3.1 Harnessing Symbolic Knowledge Extraction and Utilization for Informed Decision-Making

Ya Wang, Adrian Paschke | Fraunhofer FOKUS

Formal Methods and Knowledge Representation

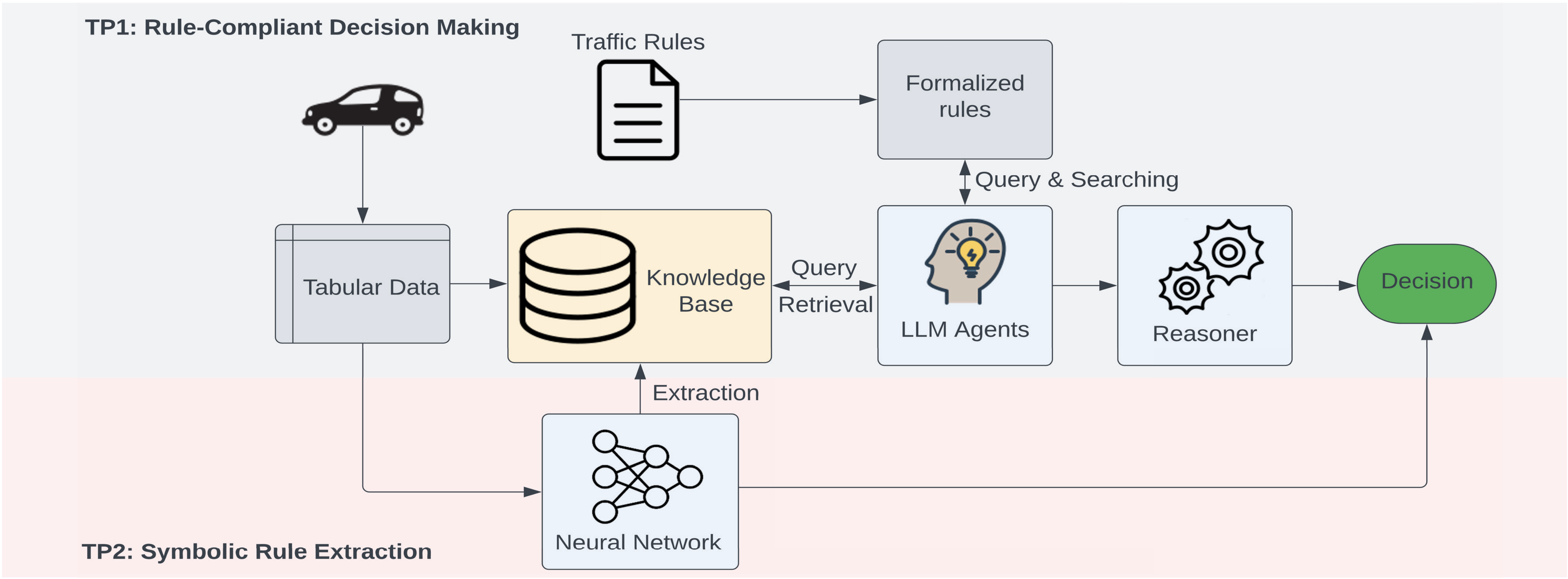

Figure 1: The neuro-symbolic architecture for rule-compliant decision making and rule extraction in autonomous driving (© Fraunhofer FOKUS)

Introduction

Decision-making is a critical and safety-essential module within autonomous driving systems, necessitating not only an in-depth understanding of traffic situation but also skilled reasoning based on world and normative knowledge. Neural networks, though widely used, present issues in reliability and safety. We propose a neuro-symbolic architecture to address these challenges, enabling traceable, rule-compliant decisions and explaining 'black-box' models with interpretable symbolic rules.

Architecture

Our architecture (Figure 1) consists of two main components using tabular driving data from perception modules and communication systems:

- Rule-Compliant Decision Making (Upper Part): Employs LLM agents to query relevant driving data and traffic regulations, using symbolic reasoners to derive actions.
- Symbolic Rule Extraction (Lower Part): Involves training neural networks on driving data for scenario classification and extracting symbolic rules to understand model decisions.

LLM Agent Assisted Rule-Compliant Decision Making

Driving Scenario Generation:

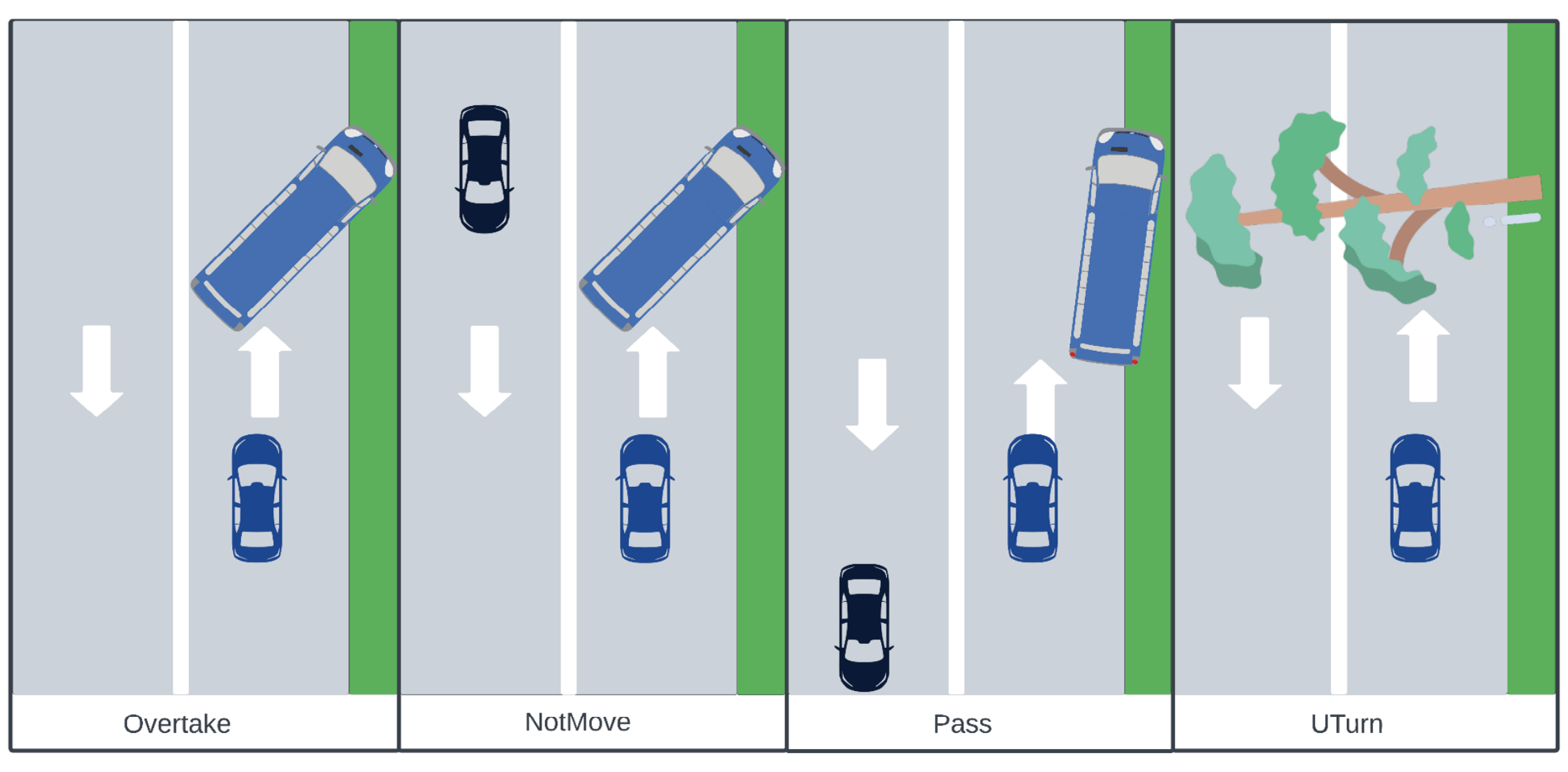

Figure 2: Exemplary illustrations of driving scenarios classified by driving actions (©Fraunhofer FOKUS)

Rule Formalization and Searching: leverages LLMs for converting rules into first-order logic and enhances semantic search of related rules using text embeddings.

Knowledge Base (KB): transforms tabular data into a structured KB through ontology.

Data Retrieval and Reasoning: LLM agents suggest actions from tabular data, verified against rules and facts in the knowledge base using symbolic reasoning.

Symbolic Rule Extraction

Decompositional rule extraction methods approximate neural network (NN) behaviors by translating NN structures and activations into symbolic rules. Our novel approach, EDICT (Extracting Deep Interpretable Concepts using Trees), enhances rule interpretability, extraction speed, and model fidelity beyond current methods.

- Core Design: utilizes a hierarchy of decision trees to approximate NN behavior and introduces new predicates reflecting activated concepts in NN, improving rule extraction speed and interpretability.

$$\bigvee_{i=1}^{n}\left(\bigwedge_{j=1}^{m} P(x_j^{(i)}, t_j) \rightarrow y^{(i)}\right)$$

- Evaluation: compares EDICT with existing techniques across multiple datasets and evaluates rule prediction accuracy, fidelity, extraction speed, and rule set size.

Conclusion

Symbolic knowledge is a crucial communication bridge between humans and AI systems. Our architecture boosts understand-ding of AI decisions and integrates symbolic knowledge, essential for effectively regulating AI behaviors and achieving AI alignment.

Partners

External partners

For more information contact:

ya.wang@fokus.fraunhofer.de
adrian.paschke@fokus.fraunhofer.de

KI Wissen is a project of the KI Familie. It was initiated and developed
by the VDA Leitinitiative autonomous and connected driving and
is funded by the Federal Ministry for Economic Affairs and Climate Action.

www.kiwissen.de @KI_Familie KI Familie

Funded by
the European Union
NextGenerationEU

Supported by:
Federal Ministry
for Economic Affairs
and Climate Action

on the basis of a decision
by the German Bundestag