

# Explainably Safe Reinforcement Learning

## Summary

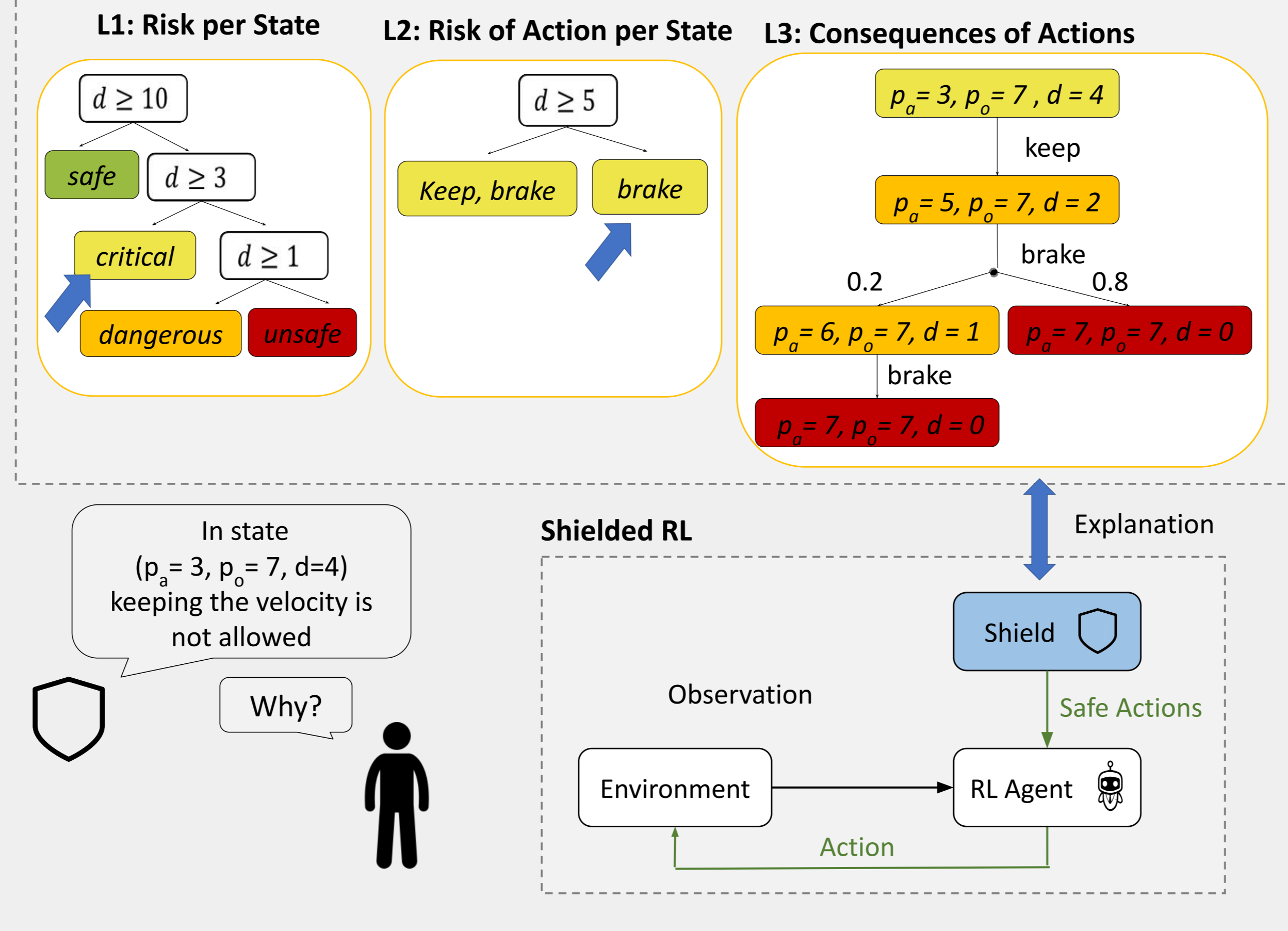
**Shield:** precomputed allowed actions (given a safety specification)

- ⊕ Safe reinforcement learning (RL)
- ⊖ Not interpretable

**Explainable** representation of shields through hierarchical decision trees:

- **L1** (Risk per state): Risk classification of states into safe, critical, dangerous and unsafe
- **L2** (Risk of action per state): Decision trees differentiating allowed and disallowed actions
- **L3** (Consequences of actions): Case-based explanations highlighting consequences of executing unsafe actions

**Computation** relies on already available information and only induces minimal overhead.



## Computing Hierarchical Safety Explanations

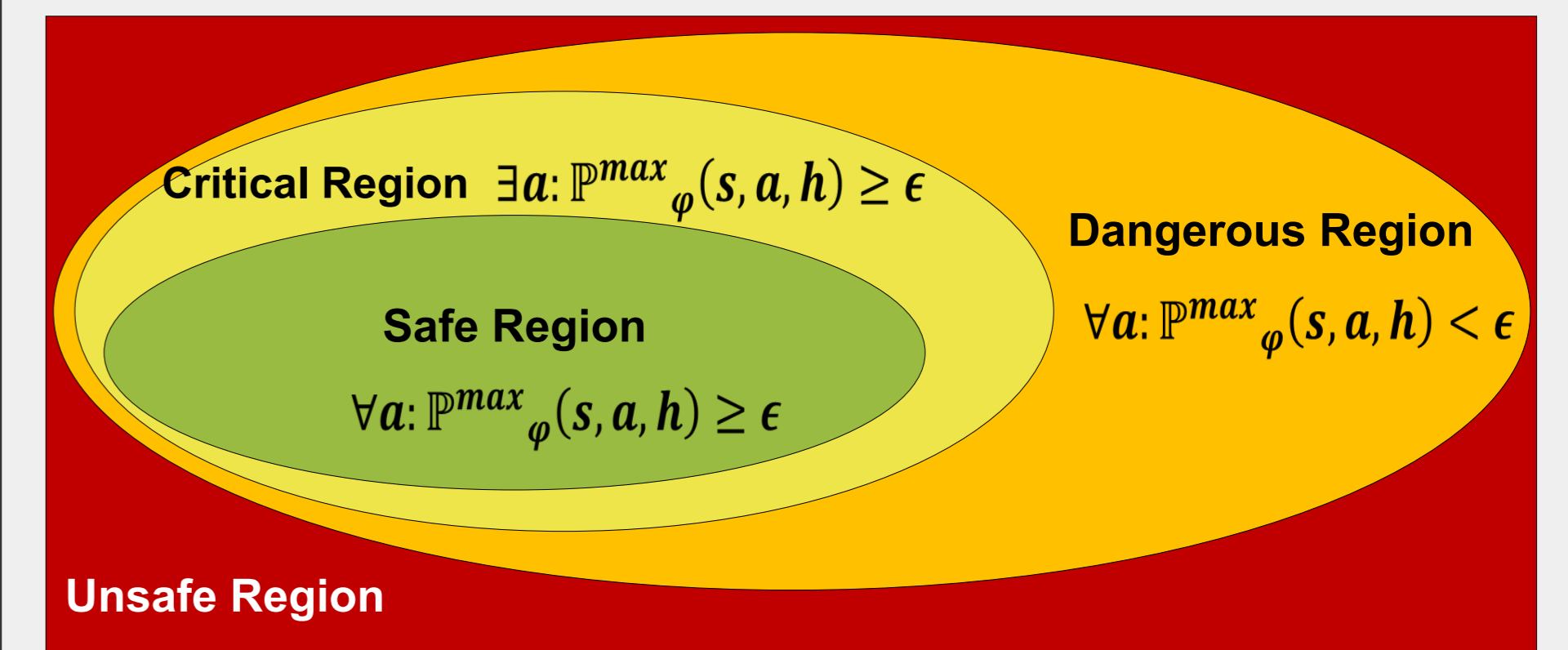
### Risk per State (Level 1)

**Risk of a state:** Minimal expected probability of reaching a state violating the safety specification

**State Partitioning:**

- **Safe states:** All actions have low risk
- **Critical states:** Some actions have low risk
- **Dangerous states:** No available action has low risk, safety specification is not yet violated
- **Unsafe states:** Safety specification is violated

**Computation:** Side product of model checking query for computing the shield.



### Risk of Action per State (Levels 2)

**Explanation** of allowed actions:

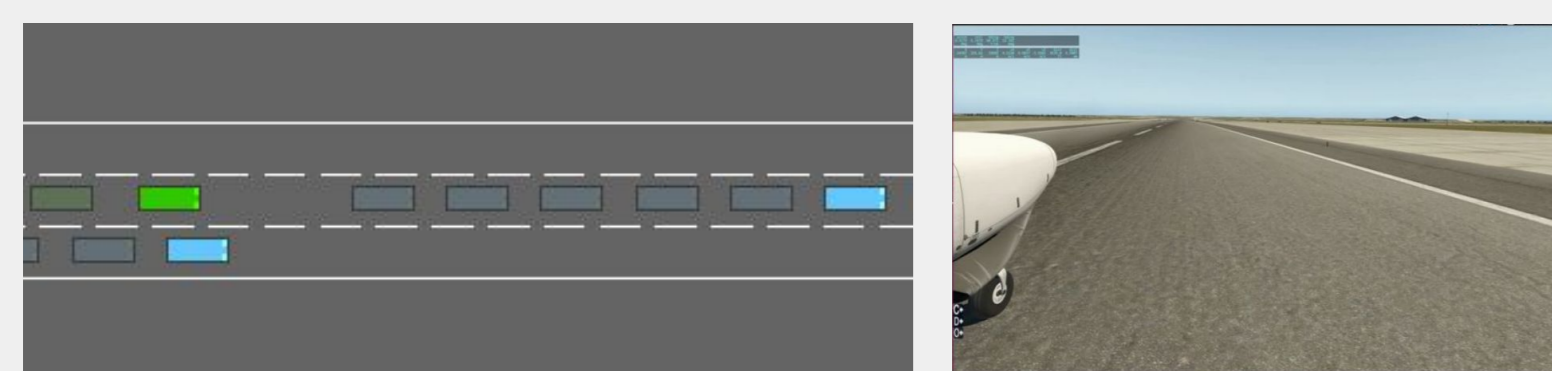
- One decision tree per critical leaf
- Maps the state to the set of allowed actions

### Consequences of Actions (Levels 3)

**Execution Tree:**

- A set of traces starting with the current state and a selected disallowed action
- Traces end in an unsafe states
- Combined probability of traces violates the safety specification

## Experimental Evaluation

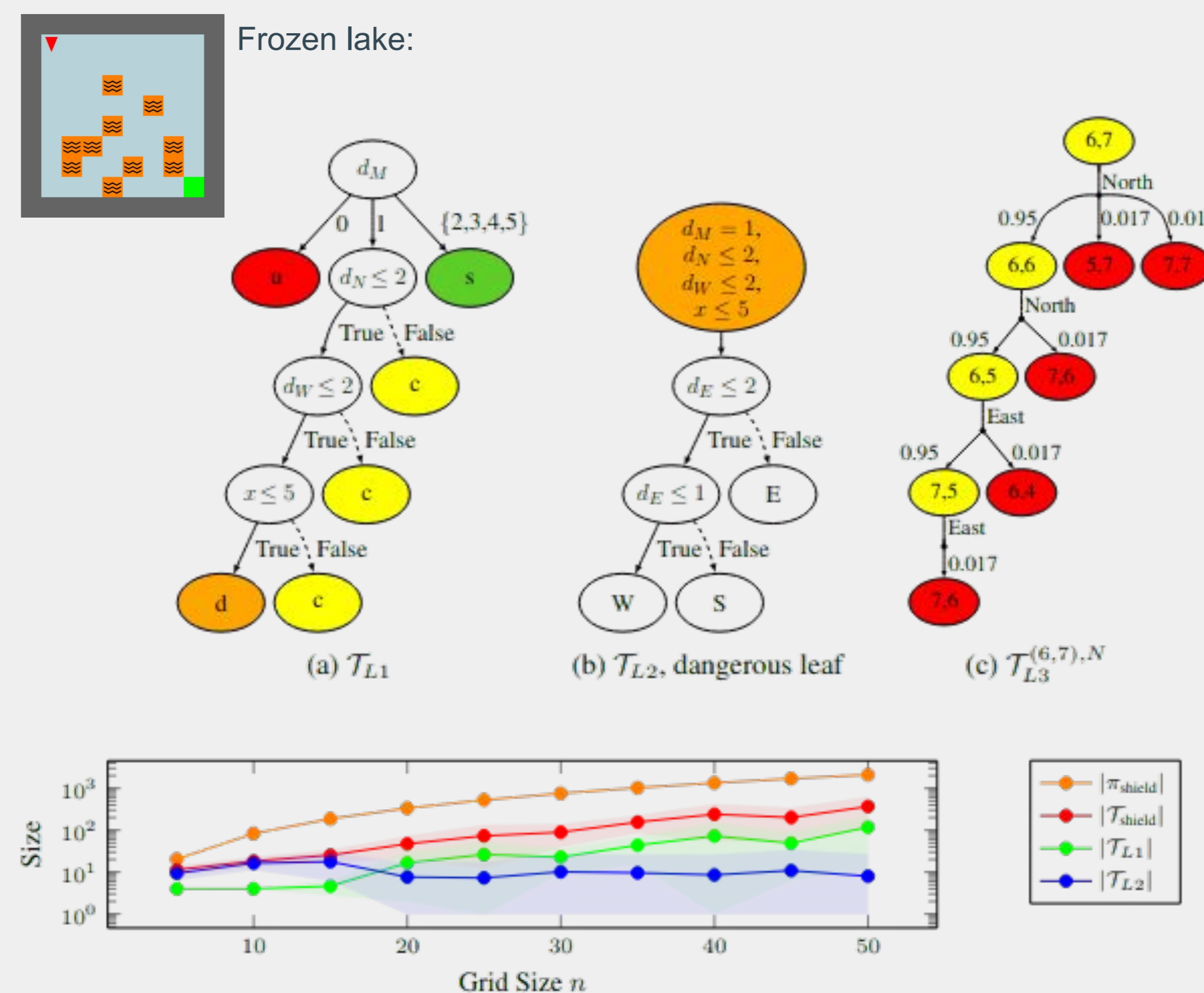


Highway

Taxiing

Tree sizes for different variants:

	$ \pi_{\text{shield}} $	$ \mathcal{T}_{\text{shield}} $	$ \mathcal{T}_{L1} $	$ \mathcal{T}_{L2} $
HW2-f	358	7	7	3
HW2-c	3553	113	9	33
HW3-f	8355	18	17	2.5
HW3-c	232216	555	57	45.1
TaxiNet	1105	39	21	6
TaxiNet	1107	37	33	5



This research has received funding from the State Government of Styria, Austria - Department Zukunftsfonds Steiermark, the European Union under Grant Agreement No. 101171844 (InOVationCS), the European Union's Horizon Europe program, Grant Agreement No. 101212818 (RobustIFAI), and the MUNI Award in Science and Humanities MUNI/1757/2021 of the Grant Agency of Masaryk University.



Co-funded by the  
European Union



ZUKUNFTSFONDS  
STEIERMARK

