-of-distribution features and unknown entities. The two scenarios were evaluated at the moment that one of the TVs performs a cut-in.

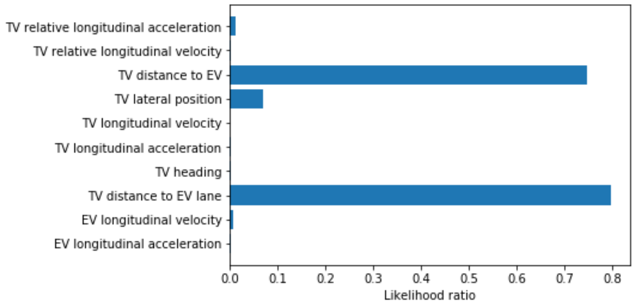


Figure 6: Likelihood ratio r of the features used to compute the feature uncertainty in the cut-in manoeuvre performed by the motorcycle (second scenario)

The two settings were first tested without the contribution of the symbolic reasoning inference, shown as Case 1 and Case 3 in Table 1. Since the Reasoner was not in place, the feature uncertainty ϕ was used as a proxy to relate the Intention Predictor to its ability to correctly perform in the given situation. For clarity, the quantity $1 - \phi$ is reported; hence, a score equal to 1 represents full confidence in the Intention Predictor output and can be directly compared with the Competence score. The threshold τ_ϕ = 0.7 was defined to establish whether it was necessary for the human driver to take over ($1 - \phi < \tau_\phi$), or the vehicle could maintain AD mode ($1 - \phi \geq \tau_\phi$). In both cases, the system decides not to maintain the AD mode, due to the high feature uncertainty, even if the scenario was safe. The absence of the Reasoner causes a lack of situational awareness: the speed of the TV was lower than the average velocities collected in the training set —— thus making the velocity an OOD feature —— but the large distance between the EV and TV is not used by the Intention Predictor to reduce the importance attributed to this quantity.

Results of the competence assessment with the Reasoner are shown as Case 2 and Case 4 in Table 1. The *Current*

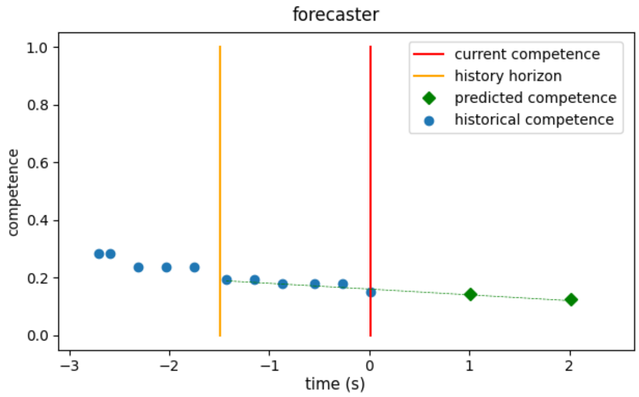


Figure 7: Future estimation of the competence corresponding to Case 4.

Competence column refers to the competence c_η as inferred by the first-order rules of the knowledge graph at the current moment. In the event that more than one vehicle was predicted to perform a cut-in, the reported value is the lowest c_η estimated among all the vehicles. The *Minimum Future Competence* $\hat{c}_{min}$ was computed converting the future doubt-embedding extrapolated by the forecaster (Fig. 7) as described in Eq. 4 and Eq. 5 ($\rho = 2$). Thus, the future competence was calculated for a prediction horizon of 2 seconds. The decision whether the vehicle should remain autonomous was performed by a thresholding function (τ_c = 0.7) on the future competence, as detailed in Eq. 6.

We found that $\hat{c}_{min}$ evaluated for Case 4 was six times lower than in Case 2. In Case 2 the threshold for takeover was never reached and the system did not hand over the autonomous control. Due to the large distance of the TV and the high visibility of the lanes, the Reasoner determined that the vehicle could stay in AD mode despite the low likelihood of the input data expressed by the average feature uncertainty. In contrast, the system decided that a takeover of the AD mode was necessary in Case 4, because the numerous sources of risk in this setting caused a low future competence. This is expressed by a competence value that is substantially lower than solely based on the feature uncertainty. Using the likelihood of the input data expressed by the feature uncertainty alone is not sufficient to correctly assess the confidence in the Intention Predictor output. This is evident by the results of Case 1 and Case 3 (Table 1), where the absence of the Reasoner fails to correctly assess the situation. In addition, the competence returned by the Reasoner shows a larger contrast between these two extreme cases than the method based on the feature uncertainty alone.

We found that the linear regression used to assess the future competence was strongly affected by small variations in the history of doubt-embeddings ρ. Thus, we do not consider that a prediction horizon higher than 2 seconds would be reliable enough to support the decision making process.

Conclusions and future work

We have presented a hybrid-AI framework for the safe application of AI functions in automated driving. The framework aggregates road observations and the results of data-driven AI computations — such as a DNN for intention prediction in our case study — into a knowledge graph. The graph is built by means of an ontology, which specifies the entities that can exist on the road, and a set of first-order logic inference rules, the latter aiming to estimate the severity level of the road situations. The knowledge graph is then compressed into a single value (embedding), stored in a working memory, and used to forecast imminent severity levels. A final decision maker modules establishes whether the vehicle should continue driving autonomously or whether the steering wheel should be handed over to a human driver or backup safety system. The knowledge graphs encode the situational awareness capabilities of the vehicle, whilst the forecasting and decision making processes realise the vehicle's competence assessment capability.

We have shown that the reasoner correctly assigns high competence to the Intention Predictor in a situation in which some features of the DNN are uncertain, but the TV poses no safety threat to the EV due to the large distance and high lane visibility. The added value of the Reasoner is also shown in a situation that contains a vehicle (in this case a motorcycle) that has never been seen before by the predictor, in an environment with important entities that require attention (in this case an entrance lane). The predictor output is unreliable in this case, potentially leading to erratic and dangerous behaviour of the EV if taken at face value. Here, the Reasoner correctly assigns a low competence to the predictor based on the presence of the motorcycle (high doubt) and the presence of the entrance lane (high importance).

These results provide a solid starting point for future investigations on situational awareness. In future work, we will extend situational awareness to the entire automated vehicle instead of a single component. In addition, the reasoner will aggregate more types of observations, for example those regarding road works or weather conditions, and its first-order logic inference rules could be parameterised via data-driven approaches instead of solely relying on domain knowledge. Combining the DNN with the knowledge graph into a graph neural network will result in a better estimation of competence, especially further into future. Graph neural networks might also aid in enhanced explainability on *why* takeover is needed.

While limited to a single function in a simulation environment, our work shows that a hybrid-AI approach to situational awareness is essential for the safe application of AI systems in automated driving.

References

Bansal, M.; Krizhevsky, A.; and Ogale, A. 2019. ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst. In *Robotics: Science and Systems*.

Brochu, E.; Cora, V. M.; and De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599* .

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Czarnecki, K. 2018a. Operational world model ontology for automated driving systems–part 1: Road structure. *Waterloo Intelligent Systems Engineering Lab (WISE) Report, University of Waterloo* .

Czarnecki, K. 2018b. Operational world model ontology for automated driving systems–part 2: Road users, animals, other obstacles, and environmental conditions,". *Waterloo Intelligent Systems Engineering Lab (WISE) Report, University of Waterloo* .

de Gelder, E.; Paardekooper, J.-P.; den Camp Olaf, O.; and De Schutter, B. 2019. Safety assessment of automated vehicles: how to determine whether we have collected enough field data? *Traffic Injury Prevention* 20(S1): S162–S170.

Deo, N.; and Trivedi, M. M. 2018. Multi-Modal Trajectory Prediction of Surrounding Vehicles with Maneuver based LSTMs. In *IEEE Intelligent Vehicles Symposium, Proceedings*, 1179–1184. University of California, San Diego, San Diego, United States, IEEE.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803* .

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* .

Endsley, M. R. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors* 37(1): 32–64.

Endsley, M. R. 2020. Situation Awareness in Driving. In Fisher, D.; Horrey, W.; Lee, J.; and Regan, M., eds., *Handbook of Human Factors for Automated, Connected, and Intelligent Vehicles*, chapter 7. CRC Press.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, 1651–1660. University of Cambridge, Cambridge, United Kingdom.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599* .

Habbema, J.; JDF, H.; Van den Broek, K.; et al. 1974. A stepwise discriminant analysis program using density estimation. .

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 41–50.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* .

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proceedings of International Conference on Learning Representations* .

Horn, A. 1951. On sentences which are true of direct unions of algebras. *The Journal of Symbolic Logic* 16(1): 14–21.

Koopman, P.; and Wagner, M. 2016. Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety* 4(1): 15–24.

Kristiadi, A.; Hein, M.; and Hennig, P. 2020. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In Daumé III, H.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5436–5446. Virtual: PMLR.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553): 436–444.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* .

McAllister, R.; Gal, Y.; Kendall, A.; van der Wilk, M.; Shah, A.; Cipolla, R.; and Weller, A. 2017. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 4745–4753. University of Cambridge, Cambridge, United Kingdom.

Meyer-Vitali, A.; Bakker, R.; van Bekkum, M.; Boer, M. d.; Burghouts, G.; Diggelen, J. v.; Dijk, J.; Grappiolo, C.; Greeff, J. d.; Huizing, A.; et al. 2019. Hybrid ai: white paper. Technical report, TNO.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Okuda, R.; Kajiwara, Y.; and Terashima, K. 2014. A survey of technical trend of ADAS and autonomous driving. In *Proceedings of Technical Program - 2014 International Symposium on VLSI Technology, Systems and Application, VLSI-TSA 2014*. Renesas Electronics Corporation, Tokyo, Japan.

Paardekooper, J.-P.; van Montfort, S.; Manders, J.; Goos, J.; de Gelder, E.; Op den Camp, O.; Bracquemond, A.; and Thiolon, G. 2019. Automatic Detection of Critical Scenarios in a Public Dataset of 6000 km of Public-Road Driving. In *Enhanced Safety of Vehicles*, 1–8.

Parzen, E. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33(3): 1065–1076.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 14707–14718.

Sakr, S.; Elshawi, R.; Ahmed, A. M.; Qureshi, W. T.; Brawner, C. A.; Keteyian, S. J.; Blaha, M. J.; and Al-Mallah,

M. H. 2017. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. *BMC medical informatics and decision making* 17(1): 174.

Thill, S.; Hemeren, P. E.; and Nilsson, M. 2014. The apparent intelligence of a system as a factor in situation awareness. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2014*, 52–58. RISE Viktoria, Gothenburg, Sweden, IEEE.

van Harmelen, F.; and ten Teije, A. 2019. A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems. *arXiv.org* .

Vellinga, N. E. 2019. Automated driving and its challenges to international traffic law: which way to go? *Law, Innovation and Technology* 11(2): 257–278.

WHO. 2018. Global status report on road safety 2018.

Zhao, L.; Ichise, R.; Yoshikawa, T.; Naito, T.; Kakinami, T.; and Sasaki, Y. 2015. Ontology-based decision making on uncontrolled intersections and narrow roads. In *2015 IEEE intelligent vehicles symposium (IV)*, 83–88. IEEE.