



**From Logs to Logic:  
Learning and Evaluating Interpretable Representations of  
Behavior for Autonomous Systems**

av

**Simona Gugliermo**

**Akademisk avhandling**

Avhandling för teknologi doktorsexamen i datavetenskap,  
som kommer att försvaras offentligt  
Fredag den 12 december 2025 kl. 13.00,  
Hörsal L2, Örebro universitet

Opponent: Professor Roman Barták  
Charles University  
Prague, Czech Republic

Örebro universitet  
Institutionen för naturvetenskap och teknik  
701 82 ÖREBRO

# Abstract

Autonomous systems are increasingly being deployed across various real-world domains, such as fleets of self-driving vehicles, robotic warehouses, and delivery services using unmanned aerial vehicles. These systems are required to operate with high reliability and predictability, to adapt continuously to changing conditions, and to remain accountable to human supervisors. To achieve these objectives, autonomous systems need explicit, formal representations of their behavior that facilitate task planning, system verification, and human oversight.

In current industrial practice, such representations, whether for task-level control or action planning, are typically engineered manually. While hand-crafted representations can be precise, their development is labor-intensive and difficult to scale. Learning-based approaches offer a promising alternative by extracting behavioral representations from execution data. However, they often make unrealistic assumptions, such as access to simulated environments or large volumes of high-quality training data. Moreover, they fail to simultaneously achieve all the three critical objectives, that is reliability, adaptability, and interpretability. Therefore, there is a clear need for methods capable of efficiently learning accurate, interpretable representations under realistic conditions.

In this thesis, we address the problem of learning interpretable representations of system behavior from execution traces - sequences of observed actions and state transitions generated during the operation of autonomous systems. Learning from such traces is appealing because they are readily available from system logs and provide direct evidence of how a system behaves in realistic, often complex environments. The overarching goal is to derive representations that not only support automated planning but also enhance human understanding and oversight.

Two distinct types of behavior representation are explored: Behavior Trees (BTs) and STRIPS-style planning domains. For each, a novel method to automatically construct representations from execution traces is proposed. Specifically, for BTs, we introduce a method that combines Boolean logic, leveraging algorithms originally developed in circuit theory, with decision tree learning to induce structured, interpretable behavior representations. To assess the interpretability of BTs, a user study is conducted to examine how such representations are perceived by human users. The study identifies key features that influence user comprehension, contributing empirical evidence to a space that has traditionally lacked systematic analysis. Furthermore, a structured evaluation method for BTs along with quality metrics and design principles is presented, addressing the current lack of guidance for assessing BT quality beyond functional performance.

For STRIPS-style domains, we introduce a novel learning framework to construct symbolic action representations directly from execution traces, even in the presence of noise. In addition to the learning algorithm, a systematic methodology is proposed for evaluating learned planning domains through structural and task-based analysis, thereby addressing a critical gap in current practice and thus responding to the growing need for rigorous assessment methods.

The results demonstrate that it is possible to extract interpretable representations of autonomous behavior from noisy data. The proposed methods enable the transition from raw execution traces to structured representations that can support planning, validation, and human-in-the-loop systems. By advancing methods for learning, interpreting, and evaluating learned behavior representations, this work contributes to the development of autonomous systems that are both operationally effective and intelligible to human stakeholders.