hri 2025 Melbourne Workshop on Explainability for Human-Robot Collaboration

Interpretability Analysis of Symbolic Representations for Sequential Decision-Making Systems

Pulkit Verma, Julie A. Shah

pulkitv@mit.edu, julie_a_shah@csail.mit.edu

Acknowledgement: This research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Why Interpretability Matters?

- Robots increasingly work alongside humans in various environments
- Trust and transparency are critical for effective collaboration
- Humans need to understand robot decision-making processes
- Gap in consolidated research on interpretability for sequential decision-making

Sequential Decision-Making (SDM) Systems

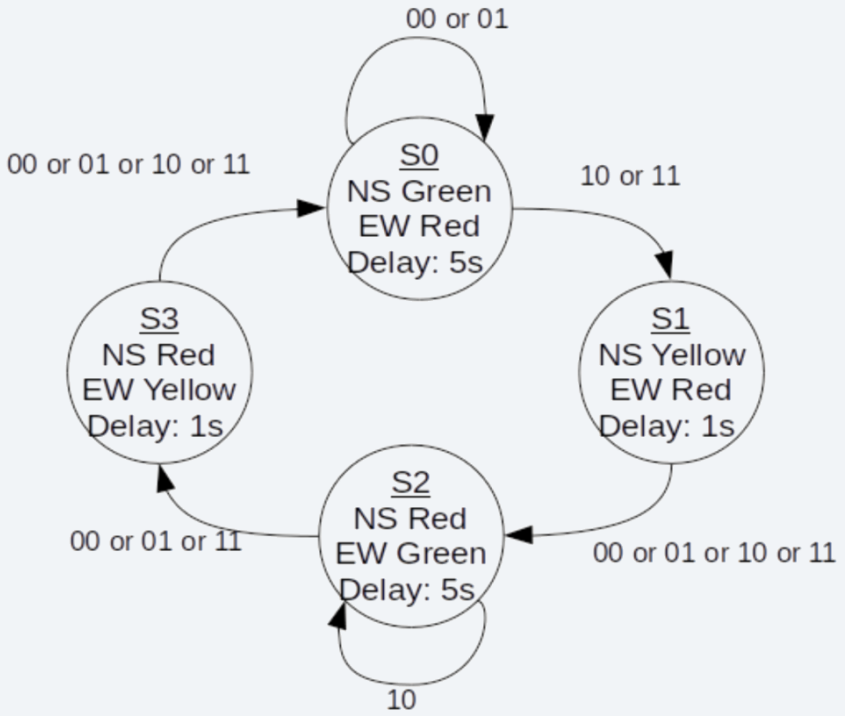
- Make series of decisions over time, where each decision influences future states
- Often modeled as Markov Decision Processes (MDPs)
- Unique interpretability challenges:
  - Temporal dependencies
  - Complex state spaces
  - Trade-offs between interpretability and performance

Symbolic Representations for Interpretability

1. Finite State Machines (FSMs)
2. Temporal Logic
3. Decision Trees
4. Program Synthesis
5. Planning Domain Definition Language (PDDL)
6. Rule-Based Systems

Finite State Machines (FSMs)

- States, transitions, and associated actions
- Strengths:
  - Intuitive graphical representation
  - Transparent computation
  - Effective for systems with clear decision points
- Limitations:
  - Struggle with nuanced continuous reasoning
  - Scalability challenges (state explosion)
  - Not suited for high-dimensional spaces



Planning Domain Definition Language (PDDL)

- Predicates (representing states) and actions
- Strengths:
  - Explicit and modular representation
  - Human-readable syntax
  - Clear traceability of plan steps
  - Can be visualized as graphs for enhanced interpretability
- Limitations:
  - Manual domain specification is labor-intensive
  - Scalability challenges in complex environments

```
(:action pick-object
  :parameters (?ob)
  :precondition (and
    (handempty)
    (onshelf ?ob))
  :effect (and
    (not (handempty))
    (not (onshelf ?ob))
    (holding ?ob))
)
```

Decision Trees

- Tree-like structure with decisions at each node

- Strengths:
  - Decompose predictions into feature contributions
  - Clear step-by-step decision process
  - Intuitive hierarchical structure

- Limitations:
  - Susceptible to overfitting with increased depth
  - Struggle with linear relationships
  - Difficulty handling online updates

Other Symbolic Approaches

1 Temporal Logic

- Formal language for temporal properties
- Powerful for verification but complex for non-experts

2 Program Synthesis

- Generates readable programs
- Strong interpretability but computational challenges

3 Rule-based Systems

- Knowledge encoded as "if-then" rules
- Explicit and transparent structure but scalability issues

Hybrid Representations

1 Causal Models

- Represent cause-and-effect relationships
- Explain the "why" behind decisions
- Limited to domains with well-defined causality

2 Neuro-Symbolic Integration

- Combines neural networks with symbolic reasoning
- Balances performance and interpretability
- Still an emerging field with implementation challenges

Classification Framework

Dimension	Description
Formalism Type	Symbolic, Subsymbolic, or Hybrid
Interpretability Level	How easily humans understand the representation
Temporal Expressiveness	How well time-related concepts are represented
Abstraction Level	Level of detail in the representation
Explanation Type	How explanations are generated
Domain Specificity	General vs. domain-specific applicability
Human Interaction	How humans interface with the representation

Classification Framework

Representation	Interpretability Level	Temporal Expressiveness	Abstraction Level	Explanation Type	Domain Specificity	Human Interaction
Markov Decision Processes (MDPs)	Medium	Discrete Time	Low-Level	Global	General	Indirect
Finite State Machines (FSMs)	Medium	Discrete Time	Low-Level	Global	General	Direct
Decision Trees	High	N/A	Low-Level	Global	General	Direct
Rule-Based Systems	High	N/A	Low-Level	Global	General	Direct
Temporal Logic (LTL, STL, etc.)	Medium	Continuous Time	Low-Level	Global	Domain	Indirect
Program Synthesis	Low	N/A	High-Level	Global	Domain	Indirect
Planning Domain Definition Language (PDDL)	Medium	Discrete Time	High-Level	Global	Domain	Indirect
Causal Models	Medium	N/A	Multi-Level	Global	General	Indirect
Neuro-Symbolic Integration	Low	N/A	Multi-Level	Global	General	Indirect

Conclusion and Future Directions

- No single representation offers a universal solution
- Choice depends on application requirements
- Future research directions:
  - Scalable methods for extracting interpretable representations
  - Standardized evaluation metrics
  - Hybrid approaches balancing transparency and performance
  - User-centric design of interpretable systems

Conclusion and Future Directions

- No single representation offers a universal solution
- Choice depends on application requirements
- Future research directions:
  - Scalable methods for extracting interpretable representations
  - Standardized evaluation metrics
  - Hybrid approaches balancing transparency and performance
  - User-centric design of interpretable systems

✉ pulkitv@mit.edu, julie_a_shah@csail.mit.edu