Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

Christian Müller, Yufeng Liu, Yi Wang, Kevin Baum, Rudraksh Bhawalkar, Hendrik Purwins

Responsible AI in the automotive industry:

Techniques and use cases

Executive Summary

Artificial Intelligence (AI) is increasingly integrated into our lives, with advanced systems like Generative AI and AI-driven services becoming more efficient, user-friendly, and feature-rich. As the demand for AI capabilities such as autonomous reasoning and complex image generation grows, so does the need for extensive training data and system complexity. The human ability to monitor and oversee the system's functionality becomes limited. Depending on the kind of training data or the coding practices, the trustworthiness of AI system could be challenged.

To ensure that AI systems are responsible in nature and can be trusted, Accenture and DFKI have joined forces to examine the implications of Responsible AI in the automotive industry. This paper thus outlines **Accenture's** principles alongside their application in enhancing the safety of Autonomous Driving, particularly emphasizing the need for organizations to achieve higher **AI maturity** to realize the full potential of AI. To support this advancement, we delve deeper into **neuro-explicit AI** as an example of differential AI on the algorithmic level, which has the potential to address some of the ethical and safety challenges in the autonomous driving industry. To present these approaches in a real-world setting, we identified **7 use cases from the automotive industry** for which we discuss how Responsible AI can be achieved. We conclude that the need to take action in order to ensure robust and ethically sound systems is more pressing than ever – an objective to which neuro-explicit AI methods can vastly contribute – and are dedicated to contributing to this with deep technological understanding and value commitment.

Contents

Introduction

The past decade has been marked by the commercial advancement and exploitation of Deep Learning, which has transformed AI from a niche science into a socially relevant "mega technology." However, despite all the unprecedented successes in autonomous driving (AD) and many other areas, including large language models, the limitations of pure deep learning as it has been applied so far have become all too apparent. Against this background, there is much to suggest that the coming decade could be a decade of an altogether new kind of artificial intelligence, which can be called Trustworthy AI (TAI) or Responsible AI (RAI).

To understand why a new era of AI needs to dawn right now, we go back to 2006/2007, when scientific competitions in autonomous driving and in almost all relevant areas of image and video analysis, as well as text and language processing, began to see quantum leaps in the performance of AI systems. A new technology, which had become possible primarily due to new types of hardware (graphics processors made usable for general computing operations) and the availability of large amounts of data (from social media and similar sources), was to become known to a broad public as Deep Learning in the years that followed. Deep Learning has played an overwhelming part in the commercial success of AI, so much so that it has since been traded as a solution to almost any problem for which sufficient data is available.

Then, as problems emerged such as the lack of internal representation of meaning (explainability issue), susceptibility to changes in the input signal (robustness issue), lack of transferability to cases not covered by the data (generalization issue), and last but not least, the big data hunger itself (sustainability issue due to large energy consumption in training), companies began to lower their expectations. According to the forecasts made a few years ago, autonomous driving should be a widespread reality on the streets today. Instead, it emerges in few selected hot spots, still being rather experimental as news from San Francisco suggest, where robotaxis apparently disturb traffic and even block firefighting operations (Sims, 2023). This failure to meet expectations is accompanied by accidents resulting in damage to persons (Shepardson, 2023).

Accenture Tech Vision Research (2022) shows that only 35% of global consumers trust how organizations are implementing AI, and 77% feel that these organizations must be held accountable for their misuse of the technology. As businesses increasingly rely on AI to drive their operations, they must be mindful of new regulations and take steps to ensure compliance. This is where Responsible AI comes in, providing a framework for organizations to use AI in a manner that is ethical and transparent, thereby promoting trust and accountability.

Responsible AI

Responsible AI is an approach to designing, building, and deploying AI systems that prioritizes safety and ensures that these systems are developed and used in an ethical, transparent, and fair manner.

This involves a number of different considerations, including the potential impact of AI systems on people and society, the need to ensure that AI systems are built with a focus on privacy and data protection, and the importance of ensuring that these systems are not used in a way that perpetuates bias or discrimination. In practice, Responsible AI covers a number of different approaches and techniques that are designed to promote these goals. For example, Responsible AI may involve developing AI models in a way that is transparent, so that customers and other stakeholders can understand how the model was created. It may also include using techniques like explainability and interpretability to help people understand how the model is arriving at its predictions, and to ensure that these predictions are based on valid, non-discriminatory data.

In addition, Responsible AI also encompasses a focus on ethical considerations, such as ensuring that AI models are not used in a way that violates individual privacy or undermines human autonomy. In addition, AI models should only involve well-considered, justified biases, for example to balance unavoidable trade-offs or to equalize existing injustices (Wachter et al. 2021). This may include developing AI systems that are designed to protect sensitive data, or that are designed to provide individuals with greater control over how their data is used. Ultimately, the goal of Responsible AI is to build AI models that can be justifiably trusted by both customers and society more broadly. By ensuring the ethical and transparent development and use of AI, companies can build trust in their models and scale them with confidence, knowing that they will not have a negative impact on individuals or society.

Responsible AI can be regarded at different levels, including the organization level, system level, and algorithmic level. At the organizational level, Responsible AI involves establishing ethical guidelines and best practices for the development and deployment of AI technologies within an organization. At the system level, Responsible AI involves ensuring that AI systems are safe, transparent, explainable, and fair. Finally, at the algorithmic level, Responsible AI encompasses developing algorithms that are accurate, robust, and, more generally, trustworthy. In this article, we will focus on Responsible AI on the algorithmic level. Specifically, we will describe the emerging class of algorithms called ‘neuro-explicit AI’ as a method to solve some of the problems of pure deep learning. Neuro-explicit AI combines the strengths of neural networks and symbolic reasoning to enable more robust and explainable AI models, making them more suitable for certain applications where trust and interpretability are critical.

Critics of neuro-explicit AI argue that scaling deep learning is the most effective approach in the current era of large language models and their recent successes. However, while this technique has delivered exceptional outcomes in natural language processing, where large pretrained transformers have achieved groundbreaking performance in various tasks, it remains debatable whether it is the optimal approach for all AI applications. Especially for the automotive industry, where physics plays a significant role, neuro-explicit AI is a more promising approach as it allows to model the underlying physics of the system explicitly and, thus, can deliver superior accuracy, better generalization, and more interpretability than traditional deep learning approaches – with less data. However, it's important to note that neuro-explicit AI is just one example of how differentiation can be achieved on the algorithmic level. There are other key capabilities required to achieve high performance for customers, shareholders, and employees. This includes foundational capabilities like data platforms, architecture, and governance. AI Achievers, the group that has advanced their AI maturity enough to achieve superior growth and business transformation, are those that have mastered a set of key capabilities in the right combination. Hence, while neuro explicit AI is a promising development in the field of AI, achieving responsible and mature AI requires a holistic approach that goes beyond algorithmic differentiation.

Accenture's Responsible AI Principles

A number of organizations and experts have developed guidelines for Responsible AI. But even the the most widely recognized principles, which were developed by **the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG),** need refinement.

These principles provide a framework for the development and deployment of AI systems that are human-centric, transparent, and accountable, and that respect fundamental rights and values. The seven key requirements include human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability (AI HLEG, 2020).

The seven key requirements of Responsible AI developed by the AI HLEG provide a valuable starting point for ensuring that AI is developed and used in a responsible and trustworthy manner. For industrial context, Accenture has established Responsible AI principles (Accenture, 2024), focusing on these 7 areas: human by design, fairness, transparency, explainability and accuracy, safety, accountability, compliance, data privacy & cybersecurity, sustainability.

01 Human by design

Human by design is a principle of Responsible AI that emphasizes understanding and managing AI's impact on humans. This principle focuses on creating technology that enhances human capabilities, productivity, and creativity while ensuring that its effects on people are identified and responsibly managed. As technology advances, it's crucial to ensure AI systems are designed to be intuitive, user-friendly and capable of understanding human intent. This involves developing AI-powered tools that mimic human reasoning, provide personalized interactions, and integrate seamlessly into daily life. By making technology more humanity-centered, enterprises can unlock new levels of human potential, productivity, and creativity, leading to innovative solutions that enhance both individual and organizational capabilities. The goal is to reduce friction between humans and technology, creating tools that not only augment human abilities but also reflect our natural ways of interacting with the world while respecting fundamental human rights and minimizing negative impact on humans.

02 Fairness

The principle of fairness is a key aspect of Responsible AI. It involves ensuring that AI systems are designed and deployed in a manner that produces reasonable and just outcomes for all individuals affected by the algorithm, regardless of their gender, ethnicity, or other demographic factors. The goal is to eliminate any discriminatory or unfair treatment that might arise from the use of the algorithm. However, achieving fairness in AI is not always straightforward. While theoretically bias-free algorithms may exist, they can still be applied in a manner that is unreasonable and unjust. Similarly, biased algorithms may be employed in a way that produces fair results. Thus, achieving fairness requires a contextual approach that takes into account the specific application context in which the AI system is being used. This is a challenging task that requires careful consideration of a wide range of factors, including the data used to train the algorithm, the design of the algorithm itself, and the social and ethical implications of the system's use. Additionally, achieving fairness may involve addressing accessibility, incorporating universal design principles, and consulting affected stakeholders (AI HLEG, 2019).

03 Transparency, Explainability and Accuracy

Building trust is a key element of Responsible AI. One way to achieve this is to disclose AI use where appropriate and ensure all can understand and evaluate AI outputs and decision-making processes. Transparency must be maintained throughout the entire lifecycle of products and services that use AI. Transparency may also require AI to be explainable, traceable. Responsible AI requires a certain level of interpretability so that varying degrees of context-dependent certainty can be taken into consideration. Explanations go one step further, rendering the inner workings of the AI models to humans who can review the system before it is approved, control it while it is working, or investigate it after an incident. Explanations can also be rendered to programmers or data scientists during the design phase of the model. While black box models like neural networks are difficult to interpret, white box or grey box models are either explainable by design or easier to make explainable. In this article, we use the example of neuro-explicit models, which are among the alternatives to black box models that can be made more explainable.

Data and model selection are critical for ensuring the accuracy, fairness and efficiency of an AI system. When selecting the right data for a model, it is important to consider whether the data adequately represents the target population to avoid data biases that may lead to unfair outcomes. Additionally, selecting the right model requires understanding the problem domain, the model's limitations, and the intended use of the model. A model that works well in one context may not be suitable for another. For instance, a model trained on data from one country may not be effective when used in another country due to differences in cultural, linguistic, or economic factors. Therefore, careful consideration of context and appropriate controls must be applied when selecting data and models. Furthermore, in times of generative models, care must be taken to ensure that questions aimed at facts are answered in a factual manner and not based on hallucinations and confabulations. The challenge here, however, is to avoid answering questions that cannot be considered to have a universally accepted answer and to avoid answering them with an overly speculative worldview and thus intervening in the formation of public opinion, for example. It is also important to avoid paternalistic design decisions in connection with creative tasks without losing sight of fairness. All of this requires transparency to ensure that the data and models chosen are sound and appropriate for the problem at hand, and to allow for reasoned contestation in the spirit of a feedback culture designed for iterative improvement.

These perspectives are crucial not only for AI practitioners but also for users, who should be informed when they are interacting with an AI system and made aware of its capabilities and limitations (AI HLEG, 2019).

04 Safety

AI can do harm to human health, well-being, life and property when misused. In the times of Generative AI, AI can also amplify the risks such as production of harmful content and misinformation. It is essential to evaluate potential safety concerns and take action to mitigate harm when deploying AI.

To ensure safety, one important aspect to consider is robustness, which refers to the ability of an AI system to perform accurately and reliably under all conditions for which it was designed. This includes withstanding different types of perturbations in input data, whether due to natural causes such as changes in weather or time of day, or due to adversarial attacks. The importance of robustness is related to the concept of generalizability, which refers to the ability of an AI system to perform well on data that was not part of its training set. Achieving robustness, and thereby safety, is crucial to ensure that AI systems remain reliable in the face of unexpected or changing conditions, covering all situations within the operational design domain (ODD) of the AI system.

05 Accountability

To ensure Responsible AI, it is essential to establish cross-domain governance structures that promote transparency and accountability in the development and deployment of AI technologies. Such structures help to identify clear roles, policies, expectations, and responsibilities for all stakeholders, building internal confidence and trust in AI technologies. The question of accountability is central to Responsible AI, and it is essential to understand who is responsible for the decisions made by AI systems. The level of human involvement in producing outcomes must also be well-defined, ensuring that humans remain in control of AI systems and can intervene when necessary. Governance structures should also include policies and procedures for auditing, monitoring, and addressing issues that may arise during the lifecycle of AI products and services, ensuring human oversight. Tensions may arise between different principles, even though they are closely intertwined. It is advisable to evaluate, reason, and document inevitable trade-offs. Additionally, the appropriateness of decisions should be continually reviewed and adapted as needed (AI HLEG, 2019). By establishing such structures, organizations can ensure that their AI systems are transparent, accountable, and trustworthy, enabling them to make more informed decisions while building public trust in AI.

06 Compliance / Data Privacy / Cybersecurity

Ensuring compliance, data privacy and cybersecurity is an essential aspect of Responsible AI, especially when processing personal or sensitive data. A privacy and security-first approach must be implemented on the system and process levels to make sure that the AI system complies with legal and regulatory requirements. It is important to take appropriate measures to secure sensitive data to avoid unauthorized access, theft, or disclosure. Additionally, it is crucial to make the AI algorithm robust against attacks. Attacks can occur in various forms, such as membership inference attacks, in which an attacker can reconstruct sensitive personal data directly from the model by exploiting certain characteristics in the inner workings of the algorithm. To mitigate such attacks, privacy-preserving techniques such as differential privacy, secure multiparty computation, and homomorphic encryption can be used. A privacy and security-first approach to AI is crucial to ensure that personal data is protected and used only for its intended purposes while building trust among stakeholders.

07 Sustainability

Sustainability is an essential aspect of Responsible AI. It includes designing and implementing AI systems that meet the needs of current generations without compromising the ability of future generations to meet their own needs. Sustainability encompasses the economic, environmental, and social dimensions of AI. AI systems should be designed with the goal of reducing energy consumption and minimizing their environmental impact while still delivering efficient and effective results. One aspect of sustainable AI is frugal AI, which focuses on reducing the overall energy consumption of AI systems while still maintaining their performance. This encompasses optimizing the AI lifecycle from training to deployment, including energy-efficient hardware design, and developing new algorithms and optimization methods to reduce energy consumption. By prioritizing sustainability in the design and deployment of AI systems, we can help to minimize the environmental impact and ensure that resources are available for future generations to use and benefit from.

These principles are not mutually exclusive but rather intersect and complement each other in various ways. For example, consider the overlap between the principles of security and safety, particularly in the context of adversarial machine learning. When an AI system is designed to be robust against naturally occurring perturbations, such as changes in weather or time of day, this is primarily a matter of safety, ensuring that the system operates reliably under diverse conditions. However, if the system must also defend against deliberate attacks by malicious actors, such as data poisoning or adversarial examples, this falls under security. Despite whether there is an actual adversarial attack or not, the mitigation methods for enhancing robustness against these perturbations are often similar, if not the same. Both scenarios require developing resilient algorithms that can maintain their performance in the face of unexpected or maliciously crafted input variations. This intersectionality extends to other principles as well; for instance, ensuring transparency and explainability can enhance accountability, and prioritizing fairness can support human-centric design by ensuring equitable outcomes for all users. Recognizing these overlaps helps in creating a holistic approach to Responsible AI, where multiple principles work together to build trustworthy and effective AI systems.

AI Maturity and Responsible AI

Most organizations are barely scratching the surface when it comes to making the most of AI's full potential and their own investments. Only 12% of firms have advanced their AI maturity enough to achieve superior growth and business transformation, according to Accenture's research (2022). These "AI Achievers" can attribute nearly 30% of their total revenue to AI, on average. The share of AI Achievers will increase rapidly and significantly, more than doubling from the current 12% to 27% by 2024 (Accenture, 2022).

AI maturity is defined as the degree to which organizations have mastered AI-related capabilities in the right combination to achieve high performance for customers, shareholders, and employees.

AI maturity is related to Responsible AI in several ways. AI maturity measures an organization's level of proficiency in using AI to drive growth, innovation, and customer experience. To achieve AI maturity, companies need to prioritize their investments in AI and talent, industrialize AI tools and teams, and use AI responsibly from the start. Responsible AI is crucial in achieving AI maturity, as it involves ensuring that AI systems are designed and deployed in a way that is ethical, transparent, and fair, taking into account their impact on individuals, society, and the environment. Moreover, AI Achievers are more likely to prioritize the responsible use of AI, reducing their greenhouse gas emissions, using natural resources economically, and developing strong relationships with customers. Therefore, achieving Responsible AI practices is an essential component of advancing AI maturity and driving growth, innovation, and customer experience.

One way to achieve "differentiation AI” capabilities is to focus on the algorithmic level. Specifically, we will focus on one of the latest trends in AI called neuro-explicit AI, which involves the integration of deep learning and explicit human knowledge. We will explore how this approach can be used to achieve differentiation on the algorithmic level and support Responsible AI practices. However, this is only one aspect of the comprehensive AI-maturity framework developed by Accenture. In addition to algorithmic capabilities, the framework also includes other critical areas such as data and AI platforms, organizational strategy, talent, and culture. Achieving high performance in all these principles is key to unlocking AI's full potential and achieving sustainable growth.

Neuro-explicit AI

Deep neural models learn from data and are highly adaptable. In contrast, explicit models capture knowledge about a task or domain using symbolic representations or differential equations, for example, in a way that is more interpretable and can be more data efficient. Neuro-explicit models combine neural and explicit elements and take on the complementary strengths of both.

So far, we have explored Responsible AI from the perspective of its overarching goals, such as safety, security, transparency, accountability, and fairness. In this section, we will focus on a specific approach to achieving Responsible AI on the algorithmic level, namely the use of neuro-explicit methods. These techniques involve incorporating insights from explicit knowledge representations and reasoning into the design and implementation of AI systems, with the aim of creating more explainable and interpretable models that are better aligned with human cognition. In this approach, knowledge is represented using a set of symbols, physical equations, and mathematical formulas that are designed to capture the underlying structure and relationships of the knowledge.

The debate between connectionists and advocates of hybrid AI has been ongoing for decades. Connectionists believe that a neural network's ability to learn from raw data without the need for explicit knowledge representation is the key to achieving the best performing AI. Jan LeCun and Gary Marcus are two prominent figures in the debate over the best approach to building AI systems. While LeCun believes that deep learning and connectionist methods are the key to building truly intelligent machines, Marcus argues that AI should be built using a hybrid approach that combines symbolic reasoning with statistical learning. We argue that neuro-explicit AI is a leap forward towards more advanced and trustworthy AI systems. By incorporating subtle variants such as retro-fitting from explicit knowledge sources, neuro-explicit AI can improve the interpretability of AI models and provide a more comprehensive understanding of the problem domain.

We believe that by using neuro-explicit methods, we can make significant strides towards achieving Responsible AI that is not only effective, but also ethical and trustworthy. First, neuro-symbolic AI can provide more explainable results than purely neural network-based AI systems. Neural networks can be difficult to interpret and understand, which can make it challenging to identify and address potential biases or other issues. By incorporating symbolic reasoning, which uses a more rule-based approach, neuro-explicit AI can provide more transparency into how a decision was made. Second, neuro-explicit AI can help to mitigate bias in AI systems by allowing for the integration of domain-specific knowledge into the decision-making process. This can help to ensure that AI systems are making decisions that align with ethical and social considerations. Finally, neuro-symbolic AI can help to ensure that AI systems are more robust and reliable.

For example, neuro-explicit reinforcement learning can be used to train autonomous agents to perform specific tasks. In general, physics-informed neural networks can be trained using loss or reward functions that incorporate both the experimental data and physical principles. For example, the loss function can be designed to minimize the difference between the predicted and actual deformations, while also penalizing any violations of known physical principles, such as conservation of energy or momentum. Consider a robot that needs to learn how to balance a ball on a platform. A neuro-explicit reinforcement learning algorithm can be used to train the robot to balance the ball using physics-motivated rewards instead of rewards that only take the data into account. The algorithm combines neural networks with symbolic reasoning to enable the robot to learn and reason about the physics of the task.

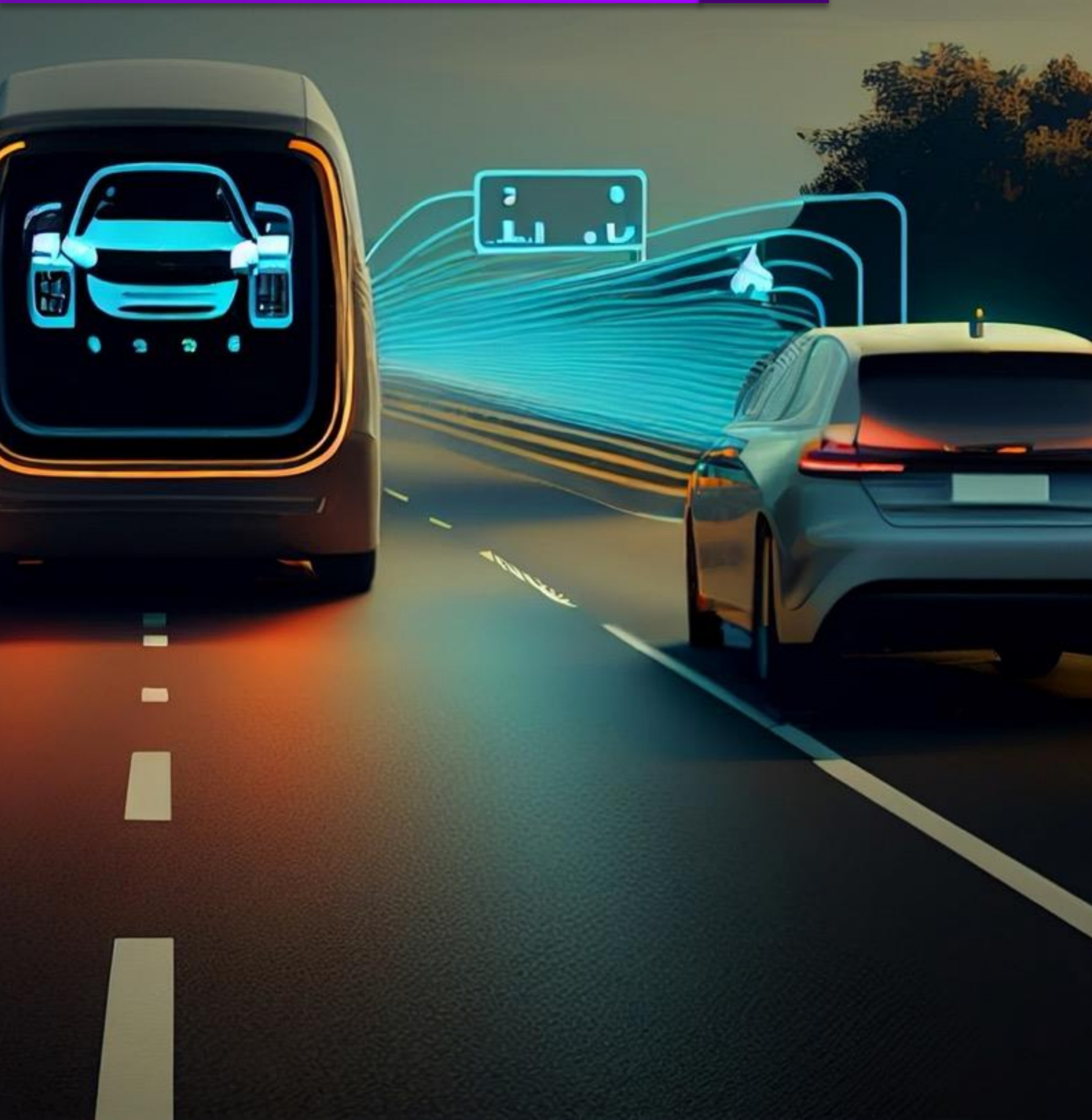
Retrofitting large neural nets from explicit knowledge sources is a technique used in neuro-explicit AI systems to improve the performance and interpretability of large-scale foundation models. The process involves taking a pre-trained neural network and incorporating additional knowledge sources, such as rules, constraints, or ontologies, to refine the network's outputs. The explicit knowledge sources can help the network learn more accurately by constraining the solution space, improving the network's generalization abilities, and providing domain-specific information. By retrofitting explicit knowledge sources to neural nets, hybrid AI systems can achieve higher accuracy, better interpretability, and more robustness than pure deep learning models. This technique can also help AI models comply with regulatory or ethical constraints and allow domain experts to ensure that the model's decisions align with domain-specific rules and constraints.

Opponents of neuro-explicit AI often present a range of counterarguments, focusing on several key issues. One such issue is the difficulty in constructing explicit knowledge representations that can be effectively integrated with neural networks. This challenge stems from the complexity of representing intricate knowledge in a way that can be easily integrated, which in turn may limit the system's performance. Another issue is the potential lack of scalability, as the system's performance may deteriorate when tackling larger and more complex problems that require an increased amount of knowledge and reasoning. The increased complexity of neuro-explicit AI can also make it more challenging to implement and maintain, while also increasing the system's computational requirements, thereby limiting its scalability. Finally, despite promising results in certain applications, neuro-explicit AI has not achieved the same level of success as pure deep learning approaches in many domains, which may be attributable to the aforementioned challenges.

The automotive industry, however, presents a unique case where neuro-explicit AI's benefits are likely to outweigh its drawbacks. In the context of autonomous driving, the influence of physics is significant, and modular architectures based on explicit models of the underlying physical systems can enable more accurate and efficient decision-making. Furthermore, neuro-explicit AI's ability to incorporate domain-specific knowledge and reasoning can mitigate the risk of bias in decision-making, improving the overall safety and performance of autonomous vehicles. Above all, the automotive industry has long been a paragon of modular architectures, consistently demonstrating the advantages of integrating separate, specialized components to create a cohesive and efficient system. This legacy of modular design has laid a strong foundation for the application of neuro-explicit methods, which inherently embrace a similar approach. By utilizing distinct, interconnected modules, neuro-explicit methods allow for a more streamlined and targeted learning process that aligns well with the established modularity within the automotive sector. This modular approach stands in stark contrast to monolithic end-to-end learning systems, which can struggle to adapt to the complex, multi-faceted nature of automotive design and functionality. Therefore, we argue that for the automotive industry, neuro-explicit AI presents a compelling case for its use, and its potential benefits make it a valuable area for further research and development.

To put our position into perspective, it is crucial to recognize that while neuro-explicit AI presents a promising approach for achieving differentiation AI capabilities on the algorithmic level, it should not be viewed as the sole solution for achieving AI maturity and Responsible AI. As previously discussed, AI maturity requires a range of capabilities, including foundational and differentiation AI capabilities, as well as the development of talent and culture. Furthermore, Responsible AI requires organizations to address ethical considerations surrounding data governance, trust, and legality, and to remain attentive to new and emerging regulations. Consequently, a comprehensive approach to AI maturity and Responsible AI must account for these various aspects beyond the algorithmic level.

Responsible AI in the Automotive Industry



Importance of AI for Autonomous Driving

The Automotive Industry uses large capacities in developing and deploying ML/AI technologies throughout its entire value chain: from supply chain resilience to computer vision guided fault detection during production, to stock optimization, advanced driver assistance systems and to autonomous driving.

These technologies have the potential to revolutionize the way we travel and to improve road safety significantly. However, the use of AI in the domain of autonomous driving also raises several concerns, such as the potential for accidents and privacy issues. Therefore, by examining Responsible AI in the context of the automotive industry, we can gain valuable insights into how to create AI systems that are not only safe and secure, but also ethical, transparent, and accountable. This makes the autonomous driving domain an ideal case study for exploring the challenges and opportunities associated with Responsible AI.

One of the most significant concerns with autonomous vehicles is the potential for accidents. In 2018, an Uber self-driving car struck and killed a pedestrian in Arizona, highlighting the need for improved safety features and testing protocols in autonomous vehicles (Levin & Wong, 2018). In 2017, a self-driving shuttle bus in Las Vegas was involved in an accident after it collided with a delivery truck in a parking lot during its first hour of operation (Lee, 2017). There have also been instances where autonomous vehicles misinterpret road signs or markings, leading to unsafe driving conditions. In 2016, a Tesla Model S crashed into a tractor-trailer because the vehicle's Autopilot system failed to recognize the white side of the trailer against a bright sky (Yadron & Tynan, 2016). Adverse weather conditions such as heavy rain, snow, or fog can affect the performance of AI-based autonomous driving systems. For example, there have been reports of autonomous vehicles encountering difficulty during heavy snowfall, as the sensors and cameras can become obstructed or disrupted by the snow. Additionally, fog and other adverse weather conditions can also pose challenges for autonomous vehicles, as visibility is reduced, and the accuracy of sensors can be compromised.

Further, generative AI, on the other hand, could allow to explain the vehicle's complicated and possibly sometimes unexpected decisions to a passenger in that vehicle. This should occur in a way that abstracts from technical details and thus makes it easy to understand. However, it must be ensured that the passenger can actually rely on these explanations and is not lied to – as a New York City chatbot did (Orland, 2024).

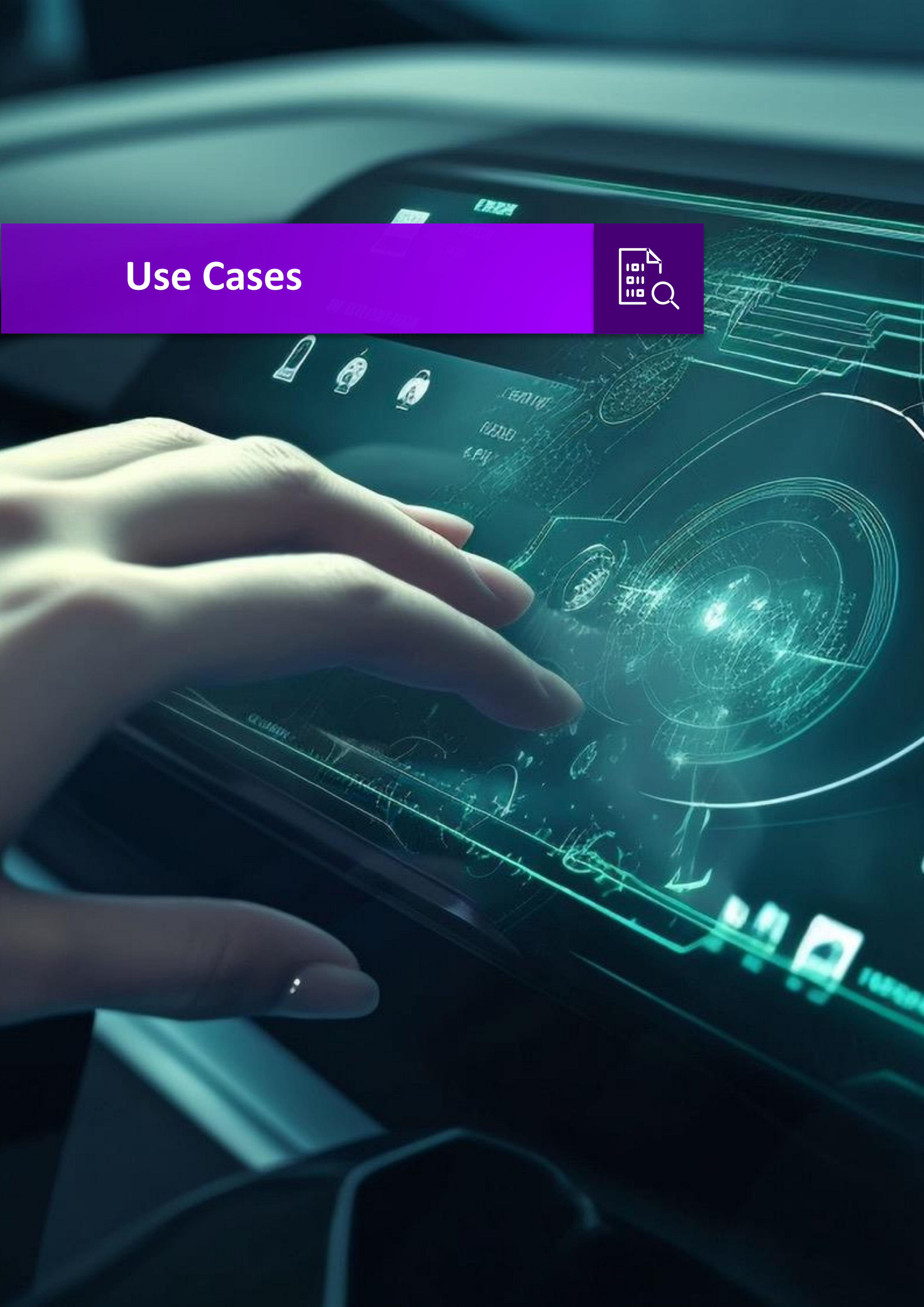
Deep learning is a powerful tool for creating AI models, but it has its limitations. At its core, deep learning is based on complex statistical correlations between inputs and outputs, without any notion of meaning. As a result, there are no guarantees or controls on what is learned, and the quality of the model can only be assessed through coarse-grained introspection or observation of its behavior. To overcome the limitations of deep learning, several measures have been proposed in the automotive domain. One such measure is to use better architectures that incorporate more meaningful features into the model. Another approach is to add redundancies in critical components to improve the model's robustness. Additionally, non-AI solutions can be used where possible to reduce the reliance on deep learning.

Beyond Safety and Security

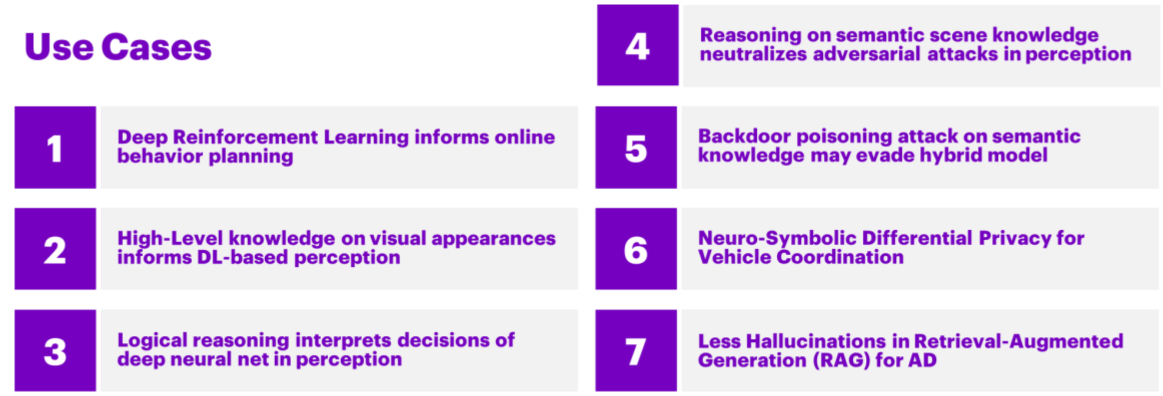
While safety and security are undoubtedly the most critical factors for Responsible AI in autonomous driving, they are not the only considerations. Other vital criteria for Responsible AI include privacy protection, avoidance of biases, and accountability maintenance. In the automotive industry, Responsible AI entails safeguarding the privacy of drivers and passengers by ensuring that data collected by these systems is only used for legitimate purposes and is properly secured. It also involves designing and deploying AI systems in an ethical manner that avoids discriminatory outcomes, respects the rights of drivers and passengers, and ensures that the benefits of AI are distributed equitably across society. Unfortunately, in the past, there have been several incidents that highlight the challenges of achieving Responsible AI in autonomous driving.

Autonomous driving systems have raised several concerns around privacy, fairness, and ethics. One major concern is the amount of data collected by these systems, including personal information such as location data, driving habits, and biometric data. This raises questions about how the data is collected, stored, and shared, as well as its potential for misuse, as demonstrated by a 2019 data breach of Didi Chuxing in China (Reuters, 2022). Another concern is the potential for bias and discrimination in AI-based autonomous driving systems. For instance, a study by the Georgia Institute of Technology found that these systems are more likely to hit dark-skinned pedestrians than light-skinned pedestrians due to differences in the way that the systems recognize and respond to different skin tones (Wilson et al., 2019). Additionally, there are ethical concerns around programming these systems to make ethical decisions in the event of an accident or emergency, as demonstrated by the above-mentioned Uber accident. Achieving Responsible AI in autonomous driving systems requires addressing these privacy, fairness, and ethical concerns through proper regulation, oversight, and ongoing research and development.

Use Cases



In this section, we present seven use cases that illustrate the potential of neuro-explicit AI in the context of autonomous driving. These use cases have been selected to showcase examples of how neuro-explicit AI can be applied to perception and behavior in autonomous driving systems to increase safety, security, privacy, and sustainability. The use cases are all based on the concept of neuro-explicit AI, which combines neural networks with explicit knowledge representations to create more effective and efficient AI models. Two of the use cases are described in more technical detail, providing a deeper understanding of the underlying principles and techniques involved. The other six use cases are described more generally, highlighting the potential benefits of neuro-explicit AI in various aspects of autonomous driving. It is important to note that these use cases are intended to serve as a mind opener, offering insights and ideas rather than final solutions to all the challenges and opportunities in this field.



Use Cases

1 Deep Reinforcement Learning informs online behavior planning

2 High-Level knowledge on visual appearances informs DL-based perception

3 Logical reasoning interprets decisions of deep neural net in perception

4 Reasoning on semantic scene knowledge neutralizes adversarial attacks in perception

5 Backdoor poisoning attack on semantic knowledge may evade hybrid model

6 Neuro-Symbolic Differential Privacy for Vehicle Coordination

7 Less Hallucinations in Retrieval-Augmented Generation (RAG) for AD

Figure 1: Overview of the 7 use cases from the automotive industry that are discussed in this paper.

Deep Reinforcement Learning informs online behavior planning

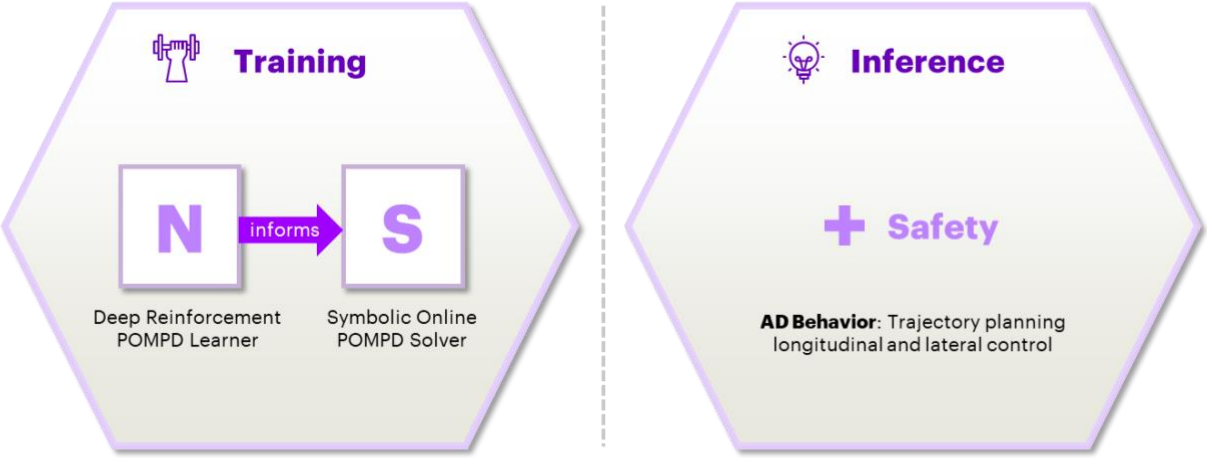


Figure 2: Deep Reinforcement Learning informs Online Planning in autonomous driving: Combining neural and symbolic models for safer decision-making in uncertain and dynamic environments. The "N" stands for a neural model and "S" for symbolic, semantic, explicit.

Reinforcement Learning and symbolic Online Planning have been utilized in the autonomous driving industry to train and develop intelligent vehicles that can make decisions in real-time situations. This use case shows how Deep Reinforcement Learning informs Online Planning of the car's behavior. It integrates the neural models (N) and the symbolic model (S) during training, potentially leading to safer situations at inference time.

Driving can be described as a Partially Observable Markov Decision Process (POMDP) problem, where the state of the system (e.g., the position and speed of other vehicles, traffic signals, and road conditions) is partially observable, and the decision to be made (e.g., changing lanes or slowing down) depends on the current state and the anticipated future state of the system. In other words, driving is a complex decision-making problem that involves reasoning about uncertain and dynamic environments. Online POMDP solving involves using Bayesian inference to update the probability distribution over the possible states of the system, based on observations made by the autonomous vehicle. This approach can be computationally efficient and effective for systems with a small number of states, but it may struggle to handle the large and complex state spaces that are often present in autonomous driving. On the other hand, POMDP solving using Deep Reinforcement Learning involves training a neural network to approximate the optimal policy for a given state of the system. This approach can handle large and complex state spaces, but it requires large amounts of data and computing power for training the neural network. It can also be challenging to ensure that the trained model is safe and reliable, particularly in complex and uncertain environments.

Pusse and Klusch (2019) developed a neuro-explicit variant that deeply entangled online POMDP planning with Deep Reinforcement Learning (see Figure 2). Their approach involved using a Deep Reinforcement Learning algorithm to learn a policy for a set of common driving scenarios, while simultaneously using an online POMDP solver to handle the uncertainties and variations in the driving environment. The Deep Reinforcement Learner informed the online planner by providing feedback on the policy and updating the online POMDP solver with additional information about the driving environment. This approach overcame the disadvantages of both online POMDP planning and Deep Reinforcement Learning. Pusse and Klusch tested their approach using the German In-Depth Accident Study (GIDAS) database, which contains detailed information about real-world traffic accidents. Their approach achieved the best performance among several state-of-the-art methods, demonstrating its effectiveness in handling the complex and uncertain nature of driving as a POMDP problem.

A Deeper Look

The HyLEAP method (Pusse & Klusch, 2019) is a combination of two approaches, the online POMDP planner IS-DESPOT (Luo et al., 2018) and the Deep Reinforcement Learner NavA3C (Mirowski et al., 2016), for approximating the collision-free navigation problem of self-driving cars. In its core, NavA3C is a Long Short-Term Memory (LSTM) Model, which is a type of recurrent neural network (RNN) architecture. RNNs are a class of artificial neural networks designed for processing sequential data, making them well-suited for tasks that involve time series. LSTMs, in particular, are an improvement over traditional RNNs, as they can learn long-range dependencies in the input data more effectively. In the context of the DRL network of HyLEAP, the NavA3C architecture is used, which is an Asynchronous Advantage Actor-Critic (A3C) algorithm for navigation tasks. In this architecture, a single LSTM layer is incorporated to improve the model's ability to learn temporal dependencies in the environment. By using only one LSTM layer, the model achieves a balance between learning complex patterns and maintaining faster execution, which is critical for real-time decision-making in navigation tasks. The HyLEAP method integrates the advantages of both approaches by using the NavA3C network as a critic to guide the online approximated POMDP planning. The overall HyLEAP architecture is coupled with the OpenDS driving simulator (OpenDS, 2016).

HyLEAP is trained in two consecutive phases. In the first phase, IS-DESPOT determines its action policy using the coupled NavA3C network with fixed, initially random weights, for evaluating its belief tree construction at each time step of the scene simulation and executes the selected actions in the scene. After simulation of the scene, the NavA3C network then gets trained to update its weights during its action policy computation according to the received total discounted reward, such that its experience-based evaluation feedback to IS-DESPOT in the next scene simulations can improve.

The observation input from IS-DESPOT is the car intention RGB image, which is passed through two convolutional network layers, conv1 and conv2, and then fed into a fully connected layer. The output of the fully connected layer is concatenated with the last received reward, current car velocity, and last executed action by IS-DESPOT in the simulated scene, and then fed into the LSTM network layer with forget gates together with the history. The output of the LSTM layer is reduced to size one for the estimated value V and to |A| outputs for the DRL action policy πDRL based on its weights. The network is trained to minimize the error between the received total discounted reward and the predicted value V and the error between the neural network policy πDRL and the APPL policy of IS-DESPOT for each scene simulation point in time.

The trained HyLEAP car is executed in a simulated GIDAS accident test scene, which is one run-through of IS-DESPOT with integrated evaluation by the trained NavA3C network with input from the path planner and the driving simulator OpenDS. Training of HyLEAP is performed in two phases per traffic scene simulation in OpenDS. Phase 1 is network-guided action planning in traffic scene simulation. At each point in time t of one traffic scene simulation in OpenDS, IS-DESPOT constructs a belief tree with root using the neural network as guidance, and then outputs a policy πAPPL from which an action is sampled for execution. This process is repeated until the scene simulation ends. Phase 2 is HyLEAP NavA3C network training. After simulation of the considered traffic scene ends, the NavA3C network gets trained for this episode to improve on its evaluation of the APPL policy created by IS-DESPOT according to the total discounted rewards received.

The empirical evaluation uses a benchmark OpenDS-CTS 1.0 consisting of about 38,000 accident scenes simulated in OpenDS with 9 different single pedestrian-car accident scenarios to evaluate three self-driving methods: NavA3C-p, IS-DESPOT-p, and HyLEAP. The evaluation metrics are defined for pedestrian safety as the number of crashes, time-to-goal, and smoothness of driving. The study shows that HyLEAP outperforms IS-DESPOT-p and NavA3C-p in most of the GIDAS accident scenarios on the simulated test drive regarding safety.

Furthermore, the study also examines the behavior of self-driving methods in situations with unusual pedestrian movement patterns and multiple pedestrians crossing the street.

The findings of the study demonstrate the effectiveness of a neuro-explicit variant for addressing the challenges of driving as a POMDP problem. This approach is not only applicable to autonomous driving but can also be transferred to other planning. By combining the strengths of both approaches, this neuro-explicit variant offers a promising solution for developing safe, reliable, and effective autonomous systems that can adapt to complex and uncertain environments. As the field of AI continues to evolve, neuro-explicit approaches like this can play an essential role in enabling autonomous systems to reason about complex and uncertain environments, ultimately advancing the field of AI and the development of responsible and trustworthy autonomous systems.

High-Level knowledge on visual appearances inform DL-based perception

In the previous use case, we discussed how Deep Reinforcement Learning can inform Online Planning to enhance safety in autonomous driving behavior. Now, we will focus on another use case that utilizes high-level knowledge on visual appearances to inform DL-based perception. In this use case, the symbolic model (S) informs the neural model (N) during training, creating a link between the two models. The main objective is to enhance the safety of autonomous vehicles by improving the accuracy of their perception of the environment. By incorporating high-level knowledge of visual appearances, we can enable the vehicle to make more informed and accurate decisions, ultimately leading to a safer driving experience.

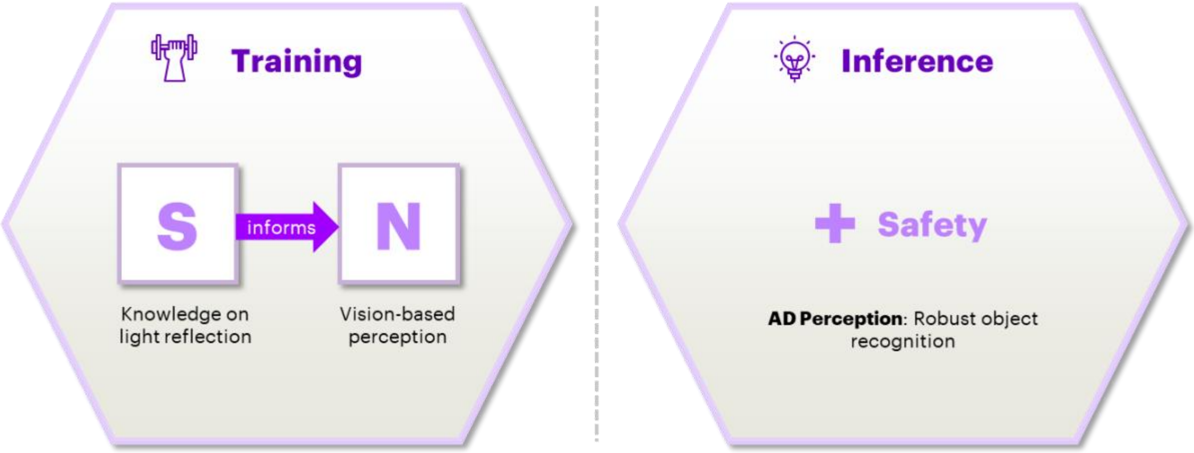

Training
S
informs
N
Knowledge on light reflection
Vision-based perception
Inference
Safety
AD Perception: Robust object recognition

Figure 3: Improving autonomous driving safety with neuro-explicit methods: Incorporating high-level knowledge of visual appearances to inform DL-based perception, enhancing robustness against perturbations, and adapting to different environments.

This use-case shows a way, how perception in autonomous driving can be made more robust by neuro-explicit methods. The core idea, based on Muggleton et al. (2018), is to create background knowledge that describes some of the physical properties of light propagation. A ball, for example, would reflect sunlight in a distinct kind of way, having the lighter areas in the top hemisphere and the shadowy areas in the bottom one. Additional knowledge about the available light sources would make the reflection hypotheses about the objects even stronger. This background knowledge could be used in various ways, for example as an additional information source during training. We could assume that the neural model will end up being more robust against certain kinds of perturbations. For example, if a ball would be partially occluded, it could in some cases still be recognized as a ball, because the distinct light reflections are still present. The same would be possible if an adversarial perturbates the input signal (for example by putting special stickers on the object). If the original neural perception model was to be unchanged for whatever reason, we could still use the knowledge-based approach for plausibilization. It could then create the hypotheses of certain objects corresponding to their respective light reflection patterns being present in the scene, and either confirm or contradict the neural result. Supervisory control architectures could then deal with the context-dependent fusion of the results.

In a related research project, a research team from the University of Illinois and Stanford University developed a Deep Reinforcement Learning-based approach for robots to navigate crowded environments without colliding with people. The robots analyze human behavior in their surroundings to determine potential obstacles, even if they are hidden. This approach treats humans as sensors, and the team's main insight is that by observing interactive human behavior, they can draw inferences about the spatial environment. Similarly, the visual appearance use case involves incorporating high-level knowledge of visual appearances to inform DL-based perception in autonomous vehicles. Both approaches highlight the importance of leveraging additional information beyond what can be learned from purely direct image-based models to enhance the accuracy and robustness of AI systems. The robot navigation approach can be applied to various industries where robots operate in busy environments, such as supermarkets and airports. The research team plans to refine and expand their method to other applications, including assistive and warehouse robots.

While we assume that neuro-explicit approaches are superior to pure neural approaches with respect to the traits of Responsible AI (here: robustness), it must be made very clear that it is not impossible to attack them, too. The logic of a backdoor poisoning attack can be in principle also applied to the approach described above. Instead of looking for ways to efficiently manipulate the input data in order create a backdoor and to make the system evade, when corresponding stimuli are present, the adversarial would do the same on the knowledge base. In this case, it could introduce light reflection patterns that do not (or very rarely) occur in real environments, later introducing corresponding objects. We explore this possibility in Section "Backdoor poisoning attack on semantic knowledge may evade hybrid model".

Logical reasoning interprets decisions of deep neural net in perception

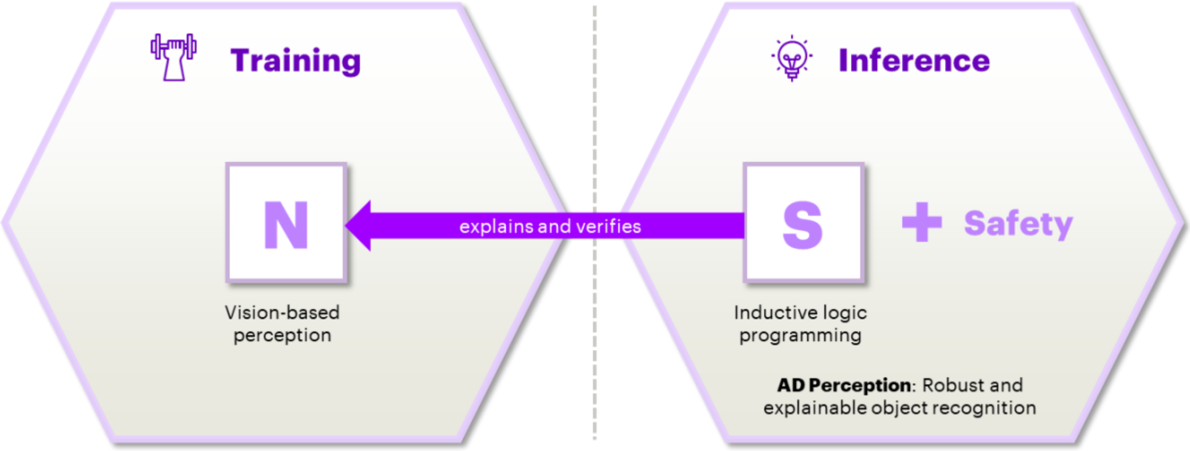

Training
N
Vision-based perception
explains and verifies
Inference
S
Inductive logic programming
+ Safety
AD Perception: Robust and explainable object recognition

Figure 4: Enhancing interpretability in deep learning-based perception systems: Combining symbolic reasoning with neural networks to improve reliability, explainability, and safety in AI systems, using description logics and Semantic Web technologies for meaningful explanations.

We will now examine another use case of how logical reasoning can help in interpreting the decisions made by deep neural networks in perception. In this use case, the symbolic model (S) is used to verify and interpret the decisions made by the neural model (N) at inference time. This approach allows for the integration of logical reasoning in the decision-making process of deep learning-based perception systems, enhancing their safety and reliability. By using symbolic reasoning to verify the decisions made by the neural network, we can increase the interpretability and explainability of the system, allowing for more effective debugging and error correction. This use case demonstrates the potential of combining deep learning with symbolic reasoning to create more reliable and Trustworthy AI systems.

Sarker at al. (2017) discuss the issue of neural networks being considered as black boxes and not providing direct indications of how they reached their output, which is problematic for safety-critical applications. While rule extraction has been pursued as a means of explaining neural network behavior, it is limited in terms of generating human-understandable explanations. Novel deep learning architectures have attempted to retrieve explanations, but often only for computer vision tasks. The paper proposes a new paradigm that goes beyond the propositional paradigm and targets the problem of explaining neural network activity directly, rather than just the qualities of the input. The paradigm leverages advances in knowledge representation on the World Wide Web, specifically from the field of Semantic Web technologies, which utilize knowledge graphs and type logics called ontologies.

The paper discusses the use of description logics, a major paradigm in knowledge representation within artificial intelligence, as the foundation for the W3C Web Ontology Language OWL. Description logics are a decidable fragment of first-order predicate logic, with unary predicates referred to as atomic classes and binary predicates as roles. ALC, a fundamental description logic, serves as the basis for OWL. The paper also introduces DL-Learner, a machine learning system inspired by inductive logic programming, which constructs class expressions to include all positive examples while excluding negative examples.

The authors detail experiments conducted using DL-Learner to provide explanations for classifications based on background knowledge. The experiments involve the use of the ADE20K dataset, which contains pre-classified images of scenes, and the Suggested Upper Merged Ontology (SUMO) for background knowledge. The results demonstrate the dependency of explanations on the conceptualizations encoded in the background knowledge and show the potential of the approach for providing human monitors with explanations for classifications that may help identify misclassifications or missed elements.

Reasoning on semantic scene knowledge neutralizes adversarial attacks in perception

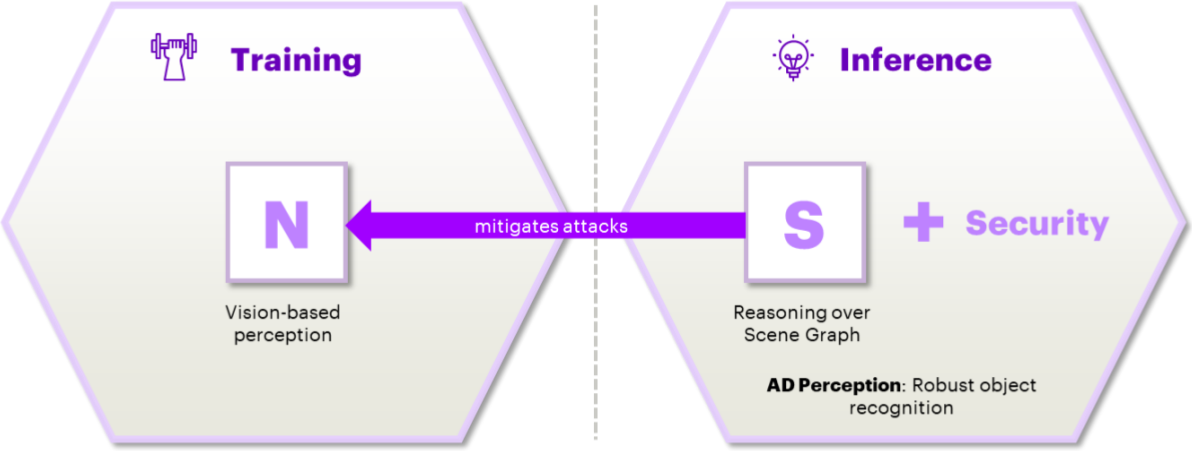

Training
Inference
N
mitigates attacks
S
+ Security
Vision-based perception
Reasoning over Scene Graph
AD Perception: Robust object recognition

Figure 5: Adversarial attacks on autonomous driving perception: Highlighting the need for robust, secure systems and the potential of neuro-explicit approaches with knowledge-base plausibilization to mitigate data vulnerabilities and improve safety.

Autonomous driving perception is vulnerable to adversarial attacks, where an attacker can manipulate the input to the system to cause it to make incorrect or unsafe decisions. These attacks can occur by modifying the physical environment or by adding noise to the sensor data that the autonomous vehicle uses to make decisions. For example, an attacker could use stickers or other visual cues to make a stop sign appear as a speed limit sign, causing the vehicle to ignore the stop sign. These attacks can be challenging to detect, and their potential consequences could be severe, highlighting the need for developing more robust and secure perception systems in autonomous vehicles.

In 2022, Yang et al. proposed a method to mitigate the potential data vulnerability in autonomous driving perception systems by making the perception neuro-explicit and adding a knowledge-base plausibilization layer. This approach involved using a neural network to extract features from sensor data and map them to a symbolic representation, which can then be checked against a knowledgebase of plausible scenarios.

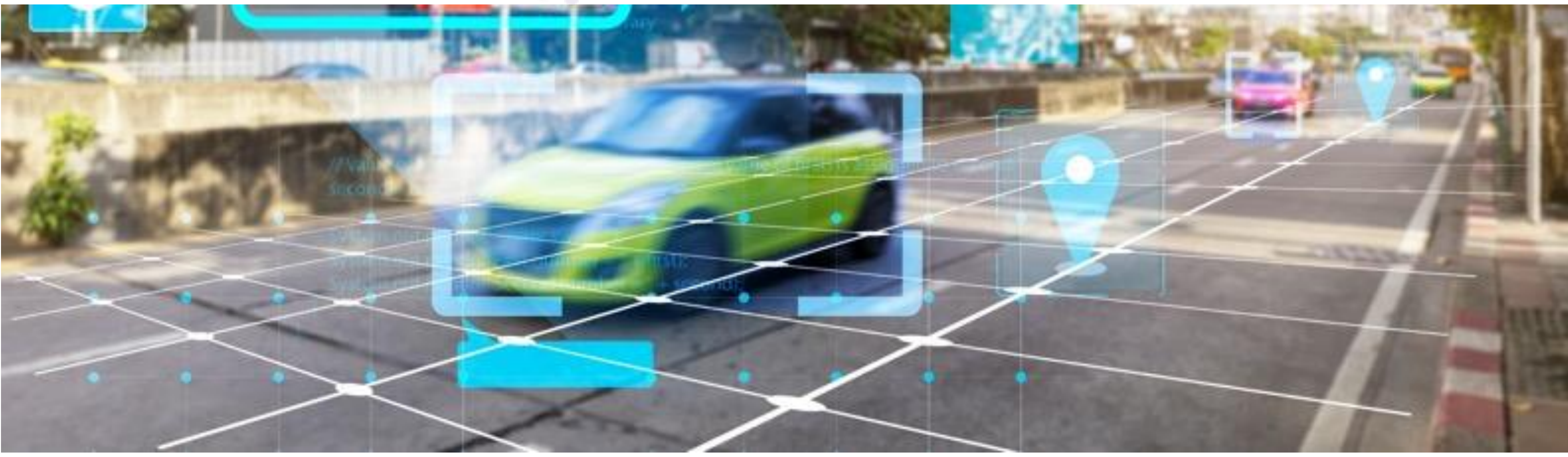
A Deeper Look

Yang et al. (2022) discuss the vulnerability of deep learning models to small perturbations in the input, which can significantly alter model predictions. Various methods have been proposed to defend against these attacks, such as training with adversarial examples and certifiable defenses. However, these methods have limitations and are vulnerable to carefully designed adversarial attacks. The authors suggest that human-like reasoning, which relies on commonsense knowledge and contextual information, could be a natural defense mechanism against such attacks. The logic adversarial defense (LOGICDEF) is proposed as an interpretable defense framework that utilizes a scene graph to mine commonsense knowledge and represent it as logic rules through inductive logic programming (ILP). These rules are integrated into the deep classifier as constraints using posterior regularization (PR) to construct a defense model. LOGICDEF can also incorporate existing commonsense knowledge base and utilize natural object taxonomy for curriculum learning to improve its defense performance with fewer labeled data.

The defense mechanism proposed utilizes a scene graph representation of the objects in the image, which includes relations and attributes of the objects. The defense model constructs a set of inductive logic rules based on the scene graph to interpret the objects' relationships and properties in the real world. The defense model integrates these rules as constraints into the deep learning classifier using posterior regularization, aiming to improve the classifier's robustness against adversarial attacks. LOGICDEF aims to improve the robustness of an object classifier against adversarial attacks. The construction of the defense model involves addressing three main challenges: scene graph generation, rule mining and defense, and knowledge integration. In the scene graph generation step, the goal is to generate a scene graph that includes object relations for reasoning with commonsense. The scene graph is generated by collecting ground-truth graphs from the Visual Genome dataset or with a pre-trained scene graph generator, such as Graph R-CNN.

The rule mining and defense step involves collecting a set of logic rules to be used for defense. The defense model utilizes the framework of restricted first-order logic (FOL) to represent the defense. Logic entailment rules are defined to encode the knowledge that an object belongs to a certain class if and only if certain conditions are true in the scene graph. The defense rule set is collected through rule mining and human-generated prior knowledge. In the knowledge integration step, the rule set is incorporated into the object classifier in a principled manner using posterior regularization (PR) technique. PR converts rule knowledge into constraints on the posterior distributions of the classifier model. This ensures that the inference is robust against attack. LOGICDEF is constructed by addressing these challenges, and it incorporates prior knowledge from various sources like ConceptNet for better robustness. It is trained using a differentiable ILP model and a curriculum-based learning approach to learn logic rules for objects. The defense model is constructed by integrating the rule set with the object classifier using PR, ensuring that the inference is robust against adversarial attacks.

The use case presented by Yang is another example of how safety and security merge in the field of adversarial machine learning, as it addresses the critical challenge of developing more robust and secure perception systems in autonomous driving and other domains. By mitigating the potential data vulnerability of autonomous driving perception systems against adversarial attacks, this approach ensures that the system can make accurate and safe decisions in complex and uncertain environments, ultimately enhancing the safety and security of the overall system.



Backdoor poisoning attack on semantic knowledge may evade hybrid model

In this use-case section, we explore a scenario where the symbolic module (S) is utilized to inform and guide the neural model (N) during training, similar to the first use case. However, we specifically examine a situation where the security of the system is compromised due to a poisoning attack on the symbolic module (S). This attack results in the creation of a backdoor, which has the potential to undermine the overall reliability and trustworthiness of the model. Through this example, we aim to highlight the importance of robust security measures also on explicit knowledge base and the possible implications of such attacks on the integrity of AI systems.

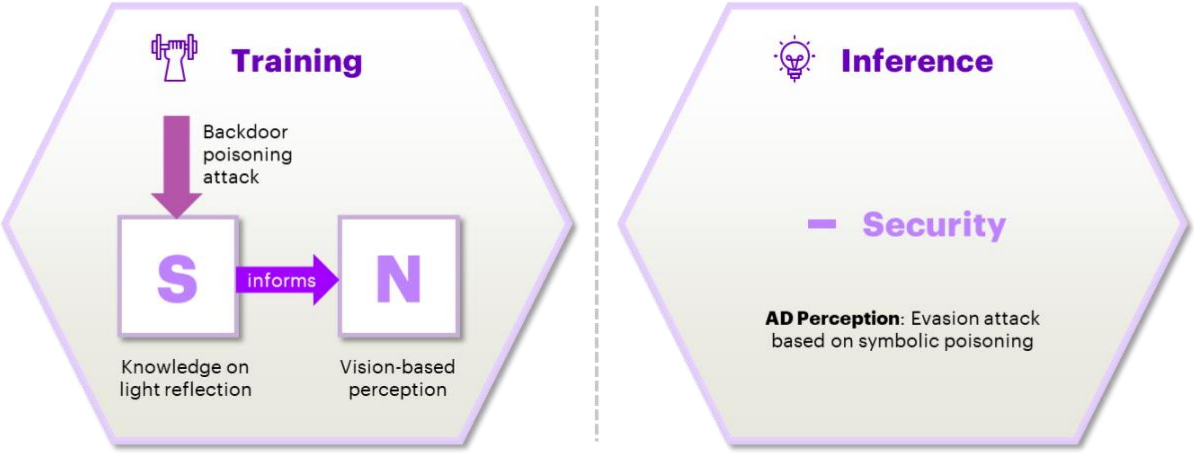

Training
Backdoor poisoning attack
S
informs
N
Knowledge on light reflection
Vision-based perception
Inference
Security
AD Perception: Evasion attack based on symbolic poisoning

Figure 6: Backdoor attack on neuro-explicit AI: A cautionary scenario emphasizing the need for robust security measures to protect autonomous driving systems' reliability and safety from compromised knowledge bases.

The second use-case above demonstrates how neuro-explicit methods can enhance the robustness of perception in autonomous driving. The approach uses background knowledge of physical properties of light propagation to improve object recognition. This knowledge can be applied during training to make the neural model more resistant to perturbations, such as partial occlusion or adversarial input manipulation. The knowledge-based approach can also be used for plausibilization, confirming or contradicting neural results in a supervisory control architecture. Although neuro-explicit methods contribute to Responsible AI traits like robustness, they are not immune to attacks. A backdoor poisoning attack could target the knowledge base, introducing rare or non-existent light reflection patterns and corresponding objects.

Backdoor poisoning attacks on autonomous vehicles are a type of cybersecurity threat where hackers insert malicious code or data into the vehicle's AI system. This is typically done during the training process of the AI model, when it learns to recognize and respond to various scenarios. The attacker introduces specific patterns or triggers, which when detected by the AI system, can cause it to behave unexpectedly or maliciously. In simple terms, it's like secretly teaching the vehicle's AI to misbehave when it encounters a specific signal, which could lead to dangerous situations or loss of control over the vehicle.

In our use case, the AI system relies on a semantic knowledge base to understand the environment and make informed decisions. In our example the knowledge base is containing information about the light reflection properties of objects in the environment, which helps the AI recognize and interpret the objects accurately. Suppose attackers decide to poison this semantic knowledge base instead of a statistical training set. They could introduce false information about the light reflection properties of specific objects or create entirely new, unrealistic reflection patterns that rarely, if ever, occur in real-world situations.

For instance, the attacker could manipulate the knowledge base to make the AI system believe that a specific light reflection pattern corresponds to a pedestrian when, in fact, it does not. Later, when the vehicle encounters an object with this manipulated reflection pattern, the AI system would falsely identify it as a pedestrian and take unnecessary evasive action, potentially causing accidents or endangering other road users. By poisoning the semantic knowledge base, the attacker is exploiting the AI system's reliance on this information, causing it to make wrong decisions based on the corrupted knowledge. This type of attack underscores the importance of ensuring the security and integrity of both statistical training sets and semantic knowledge bases in autonomous driving systems.

The scenario described above, where an attacker gains access to the background knowledge and manipulates the symbolic module in a hybrid AI system, highlights the fact that neuro-explicit methods alone are not a guarantee for AI safety and AI security. While the use of symbolic reasoning can improve the robustness and interpretability of AI systems, careful consideration of the system design is also necessary for ensuring Trustworthy AI. AI developers must implement appropriate security measures, such as access control and data privacy, and must design AI systems with the potential for attacks in mind. Furthermore, monitoring the consistency between the neural net and the symbolic module can provide an additional layer of security, which is critical for ensuring the reliability and safety of AI systems. This use case demonstrates the importance of a holistic and thoughtful approach to AI design and development that addresses both the strengths and limitations of different AI techniques to ensure Trustworthy AI.

Neuro-Symbolic Differential Privacy for Vehicle Coordination

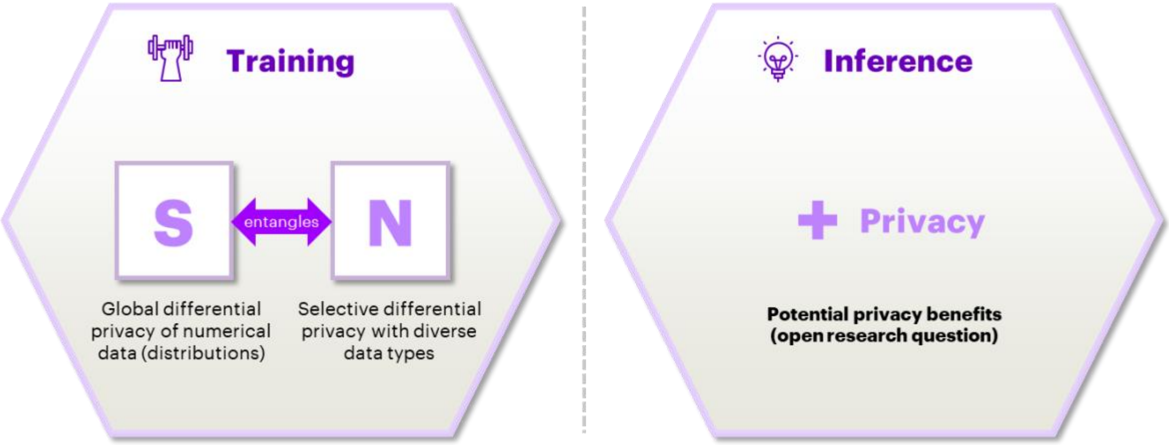

Training
S
entangles
N
Global differential privacy of numerical data (distributions)
Selective differential privacy with diverse data types
Inference
Privacy
Potential privacy benefits (open research question)

Figure 7: Neuro-explicit models in multi-vehicle coordination: Enhancing privacy through modularity, symbolic reasoning, and abstract data representations while addressing challenges in differential privacy application.

In the rapidly evolving field of autonomous driving, protecting the privacy of individual vehicles and passengers is of paramount importance. This text presents a hypothetical use case involving multi-vehicle coordination, where differential privacy and neuro-explicit models are combined to enhance privacy protection. It is important to note that this use case has not yet been validated by published studies, and it serves as a thought experiment to explore the potential benefits and challenges of applying differential privacy techniques in neuro-explicit models.

Differential privacy, a mathematical framework that provides privacy guarantees by adding controlled noise to data or computation outputs, plays a significant role in addressing privacy concerns in this scenario. Neuro-explicit models can offer superior privacy protection in multi-vehicle coordination compared to pure deep neural networks. Their modularity and interpretability enable more selective application of differential privacy, targeting specific components or data representations without compromising the overall system's functionality. For instance, a neuro-explicit model could separate the processing of raw sensor data from the high-level decision-making process, allowing differential privacy to be applied only to the decision-making component. This would enable vehicles to share abstract, symbolic information about their intended actions without revealing sensitive raw data, thereby preserving the privacy of individual vehicles and passengers.

Moreover, the symbolic reasoning capabilities of neuro-explicit models can be leveraged to further preserve privacy. By injecting noise into the logical rules or the output of the reasoning process, differential privacy can be maintained even when vehicles exchange symbolic information about their environment and intentions. This approach requires adapting existing differential privacy techniques to work with symbolic data, but it can lead to better privacy guarantees while maintaining the interpretability and utility of the exchanged information.

Additionally, neuro-explicit models often work with higher-level, abstract data representations, which can inherently offer better privacy protection than raw data used by deep neural networks. However, it is crucial to ensure that privacy guarantees are maintained when working with abstract or symbolic data. In the multi-vehicle coordination scenario, this could mean developing protocols to prevent vehicles from inferring sensitive information about other vehicles or passengers based on the abstract data they share. This is where the principles of differential privacy come into play, as they provide a rigorous method for quantifying privacy and ensuring that the output of computations remains nearly the same regardless of whether a specific individual's data is included in the dataset.

On the other hand, there are some considerations when applying differential privacy to neuro-explicit models. One potential counterargument is that, due to their modularity and the use of symbolic reasoning, these models might be more vulnerable to privacy attacks targeting specific components or data representations. Additionally, the need to adapt differential privacy techniques for symbolic data may result in increased complexity and potential challenges in balancing privacy and utility. Despite these concerns, the overall benefits of using neuro-explicit models in combination with differential privacy make them a compelling choice for preserving privacy in autonomous driving scenarios, such as multi-vehicle coordination.

In conclusion, while neuro-explicit models show promise in providing better privacy protection in autonomous driving scenarios, such as multi-vehicle coordination, the extent of these benefits compared to pure deep neural networks is still an open research question. Further investigation is needed to evaluate the effectiveness of differential privacy techniques in neuro-explicit models and to address potential challenges related to modularity, symbolic reasoning, and data abstraction.

Less Hallucinations in Retrieval-Augmented Generation (RAG) for AD

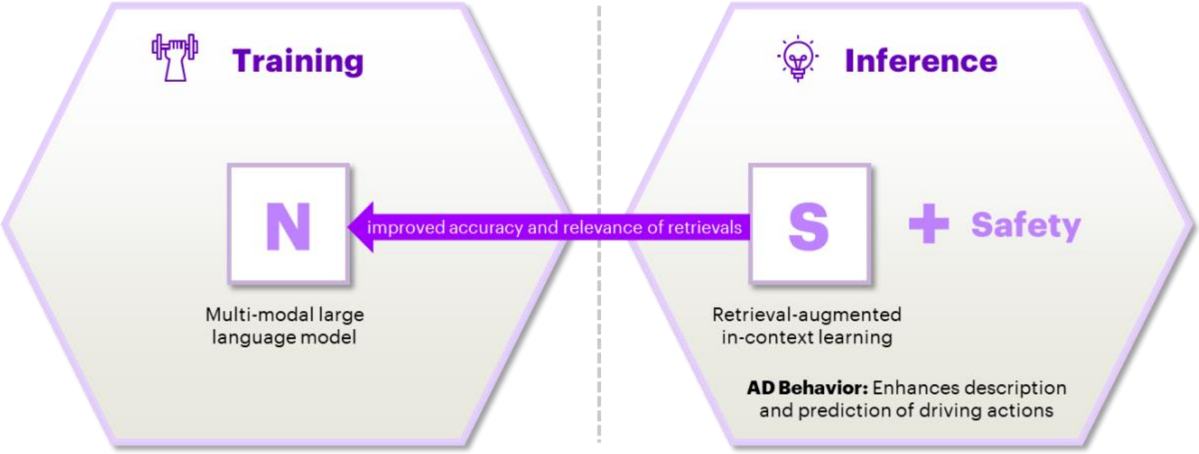

Training
Inference
N
improved accuracy and relevance of retrievals
S
+ Safety
Multi-modal large language model
Retrieval-augmented in-context learning
AD Behavior: Enhances description and prediction of driving actions

Figure 8: Multi-modal Large Language Models for AD: Reducing hallucinations for predicting AD actions, their justification and control signals (turning angle and speed) using Retrieval-Augmented Generation (RAG), thereby improving accuracy and understanding of the predictions and, ultimately, resulting in higher driving safety.

Retrieval-Augmented Generation (RAG) combines the strengths of information retrieval systems and generative models to enhance generation, most commonly in natural language processing. This method allows for the production of content that is coherent, contextually relevant, and information-rich, tailored to the specific inputs. The RAG process begins with a search through a large database to find relevant information using either keyword search tactics or more advanced neural network-driven methods. After identifying significant documents or snippets, these are integrated into the inputs of a generative model, like those based on the Transformer architecture, enabling the creation of nuanced output that incorporates knowledge from the retrieved information, thus improving the output's depth and precision.

Although RAG cannot guarantee that hallucinations will not occur, it significantly reduces the likelihood of such occurrences and increases accuracy (Shuster et al., 2021) Thus, RAG is valuable in various applications, from question answering, where it pulls in essential documents to craft detailed responses, to content creation tasks like article writing or summarization, where it adds relevant background information. It also enhances conversational AI, chatbots, and code generation by providing more informative and context-aware responses and improving accuracy and functionality through references to similar examples or documentation.

RAG can potentially also be applied in the domain of autonomous driving. The “RAG-Driver” approach (Yuan et al. , 2024), for example, is a multi-modal large language model with retrieval-augmented in-context learning capabilities designed specifically for AD. It aims to address the challenges of explainability, data scarcity, and the need for generalizability in self-driving cars. RAG-Driver integrates advanced multi-modal large language models (MLLMs) with a retrieval mechanism that augments the system's current predictions through in-context learning. This is achieved by searching for and using driving scenarios similar to the current condition, enhancing both the description and prediction of driving actions and making the system more generalizable to new deployment domains. The model initially aligns visual and language features by projecting pre-trained video embeddings into language tokens processable by the LLM. This allows for a seamless fusion of visual and textual data, crucial for accurate and explainable autonomous driving decisions. RAG-Driver's architecture includes a unified perception and planning module, leveraging a pre-trained video encoder and a cross-modality projector to align video and language embeddings efficiently. It predicts textual action explanations and numerical control signals (like speed and steering angle), relying on a memory unit built upon a hybrid vector and textual database for robust multi-modal in-context learning (ICL) during decision-making.

In the exploration of enhancing RAG such as the RAG-Driver with neuro-symbolic methods, the approach described by [cite] intertwines the precision of symbolic knowledge with the adaptability of neural networks. This integration seeks to capitalize on the strengths of both paradigms—leveraging the rich, structured information contained in knowledge graphs and the dynamic learning capabilities of neural networks to improve information retrieval. The core idea revolves around augmenting neural representations with symbolic knowledge, thus enriching the neural model's understanding and processing of information. As we argue above, symbolic knowledge, often structured in the form of entities and relationships within knowledge graphs, provides a detailed and interconnected representation of information. When these symbolic representations are integrated with neural networks, the system gains a dual ability to take advantage of both the nuanced semantics of natural language and the explicit connections within structured knowledge.

For instance, in the context of information retrieval, this neuro-symbolic approach enables the model to not only access the syntax of the multi-modal input but also to directly reason with relevant entities and their relationships from a knowledge graph. One of the methodologies for infusing symbolic knowledge into neural representations involves the annotation of input media with entities that are linked to a knowledge graph. This process, known as entity linking, allows the neural model to access a wealth of structured information that complements its learning from input. By grounding the neural representations in the reality defined by the knowledge graph, the model gains a more profound understanding of the concepts and their interrelations.

Techniques such as Graph Neural Networks (GNNs) and knowledge-infused training procedures further reinforce the neural model's ability to internalize the semantic connections between entities. This enriched understanding facilitates more accurate and relevant retrievals in response to complex queries. The neuro-symbolic approach also lends itself to improvements in reasoning about relevance and explainability in information retrieval, as in the RAG-Driver case. By harnessing both neural and symbolic reasoning capabilities, retrieval models can more effectively determine the relevance of documents –multimodal descriptions of driving scenes – to a query, incorporating both semantic nuances and explicit knowledge. Additionally, the infusion of symbolic knowledge into neural representations can aid in denoising and stabilizing the model's outputs, making the retrieval process more resilient across different domains.

Conclusions

The automotive industry has always been oriented towards functional safety, as it is critical for the protection of human lives. Therefore, safety architectures with safeguarding, redundancy, and supervisory architectures have been implemented to ensure that any faults in the system do not lead to accidents. Safeguarding refers to measures taken to prevent malfunctions or errors, redundancy involves the use of multiple components or systems that can take over if one fails, and supervisory architectures involve the monitoring and control of the system's behavior to prevent it from going beyond safe limits. These measures ensure that the system remains safe even in the face of potential errors or malfunctions.

End-to-end deep learning models for autonomous driving refer to models that take raw sensor input, such as camera images or LIDAR data, and output driving commands directly, without any intermediate processing or human intervention. These models have been suggested as a potential solution for achieving fully autonomous driving, but they have not been widely adopted by the industry for several reasons.

Firstly, the automotive industry has a long history of relying on a structured approach to develop safety-critical systems. In this approach, a system is decomposed into smaller, more manageable components that can be independently verified and tested. This enables engineers to identify and mitigate potential failures at the component level, rather than waiting until the entire system fails. End-to-end deep learning models do not fit into this culture of component-based design, as they are complex, non-linear models that are difficult to interpret and debug. As a result, the industry has been hesitant to adopt end-to-end deep learning models for safety-critical systems like autonomous driving.

Secondly, end-to-end deep learning models are often criticized for being "black boxes." That is, they produce results without providing any insight into how those results were generated. In the case of autonomous driving, this lack of transparency could be particularly problematic, as it could make it difficult to determine why a particular decision was made in the event of an accident or other safety incident. This lack of transparency could also make it challenging to build trust with regulators, policymakers, and the general public.

Finally, end-to-end deep learning models may not be reasonable or practical for autonomous driving. While these models have shown impressive performance in the language or image domain, they often require vast amounts of data and computing resources to train and run. This could make them prohibitively expensive to implement in the multisensory physical world domain, especially given the constraints of limited processing power and energy consumption in automotive systems. Moreover, generative models based on transformer architectures are prone to hallucinations and confabulations that threaten to undermine trust in the systems that utilize them.

Neuro-explicit methods have been gaining popularity in recent years, especially in the automotive industry, as they offer a good middle way between the two extremes of end-to-end deep learning and traditional model-based methods. In contrast to end-to-end deep learning, neuro-explicit methods explicitly incorporate physical knowledge into the model. This means that the model is designed to obey known physical laws and constraints, such as conservation of momentum, energy, or mass. As discussed in this article, this explicit modeling of physical laws not only makes the model more interpretable, but also ensures that the model produces physically plausible outputs. Further, both domain and world knowledge can be incorporated to reduce hallucinations in generative models and guarantee that established rules are obeyed. At the same time, neuro-explicit methods also make use of deep neural networks, which can capture complex patterns and relationships in the data that may be difficult to model explicitly. In the context of autonomous driving, neuro-explicit methods offer several advantages over end-to-end deep learning. For example, by explicitly modeling physical constraints, neuro-explicit methods can help to ensure that the system behaves safely and predictably, even in extreme or unexpected situations. This is especially important in safety-critical applications like autonomous driving, where the consequences of a failure can be severe. Furthermore, neuro-explicit methods fit well with the safety architectures that are common and proven in the automotive industry. These architectures typically involve multiple layers of redundancy and fault tolerance, which are designed to ensure that the system remains safe even in the presence of failures or errors. By incorporating physical knowledge into the model, neuro-explicit methods can help to ensure that these safety architectures are more effective and reliable.

Throughout this article, we have presented neuro-explicit AI as a promising approach for enhancing Responsible AI in autonomous driving. By combining neural networks and symbolic reasoning, neuro-explicit algorithms offer a more robust and explainable AI model, addressing some of the limitations of pure deep learning. We have provided a detailed exploration of eight use cases related to autonomous driving, each focusing on different aspects of perception and behavior. Some use cases focused on perception, such as detecting pedestrians or recognizing road signs, while others focused on behavior, such as predicting the actions of other vehicles or planning the car's route. The use cases were also grounded in principles of Responsible AI, such as AI security, safety, explainability, robustness, sustainability, fairness, and privacy. By focusing on these principles, the use cases offered valuable insights into how AI can be developed and deployed responsibly in the field of autonomous driving, with the ultimate goal of improving safety, efficiency, and convenience for drivers and passengers.

While this is just one example of differentiation on the algorithmic level, Accenture's AI maturity concept encompasses a broader spectrum of aspects. Accenture emphasizes the importance of developing AI solutions that are not only technically sound but also align with the organization's overall values and ethics. Furthermore, it emphasizes the importance of building a governance structure around AI development and deployment to ensure accountability, transparency, and trust. As AI continues to advance, it is essential that we approach its development responsibly, and Accenture's AI maturity concept provides a roadmap to achieve this objective.

As we have discussed, the development and deployment of Responsible AI technologies are critical to the success of autonomous driving in the automotive domain. At Accenture, we are committed to helping our customers make this transition from AI to Responsible AI. Our expertise in AI, data science, and ethics positions us as a trusted partner for automotive companies looking to develop Responsible AI models that address key issues such as safety, security, and privacy.

With our deep industry knowledge and innovative solutions, we are well-equipped to help our customers navigate the complex landscape of Responsible AI. Our comprehensive approach to Responsible AI helps to make solutions transparent, explainable, and trustworthy, while adhering to ethical and legal frameworks.

We are dedicated to keeping pace with the latest developments in AI and Responsible AI, ensuring that our customers are equipped with the most up-to-date technologies and methods to solve their challenges. Our commitment to Responsible AI is underscored by our ongoing investment in research and development, which enables us to stay at the forefront of this rapidly evolving field.

References

Accenture (2022). *The art of AI maturity: Advancing from practice to performance*.

https://www.accenture.com/content/dam/system-files/acom/custom-code/ai-maturity/Accenture-Art-of-AI-Maturity-Report-Global-Revised.pdf [Accessed: 21.03.23]

Accenture Tech Vision Research (2022). *Responsible AI: Scale AI with confidence.* https://www.accenture.com/es-es/services/applied-intelligence/ai-ethics-governance [Accessed: 29.07.24]

Accenture (2024). Accenture's blueprint for responsible AI. https://www.accenture.com/au-en/case-studies/data-ai/blueprint-responsible-ai [Accessed: 01.08.24]

European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). (2020). *Policy and Investment Recommendations for Trustworthy Artificial Intelligence*. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence [Accessed: 21.03.23]

European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). *Ethics Guidelines for Trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed: 02.05.24]

Lee, B. D. (2017). *Self-driving shuttle bus in crash on first day.* BBC News. https://www.bbc.com/news/technology-41923814 [Accessed: 21.03.23]

Levin, S., & Wong, J. C. (2018). *Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian.* The Guardian. https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe [Accessed: 21.03.23]

Luo, Y., et al. (2018). *Importance sampling for online planning under uncertainty.* The International Journal of Robotics Research. doi:10.1177/0278364918780322

Mirowski, P., et al. (2016). *Learning to Navigate in Complex Environments.* Computing Research Repository (CoRR). arXiv/1611.03673

Muggleton, S., et al. (2018). *Ultra-strong machine learning: comprehensibility of programs learned with ILP.* Machine Learning, Springer. doi:10.1007/s10994-018-5707-3

OpenDS. (2016). https://opends.dfki.de [Accessed: 21.03.23]

Orland, K. (2024). *NYC's government chatbot is lying about city laws and regulations*. Ars Technica. https://arstechnica.com/ai/2024/03/nycs-government-chatbot-is-lying-about-city-laws-and-regulations/ [Accessed: 04.04.2024]

Pusse, F., & Klusch, M. (2019). *Hybrid Online POMDP Planning and Deep Reinforcement Learning for Safer Self-Driving Cars*. 2019 IEEE Intelligent Vehicles Symposium (IV). doi:10.1109/ivs.2019.8814125

Reuters. (2022). *China to fine Didi more than $1 billion for data breaches: Reuters, citing sources.* CNBC. https://www.cnbc.com/2022/07/20/china-to-fine-didi-more-than-1-billion-for-data-breaches-reuters.html [Accessed: 21.03.23]

Sarker, M. K., et al. (2017). *Explaining trained neural networks with semantic web technologies: First steps.* Proceedings of the 12th International Workshop on Neural-Symbolic Learning and Reasoning. arXiv:1710.04324

Shepardson, D. (2023). *US agency, California gathering details of accident involving robot taxi and pedestrian.* Reuters. https://www.reuters.com/world/us/us-agency-california-gathering-details-self-driving-crash-2023-10-03/ [Accessed: 03.04.2024]

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567.

Sims, D. (2023). *San Francisco robotaxis are causing false 911 calls and other chaos.* TechSpot Forums. https://www.techspot.com/community/topics/san-francisco-robotaxis-are-causing-false-911-calls-and-other-chaos.278889/ [Accessed: 21.03.23]

Wilson, B., et al. (2019). *Predictive Inequity in Object Detection.* Computing Research Repository (CoRR). arXiv:1902.11097

Yadron, D., & Tynan, D. (2016). *Tesla driver dies in first fatal crash while using autopilot mode.* The Guardian. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk [Accessed: 21.03.23]

Yang, Y., et al. (2022). *LOGICDEF: An Interpretable Defense Framework Against Adversarial Examples via Inductive Scene Graph.* Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press. doi:10.1609/aaai.v36i8.20865

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). *RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model.* arXiv preprint arXiv:2402.10828.

Zhengwei, B., et al. (2022). *Hybrid Reinforcement Learning-Based Eco-Driving Strategy for Connected and Automated Vehicles at Signalized Intersections.* IEEE Transactions on Intelligent Transportation Systems. doi: 10.1109/TITS.2022.3145798

Wachter, S., et al. (2021*). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law.* West Virginia Law Review, Vol. 123, No. 3, 2021. doi: 10.2139/ssrn.3792772

Authors:

Dr. Christian Müller

Principal Researcher DFKI Autonomous Driving *christian.mueller@dfki.de*

Yufeng Liu

Responsible AI Specialist, Accenture

yufeng.a.liu@accenture.com

Yi Wang

Ind & Func AI Decision Science Analyst, Accenture

yi.m.wang@accenture.com

Dr. Kevin Baum

Head of the DFKI Center for European Research in Trusted AI (CERTAIN)

kevin.baum@dfki.de

Rudraksh Bhawalkar

Former Principal Director || Responsible AI, Accenture

Dr. Hendrik Purwins

Senior Manager Data Science, Accenture

hendrik.purwins@accenture.com

With special thanks to: Dr. Gabriel Seiberth, Dr. Ellen Hohma