"Scalable" Synthesis of Robust Strategies for Uncertain POMDPs

Murat Cubuktepe, Ufuk Topcu

The University of Texas at Austin

TEXAS
The University of Texas at Austin

autonomous
SYSTEMS GROUP

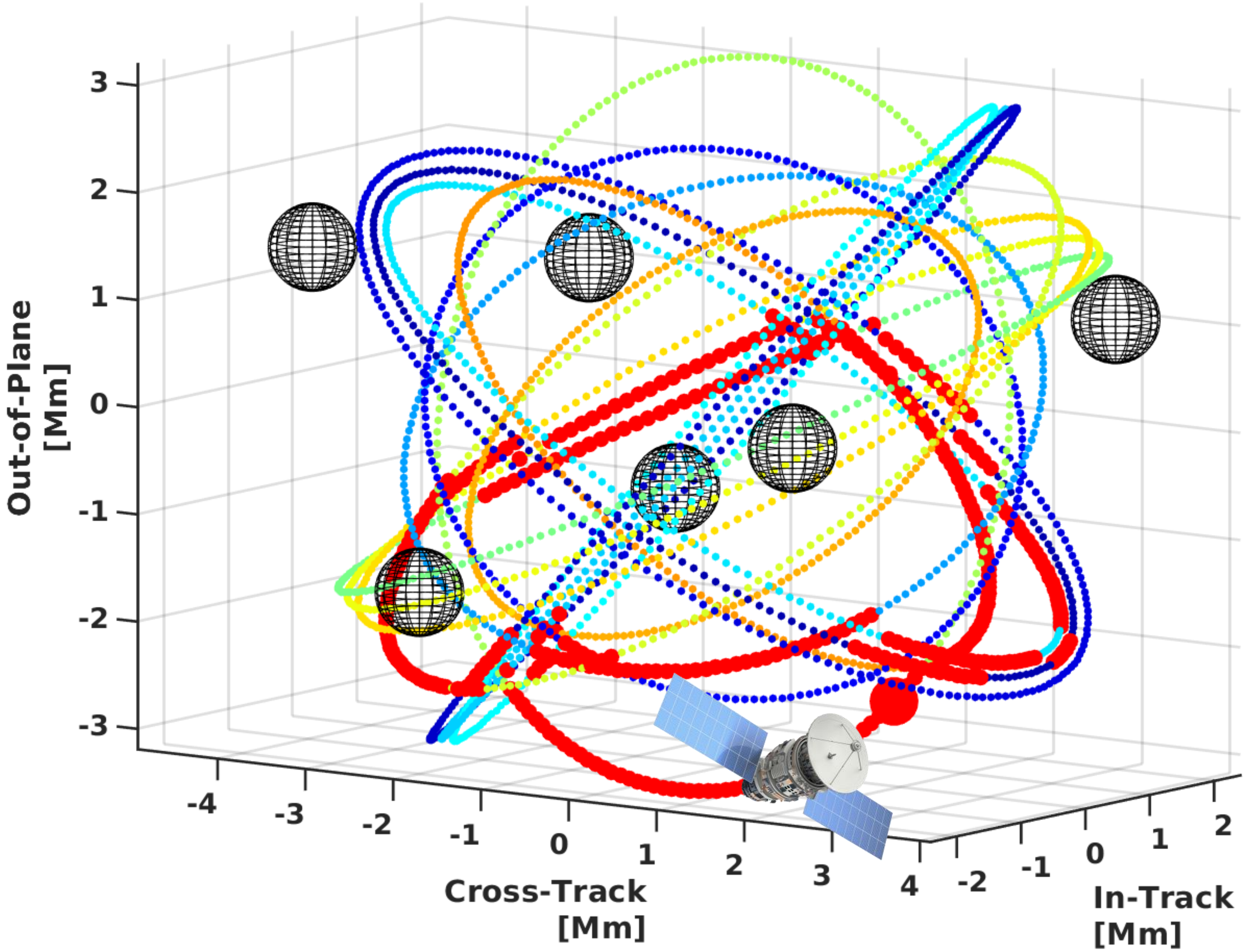
What is in this talk?

From POMDPs to
uncertain POMDPs.

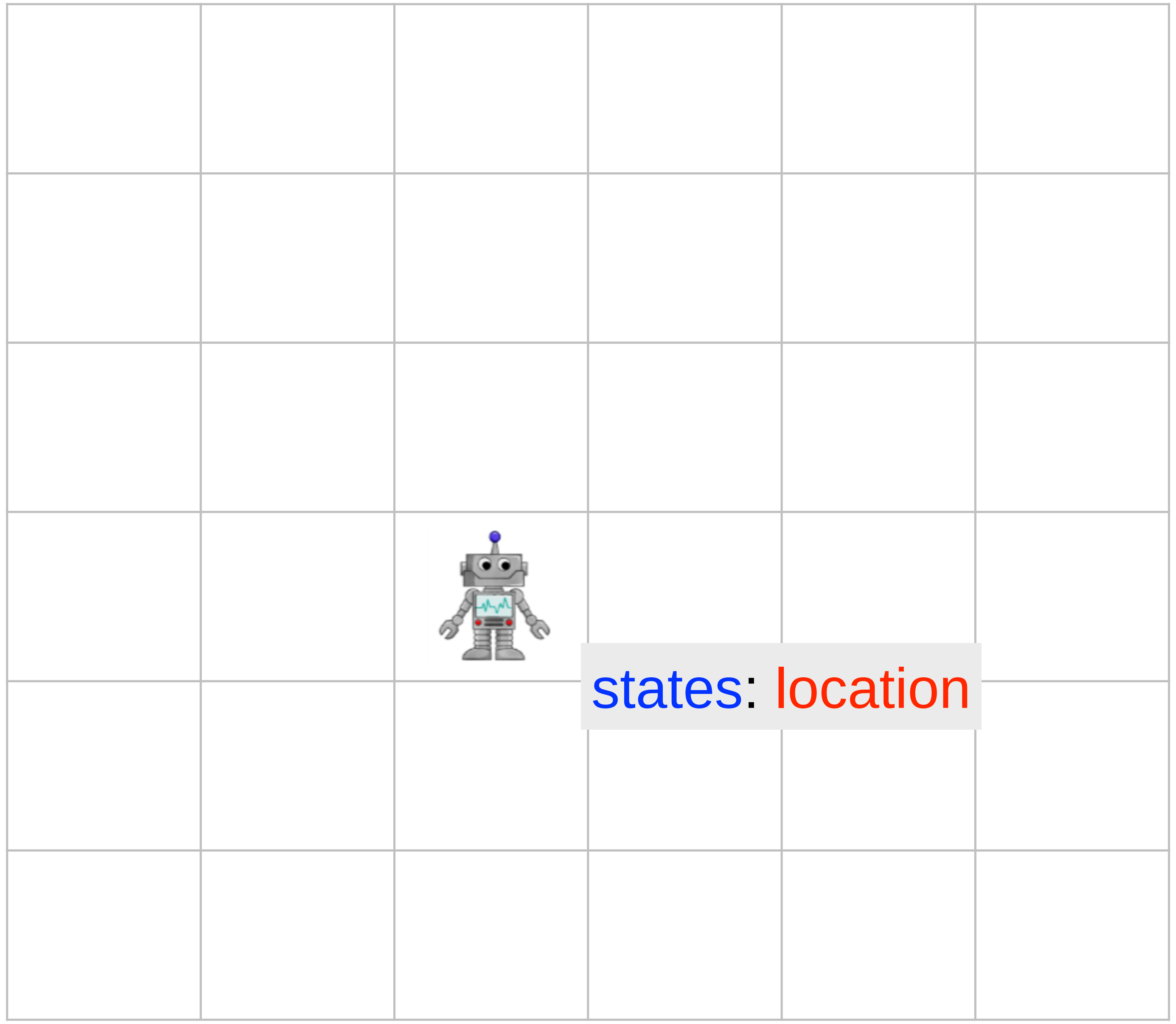
POMDPs are hard.
Uncertain POMDPs are harder.

We seem to have developed an approach that scales.

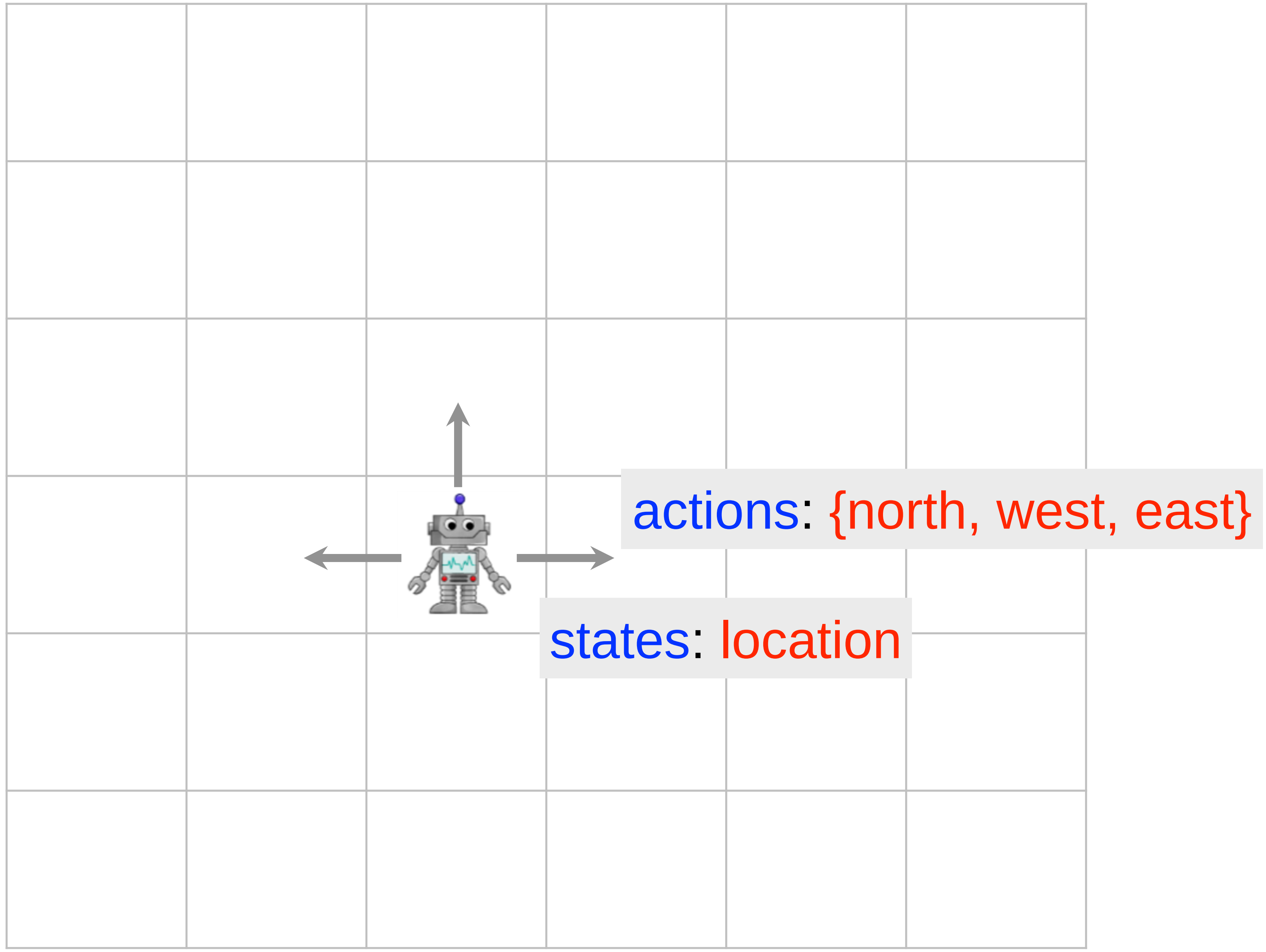
Demonstrated on spacecraft motion planning.



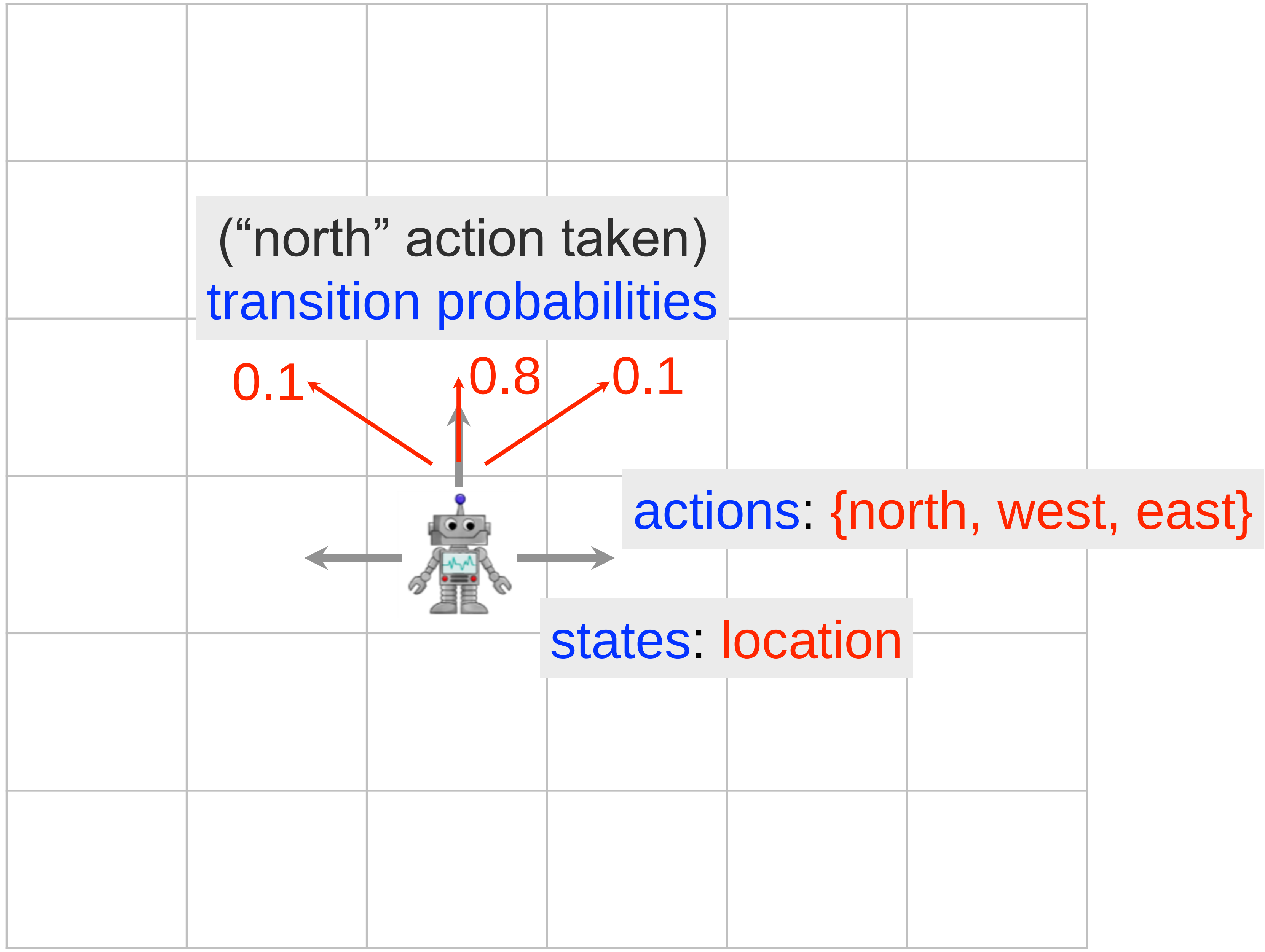
POMDPs through an example



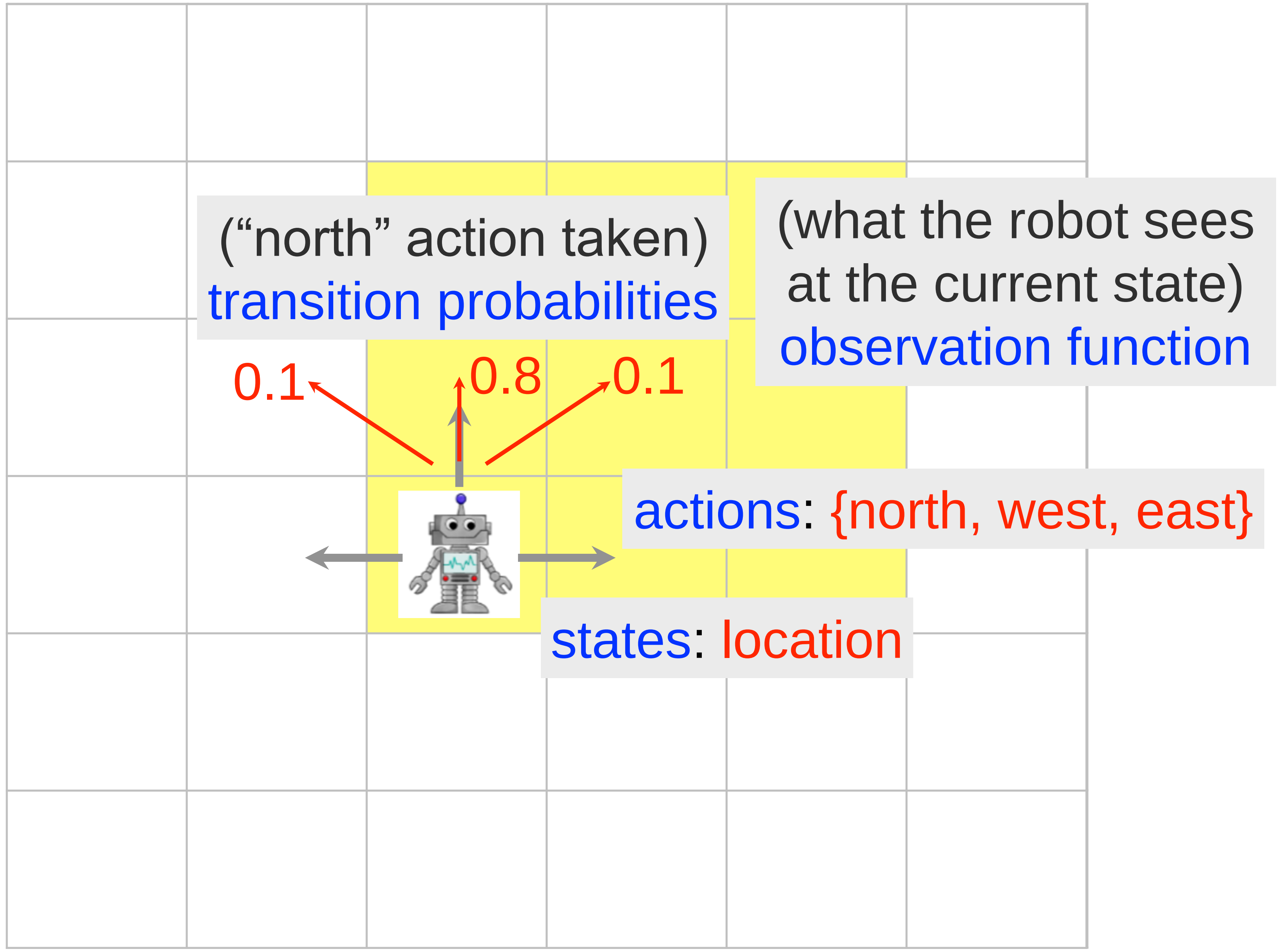
POMDPs through an example



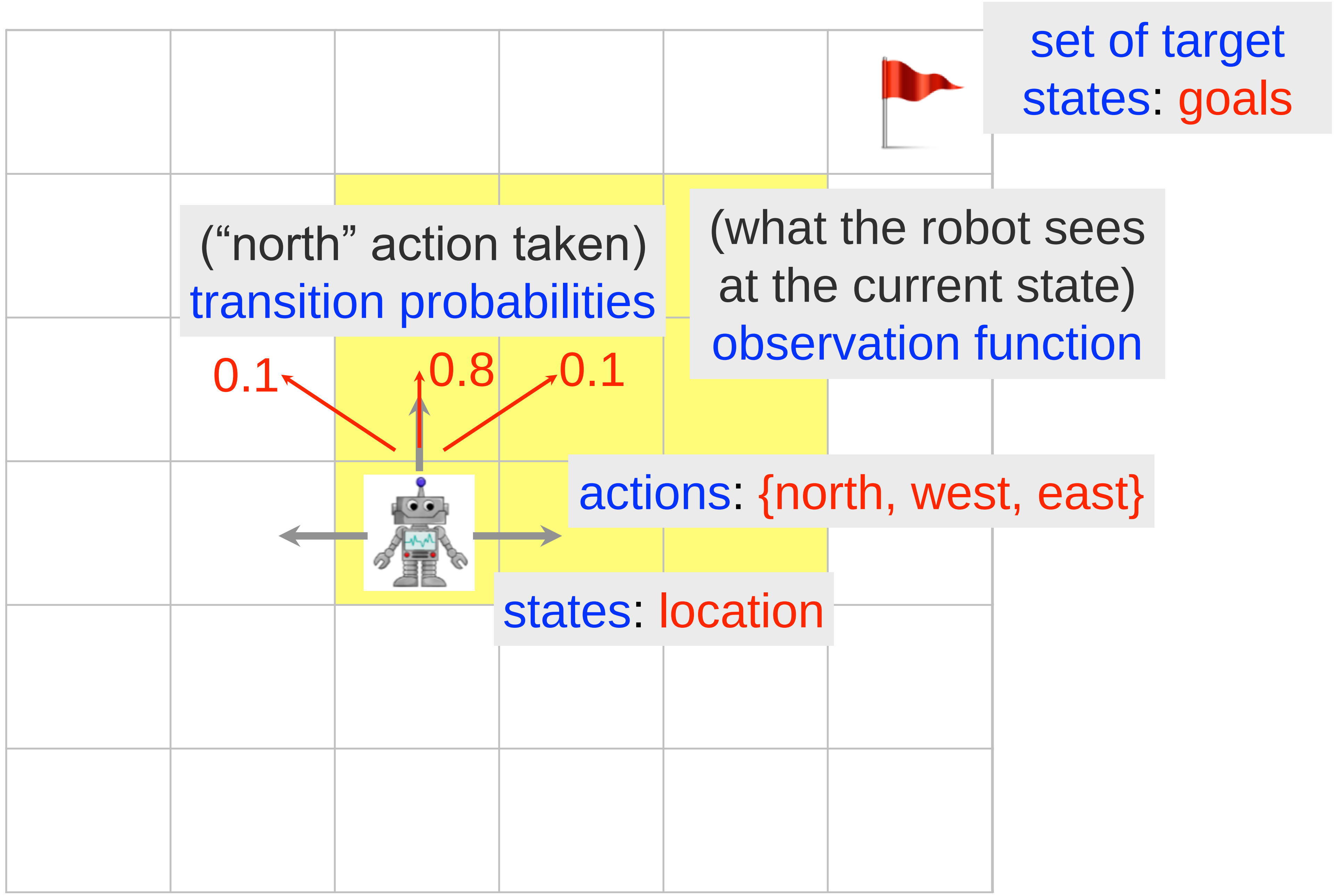
POMDPs through an example



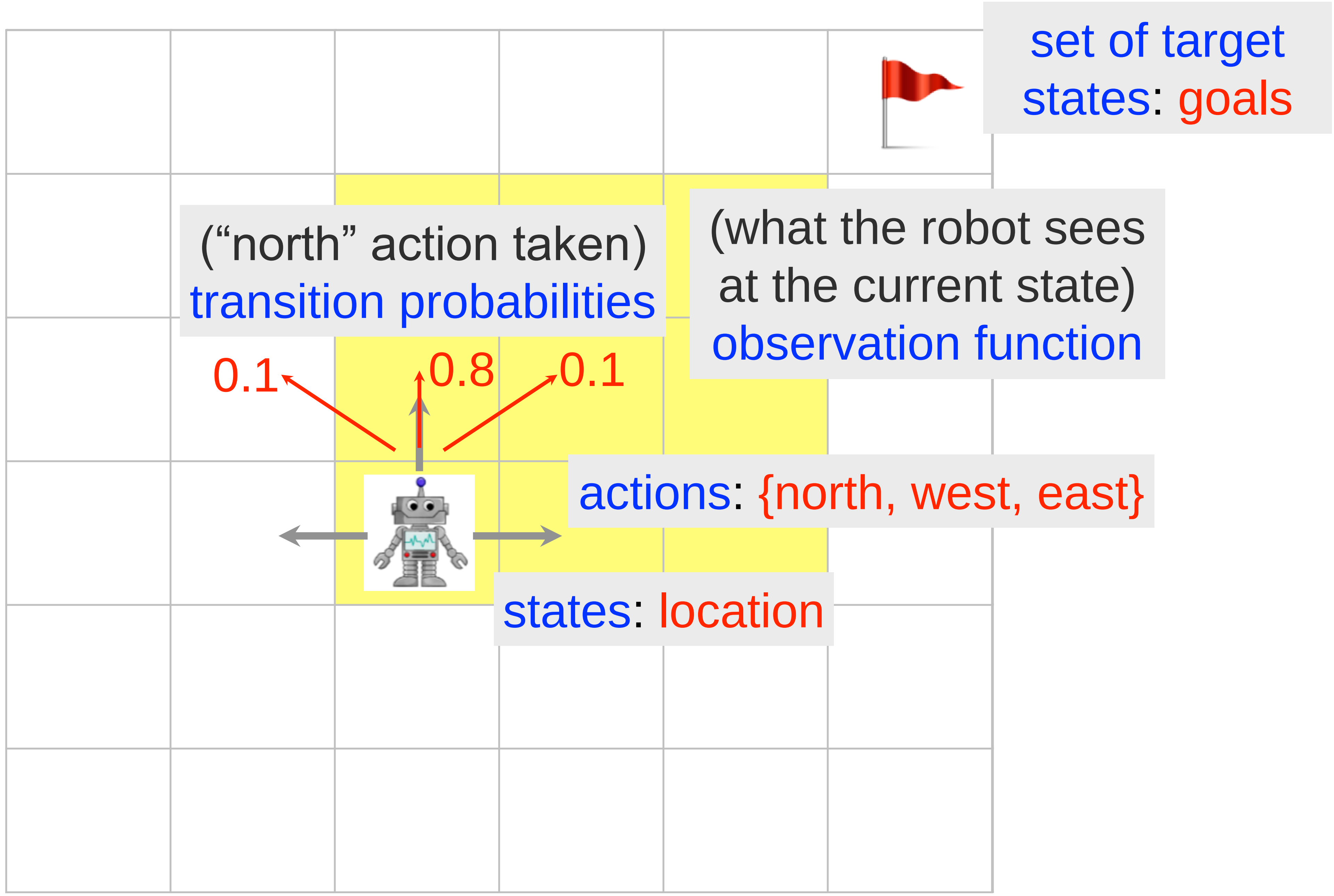
POMDPs through an example



POMDPs through an example

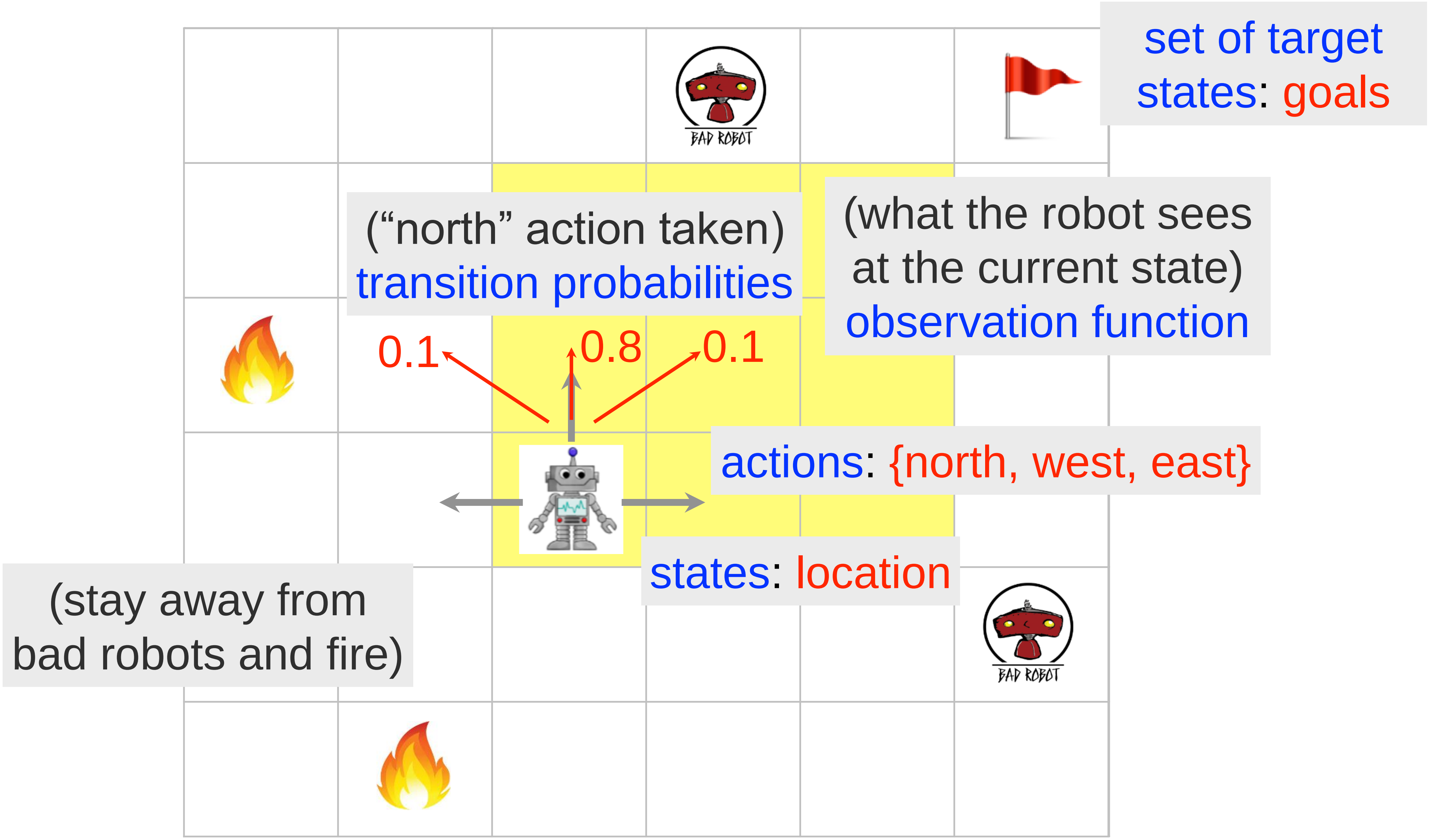


POMDPs through an example



cost: use of resources

POMDPs through an example



cost: use of resources

Partially Observable Markov Decision Processes

set of states
probabilistic transition function
observation function

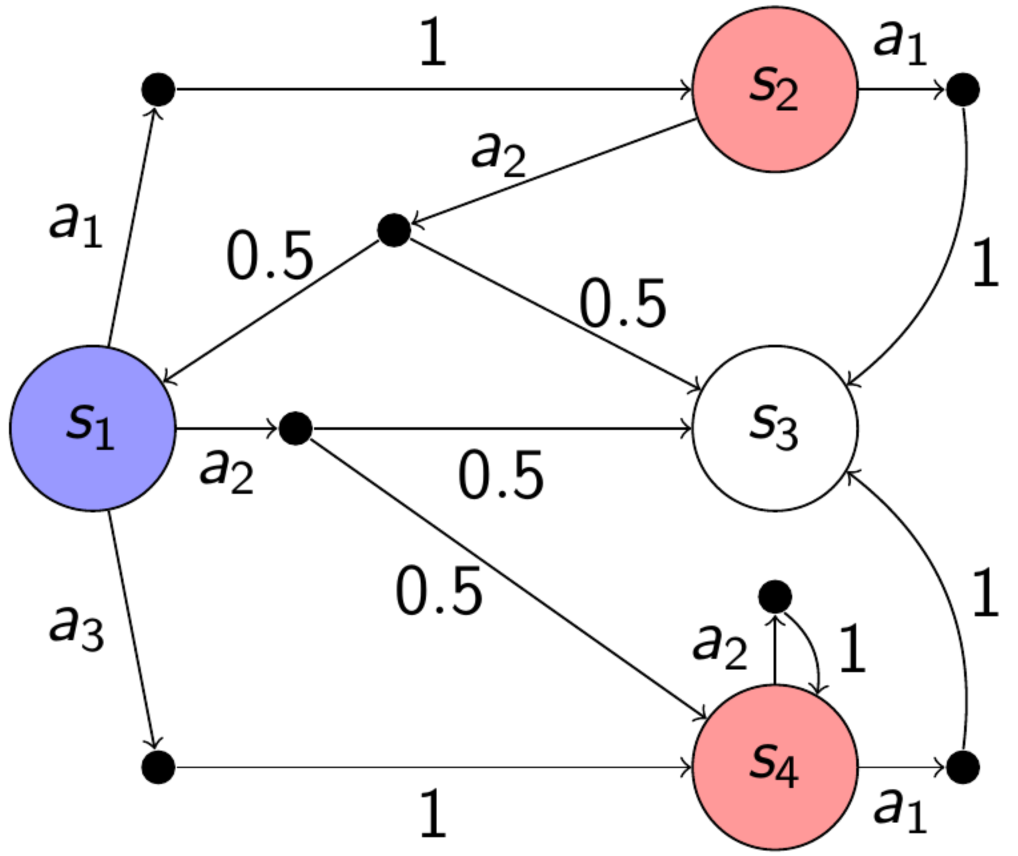
$$\mathcal{M} = (S, \quad A, \quad P, \quad Z, \quad O, \quad C)$$

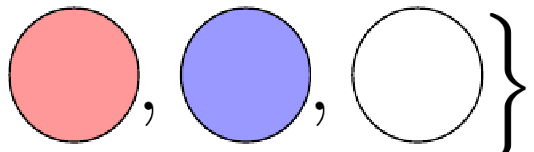
set of actions
set of observations
cost function

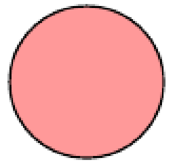
$$S = \{s_1, s_2, s_3, s_4\}$$

$$A = \{a_1, a_2, a_3\}$$

$$P(s_2, a_2, s_1) = 0.5$$



$Z =$ {}

$O(s_4) =$ 

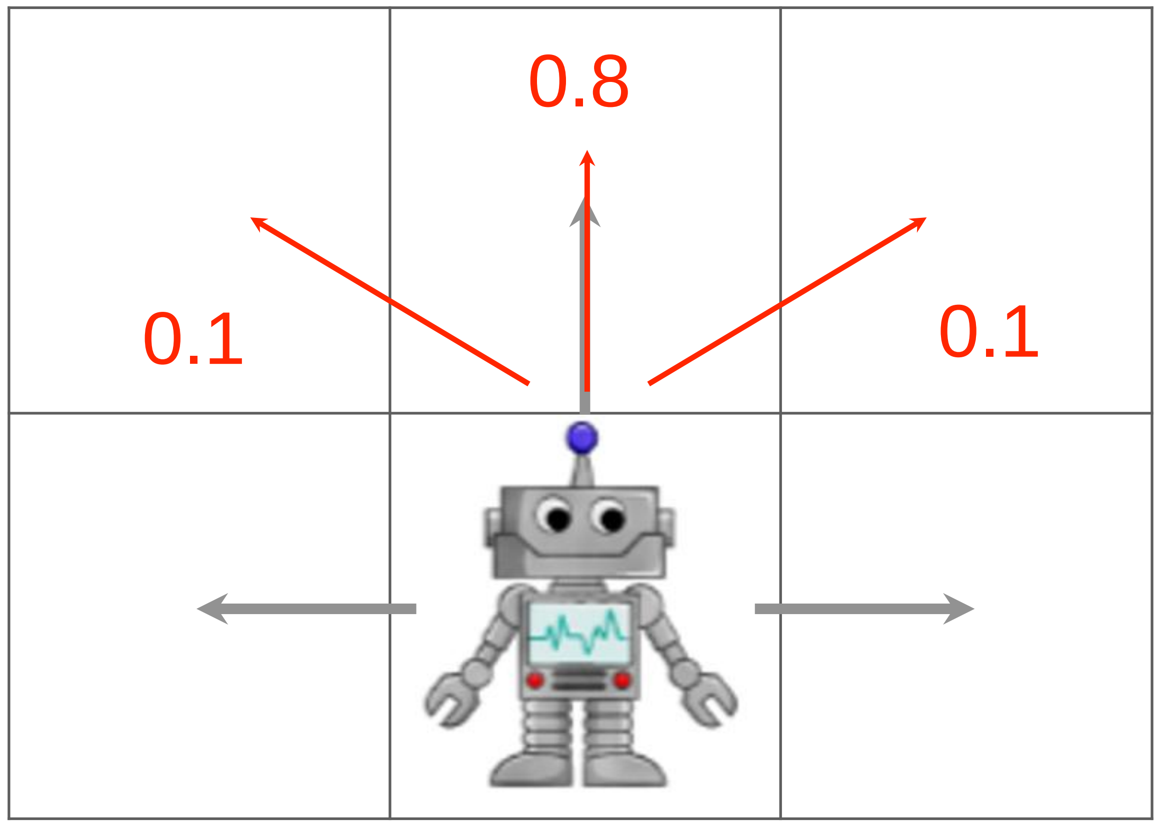
$$C(s_1, a_2) = 3$$

Uncertain POMDPs (uPOMDPs)

In a POMDP $\mathcal{M}$, transition function is assumed to be known

A uPOMDP $\mathcal{M}^{\mathcal{P}}$ extends a POMDP by allowing for probability intervals

transition function $P \in \mathcal{P}$ uncertainty set

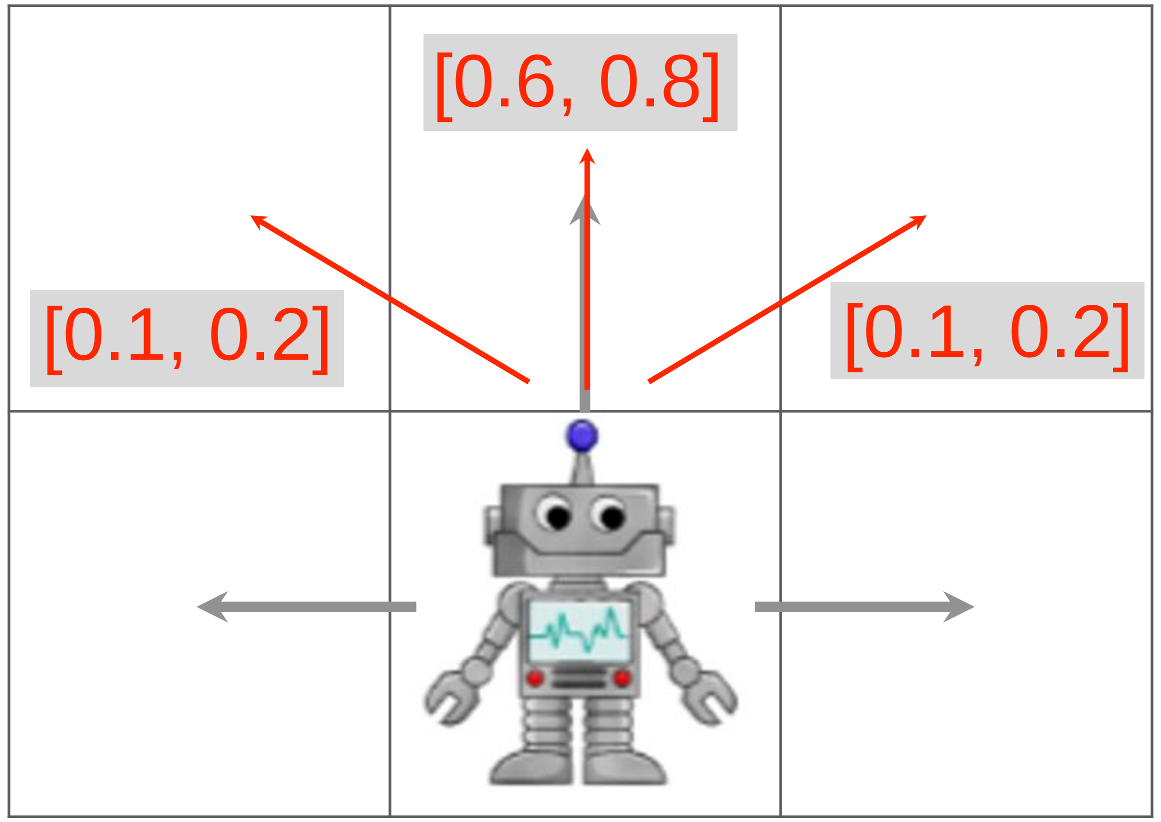


Uncertain POMDPs (uPOMDPs)

In a POMDP $\mathcal{M}$, transition function is assumed to be known

A uPOMDP $\mathcal{M}^{\mathcal{P}}$ extends a POMDP by allowing for probability intervals

transition function $P \in \mathcal{P}$ uncertainty set



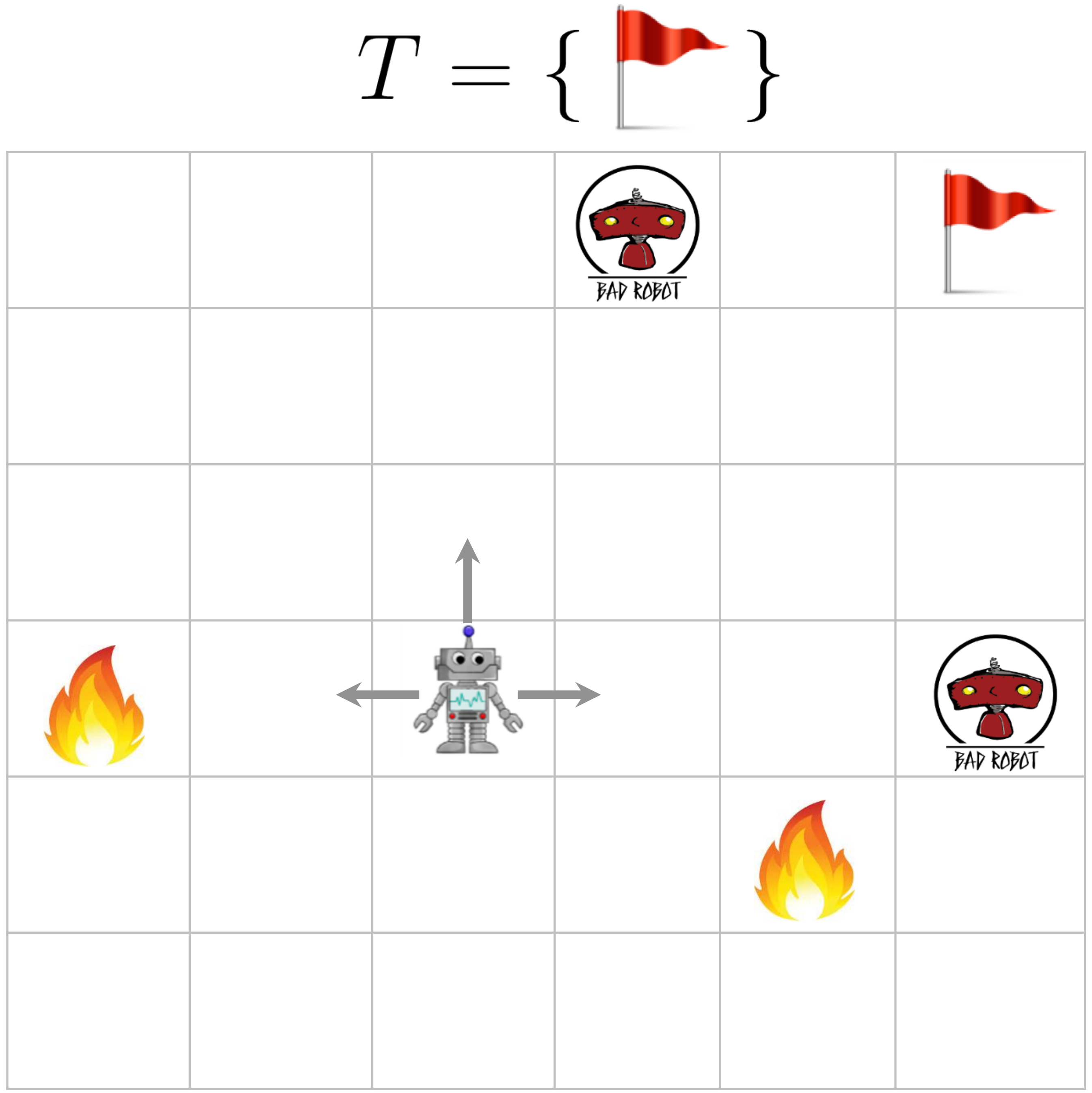
Specifications

Reachability specification

$$\varphi_{\mathrm{r}} = \mathbb{P}_{\geq\lambda}(\Diamond T)$$

the probability of reaching a state in the target set T is greater than λ

reach the flag while avoiding fire and bad robots



Specifications

Reachability specification

$$\varphi_{\mathrm{r}} = \mathbb{P}_{\geq \lambda}(\Diamond T)$$

the probability of reaching a state in the target set T is greater than λ

reach the flag while avoiding fire and bad robots

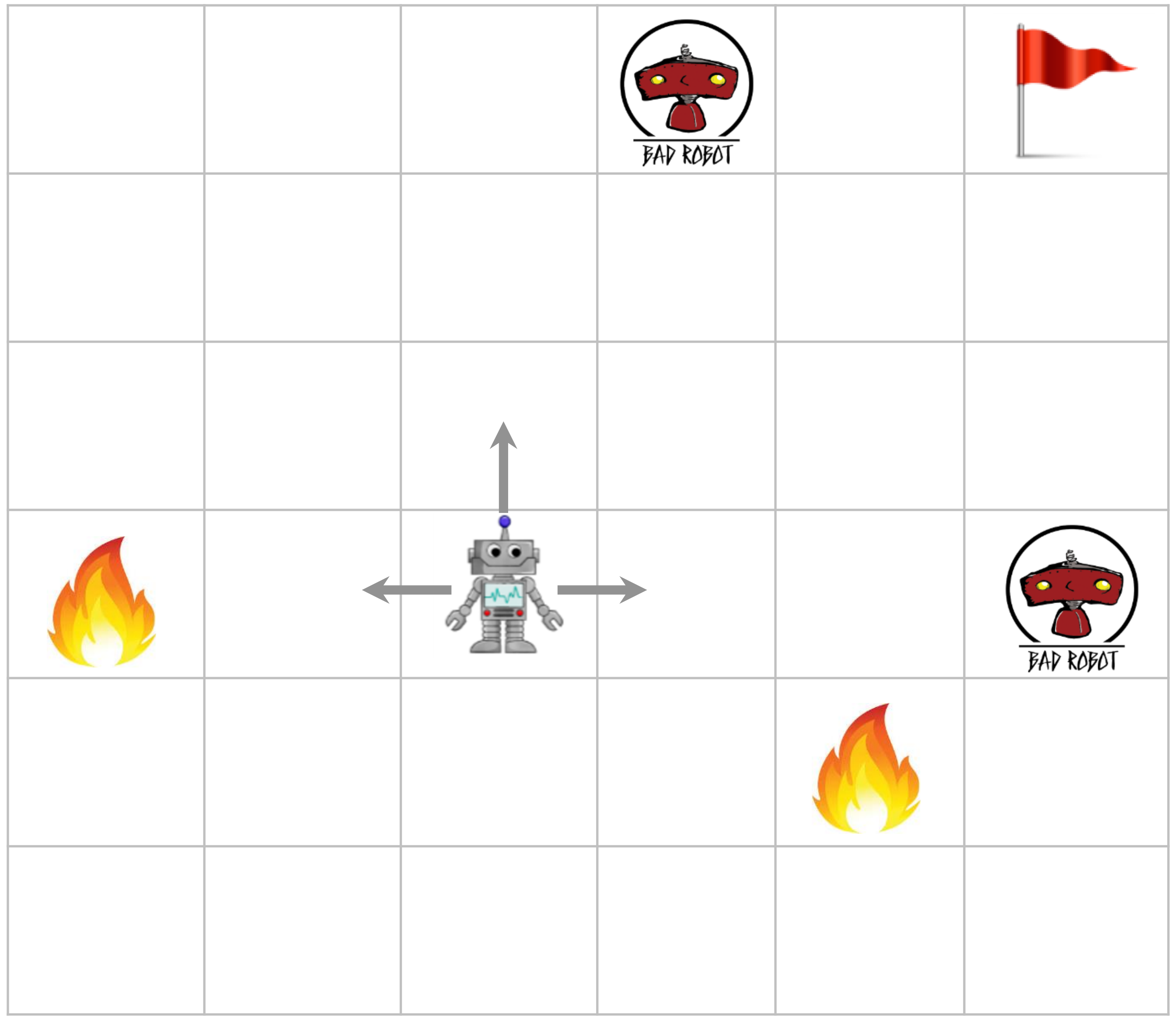
Performance specification

$$\varphi_{p} = \mathbb{E}_{\leq \kappa}(\Diamond T)$$

the expected accumulated reward before reaching a state in T is less than κ

minimize use of resources before reaching the flag

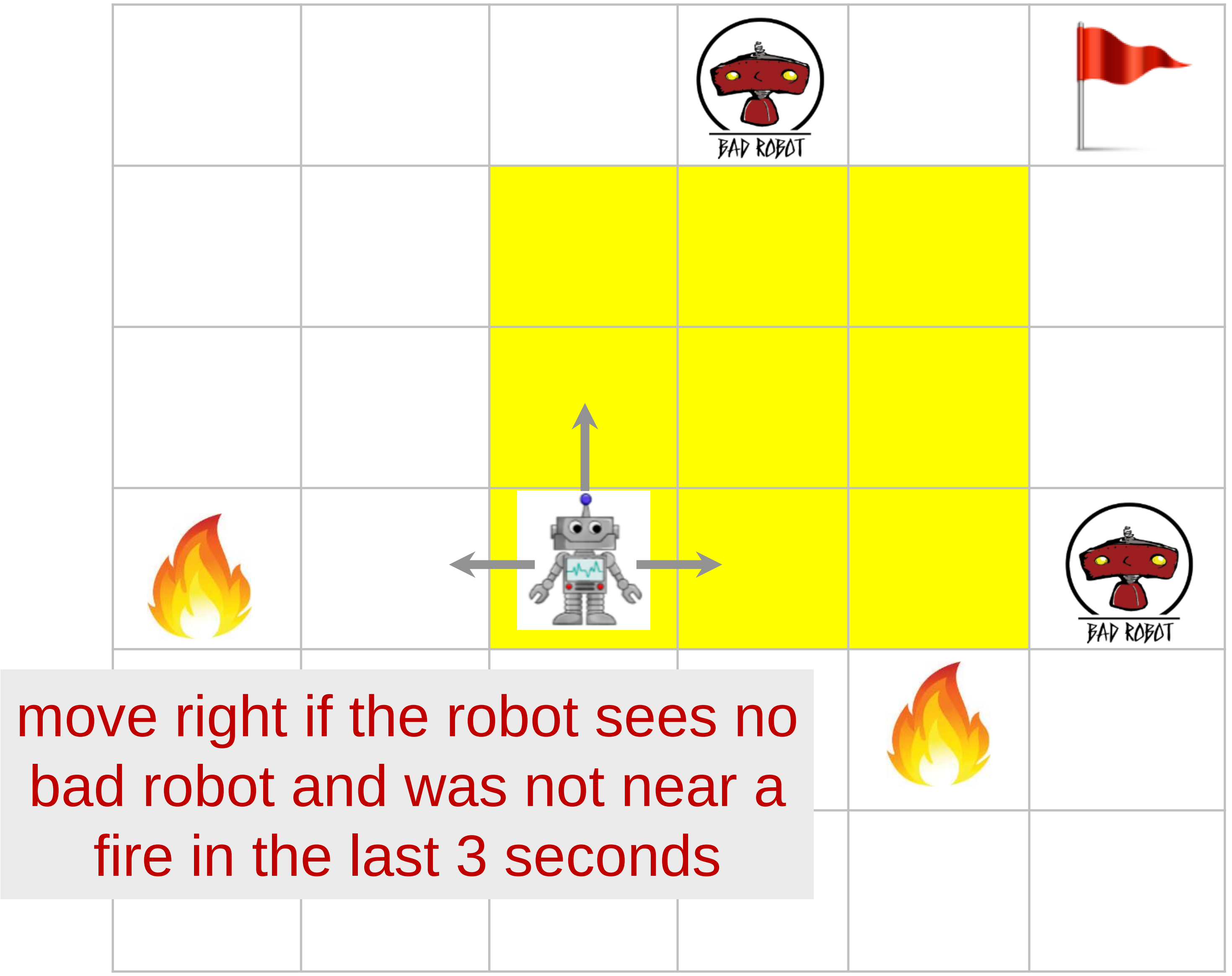
$T = \{$  $\}$



Policies

A policy σ for a uPOMDP maps sequences of observations and actions, to a distribution over actions

$$\sigma : \underbrace{(\overset{\text{observations}}{Z} \times \overset{\text{actions}}{A})^*}_{\text{memory}} \times \underset{\text{current observation}}{Z} \to \underset{\text{distribution over actions}}{Distr(A)}$$



Synthesis of Robust Policies in uPOMDPs

Given a uPOMDP $\mathcal{M}^{\mathcal{P}}$, compute a policy σ such that

transition function

induced uncertain Markov chain (uMC) $\mathcal{M}^{\mathcal{P}}_{\sigma} \models \varphi$ for all $P \in \mathcal{P}$

satisfies the specification

uncertainty set

Policy synthesis in uPOMDPs is hard(er)

Partial observability over the state of the agent makes policy synthesis in uPOMDPs computationally hard

- **Exponential** in the number of states, actions, and observations

Policy synthesis in uPOMDPs is hard(er)

Partial observability over the state of the agent makes policy synthesis in uPOMDPs computationally hard

- **Exponential** in the number of states, actions, and observations

Undecidable (policy requires infinite memory of observations)

Undecidable even when the transition function is known

The Main Ideas

Restriction to finite-memory policies
yields a decidable problem, **NP-hard** though

Dualization (of a **semi-infinite** optimization problem)
yields a **finite** (yet still **nonconvex**) problem

Linearization + verification
yields a **finite, convex** problem

Overview of the Algorithm

Problem Input

uPOMDP $\mathcal{M}^{\mathcal{P}}$

Specification φ

Candidate policy σ

Overview of the Algorithm

Problem Input

uPOMDP $\mathcal{M}^{\mathcal{P}}$
Specification φ
Candidate policy σ

Reformulate

(exact)

Semi-infinite nonconvex problem

cost variables

semi-infinite constraints **(intractable)**

$$\boxed{c_s \geq C(s) + \sum_{s' \in S} P(s, s') \cdot c_{s'} \quad \forall P \in \mathcal{P}}$$

uncountable uncertainty set

finite nonconvex constraints **(intractable)**

$$\boxed{c_s \geq \sum_{\alpha \in A(s)} \left(C(s, \alpha) + \sigma_{s,\alpha} \cdot \sum_{s' \in S} \mathcal{P}(s, \alpha, s') \cdot c_{s'} \right)}$$

policy variables

nonconvex function over the variables

Overview of the Algorithm

Problem Input
uPOMDP $\mathcal{M}^{\mathcal{P}}$
Specification φ
Candidate policy σ

$\xrightarrow[\text{(exact)}]{\text{Reformulate}}$

Semi-infinite nonconvex problem

$\xrightarrow[\text{(exact)}]{\text{Dualize}}$

Finite nonconvex problem

finite affine constraints **(tractable)**

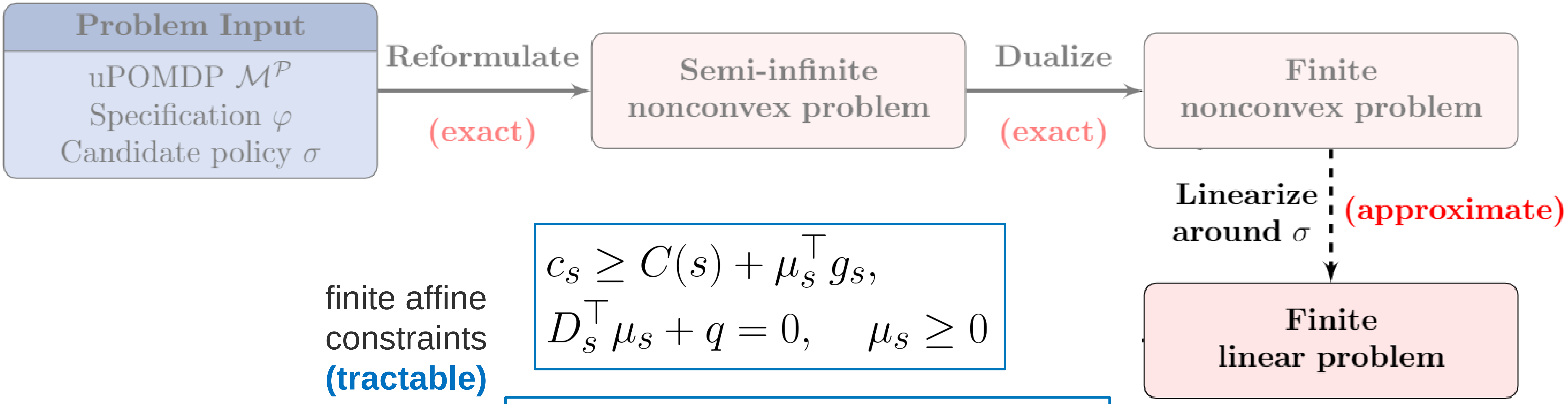
$$\boxed{\begin{array}{l} c_s \geq C(s) + \mu_s^\top g_s, \\ D_s^\top \mu_s + q = 0, \qquad \mu_s \geq 0 \end{array}}$$

finite nonconvex constraints **(intractable)**

$$\boxed{c_s \geq \sum_{\alpha \in A(s)} \Big(C(s, \alpha) + \sigma_{s,\alpha} \cdot \underbrace{\sum_{s' \in S} \mathcal{P}(s, \alpha, s') \cdot c_{s'} \Big)}_{\text{nonconvex function over the variables}}}$$

policy variables

Overview of the Algorithm



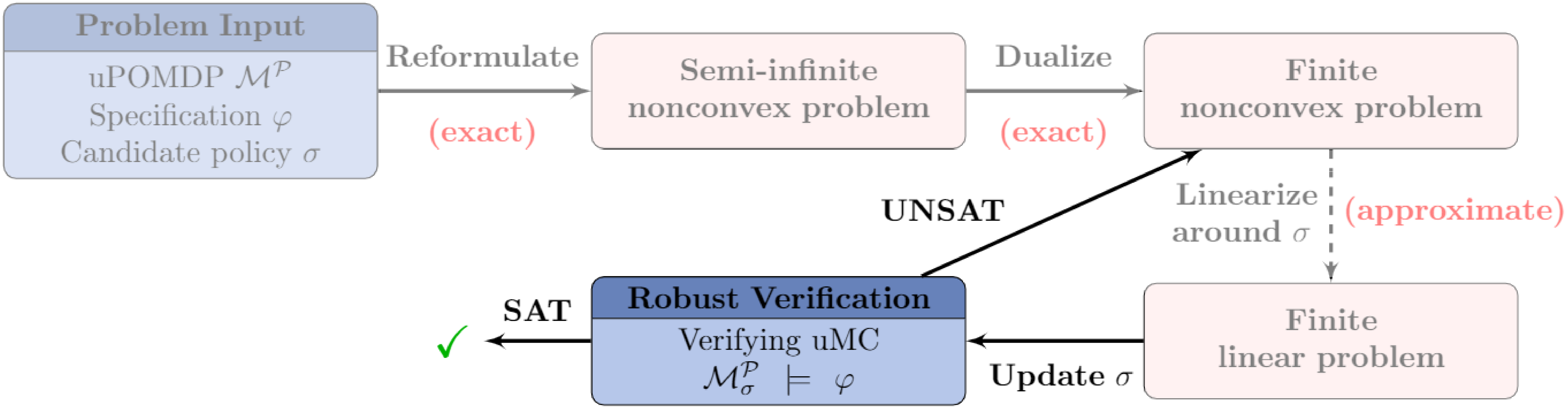
finite affine constraints **(tractable)**

$$\boxed{\begin{aligned} &c_s \geq C(s) + \mu_s^\top g_s, \\ &D_s^\top \mu_s + q = 0, \quad \mu_s \geq 0 \end{aligned}}$$

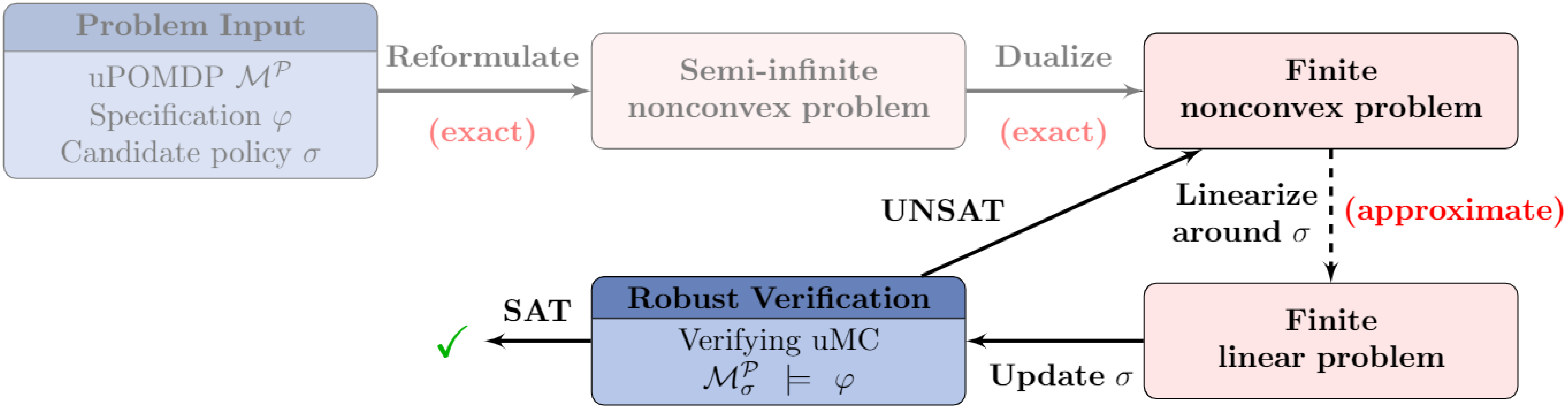
$$\boxed{c_s \geq \sum_{\alpha \in A(s)} \Big(C(s,\alpha) + h_{\text{aff}}(s,\alpha,s') \Big)}$$

affine approximation of the nonconvex function

Overview of the Algorithm



Overview of the Algorithm

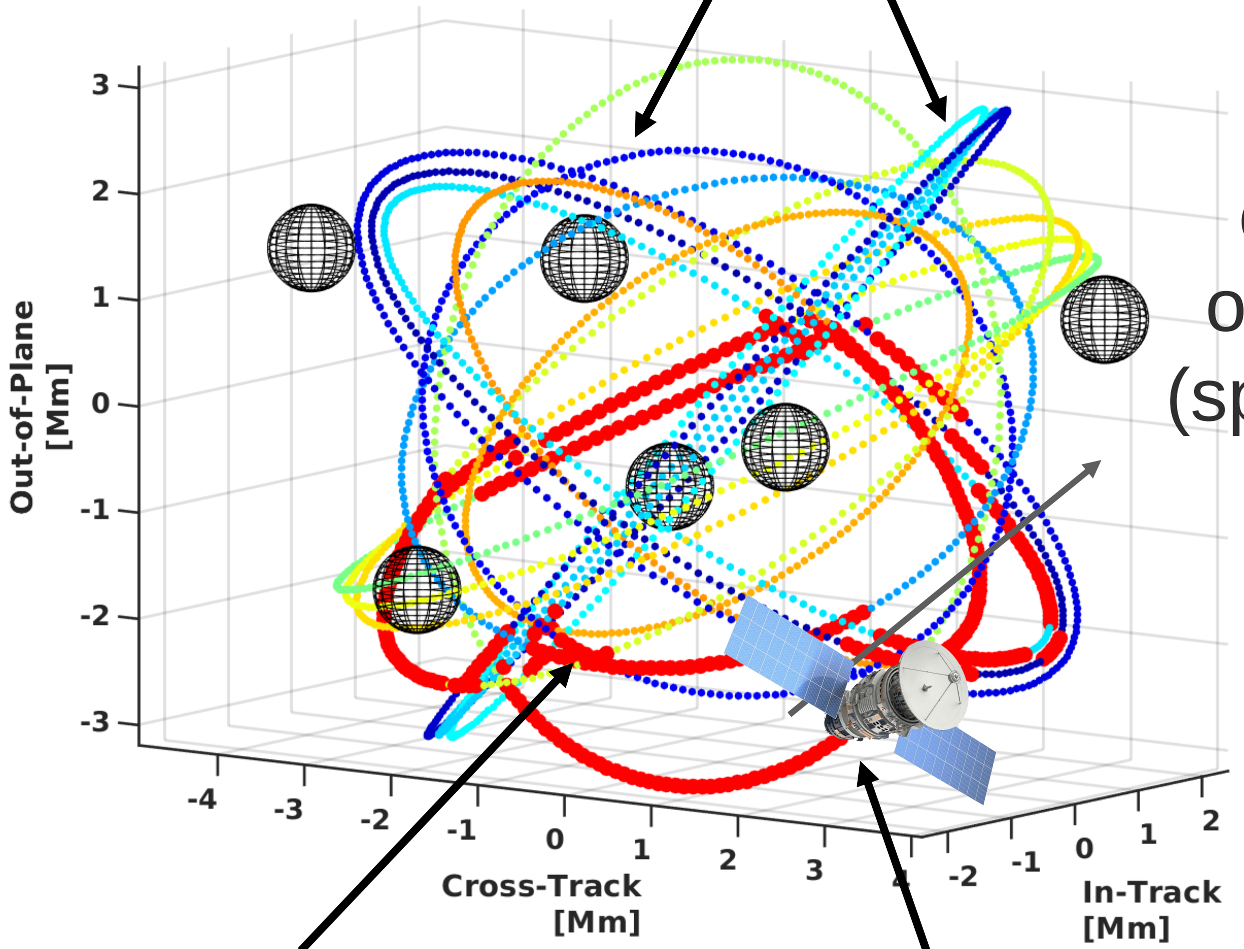


Spacecraft Motion Planning (with operator in the loop)

Problem:
Switching between orbits is possible if the orbits are close to each other

Partial observability over the current position of spacecraft, uncertainty on the location of other objects and operator

dotted curves: possible orbits

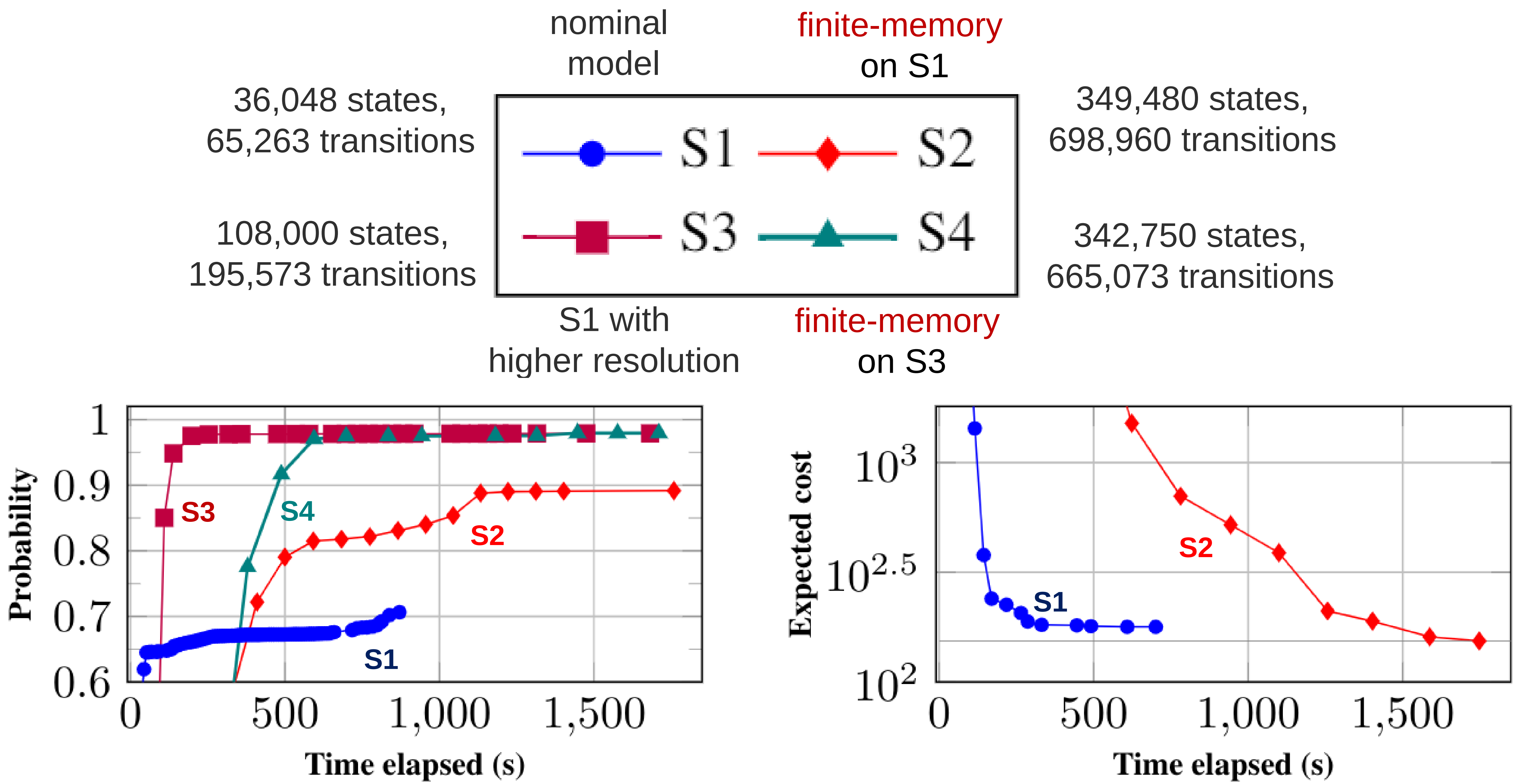


other objects (spheres)

trajectory of the spacecraft (shown in red)

spacecraft (controlled)

Results on Spacecraft Motion Planning

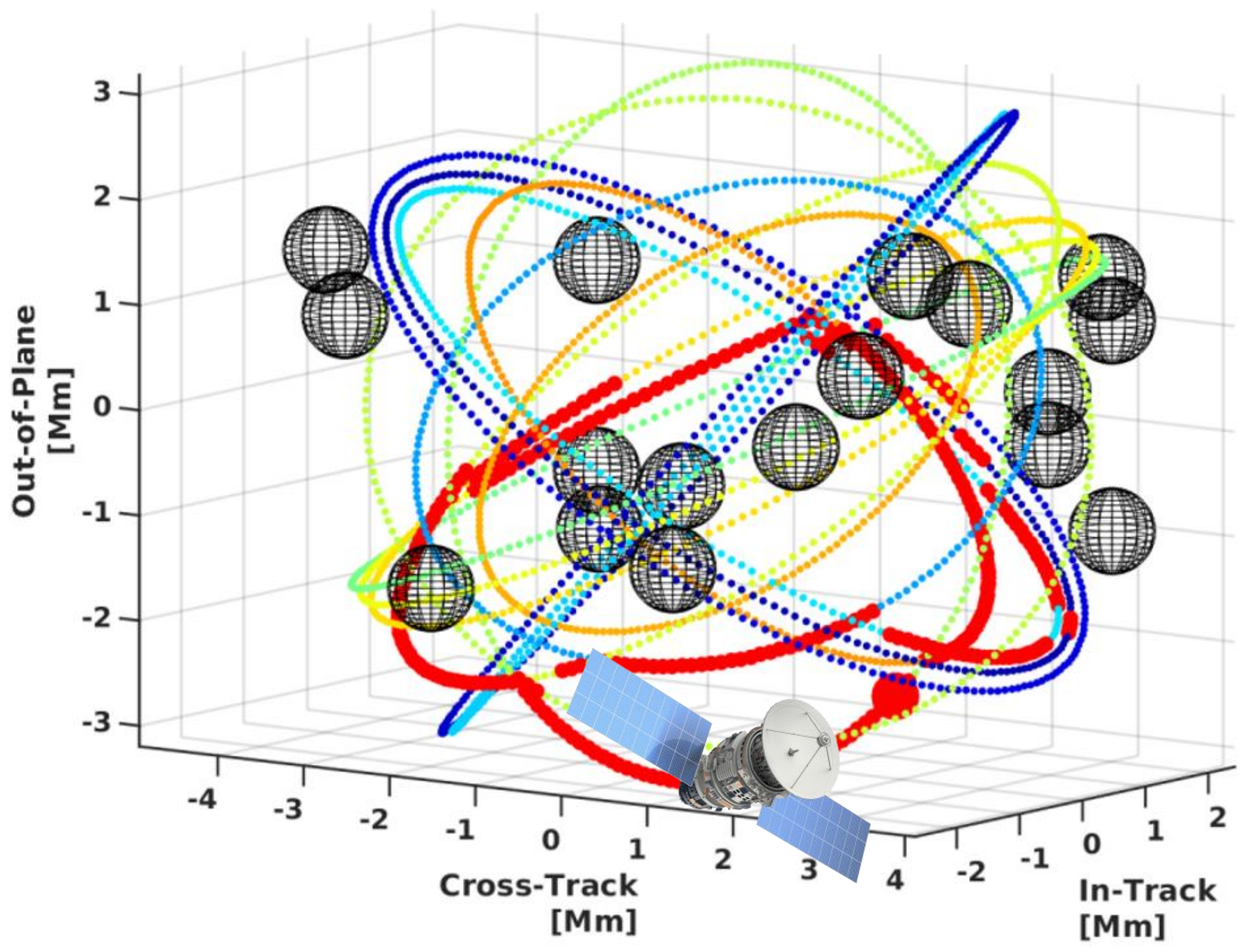


Can solve uPOMDPs with **hundreds of thousands states in minutes**

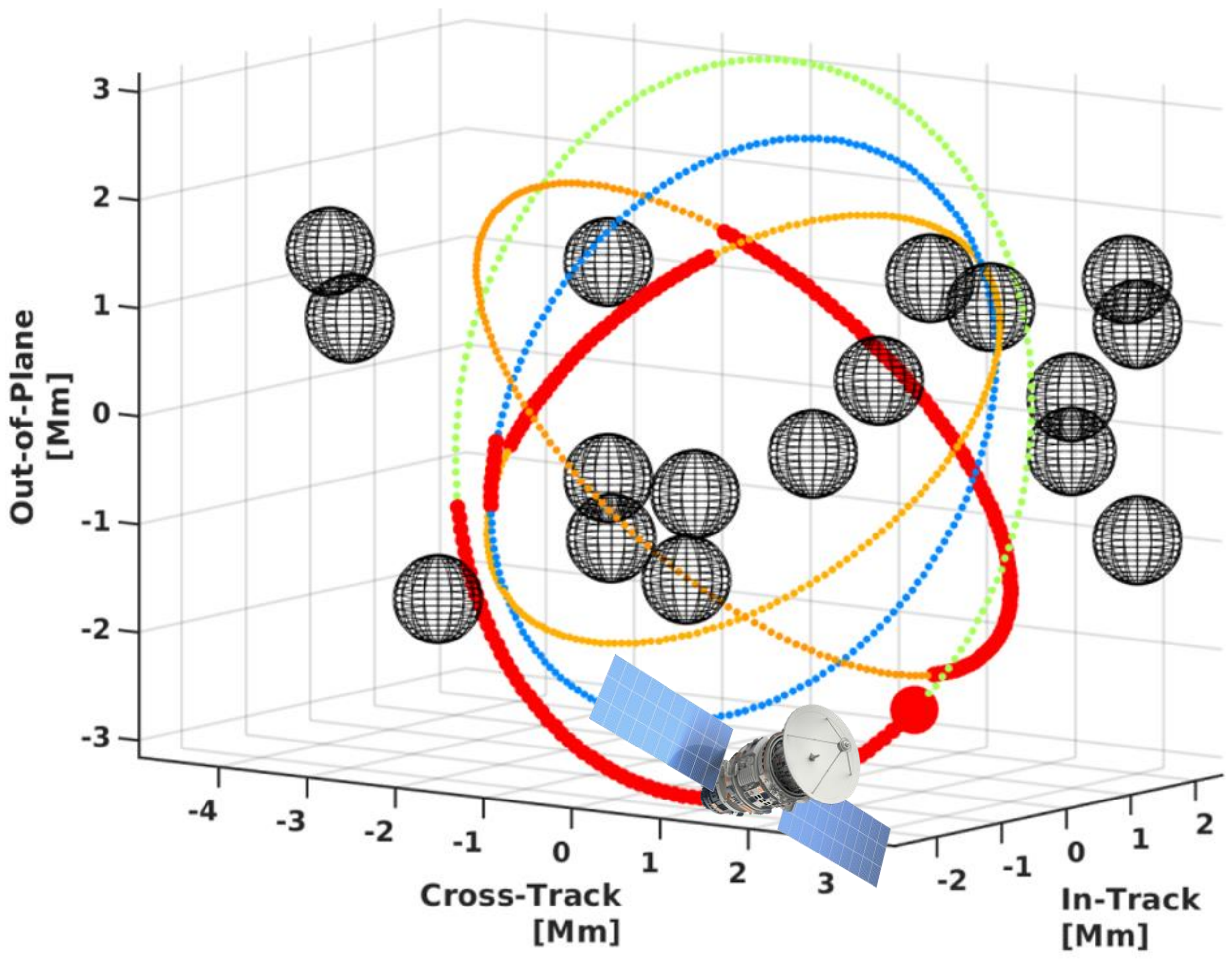
Models with memory take a **longer time to solve** but **yield better performance**

Results with Finite-Memory Policies

The memoryless policy (S1) makes more orbit changes than the policy with 5 memory nodes (S2)



memoryless policy (S1)



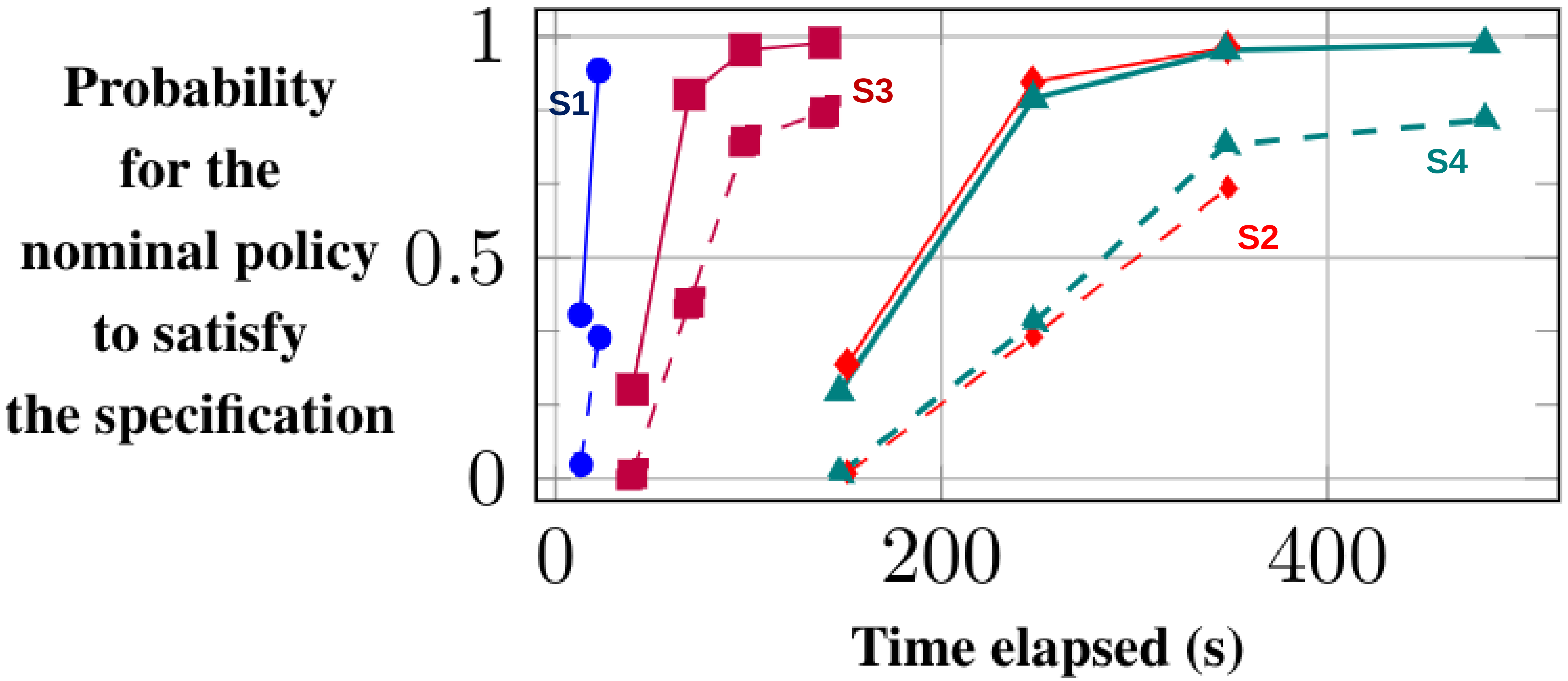
finite-memory policy (S2)

Robust Policies are Indeed More Robust

compute a policy
on the nominal POMDP
(solid lines)

AND

apply to the
uPOMDP
(**dashed lines**)



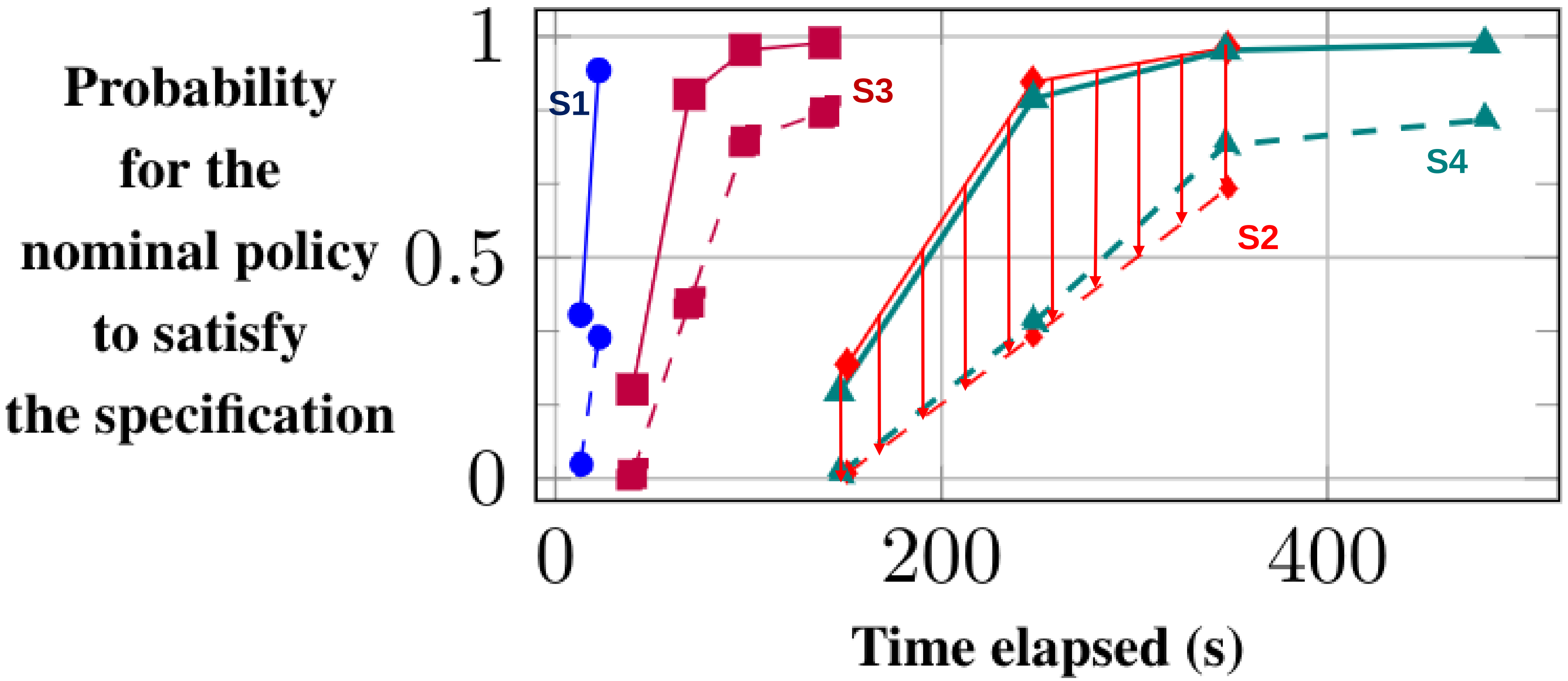
The performance of the nominal policies can **reduce up to 60 percentage points under uncertainty**

Robust Policies are Indeed More Robust

compute a policy
on the nominal POMDP
(solid lines)

AND

apply to the
uPOMDP
(dashed lines)



The performance of the nominal policies can **reduce up to 60 percentage points under uncertainty**

Conclusions and Future Work

Developed algorithms that scale to uPOMDPs that are **3 orders of magnitude larger** than previous approaches

Future work:

Uncertainty sets with correlations between different states

Incorporate these algorithms for safety in reinforcement learning