A virtual driving instructor that assesses driving performance on par with human experts

Johannes Rehm[a,b] (johannes.rehm@ntnu.no), Irina Reshodko[a,b] (irina.reshodko@way.no), Odd Erik Gundersen[a] (odderik@ntnu.no)

[a] Department of Computer Science, Norwegian University of Science and Technology, Norway

[b] Way AS, Nordre gate 11, 7011 Trondheim, Norway

Corresponding Author:

Johannes Rehm

Norwegian University of Science and Technology, Sem Sælands vei 9, 7034 Trondheim, Norway

Tel: +49 1709190294

Email: johannes.rehm@ntnu.no

A virtual driving instructor that assesses driving performance on par with human experts

Johannes Rehm[a,b], Irina Reshodko[b], Odd Erik Gundersen[a,*]

[a] *Department of Computer Science, Norwegian University of Science and Technology, Norway*
[b] *Way AS, Nordre gate 11, 7011 Trondheim, Norway*

Abstract

The advent of virtual driving instructors has the potential to revolutionize driver education by providing real-time, unbiased feedback to learner drivers. This paradigm shift aims to mitigate the innate subjectivity associated with human evaluations. Our research focused on the creation of a virtual driving instructor capable of assessing a learner driver's performance in real-time, with an emphasis on eliminating the inherent biases associated with human evaluations. Our approach involved the development of a rule-based assessment system, employing a multi-agent system based on the subsumption architecture. Each agent in the system was tasked with assessing a specific aspect of driving performance. Additionally, we utilized a knowledge graph to maintain a continuous understanding of the situational context, further enhancing the system's assessment capabilities. We posited that our system, given its methodical structure and objective rule-based framework, would be able to accurately and objectively assess various driving scenarios. Further, we hypothesized that our system's performance would be on par with expert human evaluations. The validation of our system was conducted using real driving sessions in simulators with actual students. The system was tested on various scenarios including intersections, roundabouts, and overtakes. The assessment results aligned closely with expert

*Corresponding author.
Email addresses: johanes.rehm@ntnu.no (Johannes Rehm), irina.reshodko@way.no (Irina Reshodko), odderik@ntnu.no (Odd Erik Gundersen)

consensus, showcasing the system's capacity to match the evaluative precision of human experts.

Keywords: Virtual driving instructor, multi-agent system, knowledge graph, ontology, real-time assessment, driver education, traffic situation awareness, driving simulation

---

1. Introduction

According to the World Health Organization (WHO), an estimated 1.35 million people lost their lives in road crashes in 2016 (World Health Organization, 2018). Tragically, road traffic injuries were the leading cause of death for children and young adults aged 5-29 years. Every small improvement in road safety has the potential to save many lives, which is why the European Union has set an ambitious goal of reducing road fatalities to close to zero by 2050 (European Commission, 2011).

One way to improve road safety is to improve driver education. Young drivers are especially exposed, as deaths by traffic accidents account for 23% of deaths for young adults that are between 18 and 24 years old (Gicquel et al., 2017). The rapid technology advancements and reduced hardware cost mean that it is viable to provide driver education in simulators with virtual driving instructors (VDI). Driving simulators enable safe learning for both the driver and the environment, which means that risky situations can be experienced without any concerns for safety. VDIs on the other hand can help to both personalize and standardize driver education. As the virtual environment can be changed dynamically, education can be personalized by presenting driving situations that will optimize learning for individual students. The VDI can choose traffic situations that will train the skills where the student has most to gain and improve learning reducing the complexity of a situation by removing traffic if the student is stressed, as stress reduces the student's capacity for learning (Vogel & Schwabe, 2016). Also, feedback can be personalized, as it can be tailored to the learning style that suits the student best. A VDI enables standardization of driver education as a student's skill level can be quantified, which is not practically possible when a driver instructor sits inside a car with a student driving in real-world traffic. Finally, a VDI can ensure that a driver has experienced all types of situations. This includes scenarios that cannot be safely triggered in real-world traffic, ensuring that students master them to the desired degree.

Fully automating driver education is also advantageous as it decreases the cost of training. The cost-driver of high-quality driver education, so-called graduated drivers licensing, which has shown to be effective in reducing accidents O'Neill (2020), is that it is done one-on-one with one instructor per student. Automation reduces labor costs, which are particularly high in countries where graduated driver licensing is most prevalent.

In this paper, we describe a VDI that can sense the simulation environment, analyze traffic situations, and provide feedback to both driving students and instructors. This system is composed of two sub-systems: 1) the assessment system and 2) the tutoring system. The assessment system evaluates the driving performance of the student, determining what was done correctly and what was done incorrectly, while the tutoring system uses this information from the assessment system for providing feedback to the student. Additionally, the tutoring system is responsible to provide situation awareness and decision support to a driving instructor who can monitor several students in parallel.

The main focus of this paper is the assessment system. The assessment system is largely designed as a rule-based system, a type of artificial intelligence system that uses predefined rules or heuristics to make decisions (Shu-Hsien Liao (2005)). This is different from the learning-based methods commonly used in intelligent agents today. Rule-based systems can be challenging when it comes to mapping the sensed environment to a symbolic representation (Wooldridge (2009)). In contrast, learning-based methods such as reinforcement learning (Ndousse et al. (2021), Baker et al. (2019)) and imitation learning (Onishi et al. (2019), Codevilla et al. (2018)) are well-suited to handle uncertainties, which are common in real-world problems. However, these methods have two crucial disadvantages for the assessment system.

First, training machine learning models requires a large amount of labeled data (Hestness et al. (2019)). In the case of imitation learning, not only is raw driving data required, but also additional ground truth data generated by human experts evaluating each driving situation (Ross et al. (2011)). Second, learning-based methods are black box systems (Rudin (2019)). In the case of the

VDI, explainability is key. The VDI must not only output performance scores for certain driving skills, but also provide explanations (Gunning & Aha (2019)) for why the performance was assessed as high or low, and what the student needs to improve. A rule-based system can easily trace back the reasoning behind its decisions, whereas making the decisions of black box systems explainable adds a lot of complexity (Kim et al.; Goodman & Flaxman (2017)).

The research question we seek to answer is: *Can we develop a hybrid AI system, combining the advantages of both rule-based and learning-based methods, that can assess traffic situations with performance on par with human experts?*

Our contributions are as follows:

- We present an innovative driving assessment system integrating a multi-agent framework with a knowledge graph that evaluates driving students in real-time.
- We perform a thorough evaluation of the VDI where its performance in evaluating advanced traffic situations is compared to the performance of human experts showing that the VDI is on par with human experts.

In previous publications, we have presented the architecture design of the assessment system (Sandberg et al., 2020), the deployment and the student's experience of using the simulator (Rehm et al., 2024), a method for utilizing driving context to improve annotation efficiency of driver gaze image data (Rehm et al., 2021) and automatic generation of driving lessons in the virtual world (Bjørnland et al., 2024).

2. Related work

Virtual driving instruction represents a rich and varied field of study, with diverse methodologies and approaches being employed to enhance the learning experience. One such study by Ropelato et al. (2018) involves the use of an intelligent tutoring system (ITS) incorporated into a virtual reality head-mounted display-based driving simulator.

This adaptive system delivers new activities to learners, optimized for maximizing their expected learning progress. The ITS primarily assesses basic driving skills, such as maintaining lane discipline and regulating constant speed. The evaluation of this system, however, is principally done through participant feedback from a questionnaire, which focuses on the level of immersion experienced by the participants and whether the simulator induced so-called simulator-sickness Kolasinski (1995). While this study makes a valuable contribution to the field, it lacks extensive details about the exact mechanisms employed in the skills assessment, as well as the comprehensive evaluation of the intelligent tutoring system itself. Consequently, a detailed exploration of these aspects could offer avenues for further development and refinement in the realm of virtual driving instruction.

Another study that is closely aligned with our work is Weevers et al. (2003). They also concentrate on the concept of a VDI. This study, akin to ours, adopts a multi-agent system for the design of its intelligent tutoring system. This system is composed of three distinct agents: the situation agent, the presentation agent, and the curriculum agent, each playing a crucial role within the ITS. However, the research provides only a high-level summary of the system without delving into the specifics of the individual agents or their operations. Furthermore, it lacks empirical validation, as no experiments are presented to substantiate the effectiveness of the approach. This absence of detailed operational information and empirical data highlights areas where further research could significantly contribute to the development of virtual driving instruction.

Sharon et al. (2005) diverges somewhat from our approach, targeting their research towards the continued education of already licensed drivers, rather than novice learners. Their system, CarCoach, is implemented in an actual vehicle environment rather than a simulator. The effectiveness of CarCoach was subsequently evaluated in a study by Arroyo et al. (2006), who assessed participants' levels of frustration in response to feedback from the system. While this evaluation approach lends insight into user experience, it does not extensively explore more objective performance metrics.

Performance assessment in simulators often involves both qualitative and quantitative measures. However, the reliance on human input for qualitative measures complicates automation. Therefore, robust quantitative measures are critical, as Ekanayake et al. (2011) demonstrate. They propose a competency formula based on quantifiable outcomes and effort. 'Good speed' is a positive outcome, off-road driving, and crashes are negatives, and effort is gauged via physical pressure on the hand controller. The study involved seven licensed drivers in two racing simulator sessions, correlating their quantitative approach with a qualitative assessment by an experienced driver. While this underscores the viability of quantitative assessment, the study's methodology poses some challenges when contrasted with our objectives. Favorable and unfavorable outcomes in regular, non-racing driving could be more nuanced than speed or crashes, incorporating factors like yielding errors or unsteady lane adherence. Furthermore, typical driver competency evaluations do not involve effort metrics. Lastly, a more reliable qualitative assessment would be conducted by seasoned professionals.

In a study of simulator efficacy, de Winter et al. (2009) demonstrated a meaningful connection between simulator performance and subsequent results on actual on-road driving tests. They found that learners making fewer steering errors in the simulator were more likely to pass their first on-road driving test, signifying that skills honed in a virtual environment could transfer to real-life driving. Yet, despite these promising findings, they underscored the need for more extensive research to affirm the reliability and validity of simulator training, highlighting an area that our work aims to address.

De Winter et al. (2012) provide an overview of the potential of driving simulation technology, citing its benefits and drawbacks. Simulators' controllability, reproducibility, and safety are advantageous, but their limited fidelity, scarce validation research, and potential for user discomfort present challenges. With the anticipated rise of affordable virtual-reality applications, the need for addressing these challenges, particularly the validation aspect, is evident. Our work targets this unmet need, aiming to develop a rigorously validated, auto-

mated assessment system.

In the preceding discussions on virtual driving instruction, several techniques and methodologies have been highlighted. Approaches range from intelligent tutoring systems and multi-agent systems to using actual vehicle environments for feedback. These methodologies are integral to our goal of creating a well-validated, automated assessment system for virtual driver education.

On a technical front, it's worth noting the prominent role of ontologies (Uschold & Gruninger (1996)) and rule-based multi-agent systems (Dorri et al. (2018)) in such applications. These methodologies, also used in autonomous driving and ADAS applications, provide the required higher-level reasoning for achieving situation awareness (Endsley, 1995).

The study by Hülsen et al. (2011b) presents an innovative method for creating comprehensive traffic situation descriptions for advanced driver assistance systems. Leveraging description logic as a knowledge representation language, their model utilizes logical reasoning to interpret intricate traffic situations and associated rules. This strategic application of logic offers an advanced way to comprehend and navigate complex traffic scenarios. They test this system in a real-time simulation framework in Hülsen et al. (2011a). Buechel et al. (2017) developed an ontology-based system to model traffic scenarios semantically. The system leverages the OWL 2 Web Ontology Language and the Pellet reasoner to encapsulate key traffic aspects. The authors further link the traffic situations to their corresponding rules, thereby facilitating enhanced situational awareness and decision-making in autonomous vehicles. Suryawanshi et al. (2019) present an ontology-based approach for modeling dynamic road data in Advanced Driver Assistance Systems. Their system uses dual-ontologies and SWRL rules to flexibly process map data, and they validate three techniques for predicting a vehicle's most probable path using a simulation, demonstrating efficient query response times and manageable system complexity.

Regele (2008) presents a novel abstract world model for traffic coordination in autonomous driving, distinguishing between trajectory planning and traffic coordination, and addressing the latter via a hierarchical, graph-like network of

lanes, vehicles, and objects, thereby simplifying high-level control development and enabling efficient traffic management in conflicting scenarios such as intersections and multi-lane roads. Halilaj et al. (2021) propose a knowledge graph (KG)-based approach for representing and utilizing information relevant to traffic situations in driver assistance and automated driving systems. The approach is evaluated using a synthetically generated dataset and compared with traditional vector-based feature representations. The results show the advantages of the KG-based approach in situation classification tasks, indicating its potential for improving interaction modeling and decision-making in autonomous vehicles. Zamora et al. (2017) introduce a rule-based multi-agent architecture for an intelligent ADAS system assisting a driver in an urban environment. This proposal has later been implemented and experimentally validated in Sipele et al. (2018). Both works are based on the multi-agent approach described in Gutierrez et al. (2014). Ontologies are also widely used in traffic management systems like *BeAware!* which is discussed in multiple publications. Baumgartner et al. (2014, 2010) include spatio-temporal reasoning concepts in their situation awareness framework. Salfinger et al. (2014) extended this framework by tracking the evolution of situations. Ontologies, knowledge graphs, and multi-agent systems are not the only way to get an understanding of traffic situations. Vacek et al. (2007) use case-based reasoning to interpret traffic situations. Platho et al. (2012) assess complex traffic situations by checking at any moment in time in which configuration a car is. They use a Bayesian network to classify in which configuration the reference entity is currently.

However, there is a lack of research in the field of AI driving education in the recent years. Related fields that are more active are self-driving cars (Badue et al., 2021) and AI in classroom education (Huang et al., 2021). The existing research on ITS for driving education reveals gaps and unmet needs, especially around the detailed mechanisms of skill assessment, comprehensive evaluation of tutoring systems, and empirical validation of simulator training. Our work, which seeks to develop an automated, rigorously validated, and explainable assessment system for virtual driver education, aims to address these challenges

and contribute to the ongoing evolution of virtual driving instruction and traffic understanding systems.

3. Architecture of the Virtual Driving Instructor

3.1. Overview

The VDI is a hybrid AI system that is integrated seamlessly with a traffic simulator. While currently confined to simulated environments, it is theoretically feasible to implement the VDI in a real-world vehicle, navigating public roads. This implementation would, of course, hinge on the presence of specific technical components: a sophisticated perception system for real-time understanding of the environment (Badue et al. (2021)), high-definition (HD) maps (Bao et al. (2022)) for in-depth environmental representations, and potentially reliable vehicle-to-everything (V2X) communication technology (Noor-A-Rahim et al. (2022)) for an enhanced, information-rich situational awareness. In the following sections, we delve into the specifics of the existing simulator system and explore how the VDI interacts within this virtual traffic environment.

The simulator setup is shown in Figure 1. A real car is mounted on a motion platform and the simulation visuals are displayed at 360° around the car by off-the-shelf projectors. This setup can be classified as a simulator type with the highest fidelity level according to Allen et al. (2007).

In Figure 2, an overview of the complete system is given. The traffic simulator is composed of three main components. The Virtual Reality (VR) environment encompasses both the simulation engine and the traffic manager. It uses the Unity game engine (Unity Technologies (2022)) which provides a robust platform for designing realistic virtual environments. All visual rendering takes place within this VR environment, ensuring immersive and engaging visuals for users. The traffic manager is responsible for the creation, deletion, and movement of all dynamic objects, such as vehicles and pedestrians, within the simulation. In addition to controlling object interactions, the traffic manager maintains real-time data for each object in the dynamic environment, storing

it as time-series data in a shared memory. This information is updated at regular intervals, ensuring the system remains synchronized and responsive to the evolving virtual landscape.

The student sits in a real car and operates it using the pedals and steering wheel. The car's control data, including pedal inputs and steering wheel movements, is made available to the traffic manager by incorporating it into the shared memory for time-series data as shown in Figure 2. This seamless integration ensures accurate and responsive synchronization between the student's actions and the virtual environment.

A camera is positioned in the car to capture the student's actions, with its video stream shared with the VDI. This setup allows the VDI to monitor the student's gaze, providing valuable insights into where their attention is focused during the simulation.

Every 20ms the VDI copies a snapshot of the time-series data and updates the knowledge graph in real-time, which is described in more detail in subsection 3.2.

The VDI communicates with the student through a multimodal interface (Philippe et al. (2020)) that includes text, graphics, and audio. The VDI autonomously decides what feedback to provide, when to provide it, and in what form. Timing is critical in this process, as students can be stressed while driving, which can reduce their ability to absorb information due to an increased level of cognitive load (Sweller et al. (2019)).

Figure 1: High-fidelity traffic simulator developed at Way AS. A real car is mounted on the motion platform in the center. Surrounding the car is a projection screen wall with six different view channels projected at the driver's front.

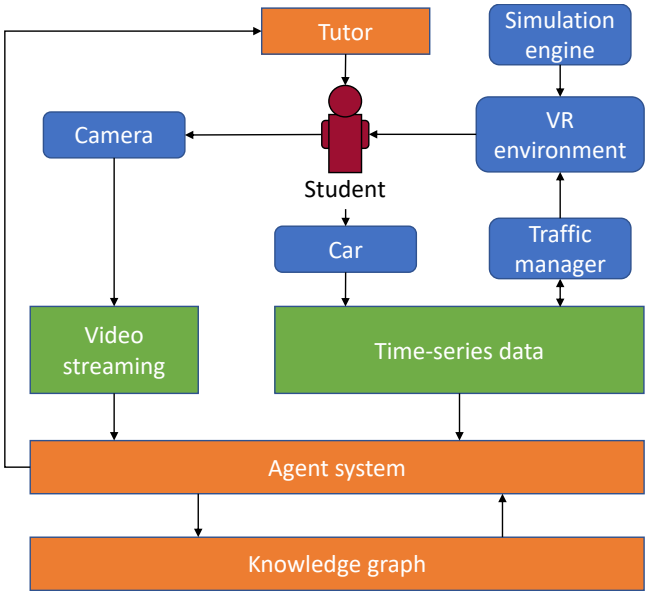

Figure 2: Overview of the system containing the driving student in red, all the simulator components in blue, shared data in green and the VDI components in orange

3.2. Knowledge graph

To provide comprehensive explanations of traffic situations, our VDI necessitates a thorough understanding of the environment. To accomplish this, we have designed a knowledge graph. Grounded in the conceptual framework detailed in Hogan et al. (2021), we adopted a property graph model. In this model, both nodes and relationships are characterized by attributes, which enables the VDI to maintain comprehensive situational awareness. We are using an ontology to provide a structured vocabulary for the knowledge graph. The ontology is based on Sandberg et al. (2020) and the core ontology for situation awareness (Matheus et al. (2003)). This structure allows us to account for changes in the values of attributes and relationships over time as the knowledge graph evolves.

This evolving knowledge graph, furnished by our ontology, forms the base for representing situations and situation objects, which are the central entities in our design. Mirroring the framework laid out in Matheus et al. (2003), every object is a node in the graph, and a situation constitutes a sub-graph. Each assessment made by the VDI is based on a situation, such as approaching and passing through an intersection. A situation has a start and end time and is comprised of all objects that are relevant to the situation. A pre-filtering step is required to identify these objects, as only those that can influence the assessment of the situation are considered relevant. For example, a car on a lane that has no connection to the intersection is not considered relevant to the intersection situation. This strategic exclusion ensures an efficient, focused, and real-time assessment process.

The assessment system performs its reasoning once the situation is over, allowing it to have full knowledge of the situation and avoiding any uncertainties related to predicting future states. In the traffic simulator, the VDI only needs to react to the student's actions, which makes the task of developing such an agent simpler than developing a self-driving agent. A self-driving car needs to plan ahead and anticipate any future dangers (Montemerlo et al. (2008); Suh et al. (2018)), whereas the VDI only needs to plan when to provide feedback. A situation can also be divided into sub-situations, allowing the tutoring system

to react to mistakes made by the student earlier in the situation, without having to wait until the complete situation is over.

The complete knowledge graph is too big to be described in full detail here. Figure 4 shows a subset of the ontology, focusing on situation objects typically found in an intersection or roundabout scenario. Situation objects can have attributes, such as the type of a road sign, and can have relationships with each other, such as a lane and an intersection having the relationship *IsIncoming*, meaning that the lane is directly connected to the intersection entry and the driving direction is going towards the intersection.

Situation objects are divided into dynamic and static objects. The static objects define a detailed road network, including roads, lanes, and road signs. This part of the knowledge graph is generated offline as it does not change during a driving session. It also includes higher-level information, such as the priority of one path through an intersection over another. The yielding rules for each path through an intersection are automatically pre-computed and stored in the knowledge graph after creating the road network using the road geometry and road signs.

The dynamic objects can appear and disappear in a situation and their information is received through the interface with the traffic simulator. A change in a dynamic object is expressed by adding a property value with a timestamp to the changed object attribute, such as the position or speed of a car. The same applies to changes in a relationship between two objects, where a property value is added to the relationship. This allows for reasoning about past situations after the situation is over, as no information is lost.

Consider, for instance, a lane change scenario as depicted in Figure 3. This situation engages various static situation objects such as lane T1_M4, lane T1_M3, and the road R_T1_M, in addition to the dynamic situation object, Ego_0 vehicle. The static objects interrelate – both lanes T1_M4 and T1_M3 belong to road R_T1_M, and lane T1_M4 lies adjacent to T1_M3 with a solid line in between the two as a lane separation. Meanwhile, dynamic relations evolve over time, exemplified by the Ego_0 vehicle initially occupying lane T1_M4 be-

fore transitioning to lane T1_M3 at the simulation timestamp of 435734ms from the simulation's onset. Drawing on this data, the agent system (as discussed in subsection 3.3) deduces a lane change action. This action is then documented within the knowledge graph as a new relationship and the system identifies a driving error as it is not allowed to cross the solid line.

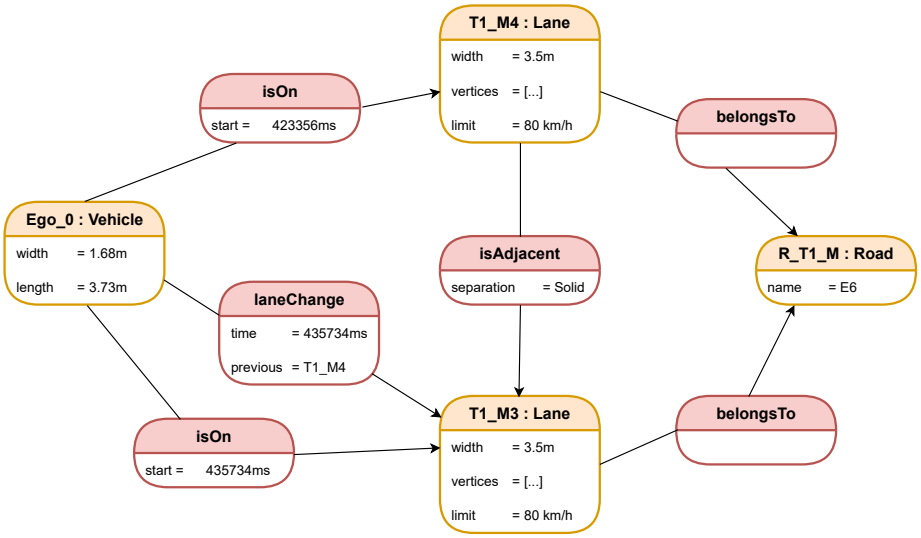

Figure 3: Instance of a property sub-graph for a lane change situation. The graph includes situation objects (Vehicle, Lane, Road) as nodes and their static and dynamic relations. The Ego_0's transition from lane T1_M4 to T1_M3 is inferred as a 'laneChange'.

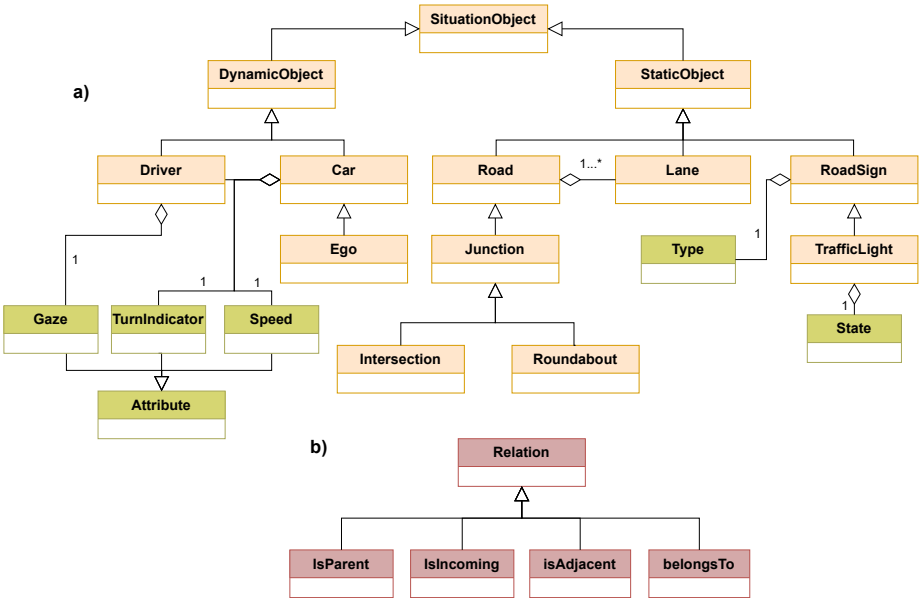

Figure 4: Ontology: a) Subset of the ontology used to assess correct behavior at intersections and roundabouts. Situation objects (orange boxes) can have multiple attributes (green boxes). b) Examples of relations between situation objects (red boxes)

3.3. Agent system

The assessment system is designed as a multi-agent system. Over 70 agents are organized in a layered structure following the principles of subsumption architecture as proposed by Brooks (1991). Each layer of agents is responsible for specific behaviors or tasks, with lower layers handling basic reactive behaviors and higher layers managing more complex tasks through the integration and interpretation of the outputs from these lower layers.

The layered structure of the subsumption architecture allows for the easy modularization of tasks, with each layer of agents specializing in certain behaviors. This greatly simplifies the problem-solving process, as each agent can focus on a specific subset of tasks without needing to be aware of the entire system. Moreover, new functionalities can be introduced seamlessly as new agents, further enriching the system's capabilities without causing disruptions. In addition to facilitating system growth and adaptability, this architecture enhances the

overall system robustness, as the independent operation of each agent mitigates the risk of system-wide failures.

In this system, agents operate within a game loop-style execution process, executing updates in serial on the main thread. Some computationally heavy agents, such as the gaze estimation agent, run in parallel on their own threads. Once an agent completes its tasks, it updates the knowledge graph with new results, if available.

The knowledge graph serves as a shared representation of the environment and the agents' internal states. Agents can access and modify this knowledge graph, enabling them to communicate, coordinate, and collaborate. This approach allows agents to share information and work together to achieve complex behaviors.

To illustrate, the TurnIndicatorAgent, a lower-level agent, processes sensor data related to turn signaling. Higher-level agents, such as the JunctionApproachingAgent, then collate and analyze data from the TurnIndicatorAgent, the LanePositioningAgent, the GazeEstimationAgent, and others. In doing so, they effectively assess and provide feedback on a student's performance at an intersection.

Our design shares similarities with a traditional blackboard system, with agents accessing a shared knowledge source, thus fostering collective problem-solving. This approach effectively circumvents bottlenecks, even with numerous agents, while preserving a common domain understanding. Parts of this architecture are illustrated in Figure 5. Here, the central role of the blackboard-like system (Erman et al. (1980)) that facilitates communication between all agents is highlighted. It manages the producers and consumers of all the data, ensuring there is only one producer for each type of data, and agents can subscribe as consumers to the data they need.

However, the traditional blackboard architecture has some shortcomings as mentioned e.g. by Nwana et al. (1996). To avoid these pitfalls, our implementation ensures:

- No master agent controls access to the blackboard.
- Concurrency and robustness are preserved, as every part of the knowledge graph has a unique producing agent.
- Information is exclusively appended to the knowledge graph. Deletions occur only after all consuming agents have processed the data set for removal.
- The controller's role is limited to defining agent roles as producers or consumers, avoiding subsequent operational overhead.

The RoadNetworkAgent plays a crucial role in the overall system. It uses a digital street map of the simulated driving environment from the static part of the knowledge graph to track the real-time location of the ego car and other traffic participants. This real-time locational information is then relayed back to the dynamic part of the knowledge graph, where higher-level agents integrate this information to assess complex tasks, such as the vehicle's position within a lane or yielding behaviors.

In this assessment system based on subsumption architecture and knowledge graphs, complex behaviors and system-level intelligence emerge from the interactions between simpler agents at different layers. This layered, emergent approach enables the system to provide detailed information, to be used to create context-specific feedback and instruction.

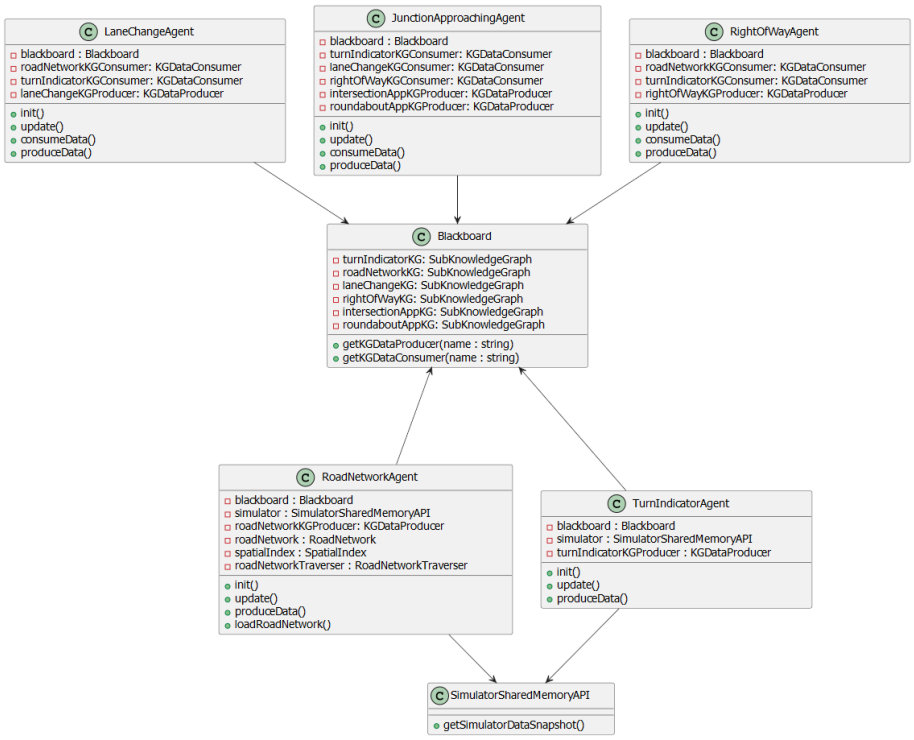

Figure 5: The architecture of the Assessment System. The system is comprised of a multitude of agents each responsible for different traffic situation awareness aspects: RoadNetworkAgent, TurnIndicatorAgent, LaneChangeAgent, JunctionApproachingAgent, and RightOfWayAgent, along with many others in the complete system. Each of these agents interfaces with the Blackboard, a central component that hosts the knowledge graph and facilitates the provision of SubKnowledgeGraphs for data producers and consumers. The Blackboard effectively governs concurrent data access, ensuring seamless information exchange. The agents serve specific roles either as KGDataProducers, contributing data to a subgraph, or as KGDataConsumers, extracting information from a subgraph. In addition, the RoadNetworkAgent and TurnIndicatorAgent are able to access the most recent simulation data through the SimulatorSharedMemoryAPI, allowing real-time interaction with the simulated environment.

3.3.1. Detailed Descriptions of Agents

This section provides an overview of specific agents discussed in this paper, selected for their relevance to our study. Each agent plays a vital role in monitoring and assessing different aspects of the student's driving performance. Let

us delve into the specific responsibilities and functionalities of each agent in Table 1.

Agent name	Description
TurnIndicator	A straightforward agent responsible for monitoring the activation status of the Ego vehicle's turn signals.
Braking	An agent dedicated to examining the smoothness and appropriateness of the braking actions executed by the student.
GazeEstimation	This agent interprets the student's head pose and gaze direction captured by the driver-facing camera. It uses this data to evaluate the student's attention to mirrors and blind spots.
Stop	This agent assesses the student's stopping performance at designated stop positions. Its assessment criteria include the student's accuracy in stopping at the stop line and the smoothness of braking, as informed by the BrakingAgent.
RoadNetwork	A key agent that identifies the positions of all traffic participants on the map and calculates distances along lanes to nearby vehicles.
LanePositioning	This agent is in charge of evaluating the student's lateral positioning within the lane, including whether the student is on the left, in the middle, or on the right of the lane.
RightOfWay	This agent inspects if the student correctly yielded when needed while entering the intersection, thus checking for potential right-of-way violations.
IntersectionPath	This agent determines the path that the student chose while traversing the intersection.

JunctionApproaching	The assessments of all agents relevant for the phase of approaching intersections or roundabouts are aggregated by this agent.
JunctionPassing	The assessments of all agents pertinent to the phase of passing through intersections or roundabouts are accumulated in this agent.
Junction	This agent integrates the results of the JunctionApproachingAgent and JunctionPassingAgent.
Tutoring	This agent takes into account all the assessments provided by the higher-level agents and determines the most suitable feedback to offer to the student.

Table 1: List of agents discussed in this work and their description

3.4. Collaborative development with traffic instructors

The development of our intelligent driver assessment system was a collaborative endeavor between our research team and a group of professional traffic instructors. This approach was chosen to ensure that the system is not only technologically sound but also pedagogically effective and relevant in the context of real-world driver education.

Traffic instructors possess a deep understanding of the complexities of driving skills, which are a blend of cognitive, motor, and perceptual abilities. They also have vast practical experience in observing, assessing, and instructing student drivers. As such, their input was invaluable in defining what to assess and how to assess it, grounding our system in the realities of driving instruction.

The instructors were actively involved in several key stages of system development. Firstly, they helped to identify the crucial driving skills and behaviors that the system should evaluate. Their expertise informed the system's focus on elements such as lane positioning, turn signaling, braking, overtaking, and approaching and passing intersections.

Secondly, they guided the development of the evaluation criteria and scoring

methodologies for each skill. For instance, they provided insights into the subtleties of assessing a student's performance in complex maneuvers like overtaking or navigating intersections. Their understanding of common student mistakes and difficulties also helped to calibrate the system's sensitivity to specific errors and omissions.

Lastly, the instructors played a significant role in the system's testing and refinement phases. They participated in the evaluation of the system, provided feedback on its accuracy and effectiveness, and helped to correct potential erroneous assessments.

The collaboration with traffic instructors ensured that our system is designed to accurately mirror the nuanced process of human driver assessment. It's a crucial aspect of our approach, reinforcing the system's potential to serve as an effective tool in driver education.

4. Reasoning example: Intersection situation

In order to elucidate how the VDI processes and interprets traffic situations, we describe a specific example where the VDI is tasked with assessing a driver's behavior at an intersection. Intersections are one of the most complex scenarios that drivers face on public roads. They can be controlled by signs, or traffic lights, or be uncontrolled, in which case vehicles have to yield to traffic coming from the right (United Nations (1968)). Drivers are expected to not only respect yielding rules but also to observe traffic and signal their intentions.

For example, signaling the intention to take a particular path through an intersection requires the use of turn signals, as well as driving in the correct lane and positioning the vehicle correctly within the lane. The VDI needs to assess if the driver is following these procedures. The VDI can make this assessment by using the information in the knowledge graph about the situation and the situation objects, such as the road signs, traffic lights, and the positions and speeds of other vehicles.

By dividing the correct behavior at intersections into three sub-situations

(approaching, passing, and exiting), the VDI can assess the student's driving performance in a more granular and detailed manner. This allows for a more accurate assessment of the student's performance and gives the ability to provide targeted feedback to improve their driving skills. For example, when approaching the intersection, the student driver has to

- observe the traffic ahead and look in the rearview mirror,
- indicate the path through the intersection by placing the car on the correct lane or on the correct side of the lane and by turn signaling,
- adjust the speed including smooth braking and stopping at the correct stop position if needed,
- respect the right of way of other vehicles but also enter the intersection in time when the gap is big enough or the ego car has priority.

4.1. Yielding violation

A yielding violation can occur when two road users cross an intersection on different overlapping paths, with one path having priority over the other. The intersection paths, which are defined in the ontology as static situation objects are derived from the lane situation object. These paths consist of a set of lane vertex points and are connected to a lane leading into the intersection and to a lane leading out of the intersection. The right-of-way can be inferred either from road signs or through the priority to the right rule, with the priorities already defined in the knowledge graph as described in subsection 3.2.

Thus, we can define a yielding violation as follows. Consider two intersection paths, p_a and p_b, where p_b has priority over p_a, as defined by the relation $hasPriority(p_b, p_a)$. An illustrative depiction of such a situation is presented in Figure 6. Suppose that a car, c_A, is on p_a and another car, c_B, is on p_b. Let $\mathbf{T}$ be the set of time points during which the environment was sampled when c_A was crossing the intersection path p_b. c_A violates the right-of-way towards c_B if there is a collision risk between the two cars at any time point in $\mathbf{T}$, or if c_B was

required to slow down to avoid a collision risk. It is important to note that this definition is general and applies to any number of intersection paths and traffic participants. The assessment system must evaluate possible yielding violations for all possible combinations of intersection paths and traffic participants.

Hence, the assessment of potential yielding violations necessitates the identification of traffic participants on intersecting paths within the context of the situation and the determination of the time frame $\mathbf{T} = t_0, t_1, ..., t_N$. After that, the collision risk must be estimated at every time step within $\mathbf{T}$.

To assess possible yielding violations of the Ego car c_{EGO}, the first step is to query the knowledge graph for all other road participants $v_i \in \mathbf{V}$ that are approaching or crossing the same intersection as c_{EGO} is entering at t_0. This allows for filtering out most of the traffic participants and only considering the relevant ones.

$$\begin{aligned} \forall v_i \in \mathbf{V} \; & \mathit{isTrafficParticipant}(v_i) \wedge \mathit{isIntersection}(s) \\ & \wedge \mathit{isEntering}(c_{EGO}, s, t_0) \\ & \wedge (\mathit{isApproaching}(v_i, s, t_0) \vee \mathit{isCrossing}(v_i, s, t_0)) \\ & \Rightarrow \mathit{atSameIntersectionWhileEntering}\,(v_i, c_{EGO}, s, t_0)\,. \end{aligned} \quad (1)$$

Equation 1 represents a first-order logic rule that can be read as follows: For every vehicle v_i within the set of vehicles $\mathbf{V}$, if v_i is recognized as a traffic participant and is either approaching or already crossing the intersection s at the same time t_0 as the Ego vehicle c_{EGO} is entering s, then v_i is considered to be at the same intersection while c_{EGO} is entering.

This method enables us to filter out most of the traffic participants in the environment and restricts further processing to only the ones relevant in the situation. After filtering out irrelevant road participants using rule (1), the RightOfWayAgent checks if the traffic participants v_i that are at the same intersection $v_i \in \mathbf{V}_s$ like c_{EGO} are actually on conflicting paths and if c_{EGO} potentially needs to yield for any of them as they are on an intersection path

with priority over the path of c_{EGO} using rule (2)

$$\forall v_i \in \mathbf{V}_s, \forall t_j \in \mathbf{T}$$
$$atSameIntersectionWhileEntering(v_i, c_{EGO}, s, t_j)$$
$$\wedge\, isOnIntersectionPath(c_{EGO}, p_{EGO}, t_j)$$
$$\wedge\, isOnIntersectionPath(v_i, p_{v,i}, t_j) \wedge isConflicting(p_{EGO}, p_v)$$
$$\wedge\, hasPriority(p_v, p_{EGO}) \Rightarrow hasToYield(c_{EGO}, v_i, s, t_j). \quad (2)$$

Equation 2 can be read as follows: For each vehicle v_i at the same intersection as the Ego vehicle c_{EGO} at any time t_j within the set of time points $\mathbf{T}$, if both c_{EGO} and v_i are on their respective intersection paths at time t_j and if the paths are in conflict with each other, with v_i's path having priority, then the Ego vehicle c_{EGO} must yield to the vehicle v_i at the intersection.

All these inferences can be made by querying the knowledge graph and performing logical reasoning. However, to determine if c_{EGO} has violated the right of way of any of the other traffic participants v_i, the collision risk needs to be calculated

$$\forall v_i \in \mathbf{V}_y, \forall t_j \in \mathbf{T}\ hasToYield(c_{EGO}, v_i, s, t_j)$$
$$\wedge\, hasCollisionRisk(v_i, c_{EGO}, t_j, t_0)$$
$$\Rightarrow hasViolatedRightOfWay(c_{EGO}, v_i, t_j). \quad (3)$$

This rule asserts that if c_{EGO} is obligated to yield to the vehicle v_i at the intersection and a collision risk is present, then a right-of-way violation against the vehicle v_i by c_{EGO} is confirmed at time t_j.

To be able to infer if there is a collision risk between vehicles by symbolic reasoning, the problem needs to be translated into symbols. This process is described in the following. The conflicting vehicle v_i could actively avoid a collision risk by braking down when it sees c_{EGO} entering the intersection. In that case, it can still be a yielding violation of c_{EGO} even if it did not lead to a collision. Thus, we predict the velocity of v_i on its path from the time c_{EGO}

enters the intersection till it leaves the intersection while being not influenced by c_{EGO}. We calculate the collision risk of this predicted path with the actual path of c_{EGO}. For simplicity, in this work, we just assume that v_i keeps the current velocity constant. In the next step, the RightOfWayAgent checks the position of c_{EGO} for all time points T_c in which c_{EGO} is on the intersection path of v_i, $p_{v,i}$. We also check the time points T_v when v reaches the positions of c_{EGO} in T_c. A collision risk is present if

$$\exists j, t_{v,j} \quad - \quad t_{c,j} \quad < \quad d_c \quad \Rightarrow \quad \mathit{hasCollisionRisk}(\mathbf{v_i}, c_{EGO}, t_j, t_0) \quad (4)$$

while d_c is the minimum time difference required between the time c_{EGO} is at a certain position ($t_{c,j}$) and v_i is predicted to reach this position ($t_{v,j}$). In our experiments, we set d_c to 3 seconds. Similarly, the set of rules are defined for all other situation assessments.

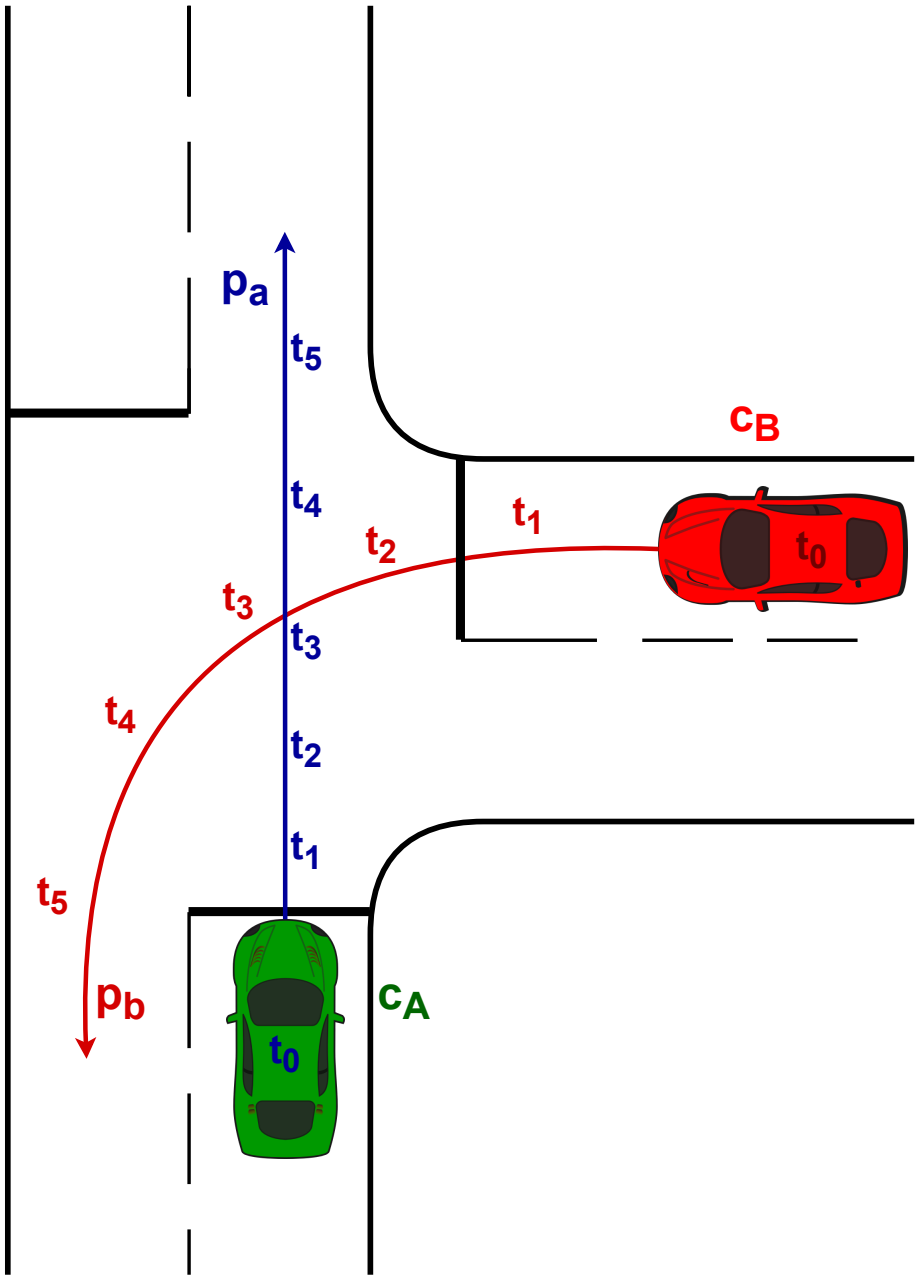

Figure 6: Depiction of an intersection situation where the green vehicle, denoted as c_A on intersection path p_a, is required to abide by the 'priority to the right' rule. Therefore, it must yield to the red vehicle c_B on the conflicting intersection path p_b to avert a potential collision. The collision risk is projected to occur at time point t_3 if the rule is not adhered to.

4.2. Intersection assessment timeline

As mentioned earlier, driving maneuvers are assessed within specific situations, such as an intersection situation. An intersection situation begins when the Ego vehicle approaches the intersection and ends after it exits. The entire process is illustrated in Figure 7. The IntersectionPathAgent infers the start and end times for this situation. When the Ego vehicle is 50 meters away from the next intersection, the agent signals that it is approaching the intersection. The agent also signals when the Ego vehicle crosses the intersection stop line and enters the intersection. Once the Ego vehicle exits the intersection, this event is signaled as well.

The exact path through the intersection is determined by the IntersectionPathAgent when the Ego vehicle exits the intersection. At this point, the agent can be certain of the specific path that the vehicle has taken and that it should have taken, even in cases where lanes partially overlap. This information is deduced by querying the knowledge graph, which contains details about which intersection paths connect the entry and exit lanes. Once the correct path is known, the system can then accurately infer the correct entry lane and the appropriate positioning within this lane, along with other factors such as the required direction for turn signaling.

This is where the high-level reasoning of the JunctionApproachingAgent comes into play. This agent interacts with the knowledge graph to gather information from a plethora of other agents. Whenever the JunctionPassingAgent detects that an intersection has been crossed, it updates the knowledge graph with relevant information. The JunctionApproachingAgent then queries the knowledge graph to obtain data from other agents involved in approaching an intersection or roundabout, such as turn indicator data from the TurnIndicatorAgent, lane, and lane positioning data from the LanePositioningAgent, braking data from the BrakingAgent, yielding data from the RightOfWayAgent, and gaze estimation data from the GazeEstimationAgent.

Using this data gathered from the knowledge graph, the JunctionApproachingAgent performs logical reasoning to determine the successes and mistakes made

by the student in the junction-approaching situation. This explanatory data is then added to the knowledge graph, where it can be consumed by the TutoringAgent to provide appropriate feedback.

Given this structure, the JunctionApproachingAgent and other agents wait until the IntersectionPathAgent signals the intersection exit before assessing the student's correct or incorrect behavior while approaching the intersection. For example, the RightOfWayAgent checks for yielding violations.

Other agents that do not require information about the exact intersection path assess the situation immediately. For instance, if the Ego vehicle stops before the intersection, the StopAgent evaluates how accurately the vehicle stopped at the stop line and how smooth the braking was.

Ultimately, the JunctionApproachingAgent combines all this information to calculate an overall performance score for the student. A similar process is followed by the JunctionPassingAgent. The JunctionAgent then combines the results from both the JunctionApproachingAgent and JunctionPassingAgent to obtain an overall junction performance score.

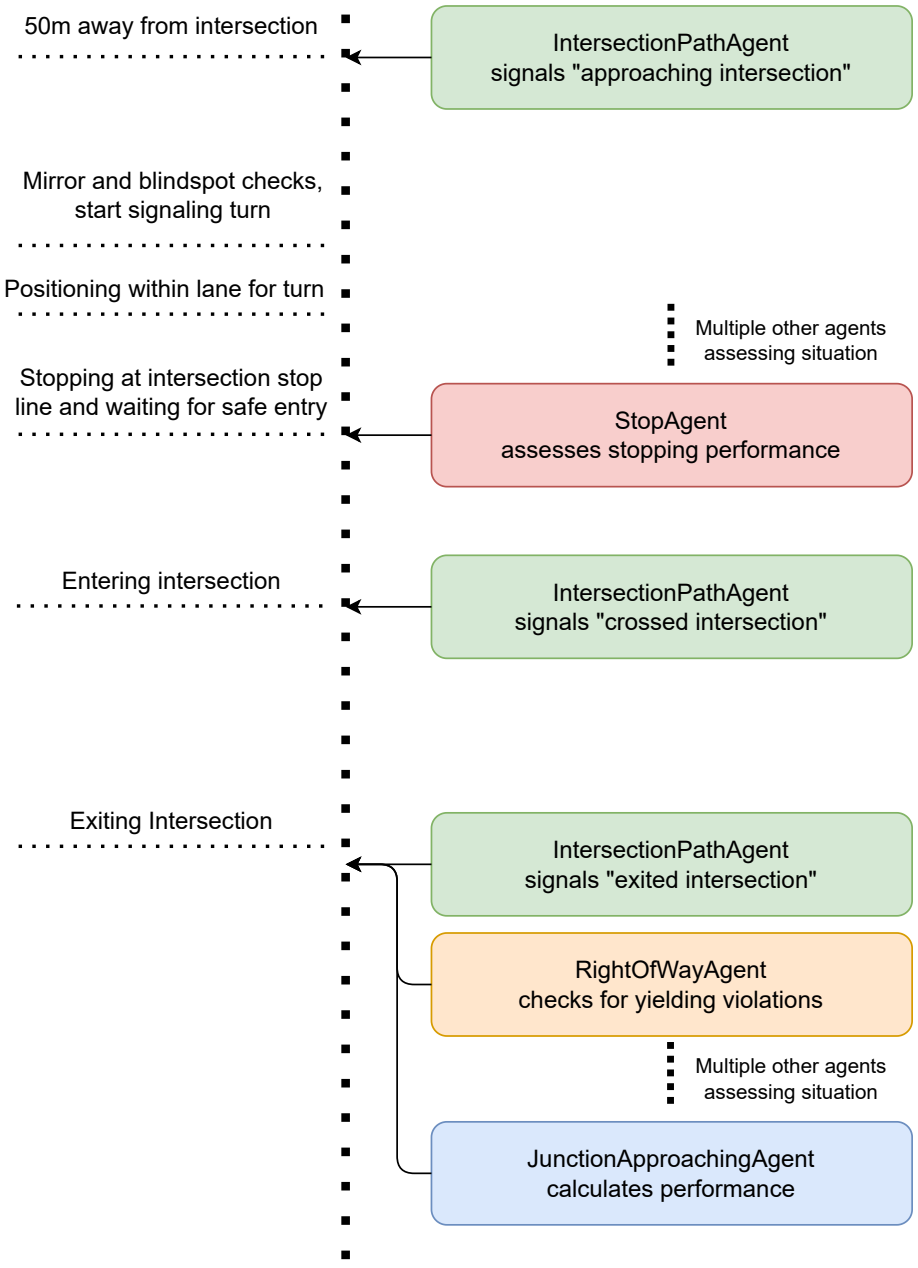

Figure 7: Intersection assessment timeline

5. Experiments and results analysis

In this section, we systematically analyze the experiments conducted and the results obtained. First, we present our experimental procedure, outlining the structure and components of the driving sessions. Then, we delve into the experts' evaluation procedure, explaining how experts assessed student performances and reached a consensus. We further perform a comparison between the expert consensus and the outputs of our VDI's automated assessment system, followed by an in-depth look into the assessment system agreement with single experts. We conclude with a discussion on the challenges for students and the level of consensus achieved amongst expert evaluators.

To the best of our knowledge, no existing work validates the VDI as thoroughly as we do in this paper. This lack of comprehensive validation in existing literature limits our ability to make direct comparisons with other studies.

5.1. Experimental procedure

Our system was put to the test with real students driving in the simulator. To guarantee the reproducibility of our experiment and allow for offline data processing, all driving sessions were recorded. The evaluation of our system was subsequently conducted on the resulting dataset derived from these recordings. The experimental setup employed a replay system capable of playing back these recorded driving sessions, effectively emulating the live system's functioning.

The participants of the experiment were students at Way AS, a commercial driving school in Norway that also develops its own high-fidelity simulators. The experiment included participants with varying levels of experience in both simulator driving and real-world driving. This diversity enabled us to evaluate our system across a range of driving proficiencies and a wide array of distinct driving errors.

Each participant drove three different lessons in the simulator:

- **Overtakes**: A circular track with a two-way street that features long stretches for overtaking other cars, as well as curves and hills that obstruct

the view. The experiment includes 18 recorded overtake sessions with a total of 74 overtake situations. Each situation was evaluated by an average of 3.5 experts.

- **City driving**: An artificial city scenario with complex multi-lane intersections that are primarily controlled by traffic lights. The experiment includes 18 recorded city driving sessions with a total of 157 intersection situations. Each situation was evaluated by 2 experts.
- **Town driving**: A real town in Norway was recreated in the simulation, featuring many roundabouts and smaller streets with the priority-to-the-right rule. The experiment includes 16 recorded town driving sessions with a total of 158 intersection situations and 87 roundabout situations. Each situation was evaluated by an average of 3.4 experts.

The experimental procedure was executed in the following stages:

1. **Design and planning**: This stage involved conceptualizing the experiment and outlining the procedure.
   Output: *Experiment design.*
2. **Simulator driving by students**: In this phase, students drove in the simulator, generating the driving sessions dataset that forms the basis of our experiment.
   Output: *Student driving sessions.*
3. **Evaluation by driving experts**: Experts meticulously assessed each driving session, with multiple experts evaluating each situation. The result is the expert evaluations of each driving situation for all the driving sessions. Output: *Expert evaluations.*
4. **Quality assurance**: The expert evaluations underwent a thorough review process by an expert committee, which revised expert evaluations that raised questions or doubts, culminating in the revised expert evaluations. Based on the revised expert evaluations, an expert consensus was computed.
   Output: *Expert consensus.*

5. **Optimization**: The assessment system was optimized based on the expert consensus.
   Output: *Optimized assessment system used in experiments*
6. **Experiments**: The optimized assessment system was run on all driving sessions to produce VDI assessments for all driving situations experienced by the students.
   Output: *VDI assessments*
7. **Analysis**: Finally, we computed the result metrics by comparing the VDI assessments to the expert consensus.
   Output: *Results metrics*

This methodical approach ensured a rigorous evaluation of the assessment system, thereby enhancing the reliability of our findings.

5.2. Experts evaluation procedure

The driving sessions are evaluated by different experts in the field. This includes experienced driving instructors, traffic education researchers, and official driving examiners. The driving sessions were presented to the experts in a web platform for evaluation which is based on Rehm et al. (2021). Each driving session encompasses a variety of situations, with our evaluation focusing specifically on overtaking, intersections, and roundabout scenarios. We distinguish between a 'scenario', which is defined as a specific class of driving maneuvers, and a 'situation', which is an instance of these maneuvers occurring in real-time. Every situation within these key areas is scrutinized through multiple evaluation items, which remain consistent across each unique scenario. To provide an example, an evaluation item might be whether a student has properly used their turn signal before initiating a lane change. This evaluation takes into account whether the signal was engaged at the appropriate time, discontinued suitably after the lane change was completed, and correctly indicated the intended direction. The evaluation items are evaluated as 'passed' or 'failed'. The web interface used for this process is showcased in Figure 8.

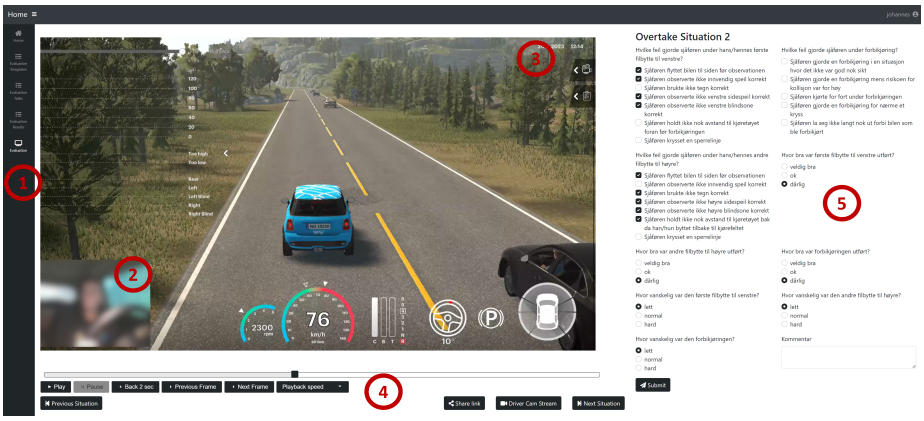

Figure 8: Screenshot of the evaluation web frontend. Features include: *1.* Navigation menu for ease of access; *2.* Camera stream showing the student's face (blurred for privacy); *3.* Live replay stream of simulation visuals; *4.* Controls for playing/pausing and navigating between different situations; *5.* Evaluation input form with checkboxes for marking mistakes, as well as other options.

Each evaluation item within each situation was assessed by multiple experts. Subsequently, we conducted a quality control review of all evaluations to identify any potential errors. There are various reasons for such mistakes, and we have listed a few of them here:

- Oversights / forgotten evaluation items: In the process of evaluating complex driving scenarios, there may be instances where experts overlook or forget certain evaluation items. This could be due to the multitasking nature of evaluation, which involves experts needing to simultaneously observe and assess multiple aspects of the student's performance.

- Mistakes while using the evaluation tool: Errors can also arise from improper use of the evaluation tool. For instance, an expert might inadvertently select an incorrect option or input erroneous data, leading to imprecise evaluations.

- Misunderstanding of evaluation items: Misinterpretations can extend beyond differing opinions on specific situations. In some cases, experts may

incorrectly attribute mistakes to evaluation items that are not entirely relevant.

For quality control, we convened a committee consisting of one technical expert and three experienced simulator driving instructors. This group collaboratively discussed each potential erroneous evaluation case and made corrections. By integrating the expert evaluations with the corrections from the committee, a consensus was reached for every evaluation item across all driving situations. The culmination of this process is the expert consensus dataset. We use this dataset as a standard to compare the assessment system's performance, as shown in Tables Table 5, Table 6, and Table 7.

5.3. Single experts agreement with expert consensus dataset

The results presented in Table 2 compare the agreement of individual experts with the consensus drawn from the expert consensus dataset. Each cell of the table comprises two values: The first number signifies the True Positive Rate (TPR), which represents the proportion of actual positive cases (i.e., driving mistakes as per expert consensus) correctly identified as such by the respective expert. The second number indicates the True Negative Rate (TNR), or the proportion of actual negative cases (i.e., correct driving maneuvers as per expert consensus) correctly identified as negatives by the respective expert.

A critical observation derived from the table is the relatively low TPR values compared to the TNR across all experts and driving contexts. This pattern suggests that experts might have overlooked a considerable number of error instances during their assessments. The intricacies and subjective nature of driving skills assessment can contribute to such oversights. Despite this, it's important to note that the experts generally align well with the consensus, particularly in terms of TNR, highlighting the overall reliability of our expert panel in identifying correct driving behaviors.

Our analysis further delved into how much experts agreed with each other on the different evaluation items. The data indicates that agreement rates are influenced by the balance of pass and fail instances, with lower consensus in

more evenly split scenarios. Moreover, items such as mirror checks pose significant challenges due to their subtle nature, suggesting the utility of eye-tracking technology for more consistent assessments. Details about this evaluation can be found in Appendix A.

We also analyzed the expert agreement in conjunction with the perceived difficulty by the experts of driving maneuvers and discovered that there is no substantial correlation between the two, suggesting that the nature of the evaluation items has a greater impact on consensus than their difficulty. Specific statistical distributions and further details of this analysis are available in Appendix B.

Name	Overtakes	Town Driving	City Driving
	TPR / TNR	TPR / TNR	TPR / TNR
Expert 1	0.73 / 0.94	0.47 / 0.97	0.46 / 0.98
Expert 2	0.80 / 0.88	0.68 / 0.94	–
Expert 3	–	–	0.57 / 0.94
Expert 4	0.58 / 0.95	0.63 / 0.95	–
Expert 5	0.66 / 0.82	0.44 / 0.98	–
Expert 6	0.60 / 0.93	0.35 / 0.97	–
Expert 7	0.90 / 0.91	0.78 / 0.91	–

Table 2: Single expert agreements with consensus from expert consensus dataset. First number is TPR (rate of agreement that it is a mistake). Second number is TNR (rate of agreement that it is no mistake)

5.4. Challenges for students

In the process of learning to drive, students face various challenges that often manifest as mistakes during their practice sessions. This section delves into the specific challenges students face in different driving scenarios, highlighting the frequency of these errors and discussing potential causes. For a detailed breakdown of these findings, please refer to Table 3.

Here, we distinguish between driving mistakes and observation mistakes. Observation mistakes encompass missed mirror checks, overlooked intersection observations, and any other errors related to visual attention. Conversely, driving mistakes constitute all other types of errors.

The highest frequency of mistakes per situation was found during overtaking maneuvers, with students committing an average of 2.4 driving mistakes and 3 observation mistakes per overtake. In contrast, in the complex city environment, students averaged only one driving mistake and 1.4 observation mistakes per intersection. Roundabouts also posed observational challenges for students, with an average of 3.3 observation mistakes and 0.9 driving mistakes per roundabout.

The relatively high number of mistakes during overtakes could be attributed to the fact that intersections and roundabouts receive more emphasis during driver's education due to their critical role in the driving examination. Overtaking maneuvers, however, are generally not part of the driver's exam requirements, leading to less practice in this area. Additionally, driving in the opposing lane induces significant stress for students, which is reflected in the fact that 2 out of the 2.4 average driving mistakes occur while the student is in the opposing lane.

Generally, the high occurrence of observation mistakes underscores the importance of incorporating reliable eye-tracking technology into such an assessment system. This would enhance the system's capability to accurately identify and address these types of errors.

The insights derived from this analysis underscore again the importance of comprehensive training and the potential value of standardized and automatic driving assessment.

	Mistakes	
Lessons	Driving	Observation
Overtakes	2.4	3.0
City Driving	1.0	1.4
Town Driving (roundabouts)	0.9	3.3
Town Driving (intersections)	0.3	1.2

Table 3: Average driving and observation mistakes per situation according to the expert consensus

5.5. Comparison of the expert consensus and the VDI assessments

The performance of the assessment system is compared to the evaluations from the Expert consensus. It is important to note that at the time of this study, a reliable eye-tracking system had not yet been implemented. Consequently, evaluation items that require the assessment of visual attention, such as intersection observation and mirror or blind spot checks, were not included in the analysis. Some evaluation items did not have a single mistake according to the evaluations in the expert consensus dataset. An example of that is the correct lane choice. The town lesson does not have intersections with multiple lanes from one road going into the intersection. So students cannot chose the wrong lane. To keep the results concise and relevant, these evaluation items are excluded, as it is not possible to calculate a true positive rate for these.

The properties of the dataset, in particular, an inherent imbalance between the mistake and no-mistake classes, renders the most basic performance metric, the accuracy, as not sufficient in our task of evaluating how close the assessment system is to the expert consensus. The accuracy metric is defined as

$$Accuracy = \frac{S_{agreed}}{S_{consensus}} \tag{5}$$

where S_{agreed} is the number of situations in which the expert consensus agreed with the evaluation of the assessment system while $S_{consensus}$ is the number of situations with expert consensus.

In cases where driving mistakes are very rare, high accuracy can be achieved by just always classifying it as no mistake. This issue is described in more detail in He & Garcia (2009). Therefore, we need to take into account all combinations of the assessment classification outcome vs expert consensus. We define a true positive (TP) as a case where the majority of the human evaluators and the assessment system evaluated an evaluation item in a specific situation as a driving mistake. False positives (FP), false negatives (FN) and true negatives (TN) are defined accordingly as depicted in Table 4

		Assessment System	
		Mistake	No mistake
Human Evaluator Consensus	Mistake	TP	FN
	No mistake	FP	TN

Table 4: Definition of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) for all the combinations where the majority of the human evaluators and / or the assessment system evaluated an evaluation item in a specific situation as a driving mistake or not.

Our research aims to evaluate the accuracy of our system in driving mistake detection while ensuring it does not erroneously identify non-mistakes as failed in the context of student driving. False-positive alerts could lead to unwarranted mistrust in the system due to students receiving incorrect negative feedback. Furthermore, such misleading alerts carry the risk of fostering incorrect learning.

This is why we look at the true positive rate and the true negative rate and not at recall and precision as they disregard the true negatives and focus only on the positive cases (Powers (2011)). They are defined as follows.

True positive rate (sensitivity)

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

True negative rate (specificity)

$$TNR = \frac{TN}{TN + FP} \tag{7}$$

A disagreement between the assessment system and the expert consensus does not automatically mean that the assessment system made a mistake. However, if the sensitivity and specificity are low and the expert agreement score is high, it would be a strong indication that the assessment system performance is poor.

An overview of the results for the city driving lesson is presented in Table 5 and its corresponding bar plot in Figure 9. The results for the town driving lesson are detailed in Table 6 and illustrated in the bar plot Figure 10. Results of the overtakes lesson can be observed in Table 7 and the associated bar plot Figure 11.

Apart from accuracy, TPR, and TNR, the tables also include the number of instances considered as failed (driving mistake) and passed (no driving mistake) for each evaluation item. The total number of instances for each evaluation item may vary slightly because instances lacking expert consensus – where experts are evenly split on whether it is a mistake – are excluded.

By optimizing the assessment system using the consensus data, we strived to align its output closely with the consensus of expert evaluations. Generally, as can be seen in Table 5, Table 6 and Table 7, the true positive rates tend to be significantly lower than the true negative rates. At first glance, this might suggest that the assessment system is primarily optimized for overall accuracy. As posited by He & Garcia (2009), classifiers optimized for overall accuracy are likely to yield high accuracy for the majority class while underperforming for the minority class.

Though we cannot completely dismiss potential bias, we contend that the primary reason for this phenomenon is that instances without mistakes are generally less ambiguous and potentially also easier to assess than those with mistakes. Given the absence of a definitive ground truth for distinguishing between mistakes and non-mistakes, we rely on expert opinions. While there

are clearly identifiable mistake cases, others fall into a grey area, regarded as mistakes by some experts and dismissed by others. If the proportion of clear non-mistake cases greatly surpasses that of ambiguous or obvious mistake cases, the former will be less ambiguous and easier to assess on average.

However, the high true negative rates show that the number of false positives is relatively low. That means our goal to have a very low number of false alarms is achieved.

Let us start diving into a few specific results in more detail. Evaluation items such as "waited too long before entering" an intersection or roundabout exhibit rather low true positive rates, as evidenced by the city driving lesson (a rate of 0.78, as per Table 5) and town driving lesson (a rate of 0.0, see Table 6). This discrepancy can be attributed to the inherent subjectivity involved in evaluating these instances – there is no definitive rule to determine when a driver has waited excessively before entering a junction. As a consequence, there is a prevalence of ambiguous cases compared to clearly identifiable mistakes, which further elucidates why none of the three mistake cases, despite achieving consensus among the experts, were identified as mistakes by the assessment system. The assessment system follows a clear rule that if a student waits for more than 5 seconds when a gap in traffic is considered sufficiently large to proceed, it is classified as a case of the student waiting too long before entering.

Moving to a different context, the evaluation of turn signaling, particularly when students change lanes to the right, has shown some interesting patterns. This aspect assesses whether drivers appropriately and timely activate their turn signals before executing the lane change back to the right at the end of an overtake. Our empirical observations indicate a noticeable frequency of student drivers initiating their right turn signal relatively late prior to transitioning back to the right lane. One potential explanation for this pattern could be the driver's heightened concentration on the ongoing events in the oncoming lane while overtaking, leading to the activation of the turn signal just prior to the actual lane change. While human evaluators might demonstrate flexibility towards this behavior due to the context, the system, conversely, does not take

into account the specific situation necessitating the turn signal, thus enforcing uniform standards.

Next, we shift our attention to the evaluation of right of way situations i.e., whether there was a yielding violation or not. Here, the assessment system demonstrated very good concurrence with the expert consensus. Among the 474 instances examined, a discrepancy was found in just 6 cases. The RightOfWay agent, tasked with this particular assessment, was not only the most complex to construct but also subjected to the most exhaustive testing. As a result, its proficiency on the expert consensus dataset is quite high.

A notable limitation of the current evaluation system pertains to speed assessment. As it stands, under-speeding is only evaluated on rural roads and highways. Determining if a driver exceeds the speed limit is a relatively straightforward assessment. However, identifying instances where learner drivers maintain speeds excessively below the norm is crucial, given its potential impact on traffic flow. The ideal speed is highly contingent on context, influenced by factors such as road width and curvature, obstructions due to buildings or parked cars, and unpredictable pedestrian movements, particularly those involving children. Moreover, this context-dependent complexity is not unique to speed assessment. Many other evaluation items in the system share this characteristic. Therefore, addressing this aspect presents a significant challenge in the development of a robust rule-based evaluation system for what could superficially be seen as straightforward items to assess.

	Acc.	TPR	TNR	#pass	#fail	ξ	ξ_f
Intersection entering:							
Turn signaling	0.91	0.86	0.93	113	36	0.92	0.70
Lane positioning	0.92	0.86	0.94	111	36	0.92	0.72
Lane choice	1.00	1.00	1.00	146	4	0.96	0.68
Smooth braking	0.80	0.95	0.78	132	20	0.93	0.63
Stop accuracy	0.96	0.87	0.99	106	31	0.92	0.61
Right of way	1.00	1.00	1.00	143	9	0.98	0.50
Waited too long	0.96	0.78	0.97	142	9	0.98	0.70
Intersection passing:							
Speed intersection	0.95	0.50	0.98	141	8	0.94	0.57
Total	**0.94**	**0.86**	**0.95**	**1034**	**153**	**0.94**	**0.64**

Table 5: Results of the **city driving lesson**. The table presents the performance metrics of all driving sessions, calculated based on the expert consensus dataset, with the exception of 'ξ' which is derived from the expert evaluations prior to the quality assurance step. 'Acc.' refers to Accuracy, 'TPR' to True Positive Rate (indicating correctly identified driving mistakes), and 'TNR' to True Negative Rate (indicating correctly identified non-mistakes). '#pass' and '#fail' represent the number of situations marked as 'passed' (no driving mistake) and 'failed' (presence of a driving mistake) respectively according to the expert consensus. 'ξ' denotes the agreement score among experts and ξ_f denotes the agreement score for all instances which have at least one fail vote.

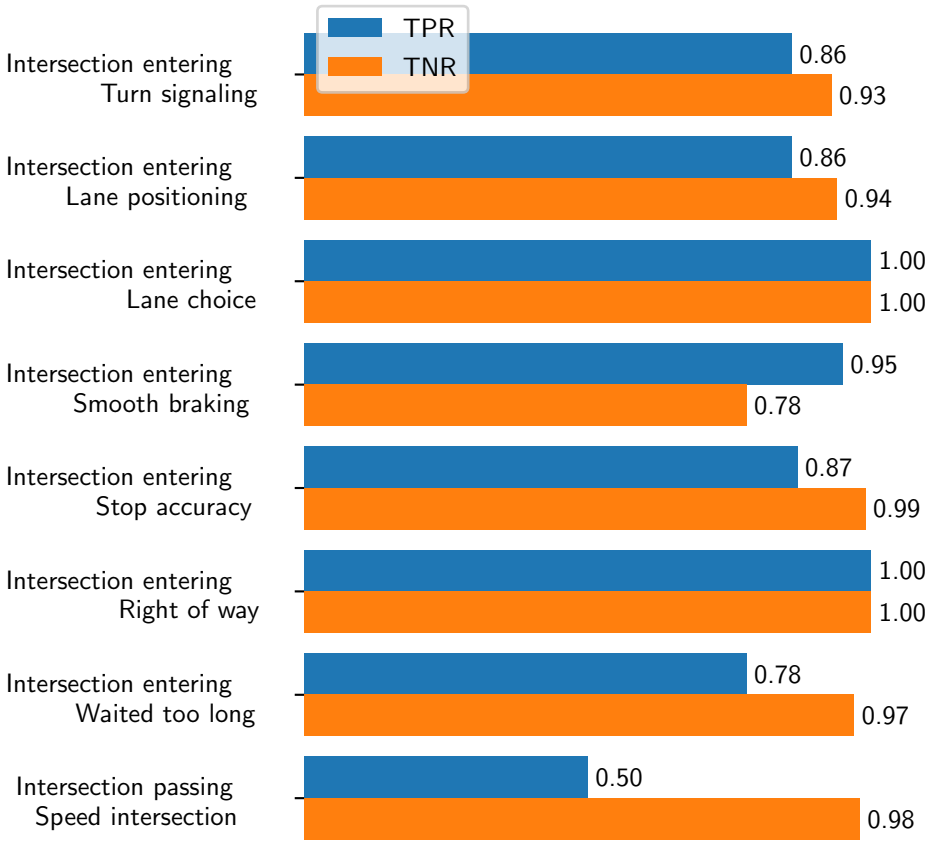

Figure 9: Bar plot of the results of the **city driving lesson**. The chart shows the True Positive Rate (TPR), indicating correctly identified driving mistakes, and the True Negative Rate (TNR), indicating correctly identified non-mistakes.

	Acc.	TPR	TNR	#pass	#fail	ξ	ξ_f
Roundabout entering:							
Lane positioning	0.85	0.83	0.85	72	12	0.91	0.71
Lane choice	0.98	0.67	0.99	81	3	0.93	0.71
Smooth braking	0.82	1.00	0.79	70	14	0.93	0.68
Right of way	0.96	0.67	0.97	80	3	0.97	0.67
Waited too long	0.96	0.00	1.00	80	3	0.95	0.69
Inside roundabout:							
Lane positioning	0.92	0.67	0.95	75	9	0.90	0.72
Lane changes	1.00	1.00	1.00	70	14	0.94	0.72
Speed	0.94	0.50	0.96	79	4	0.94	0.69
Right of way	0.98	1.00	0.97	80	3	0.97	0.74
Roundabout exiting:							
Turn signaling	0.95	0.85	0.97	70	13	0.88	0.72
Intersection entering:							
Turn signaling	0.94	1.00	0.94	148	9	0.98	0.73
Lane positioning	0.95	0.83	0.96	138	18	0.95	0.71
Right of way	0.99	1.00	0.99	148	8	0.97	0.68
Intersection passing:							
Speed	0.97	0.80	0.98	150	5	0.92	0.66
Total	**0.95**	**0.85**	**0.96**	**1341**	**118**	**0.95**	**0.70**

Table 6: Results of the **town driving lesson**. The table presents the performance metrics of all driving sessions, calculated based on the expert consensus dataset, with the exception of 'ξ' which is derived from the expert evaluations prior to the quality assurance step. 'Acc.' refers to Accuracy, 'TPR' to True Positive Rate (indicating correctly identified driving mistakes), and 'TNR' to True Negative Rate (indicating correctly identified non-mistakes). '#pass' and '#fail' represent the number of situations marked as 'passed' (no driving mistake) and 'failed' (presence of a driving mistake) respectively according to the expert consensus. 'ξ' denotes the agreement score among experts and ξ_f denotes the agreement score for all instance which have at least one fail vote.

	Acc.	TPR	TNR	#pass	#fail	ξ	ξ_f
Initiation of overtake (lane change left):							
Turn signaling	0.99	0.91	1.00	58	11	0.87	0.71
Distance vehicle ahead	0.95	0.92	0.95	61	12	0.89	0.74
Lane separation	0.97	0.00	1.00	72	2	0.99	0.77
Execution of overtake:							
Visibility	0.97	0.80	0.99	68	5	0.92	0.71
Head on collision risk	0.95	0.83	0.97	61	12	0.87	0.73
Speed	0.85	0.86	0.84	44	28	0.87	0.80
Closeness to intersection	0.97	1.00	0.96	57	12	0.90	0.73
Lateral distance	0.96	0.96	0.96	46	26	0.89	0.75
Completion of overtake (lane change right):							
Turn signaling	0.96	0.94	0.97	34	35	0.86	0.77
Distance vehicle behind	0.85	0.89	0.81	31	36	0.82	0.77
Total	**0.88**	**0.90**	**0.95**	**532**	**179**	**0.90**	**0.75**

Table 7: Results of the **overtakes lesson**. The table presents the performance metrics of all driving sessions, calculated based on the expert consensus dataset, with the exception of 'ξ' which is derived from the expert evaluations prior to the quality assurance step. 'Acc.' refers to Accuracy, 'TPR' to True Positive Rate (indicating correctly identified driving mistakes), and 'TNR' to True Negative Rate (indicating correctly identified non-mistakes). '#pass' and '#fail' represent the number of situations marked as 'passed' (no driving mistake) and 'failed' (presence of a driving mistake) respectively according to the expert consensus. 'ξ' denotes the agreement score among experts and ξ_f denotes the agreement score for all instance which have at least one fail vote.

6. Conclusion

This research aimed to develop a system capable of assessing traffic situations with accuracy comparable to human experts. Utilizing a subsumption architecture within a multi-agent system, our detailed approach employed a

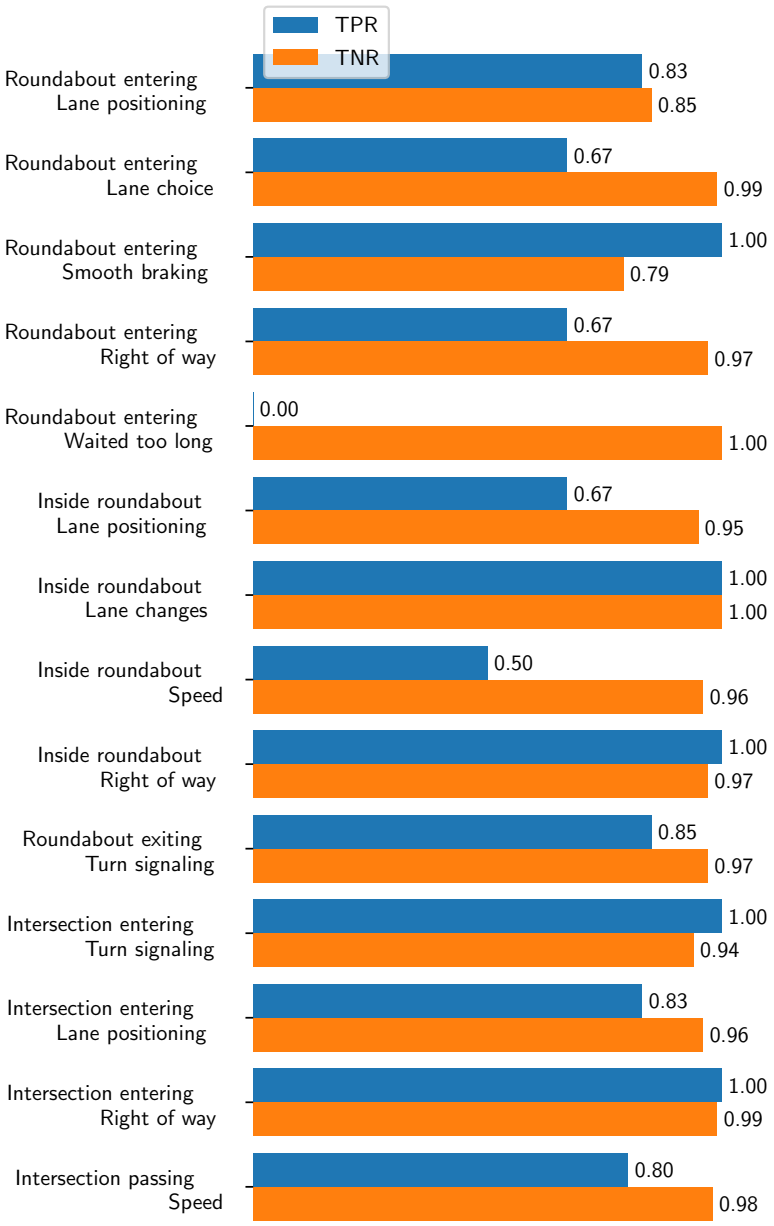

Figure 10: Bar plot of the results of the **town driving lesson**. The chart shows the True Positive Rate (TPR), indicating correctly identified driving mistakes, and the True Negative Rate (TNR), indicating correctly identified non-mistakes.

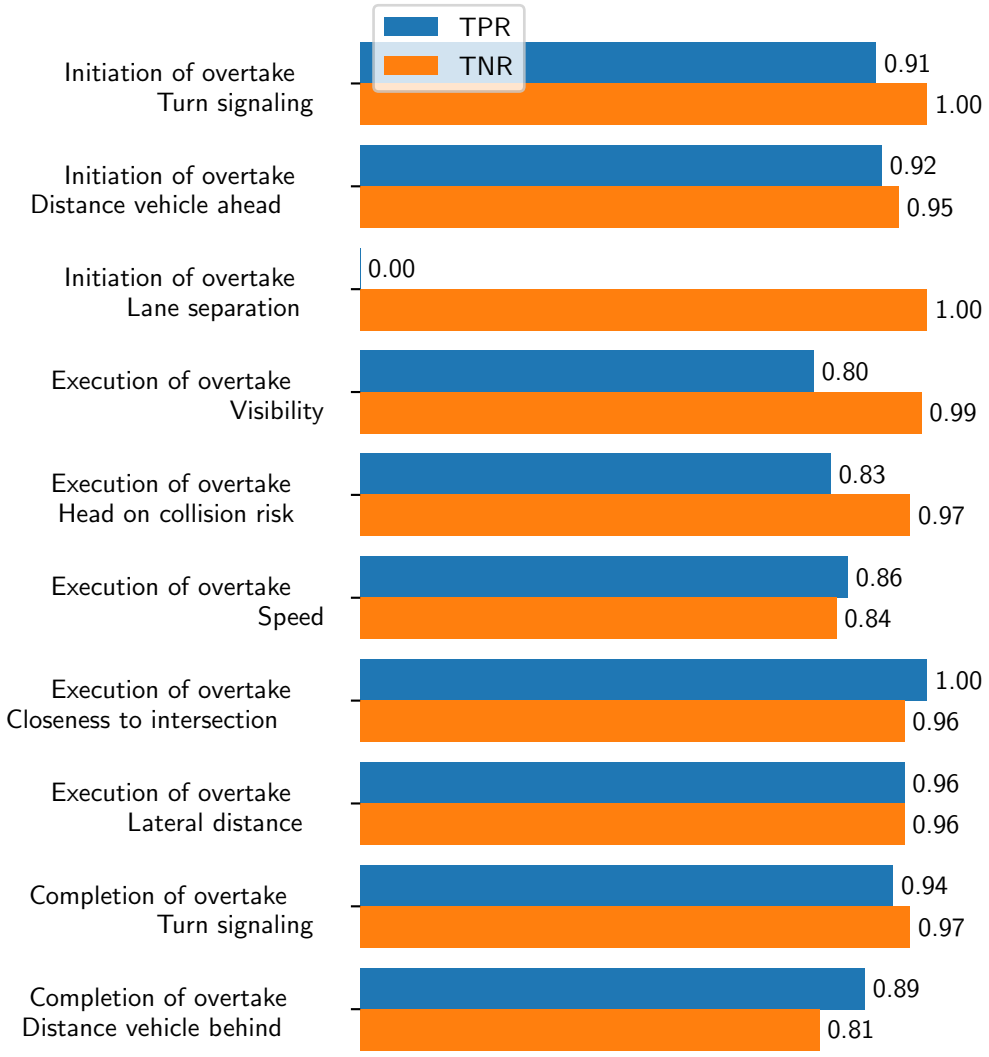

Figure 11: Bar plot of the results of the **overtakes lesson**. The chart shows the True Positive Rate (TPR), indicating correctly identified driving mistakes, and the True Negative Rate (TNR), indicating correctly identified non-mistakes.

knowledge graph for continuous situational awareness and precise driving behavior assessments.

Our validation process involved 477 driving scenarios, evaluated by 7 experts and undertaken by 21 students across three lessons. The system demonstrated a notable true positive rate of at least 0.85 and a true negative rate of at least 0.95, depending on the scenario. The results show a notable alignment with the assessments of professional driving educators, reinforcing our system's ability to accurately identify both appropriate driving actions and mistakes. These metrics answer our research question affirmatively for intersections, roundabouts, and overtaking maneuvers which are arguably among the most difficult scenarios in the driver's education. We are thereby contributing to closing the research gap in detailed evaluations and empirical validation of simulator training in driving education.

However, it's important to acknowledge inherent limitations. As with any rule-based system, our platform encounters challenges associated with rule complexity. This includes the demanding process of developing and managing a comprehensive rule set that accurately reflects the nuanced situations encountered in driving assessments. The intricacy of these rules increases with the contextual depth required for precise situation evaluation, making the system's development and maintenance both time-consuming and costly. Despite these challenges, we maintain that our approach, specifically in the context of AI-driven driving instruction, is more efficient as compared to machine learning alternatives, especially in terms of the data generation required for achieving similar levels of proficiency.

Looking forward, we plan to enhance our system with advanced eye-tracking technologies, using the latest virtual and mixed-reality headsets. This integration, along with refining our feedback mechanisms through the established assessment framework, will allow for more personalized and effective driver training. Additionally, by tracking student progress over time, we aim to develop tailored learning materials targeting individual improvement areas. These future developments are based on the strong foundation laid by our current findings,

underscoring our commitment to promoting safer driving practices through artificial intelligence. Ultimately, this research illuminates the vast potential of AI in revolutionizing traffic education and related fields.

Acknowledgements

This research was funded by Way AS and the Research Council of Norway through the TRANSPORT program under the grant agreement 296640.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Agreement on evaluations between experts

In this section, we examine the agreement between experts in their assessments of driving skills. Our analysis leverages the individual expert evaluation dataset (subsection 5.1), which encompasses the unique evaluations given by each expert.

One widely recognized measure of inter-rater reliability, Krippendorff's alpha (Hayes & Krippendorff (2007)), exhibits certain limitations within the context of our project. Specifically, our evaluation includes items that exhibit an extremely small number of failed cases, resulting in a highly skewed distribution. This imbalance in category representation significantly complicates the interpretation of Krippendorff's alpha. In particular, when a category is considerably rare, any disagreement pertaining to that category can disproportionately deflate the value of the alpha. To better accommodate these challenges, we have developed a custom **agreement score** that is more congruent with our dataset characteristics and the evaluation nuances, thus providing a more accurate representation of the expert agreement.

Firstly, all assessment values are numerically represented as follows: {'Pass': 1, 'Fail': 0}. Unassigned values are denoted as NaN. For each item instance, a consensus value is computed as the median of all expert evaluations, discarding NaNs. Subsequently, an individual agreement score $s_{\iota,\tau}$ for each item instance and for each expert is computed using the formula:

$$s_{\iota,\tau} = 1 - |c_\iota - v_{\iota,\tau}| \tag{A.1}$$

where c_ι denotes the consensus value and $v_{\iota,\tau}$ represents the value assigned by the expert. The individual agreement score ranges from 0 to 1, with 0 implying total disagreement with the consensus and 1 indicating complete agreement. The agreement scores $s_{\iota,\tau}$ are then grouped by item, and such sets contain expert agreement scores for all instances of a particular item.

Let's denote such sets Ω_{item}. Finally, the agreement score ξ_{item} for each item

is computed as the average of the distinct agreement scores:

$$\xi_{\text{item}} = \frac{\sum_{s_{\iota,\tau} \in \Omega_{\text{item}}} s_{\iota,\tau}}{|\Omega_{\text{item}}|} \tag{A.2}$$

This score offers a measure of grading difficulty for experts across different driving evaluation items, reflecting the extent of consensus among evaluations (see Table 5, Table 6 and Table 7).

As indicated by the tables, there's a strong correlation between agreement rates and the ratio of 'passed' to 'failed' cases. The lowest agreement scores are apparent for the two evaluation items in the 'Completion of Overtake' part (refer to Table 7), where the counts of 'failed' instances are roughly equivalent to those of 'passed' cases.

Conventionally, we can categorize situation evaluations into three groups: clear mistakes, clear non-mistakes, and uncertain cases. High agreement rates are expected among the experts for the first two groups and low for the latter. It seems that, for numerous evaluation items, the number of clear non-mistake instances is considerably higher than the combined count of the other two groups.

If this observation holds true, the agreement rate calculated only for instances that received at least one 'failed' vote (represented as ξ_f in the tables) should be significantly lower. In line with this conjecture, ξ_f values in the tables Table 5, Table 6, and Table 7 are indeed substantially lower than their corresponding ξ values for virtually all evaluation items, corroborating our initial premise.

This metric reveals (Figure A.12) the biggest difficulty in the assessment of driver performance for the experts: the difficulty in uniformly assessing a driver's adherence to safety protocols such as mirror checks and blind spot observations (we do not have them listed in the results tables as we do not have results on them for the assessment system). These evaluation items consistently rank among the lowest in terms of agreement score, falling below 81%. Figure A.12 also shows that the evaluation items with low agreement scores tend to cluster vertically, which suggests that the agreement score depends more on the nature

of the maneuver rather than on a specific driving session.

With respect to mirror checks and blind spot observations, the challenge of assessment is further compounded by the subtlety of the actions involved and the fleeting duration of such observations, making it inherently difficult for human evaluators to accurately appraise. The use of a highly accurate eye-tracking system will offer great potential for the automation of such tasks.

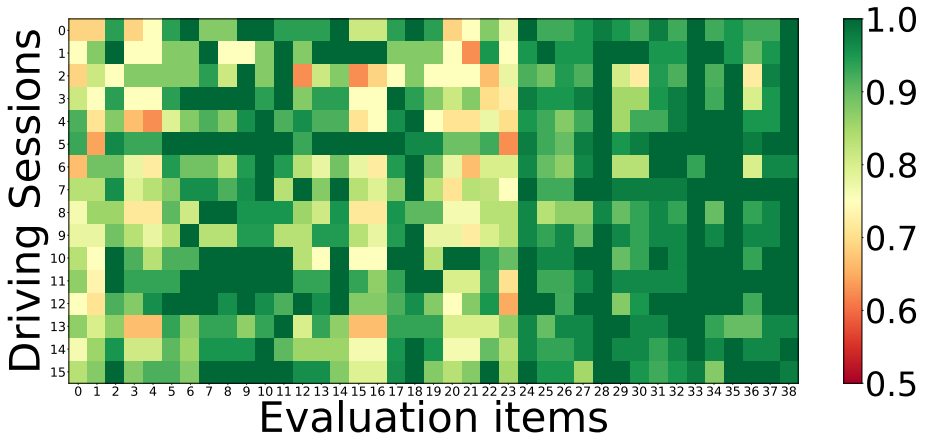

Figure A.12: Expert agreement score in the Town Driving lesson.
The horizontal axis (N_{item}) corresponds to all evaluation items relevant to the Town Driving lesson. The items are numbered for brevity, the mapping can be seen in the supplementary material.
The vertical axis corresponds to distinct driving sessions (each with a different driver).
Examples of evaluation items with low agreement score across the driving sessions are "Observing a roundabout upon approach" (N_{item}=0), "Rearview mirror: approaching a roundabout" (N_{item}=1), "Side mirror: approaching a roundabout" (N_{item}=3), "Blind spot: approaching a roundabout" (N_{item}=4), "Side mirror inside a roundabout" (N_{item}=15), "Blind spot: inside a roundabout" (N_{item}=16) and "Rearview mirror: approaching an intersection" (N_{item}=23).

Appendix B. Expert agreement and evaluation instance difficulty

One interesting point to explore is whether the difficulty of a maneuver as perceived by the experts affects the experts' agreement score distribution. An evaluation instance refers to a specific application of an evaluation item, used

to assess a learner's performance in a distinct situation during a driving session. An intuitive hypothesis would be that the more difficult an evaluation instance looks to the experts, the harder it is for them to agree on how to assess it. It is also reasonable to expect that evaluation instances, where there is no clear agreement on the difficulty, will also have a lower overall agreement score.

In addition to the initial hypothesis, examining this relationship could help identify certain blind spots or biases in experts' assessment methods. If experts consistently disagree on scores for particular maneuvers perceived as difficult, it may indicate that the criteria they're using to evaluate those maneuvers are ambiguous or subject to individual interpretation and will benefit the most from a standardized AI-based assessment system.

Furthermore, this analysis could help improve the training and evaluation processes. If certain maneuvers consistently present difficulty for both students and evaluators, these maneuvers might require a more thorough teaching approach or clearer assessment guidelines. These improvements could lead to a better understanding of the learning curve.

In this analysis, we use the Python SciPy (Jones et al. (2001–)) implementation of the A-D k-sample test to distinguish between distributions of the expert agreement score associated with different difficulty consensus classes.

The Anderson-Darling (A-D) k-sample test is a non-parametric statistical methodology designed to examine whether multiple independent samples are drawn from the same distribution. An extension of the traditional Anderson-Darling test, this variant provides an edge in identifying deviations in distributions defined by a finite number of data points (Scholz & Stephens (1987)).

Kullback-Leibler (K-L) divergence, or relative entropy, is a measure of the dissimilarity between two probability distributions (Csiszar (1975)). However, unlike the A-D k-sample test, the K-L divergence is not a statistical test and is not specifically designed to handle finite, empirical data sets. It operates asymmetrically, and its output can become infinite if a value with non-zero probability in the expected distribution has zero probability in the observed distribution. Thus, K-L divergence may not be optimal for comparing empirical

distributions derived from finite samples.

In contrast, the Kolmogorov-Smirnov (K-S) test is another non-parametric tool used for assessing whether two samples are drawn from the same distribution (Feller (1948)). Despite being a powerful tool for analyzing arbitrary distributions, the K-S test is less sensitive to deviations in the tails of the distributions, which may lead to overlooked discrepancies in these critical regions.

The interpretation of the A-D k-sample test statistics value is not straightforward. However, these can be mapped onto the p-values which can be interpreted as the probability that all samples have been drawn from the same distribution by chance.

It is important to note that the number of evaluation instances N (which is the number of situations multiplied by the number of evaluation items per situation) deemed to be of Expert difficulty is relatively low ($N_{CD} = 90$, $N_{TD} = 36$, $N_O = 44$) in all three lessons (CD = city driving lesson, TD = town driving lesson and O = overtakes lesson), and there exists a large imbalance between the rest of the classes. In the city and town driving lessons, moderately difficult ($N_{CD} = 1980$, $N_{TD} = 3782$) evaluation instances vastly outnumber Beginner ($N_{CD} = 83$, $N_{TD} = 258$), Mid-Moderate ($N_{CD} = 240$, $N_{TD} = 574$), and Mid-Expert levels ($N_{CD} = 540$, $N_{TD} = 108$) evaluation instances. In contrast, in the overtakes lesson, the Beginner-level evaluation instances ($N_O = 1133$) dominate the Mid-Moderate ($N_O = 316$) and Moderate ($N_O = 509$). The Mid-Expert level has no samples in this lesson.

The town driving lesson (Table B.8) has the most distinguishable difficulty classes with several distribution pairs achieving p=value below 0.05. The city driving (Table B.9) and overtakes (Table B.10) lessons have fewer distinguishable classes. Even in the case of apparent statistical significance of the difference between the distributions, the accuracy of the test is expected to be lower due to the large imbalance between the number of samples. Therefore, we conclude that there is no good basis to claim that the expert agreement score has a large correlation with the perceived difficulty, and instead depends mostly on the nature of the evaluation item, as suggested in Figure A.12.

	A-D k-sample test statistic	p-value
Beginner/Mid-Moderate	2.45	0.03
Beginner/Moderate	3.50	0.01
Beginner/Mid-Expert	6.30	< 0.01
Beginner/Expert	2.89	0.02
Mid-Moderate/Moderate	0.61	0.19
Mid-Moderate/Mid-Expert	4.19	0.01
Mid-Moderate/Expert	1.70	0.06
Moderate/Mid-Expert	2.75	0.02
Moderate/Expert	1.01	0.13
Mid-Expert/Expert	-1.04	Large

Table B.8: Town driving lesson: difference between agreement score distributions associated with the difficulty consensus: Beginner ($N = 258$), Mid-Moderate ($N = 574$), Moderate ($N = 3782$), Mid-Expert ($N = 108$), Expert ($N = 36$).
The calculation of the p-value is capped at 0.25, so anything above this value can be interpreted as the absence of meaningful difference. We denote such values as "Large".

	A-D k-sample test statistic	p-value
Beginner-Mid-Moderate	1.09	0.12
Beginner-Moderate	1.26	0.10
Beginner-Mid-Expert	3.49	0.01
Beginner-Expert	-0.97	Large
Mid-Moderate-Moderate	-1.30	Large
Mid-Moderate-Mid-Expert	-0.18	Large
Mid-Moderate-Expert	-0.41	Large
Moderate-Mid-Expert	2.62	0.03
Moderate-Expert	-0.35	Large
Mid-Expert-Expert	1.44	0.08

Table B.9: City driving lesson: the difference between agreement score distributions associated with the difficulty consensus: Beginner ($N = 83$), Mid-Moderate ($N = 240$), Moderate ($N = 1980$), Mid-Expert ($N = 540$), Expert ($N = 90$).
The calculation of the p-value is capped at 0.25, so anything above this value can be interpreted as the absence of meaningful difference. We denote such values as "Large".

	A-D k-sample test statistic	p-value
Beginner-Mid-Moderate	2.58	0.03
Beginner-Moderate	0.92	0.14
Beginner-Expert	0.28	Large
Mid-Moderate-Moderate	-0.25	Large
Mid-Moderate-Expert	-0.92	Large
Moderate-Expert	0.13	Large

Table B.10: Overtakes lesson: the difference between agreement score distributions associated with the difficulty consensus: Beginner ($N = 1133$), Mid-Moderate ($N = 316$), Moderate ($N = 509$), Mid-Expert ($N = 0$), Expert ($N = 44$).
The calculation of the p-value is capped at 0.25, so anything above this value can be interpreted as the absence of meaningful difference. We denote such values as "Large".

References

Allen, R. W., Park, G. D., Cook, M. L., & Fiorentino, D. (2007). The effect of driving simulator fidelity on training effectiveness. *DSC 2007 North America*, .

Arroyo, E., Sullivan, S., & Selker, T. (2006). Carcoach: a polite and effective driving coach. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 357–362).

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., & De Souza, A. F. (2021). Self-driving cars: A survey. *Expert Systems with Applications*, *165*, 113816. URL: https://www.sciencedirect.com/science/article/pii/S095741742030628X. doi:https://doi.org/10.1016/j.eswa.2020.113816.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*.

Bao, Z., Hossain, S., Lang, H., & Lin, X. (2022). High-definition map generation technologies for autonomous driving. arXiv:2206.05400.

Baumgartner, N., Gottesheim, W., Mitsch, S., Retschitzegger, W., & Schwinger, W. (2010). Beaware!—situation awareness, the ontology-driven way. *Data & Knowledge Engineering*, *69*, 1181–1193. URL: https://www.sciencedirect.com/science/article/pii/S0169023X1000090X. doi:{https://doi.org/10.1016/j.datak.2010.07.008}. Special issue on contribution of ontologies in designing advanced information systems.

Baumgartner, N., Mitsch, S., Müller, A., Retschitzegger, W., Salfinger, A., & Schwinger, W. (2014). A tour of beaware – a situation awareness framework for control centers. *Information Fusion*, *20*, 155–173. URL: https:

//www.sciencedirect.com/science/article/pii/S156625351400013X. doi:https://doi.org/10.1016/j.inffus.2014.01.008.

Bjørnland, J. F. R., Gedde, Y., Rehm, J., Reshodko, I., Børresen, & Gundersen, O. E. (2024). A virtual driving instructor that generates personalized driving lessons based on student skill level. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, *47*, 139–159.

Buechel, M., Hinz, G., Ruehl, F., Schroth, H., Gyoeri, C., & Knoll, A. (2017). Ontology-based traffic scene modeling, traffic regulations dependent situational awareness and decision-making for automated vehicles. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1471–1476). doi:10.1109/IVS.2017.7995917.

Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4693–4700). IEEE.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, *3*.

De Winter, J., van Leeuwen, P. M., Happee, R. et al. (2012). Advantages and disadvantages of driving simulators: A discussion. In *Proceedings of measuring behavior* (pp. 28–31). volume 2012.

Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. *IEEE Access*, *6*, 28573–28593. doi:10.1109/ACCESS.2018.2831228.

Ekanayake, H., Backlund, P., Ziemke, T., Ramberg, R., & Hewagamage, K. (2011). Assessing performance competence in training games. In *Affective Computing and Intelligent Interaction: Fourth International Conference,*

ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II (pp. 518–527). Springer.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64. doi:10.1518/001872095779049543.

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, *12*, 213–253.

European Commission (2011). Roadmap to a single european transport area – towards a competitive and resource efficient transport system. URL: https://www.eea.europa.eu/policy-documents/roadmap-to-a-single-european.

Feller, W. (1948). On the kolmogorov-smirnov limit theorems for empirical distributions. *Annals of Meth. Stat.*, *19*.

Gicquel, L., Ordonneau, P., Blot, E., Toillon, C., Ingrand, P., & Romo, L. (2017). Description of various factors contributing to traffic accidents in youth and measures proposed to alleviate recurrence. *Frontiers in psychiatry*, *8*, 94.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, *38*, 50–57. doi:10.1609/aimag.v38i3.2741. arXiv:1606.08813.

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*, 44–58. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2850. doi:10.1609/aimag.v40i2.2850.

Gutierrez, G., Iglesias, J. A., Ordoñez, F. J., Ledezma, A., & Sanchis, A. (2014). Agent-based framework for advanced driver assistance systems in urban environments. In *17th International Conference on Information Fusion (FUSION)* (pp. 1–8).

Halilaj, L., Dindorkar, I., Lüttin, J., & Rothermel, S. (2021). A knowledge graph-based approach for situation comprehension in driving scenarios. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18* (pp. 699–716). Springer.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*, 1263–1284.

Hestness, J., Ardalani, N., & Diamos, G. (2019). Beyond human-level accuracy: Computational challenges in deep learning. In *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming* (pp. 1–14).

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, *54*. URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

Huang, J., Saleh, S., & Liu, Y. (2021). *Academic Journal of Interdisciplinary Studies*, *10*.

Hülsen, M., Zollner, J. M., Haeberlen, N., & Weiss, C. (2011a). Asynchronous real-time framework for knowledge-based intersection assistance. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (pp. 1680–1685). doi:10.1109/ITSC.2011.6082810.

Hülsen, M., Zöllner, J. M., & Weiss, C. (2011b). Traffic intersection situation description ontology for advanced driver assistance. In *2011 IEEE Intelligent Vehicles Symposium (IV)* (pp. 993–999). doi:10.1109/IVS.2011.5940415.

Jones, E., Oliphant, T., Peterson, P. et al. (2001–). SciPy: Open source scientific tools for Python. URL: http://www.scipy.org/.

Kim, J., Rohrbach, A., Akata, Z., Moon, S., Misu, T., Chen, Y.-T., Darrell, T., & Canny, J. (). Towards explainable and advisable model for self-driving cars. *Applied AI Letters*, (p. e56).

Kolasinski, E. M. (1995). Simulator sickness in virtual environments, .

Matheus, C. J., Kokar, M. M., & Baclawski, K. (2003). A core ontology for situation awareness. In *Proceedings of the 6th International Conference on Information Fusion, FUSION 2003* (pp. 545–552). volume 1. doi:10.1109/ICIF.2003.177494.

Montemerlo, M., Becker, J., Bhat, S., Dahlkamp, H., Dolgov, D., Ettinger, S., Haehnel, D., Hilden, T., Hoffmann, G., Huhnke, B. et al. (2008). Junior: The stanford entry in the urban challenge. *Journal of field Robotics*, *25*, 569–597.

Ndousse, K. K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent social learning via multi-agent reinforcement learning. In *International Conference on Machine Learning* (pp. 7991–8004). PMLR.

Noor-A-Rahim, M., Liu, Z., Lee, H., Khyam, M. O., He, J., Pesch, D., Moessner, K., Saad, W., & Poor, H. V. (2022). 6g for vehicle-to-everything (v2x) communications: Enabling technologies, challenges, and opportunities. *Proceedings of the IEEE*, *110*, 712–734. doi:10.1109/JPROC.2022.3173031.

Nwana, H. S., Lee, L. C., & Jennings, N. R. (1996). Coordination in software agent systems. *British Telecom Technical Journal*, *14*, 79–88.

Onishi, T., Motoyoshi, T., Suga, Y., Mori, H., & Ogata, T. (2019). End-to-end learning method for self-driving cars with trajectory recovery using a path-following function. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

O'Neill, B. (2020). Driver education: how effective? *International journal of injury control and safety promotion*, *27*, 61–68.

Philippe, S., Souchet, A. D., Lameras, P., Petridis, P., Caporal, J., Coldeboeuf, G., & Duzan, H. (2020). Multimodal teaching, learning and training in virtual reality: a review and case study. *Virtual Reality & Intelligent Hardware*, *2*, 421–442. URL: https://www.sciencedirect.com/science/article/pii/S2096579620300711. doi:https://doi.org/10.1016/j.vrih.2020.07.008.

Platho, M., Groß, H.-M., & Eggert, J. (2012). Traffic situation assessment by recognizing interrelated road users. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (pp. 1339–1344). doi:10.1109/ITSC.2012.6338756.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*, 37–63.

Regele, R. (2008). Using ontology-based traffic models for more efficient decision making of autonomous vehicles. In *Fourth International Conference on Autonomic and Autonomous Systems (ICAS'08)* (pp. 94–99). doi:10.1109/ICAS.2008.10.

Rehm, J., Gundersen, O. E., Bach, K., & Reshodko, I. (2021). Utilizing driving context to increase the annotation efficiency of imbalanced gaze image data. In *Proceedings of NeurIPS DCAI Workshop*.

Rehm, J., Reshodko, I., Børresen, S. Z., & Gundersen, O. E. (2024). The virtual driving instructor: Multi-agent system collaborating via knowledge graph for scalable driver education. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*.

Ropelato, S., Zünd, F., Magnenat, S., Menozzi, M., & Sumner, R. (2018). Adaptive tutoring on a virtual reality driving simulator. *International SERIES*

on information systems and management in creative emedia (CreMedia), *2017*, 12–17.

Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 627–635). JMLR Workshop and Conference Proceedings.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. doi:10.1038/s42256-019-0048-x. arXiv:1811.10154.

Salfinger, A., Retschitzegger, W., & Schwinger, W. (2014). Staying aware in an evolving world — specifying and tracking evolving situations. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 195–201). doi:10.1109/CogSIMA.2014.6816562.

Sandberg, M. K., Rehm, J., Mnoucek, M., Reshodko, I., & Gundersen, O. E. (2020). Explaining traffic situations–architecture of a virtual driving instructor. In *International Conference on Intelligent Tutoring Systems* (pp. 115–124). Springer.

Scholz, F. W., & Stephens, M. A. (1987). K-sample anderson-darling tests. *Journal of the American Statistical Association*, *82*.

Sharon, T., Selker, T., Wagner, L., & Frank, A. (2005). Carcoach: a generalized layered architecture for educational car systems. In *IEEE International Conference on Software - Science, Technology Engineering (SwSTE'05)* (pp. 13–22). doi:10.1109/SWSTE.2005.9.

Shu-Hsien Liao (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Systems with Applications*, *28*, 93–103. URL: https://www.sciencedirect.com/science/article/

pii/S0957417404000934. doi:https://doi.org/10.1016/j.eswa.2004.08.003.

Sipele, O., Zamora, V., Ledezma, A., & Sanchis, A. (2018). Advanced Driver's Alarms System through Multi-agent Paradigm. In *2018 3rd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2018* (pp. 269–275). doi:10.1109/ICITE.2018.8492600.

Suh, J., Chae, H., & Yi, K. (2018). Stochastic model-predictive control for lane change decision of automated driving vehicles. *IEEE Transactions on Vehicular Technology*, *67*, 4771–4782.

Suryawanshi, Y., Qiu, H., Ayara, A., & Glimm, B. (2019). An ontological model for map data in automotive systems. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 140–147). doi:10.1109/AIKE.2019.00034.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*, 261–292.

United Nations (1968). Convention on road traffic.

Unity Technologies (2022). Unity. https://unity.com/. Version 2021.3.4.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, *11*, 93–136.

Vacek, S., Gindele, T., Zollner, J. M., & Dillmann, R. (2007). Situation classification for cognitive automobiles using case-based reasoning. In *2007 IEEE Intelligent Vehicles Symposium* (pp. 704–709). doi:10.1109/IVS.2007.4290199.

Vogel, S., & Schwabe, L. (2016). Learning and memory under stress: implications for the classroom. *npj Science of Learning*, *1*, 1–10.

Weevers, I., Kuipers, J., Brugman, A. O., Zwiers, J., van Dijk, E. M., & Nijholt, A. (2003). The virtual driving instructor creating awareness in a multiagent system. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 596–602). volume 2671. doi:10.1007/3-540-44886-1_56.

de Winter, J. C., De Groot, S., Mulder, M., Wieringa, P., Dankelman, J., & Mulder, J. (2009). Relationships between driving simulator performance and driving test results. *Ergonomics*, *52*, 137–153.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John wiley & sons.

World Health Organization (2018). Global status report on road safety. URL: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

Zamora, V., Sipele, O., Ledezma, A., & Sanchis, A. (2017). Intelligent agents for supporting driving tasks: An ontology-based alarms system. In *VEHITS 2017 - Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems* (pp. 165–172). doi:10.5220/0006247601650172.