

Rule-based Shielding for Partially Observable Monte-Carlo Planning

Giulio Mazzi, Alberto Castellini, Alessandro Farinelli

Università degli studi di Verona,
 Dipartimento di Informatica,
 Strada Le Grazie 15, 37134, Verona, Italy
 giulio.mazzi@univr.it, alberto.castellini@univr.it, alessandro.farinelli@univr.it

Abstract

Partially Observable Monte-Carlo Planning (POMCP) is a powerful online algorithm able to generate approximate policies for large Partially Observable Markov Decision Processes. The online nature of this method supports scalability by avoiding complete policy representation. The lack of an explicit representation however hinders policy interpretability and makes policy verification very complex. In this work, we propose two contributions. The first is a method for identifying unexpected actions selected by POMCP with respect to expert prior knowledge of the task. The second is a shielding approach that prevents POMCP from selecting unexpected actions. The first method is based on Satisfiability Modulo Theory (SMT). It inspects traces (i.e., sequences of belief-action-observation triplets) generated by POMCP to compute the parameters of logical formulas about policy properties defined by the expert. The second contribution is a module that uses online the logical formulas to identify anomalous actions selected by POMCP and substitutes those actions with actions that satisfy the logical formulas fulfilling expert knowledge. We evaluate our approach on Tiger, a standard benchmark for POMDPs, and a real-world problem related to velocity regulation in mobile robot navigation. Results show that the shielded POMCP outperforms the standard POMCP in a case study in which a wrong parameter of POMCP makes it select wrong actions from time to time. Moreover, we show that the approach keeps good performance also if the parameters of the logical formula are optimized using trajectories containing some wrong actions.

1 Introduction

Planning in partially observable environments while satisfying safety guarantees is a challenging problem. *Partially Observable Markov Decision Processes (POMDPs)* (Cassandra, Littman, and Zhang 1997) is a popular framework to model systems with uncertainty. Computing an optimal solution for POMDPs is very hard (i.e., PSPACE-complete (Papadimitriou and Tsitsiklis 1987)). However, it is possible to compute an approximate solution, and state-of-the-art algorithms achieve great performance in real-world instances of POMDPs. A pioneering algorithm for this purpose is *Partially Observable Monte-Carlo Planning (POMCP)* (Silver and Veness 2010) which uses a particle

filter to represent the belief and a Monte-Carlo Tree Search based strategy to compute the policy online. The online nature of the policy, however, makes the task of analyzing the decisions taken by POMCP very difficult (Castellini et al. 2020; Castellini, Marchesini, and Farinelli 2020; Castellini, Chalkiadakis, and Farinelli 2019). In general, with a high number of particles POMCP yields great performance, but sometimes the simulation does not properly assess the risk of certain actions, especially if the number of particles used in the simulation is limited due to engineering constraints (e.g., time limits, agents with limited computation capabilities or the model of the real environment is approximated). Moreover, in POMCP the policy is never fully computed or stored, hence it is very difficult to identify the reasons for possible unexpected decisions of the system. Explainability (Gunning 2019) is becoming a key feature of artificial intelligence systems since in these contexts humans need to understand why specific decisions are taken by the agent. Specifically, explainable planning (XAIP) (Fox, Long, and Magazzeni 2017; Cashmore et al. 2019) focuses on explainability in planning methods. The presence of erroneous behaviors in these tools (due, for instance, to the wrong setup of internal parameters) may have a strong impact on autonomous cyber-physical and robotic systems that interact with humans, and detecting these errors in automatically generated policies is very hard in practice.

In this work, we propose a methodology for generating a safety mechanism from high-level descriptions of the desired behavior of a POMCP-generated policy. In this approach, a human expert provides qualitative information on a property of the system, enriched with an indication of the expected behavior that the system should have in specific situations (e.g., “the robot should move fast if it is highly confident that the path is not cluttered, I expect this level of confidence to be above 90%”). With this information, our methodology analyzes a set of execution traces of the system and provides quantitative details of these statements by analyzing the execution of the system (e.g., “the robot moves fast if its confidence of being in an uncluttered segment is at least 93.4%”). The approach we propose formalizes the problem of parameters computation as a *MAX-SMT* problem which allows to express complex logical formulas and to compute optimal assignments when the template is not fully satisfiable (which happens in the majority of cases in

real policy analysis). This quantitative answer is then used to build a shield, namely a safety mechanism that forces the POMCP to satisfy the constraints expressed by the expert. The shield works alongside the Monte-Carlo Tree Search, by preemptively blocking some actions that, according to the rules, should not be selected in some situations. The shield is also enriched with a mechanism that quantifies how much the current belief is far from satisfying the rule. This allows some extra flexibility in filtering the beliefs that are acceptable by the shield, an important requirement for real-time algorithms that can only work with partial knowledge.

To empirically evaluate the performance of the shield, we test it in two domains, namely, the well-known *Tiger* problem and a robotic problem in which a mobile platform must move as fast as possible in a cluttered environment avoiding collisions. To test the robustness of the shield mechanism, we inject an error into POMCP by wrongly setting one of its parameters. This error subtly affects the decisions of the policy keeping almost unchanged the average performance but making some decisions non-optimal and therefore unexpected by experts (i.e., unsafe). Our experiments show that the shielding mechanism can improve performance while using a compact and expressive representation of expected properties. We also show that the use of the shield degrades performance only negligibly in practice.

In summary, the contribution of this paper to the state-of-the-art is threefold:

- we propose an SMT-based methodology that combines a logic-based description of a system with the real execution traces of a POMCP policy to create a set of rules describing expected behaviors of an agent;
- we propose a method for generating a shield from this set of rules to block unexpected actions;
- we empirically evaluate the shielding mechanism in two domains showing that it can exploit the knowledge provided by the expert to achieve higher performance than standard POMCP when its parameters are imprecise.

The paper is structured as follows. In section 2 we present related work. In section 3 provide background knowledge. In section 4 we describe our methodology, and in Section 5 we show the performance of our method. Finally, in Section 6 we summarize our contribution and we provide direction for future development.

2 Related Work

This work is mainly related to two other fields in the literature, namely, the verification of POMDPs policies and explainable planning. A possible approach for policy verification consists of encoding the POMDP problem into a standard logic-based framework, such as those presented in (Cashmore et al. 2016; Norman, Parker, and Zou 2017; Wang, Chaudhuri, and Kavvaki 2018; Bastani, Pu, and Solar-Lezama 2018), and then prove property guarantees formally using an SMT-solver. However, these frameworks present scalability problems due to the computational complexity of solving large instances of SMT. In our work, we add a shielding mechanism on top of POMCP preventing performance degradation in large problems. This is reached

by combining the logic-based representation of the problem with the highly efficient online approach provided by POMCP.

In (Zhu et al. 2019) verification is also achieved by exploiting a simplified representation of the problem provided by experts. The method verifies properties related to the safety of fully observable systems modeled by Markov Decision Processes (MDPs). It then works on a pre-trained neural network representing a black-box policy and uses a linear formula summarizing the policy behavior to use off-the-shelf verification tools. This differs from our approach because we work directly on partially observable environments, hence we explicitly consider beliefs (instead of states) in our logic-based rules describing the expected behavior of the policy. Moreover, our methodology is able to work also on subsets of actions, and it does not require rules describing the dynamic of each action.

Another approach for verifying POMCP properties is presented in (Newaz, Chaudhuri, and Kavvaki 2019), it uses *statistical model checking (SMC)* to verify that qualitative objectives, specified on possible states of the environment, are satisfied with a certain confidence level. This is different from our methodology since we specify property on beliefs instead of on states. SMC requires also a large number of particles to generate reliable results while our methodology does not have such a requirement.

The need for a shielding mechanism based on expert expectation for highly complex planning approaches is highlighted by the growing interest in the field of *eXplainable Artificial Intelligence (XAI)* (Gunning 2019), a rapidly growing research field focusing on human interpretability and understanding of artificial intelligence (AI) systems. In particular, our work is related to explainable planning (XAIP) (Cashmore et al. 2019; Fox, Long, and Magazzeni 2017; Langley et al. 2017; Anjomshoae et al. 2019) that aims at developing planning tools whose decisions are understandable by human beings. In our work, we use the high-level insight provided by the user as a guide for building rules that summarize the expected policy behavior. A particularly interesting kind of questions analyzed in XAIP is known as *contrastive questions* (Fox, Long, and Magazzeni 2017). An example is “Why have you made decision d_1 instead of d_2 , that I expect to be a better option?”. An explainable system should be able to provide a human-comprehensible answer motivating its choice and providing human-understandable evidence that it is better than the alternative option. These questions are, however, very difficult to answer in online frameworks, as POMCP, because the information required to build the answer may not be available to the agent at run time (Castellini et al. 2020). For instance, the fact that POMCP computes the policy only for beliefs encountered in its execution, and it does not generate an explicit representation of the overall policy (as other methods do with α -vectors for example) prevent the usage of methods that analyze this explicit representation of the policy to explain it. Our method does not use contrastive questions but it bases the interaction between human and planner on logical formulas that are local representations of the policy constrained by the expert’s prior knowledge and specialized by the SMT solver on previously

observed behaviors of the planner (contained in traces). A methodology focused on building explainable rules that describe as many of the decisions taken by the policy as possible is presented in (Mazzi, Castellini, and Farinelli 2021a). It identifies the decisions that do not satisfy the rule (i.e., unexpected decisions) to improve the explainability of the results. This differs from our work because it is an offline procedure that cannot be used to improve the performance of a POMCP algorithm. The integration of a rule-based explanation into a shield allows our method to identify in real-time decisions that violate user’s expectations and avoiding them improves the planner performance.

3 Background

In this section, we present a definition of the POMDP framework and the POMCP algorithm. We also briefly present the SMT problem and the MAX-SMT extension.

3.1 Partially Observable Monte-Carlo Planning

A Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, and Cassandra 1998) is a tuple $(S, A, O, T, Z, R, \gamma)$, where S is a set of partially observable states, A is a set of actions, Z is a finite set of observations, $T: S \times A \rightarrow \Pi(S)$ is the state-transition model, with $\Pi(S)$ probability distribution over states, $O: S \times A \rightarrow \Pi(Z)$ is the observation model, $R: S \times A \rightarrow \mathbb{R}$ is the reward function and $\gamma \in [0, 1]$ is a discount factor. An agent must maximize the discounted return $E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$. A probability distribution over states, called belief, is used to represent the partial observability of the true state. To solve a POMDP it is required to find a policy, namely a function $\pi: B \rightarrow A$ that maps beliefs B into actions.

In this work, we focus on *Partially Observable Monte-Carlo Planning (POMCP)* (Silver and Veness 2010) to solve POMDPs. POMCP is an *online* algorithm that solves POMDPs by using Monte-Carlo techniques. The strength of POMCP is that it does not require an explicit definition of the transition model, observation model, and reward. Instead, it uses a black-box to simulate the environment. POMCP uses a *Monte-Carlo Tree Search (MCTS)* at each time-step to explore the belief space and select the best action. *Upper Confidence Bound for Trees (UCT)* (Kocsis and Szepesvári 2006) is used as a search strategy to select the subtrees to explore and balance exploration and exploitation. The belief is implemented as a *particle filter*, which is a sampling over the possible states that is updated at every step. At each time-step, a particle (which represents a specific state of the POMDP) is selected from the filter. This corresponding state is used as an initial point to perform a simulation in the Monte-Carlo tree. Each simulation is a sequence of action-observation pairs and it collects a discounted return, and for each action, we can compute the expected reward that can be achieved. After receiving an observation, the particle filter is updated to reflect the new information available in the environment. If required, new particles can be generated from the current state through a process of *particle reinvention*. In the following, we call *trace* a set of runs

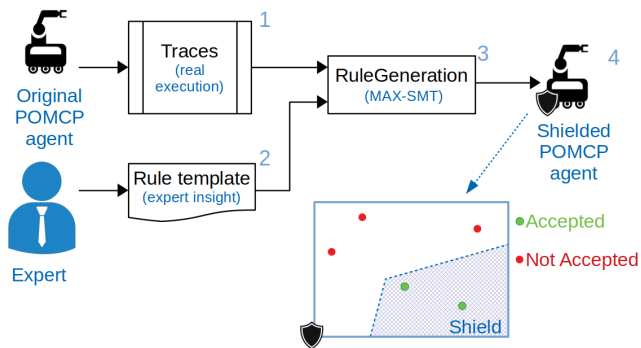


Figure 1: Methodology overview.

performed by POMCP on a specific problem. Each run is a sequence of *steps*, and each step corresponds to an action performed by the agent having a belief and receiving an observation from the environment.

3.2 SMT and MAX-SMT

The problem of reasoning on the satisfiability of formulas involving propositional logic and first-order theories is called *Satisfiability Modulo Theory (SMT)*. In our methodology, we use propositional logic and the theory of linear real arithmetic to encode the information provided by the expert, and we use Z3 (De Moura and Bjørner 2008) as an SMT-solver to compute the parameter of the shield that codify these ideas. Specifically, we encode our formulas as a MAX-SMT problem, which has two kinds of clauses, namely, *hard*, that must be satisfied, and *soft* that should be satisfied whenever possible. A model of the MAX-SMT problem hence satisfies all the hard clauses and as many soft clauses as possible, and it is unsatisfiable only when hard clauses are unsatisfiable. In our methodology, the expert provides a high-level explanation (that can include abstractions and local approximation) and it is intended to describe as many decisions as possible among those taken by the policy, hence MAX-SMT provides a perfect formalism to encode this requirement. The Z3 solver is used to solve the MAX-SMT problem (Bjørner, Phan, and Fleckenstein 2015). Subsection 4.2 presents the details of this encoding.

4 Methods

The methodology proposed in this work is summarized in Figure 1. It leverages the expressiveness of logical formulas to represent specific properties of the system under investigation, and this representation is used to automatically generate a *shield*, a security mechanism that forces the POMCP system to satisfy a set of high-level requirement.

As a first step, a logical formula with free variables is defined (see box 2 in Figure 1) to describe a property of interest of the policy under investigation. This formula, called *rule template*, defines a relationship between some properties of the belief (e.g., the probability to be in a specific state) and an action. Free variables in the formula allow the expert to avoid quantifying the limits of this relationship. These limits are then computed by analyzing a set of observed traces

(see box 1). For instance, a template saying “Do this when the confidence that the path is cluttered is at least \bar{x} ”, with \bar{x} free variable, is transformed into “Do this when the confidence that the path is cluttered is at least 0.15”. By defining a rule template the expert provides useful prior knowledge about the structure of the investigated property. This template can include strict requirements that the expert wants to satisfy (e.g., it is possible to force the rule to only use confidence values above 90%). This is combined with the real execution of a POMCP system collected into a trace. The methodology trains a *rule* (i.e., a rule template with all the free-variables instantiated) using a MAX-SMT based algorithm. This rule describes as many of the decisions taken by the agent as possible while satisfying the requirement defined in the template (see box 3). A set of rules is then used to create a *shield*, a safety mechanism that we integrate into POMCP to preemptively block actions that do not respect the details defined by the expert with the template (box 4).

4.1 Rules and Rule Templates

A rule template encodes the structure of the expected behavior of the system. It can include hard requirements that the expert expects to be true and free-variables that the methodology must instantiate. It is a set of first-order logic formulas without quantifiers explaining some properties of the policy, and has the following form:

$$\begin{aligned}
 r_1: & \text{select } a_1 \text{ when } (\bigvee_{i^1} \text{subformula}_{i^1}); \\
 & \dots \\
 r_n: & \text{select } a_n \text{ when } (\bigvee_{i^n} \text{subformula}_{i^n}); \\
 & [\text{where } \bigwedge_j (\text{requirement}_j);]
 \end{aligned}$$

where r_1, \dots, r_n are *action rule templates*. A *subformula* is defined as $\bigwedge_k p_s \approx \bar{x}_k$, where p_s is the probability of state s , symbol $\approx \in \{<, >, \geq, \leq\}$, and \bar{x}_k is a free variable that is automatically instantiated by the SMT solver analyzing the traces (when the problem is satisfiable). In general, bold letters with an overline (e.g., \bar{x}, \bar{y}) are used to identify free variable while italic letters (e.g., p, a_i) are used for fixed values read from the trace. The *where* statement is used to specify hard requirements. They can take different forms, such as the definition of a minimum value (e.g., $\bar{x}_0 \geq 0.9$) or a relation (e.g., $\bar{x}_2 = \bar{x}_3$). These are used to define prior knowledge on the domain which is used by the rule generation algorithm to compute optimal parameter values (e.g., equality between two free-variables belonging to different rules can be used to encode the idea that two rules are symmetrical).

4.2 Rule Generation

The rules used in the shield are generated using Algorithm 1, it takes as input a trace ex generated by the POMCP system that we want to shield and a rule template r . The output is an instantiation of all the free-variables of r that satisfies as many steps of the step of ex as possible. A Z3 instance (*solver*) is initialized in line 1. Hard constraints are added to force all the free-variables in the template to satisfy the probability axioms (i.e., to have value in range $[0, 1]$). Then

Algorithm 1: RuleGeneration

Data: a trace generated by POMCP ex
a rule template r
Result: an instantiation of r

- 1 *solver* \leftarrow probability constraints for thresholds in r ;
- 2 **foreach** action rule r_a with $a \in A$ **do**
- 3 **foreach** step t in ex **do**
- 4 build new dummy literal $l_{a,t}$;
- 5 $cost \leftarrow cost \cup l_{a,t}$;
- 6 compute p_0^t, \dots, p_n^t from $t.particles$;
- 7 $r_{a,t} \leftarrow$ instantiate rule r_a using p_0^t, \dots, p_n^t ;
- 8 **if** $t.action \neq a$ **then**
- 9 $r_{a,t} \leftarrow \neg(r_{a,t})$;
- 10 $solver.add(l_{a,t} \vee r_{a,t})$;
- 11 *solver.minimize(cost)*;
- 12 $goodness \leftarrow 1 - distance_to_observed_boundary$;
- 13 $model \leftarrow solver.maximize(goodness)$;
- 14 **return** *model*

in the *foreach* loop in lines 2–10 the algorithm builds a solution that satisfy the maximum number of steps in trace ex using the template r . In particular, for each action rule r_a , where a is an action, and for each step t in the trace ex , the algorithm first generates a literal $l_{a,t}$ (line 4) which is a dummy variable used by MAX-SMT to satisfy clauses that are not satisfiable by a free variable assignment. This literal is then added to the *cost* objective function (line 5) which is a pseudo-boolean function collecting all the dummy literals. Afterwards, the belief state probabilities are collected from the particle filter (line 6) and used to instantiate the action rule template r_a (line 7) by substituting their probability variables p_i with observed belief probabilities. This generates a new clause $r_{a,t}$ which represents the constraint for step t . This constraint is considered in its negated form $\neg(r_{a,t})$ if the step action $t.action$ is different from a (line 9) because the clause $r_{a,t}$ should not be true. To encode the fact that we need the maximum number of satisfiable steps, we add the clause $l_{a,t} \vee r_{a,t}$ to the problem. These clauses can be satisfied in two ways, namely, by finding an assignment of the free variables that makes the clause $r_{a,t}$ true or by assigning a true value to the literal $l_{a,t}$. To minimize the number of occurrence of the second case we use the solver to find a solution that minimize the value of the cost function. This minimization is a typical MAX-SMT problem in which an assignment maximizing the number of satisfied clauses is found. There can be more than a single assignment of free variables that achieves the MAX-SMT goal, thus the last step of the algorithm (lines 12–13) concerns the identification of the assignment which is closer to the behavior observed in the trace. This problem is solved by maximizing a goodness function which moves the free variables assignment as close as possible to the numbers observed in the trace, without altering the truth assignment of the dummy literals. Notice that this problem concerns the maximization of a real function, not the maximization of the number of sat-

isfiable clauses (as in MAX-SMT). It is solved by the linear arithmetic module of the Z3 solver.

The variable in the SMT problem are the free variables specified in the template (a constant number) and the dummy literals, that are linear on the size of the trace because the algorithm builds a clause for each step, and each clause introduces a new dummy literal. MAX-SMT is NP-hard but in practice, Z3 solves most practical instances in a reasonable time, and it provides good performance in our experiments (For example, it never takes more than one minute to compute the parameter of a rule in the more challenging case of velocity regulation).

4.3 Soft-thresholding for Rule Membership Check

It is important to retain flexibility in the shield mechanism for two reasons, namely, to allow the user to express approximate ideas (important for explainability) and to be capable of considering unexplored beliefs (important because POMCP is an online algorithm). The rule describes as many of the steps of the trace as possible, however, the MAX-SMT-based solution does not explain, in general, all the decisions taken by the policy due to the approximate nature of the template used to generate the rule. Some of these decisions are very different from the template (i.e., very different than the behavior expected by the expert), and thus should not be accepted by the shield, but others may be only slightly outside the rule boundaries (possibly because the approximation does not work well in this step). For example, if we have the rule “move at high speed when the probability of collision is below 5%”, the shield must not accept a decision to move at high speed when the risk of collision is 40%, but a belief in which we only have a 5.2% risk of collision could be considered (e.g., because other aspects that were omitted in the simplified template play a role in this decision). Moreover, the online nature of POMCP does not allow us to write a rule that describes the whole belief space due to the incomplete nature of the policy generated by this method. Training the rule using a large trace that contains many runs of the original POMCP helps in reducing this problem, but it is important to maintain flexibility toward unexplored beliefs.

To foster flexibility, we introduce a mechanism that quantifies how far a belief is from the rule, and if this difference is lower than a threshold τ we accept the belief as valid. To compute the *distance* between a new belief and the rule, we choose some beliefs that satisfy the rule as representatives of the rule itself, and we measure the *discrete Hellinger distance* (H^2) (Hellinger 1909) between the new belief (whose membership to the rule is under evaluation) and each representative. Given two discrete probability distributions P and Q over k states (in our case, the two beliefs under consideration) the Hellinger distance is defined as:

$$H^2(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2}$$

where p_i is the probability of the i -th state in P and q_i is the probability used to generate the rule of the i -th state in Q .

Algorithm 2: Shielding

Data: a belief b , a shield s , safe action a_{safe}
Result: set of legal actions \mathcal{L}

```

1  $\mathcal{L} \leftarrow \emptyset$ ;
2 foreach action  $a \in \mathcal{A}$  do
3   if  $a \notin s$  then
4      $\mathcal{L} \leftarrow \mathcal{L} \cup a$ ;
5   else if  $s.test\_constraints(b)$  then
6      $\mathcal{L} \leftarrow \mathcal{L} \cup a$ ;
7   else if  $\exists r \in s.Repr : H^2(b, r) < s.\tau$  then
8      $\mathcal{L} \leftarrow \mathcal{L} \cup a$ ;
9 if  $\mathcal{L} = \emptyset$  then
10   $\mathcal{L} \leftarrow \{a_{safe}\}$ ;
11 return  $\mathcal{L}$ ;

```

An interesting property of H^2 is that it is bounded between 0 and 1, which is very useful to define a meaningful threshold τ . The rule describes one or more constraints on the acceptable beliefs, we use these constraints to generate a set D of d possible beliefs that satisfy the rule. In other words, when the rule membership of a new belief b is checked by the shield, if the belief is out of the rule decision boundary then we compute the H^2 between b and each belief in D . Then we compare all the distances with the threshold τ . If one of the distances is below the threshold then the shield considers b as an acceptable belief for the rule.

4.4 Shielded POMCP

We integrate the shield into POMCP to preemptively prune undesired actions considering the current belief. It includes a set of rules trained as explained in Section 4.2 and a set of representative beliefs generated as described in Section 4.3. To shield the action of the POMCP, we start by building a set of *legal actions* \mathcal{L} that satisfy the logical rules and can be performed on the current belief, and we force POMCP to only consider legal actions in the first step of the simulation. After a legal action is selected, the Monte-Carlo Tree Search is performed as usual. Notice that when the original implementation of POMCP (Silver and Veness 2010) selects a particle in the simulation process, it assumes that the state encoded by the particle is the current state of the system (which for a POMDP is not observable) and thus the belief can only be considered in the first step.

With this mechanism, we can ensure that the rule of the shield is respected but we do not force POMCP to select a specific action, the best action is still decided using the regular POMCP but only among the legal ones. In more detail, as reported in Algorithm 2, for each possible action a , we consider a as a legal action if it satisfies at least one of these three conditions, namely, *i*) the shield does not define any rule (i.e., any restriction) for this action (line 3), *ii*) the current belief satisfies the constraints defined by an action rule for action a (line 5), *iii*) the Hellinger distance between the belief and the closest representative of the action rule for a is lower than the predefined threshold (line 7). These condi-

tions could result in an empty set of legal actions \mathcal{L} (i.e., if the rules are very strict). In this case, it is important to define a default safe action a_{safe} that is used when no other action is possible. While this is a domain-specific requirement, it is reasonable to assume that most domains have such action (e.g., wait and listen to gather extra data, take low-risk action that yields low rewards). The computation of legal actions is performed only once for a simulation step since the current belief does not change until a new observation is received from the real environment. Checking that the belief satisfies the constraints has a fixed cost, checking the H^2 of the representative beliefs increases linearly with the number of beliefs. As shown in Section 5, this is a negligible cost, and the reduced number of actions that must be tested (because not all actions are now legal) can also slightly reduce the execution time.

5 Results

5.1 Case Studies

We test our methodology in two domains, namely, the standard POMDP domain *tiger* and a robotic-inspired problem called *velocity regulation* in the following, in which a robot must move as fast as possible while avoiding collisions.

Tiger *Tiger* is a well-known problem (Kaelbling, Littman, and Cassandra 1998) in which an agent has to choose which door to open among two doors, one hiding a treasure and the other hiding a tiger. Finding the treasure yields a reward of +10 while finding the tiger a reward of -100. The agent can also listen (by paying a small penalty of -1) to gain new information. Listening is however not accurate since there is a 0.15 probability of hearing the tiger from the wrong door.

Velocity Regulation In *velocity regulation*, a robot travels on a pre-specified path divided into eight *segments* which are in turn divided into *subsegments* of different sizes, as shown in Figure 2. Each segment has a (hidden) difficulty value among *clear* ($f = 0$, where f is used to identify the difficulty), *lightly obstructed* ($f = 1$) or *heavily obstructed* ($f = 2$). All the subsegments in a segment share the same difficulty value, hence the hidden state-space has 3^8 states. The goal of the robot is to travel on this path as fast as possible while avoiding collisions. In each subsegment, the robot must decide a *speed level* a (i.e., action). We consider three different speed levels, namely 0 (slow), 1 (medium speed), and 2 (fast). The reward received for traversing a subsegment is equal to the length of the subsegment multiplied by $1 + a$, where a is the speed of the agent, namely the action that it selects. The higher the speed, the higher the reward, but a higher speed suffers a greater risk of collision (see the collision probability table $p(c = 1 | f, a)$ in Figure 2.c). The real difficulty of each segment is unknown to the robot, but in each subsegment, the robot receives an observation, which is 0 (no obstacles) or 1 (obstacles) with a probability depending on segment difficulty (see Figure 2.b). The state of the problem contains a hidden variable (i.e., the difficulty of each segment), and three observable variables (current segment, subsegment, and time elapsed since the beginning).

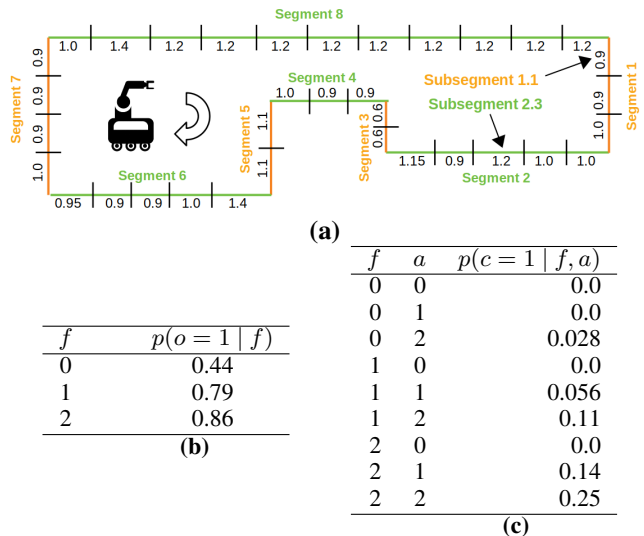


Figure 2: Main elements of the POMDP model for *velocity regulation*. (a) Path map. The map presents the length (in meters) for each subsegment. (b) Occupancy model $p(o = 1 | f)$: probability of observing a subsegment occupancy given segment difficulty. (c) Collision model $p(c = 1 | f, a)$: collision probability given segment difficulty and action.

5.2 Empirical Methodology

We implement *tiger* and *velocity regulation* as black-box simulators in the original POMCP C++ code provided in (Silver and Veness 2010). We extend the implementation with the capability of using a shield. To generate *traces*, we collect (belief, action) pairs at each step, with the belief saved as a particle distribution. We store these data using the *eXtensible Event Stream (XES)* (Acampora et al. 2017) format, a standard developed to log the executions of programs conveniently. The *RuleGeneration* algorithm (i.e., Algorithm 1) has been developed in Python. The Python binding of Z3 (De Moura and Bjørner 2008) has been used to solve the SMT formulas. To build the shields, we use rules that prove to be good explanations of the policies under investigation. We build these rules using the explainability tool presented in (Mazzi, Castellini, and Farinelli 2021a). A detailed example of the rule synthesis procedure is presented in (Mazzi, Castellini, and Farinelli 2021b). Experiments have been performed on a notebook with Intel Core i7-6700HQ and 16GB RAM. The code is available at <https://github.com/GiuMaz/ICAPS-2021-supmat>.

Error Injection To test the robustness of the shield in different scenarios, we injected an error in the POMCP implementation of the two domains. We modify the *RewardRange* parameter (called c in the following) in POMCP. This parameter defines the maximum difference between the lowest and the highest possible reward, and it is used by UCT to balance exploration and exploitation. If this value is lower than the correct one the algorithm could find a reward that exceeds the maximum expected value leading to a wrong state,

namely, the agent believes to have identified the best possible action and it stops exploring new actions, even though the selected action is not the best one. This is an interesting error because it is hard to detect, it randomly affects the exploration-exploitation trade-off without introducing any systematic (and thus, easier to identify) mistakes. This erroneous behavior is also independent of the domain used, and thus it is useful to recreate similar situations in different problems. Moreover, notice that the error we consider can happen with a relevant frequency in practical applications of POMCP, as the parameter c is not easy to evaluate automatically without exhaustively considering all the possible states of the POMDP, and thus must be tuned by the designer.

In our experimental setup, we consider several possible values of the parameter c and for each domain and value of c we create a trace using the unshielded POMCP, then we train a shield for each generated execution trace by using the methodology described in Section 4. Notice that these traces could contain errors. Finally, we run the POMCP again using the shielding and the various values of c . We evaluate the performance of the methodology by comparing the average discounted return achieved by the two versions of POMCP (original and shielded). We consider 1000 runs for *Tiger* and 100 runs for *Velocity Regulation*. The results achieved for tiger and velocity regulation are presented in subsections 5.3 and 5.4, respectively. In both cases, we generate 1000 representative beliefs and we use $\tau = 0.10$ as a threshold to check if a belief is in range.

5.3 Results for Tiger

A successful policy for *Tiger* listens to collect information on the position of the tiger and opens a door only when the agent is reasonably certain to find the treasure. From the analysis of the observation model and reward function, it is however not immediate to define what “reasonably certain” means. To investigate it, we create a rule template specifying a relationship between the confidence (in the belief) over the treasure position and the related opening action as follow:

$$\begin{aligned} r_L &: \text{select } Listen \text{ when } (p_{right} \leq \bar{x}_1 \wedge p_{left} \leq \bar{x}_2); \\ r_{OR} &: \text{select } Open_R \text{ when } p_{right} \geq \bar{x}_3; \\ r_{OL} &: \text{select } Open_L \text{ when } p_{left} \geq \bar{x}_4; \\ \text{where } &(\bar{x}_1 = \bar{x}_2) \wedge (\bar{x}_3 = \bar{x}_4) \wedge (\bar{x}_3 > 0.9); \end{aligned}$$

Action rule template r_L describes when the agent should listen, while templates r_{OR} and r_{OL} describe when the agent should open the right and left door, respectively. We also want to force the agent to only open a door if it is at least 90% sure to find the treasure behind it ($\bar{x}_3 > 0.9$). Finally, we use two hard constraints to specify that the problem is expected to be symmetric ($x_1 = x_2$ and $x_3 = x_4$).

We learn the rule parameters from a POMCP trace and we create a shield from this rule. Since this shield gives a rule for all possible actions, it is important to set a safe action as described in Section 4.4. For this domain, we use $a_{safe} = Listen$. The correct value of c is 110 because the reward interval is $[-100, 10]$. For each value of c in $\{110, 80, 60, 40\}$, we generate a trace with 1000 runs each, using a fixed seed for the pseudo-random algorithm. In each case, we use 2^{15}

particles and a maximum of 10 steps. POMCP with a correct value of c produces the optimal policy (we tested that by comparing the decisions taken by POMCP with an exact policy computed using *incremental pruning* (Cassandra, Littman, and Zhang 1997)). As shown in Table 1.a, in particular in the first row of column #SA (the meaning of each column name is defined in the table caption), the shield does not interfere with the correct policy. Lower values of c produce a lower average discounted return, as shown in column *return* of the *No Shield* section in Table 1.a. In both sections *No Shield* and *Shield* of Table 1.a, we present the average *return* and the average execution *time* of POMCP. In the *Shield* section we also present the relative increase (i.e., $RI = \frac{shielded-original}{|original|} \cdot 100$) between the original and the shielded version of the POMCP (see column *RI*). This column shows that the benefit of using a shield is greater when the number of errors increases. For the return column, we report in bold values whose difference from their no-shield counterpart is statistically significant according to a paired t-test with 95% confidence. While in the first row there is no difference between the two cases, in the other three rows the difference is statistically significant.

The average return achieved using the shield is the same in all four cases, and this is also identical to the return achieved by the correct policy. This is because in *tiger* we can write a shield that perfectly recreates the behavior of the correct policy, a goal that is difficult to achieve in real-world problems. This is particularly interesting because the shields in the cases of $c \in \{80, 60, 40\}$ are obtained by using traces generated with a POMCP implementation that does make some mistakes. As a consequence, the Execution traces contain wrong decisions. However, the combination of insight provided by the expert with the MAX-SMT-based analysis of the traces results in a shield with extremely good performances. As shown in the *time* column, in general, the presence of the shield does not noticeably impact in terms of run-time. The shield generation algorithm takes between 10 and 12 seconds to generate the shield in this case. In the last row, the original POMCP is particularly fast, this happens since the erroneous POMCP opens many doors as fast as possible, without listening, this leads to runs that achieve very low average return but ends quickly.

5.4 Results for Velocity Regulation

In *velocity regulation*, a robot moves around a predefined path in an industrial environment. As in *Tiger*, we use 2^{15} particles. In general, this leads to acceptable performance but sometimes the simulations are not good enough and the robot takes a decision that the designer considers too risky. We focus on writing a shield describing when the robot travels at maximum speed (i.e., $a = 2$). This is the most dangerous action because there is always a risk of collision involved, as explained in Table 2.c. We expect that the robot should move at speed 2 only if it is confident enough to be in an easy-to-navigate segment, but this level of confidence varies slightly from segment to segment (due to the length of the segments, the elapsed times, or the relative difficulty of the current segment in comparison to the others). To write

c	No Shield		Shield			
	return	time (s)	return	RI	time (s)	#SA
110	3.702(± 0.623)	0.066(± 0.027)	3.702(± 0.623)	0.00%	0.065(± 0.029)	0
80	3.593(± 0.632)	0.067(± 0.030)	3.702 (\pm 0.623)	3.03%	0.061(± 0.027)	4
60	3.088(± 0.673)	0.060(± 0.025)	3.702 (\pm 0.623)	19.88%	0.061(± 0.027)	121
40	-4.173(± 1.101)	0.035(± 0.017)	3.702 (\pm 0.623)	188.71%	0.052(± 0.023)	647

a) Tiger

c	No Shield		Shield			
	return	time (s)	return	RI	time (s)	#SA
103	24.716(± 3.497)	10.166(± 0.682)	26.045 (\pm 3.640)	5.38%	10.118(± 0.238)	7
90	18.030(± 3.794)	10.173(± 0.234)	22.680 (\pm 3.524)	25.79%	10.166(± 0.241)	12
70	4.943(± 5.260)	10.278(± 0.234)	8.970 (\pm 4.556)	81.46%	10.377(± 0.230)	51
50	0.692(± 5.051)	10.374(± 0.230)	1.638(± 4.525)	136.53%	10.435(± 0.336)	171

b) Velocity Regulation

Table 1: Experimental Results. The first column shows the different values of the RewardRange c . The second (third) column shows the average return (time) achieved by the original POMCP and the relative standard deviation. The *Shield* section shows the average return and time achieved by POMCP using a shield (column four and six), values in bold show a statistically significant difference with respect to the shield counterpart (according to a paired t-test with 95% confidence level). Column *RI* shows the relative increase in performance between the two original and shielded POMCP. Finally, column *#SA* shows how many times the shield alters the decision during the execution

a rule template that is compact but informative, we want the rule to be a local approximation of the behavior of the robot, thus we only focus on the current segment without considering the path as a whole. To do that, we introduce the `diff` function, which, given a belief, a segment, and a difficulty value, extracts the expected difficulty of the segment in the specified belief. We can not write a compact template:

$$r_2: \text{select action } S_2 \text{ when}$$

$$p_0 \geq \bar{x}_1 \vee p_2 \leq \bar{x}_2 \vee (p_0 \geq \bar{x}_3 \wedge p_1 \geq \bar{x}_4)$$

where $p_i = \text{diff}(\text{distr}, \text{seg}, i) \forall i \in \{0, 1, 2\}$
 $\wedge \bar{x}_1 \geq 0.9$

where $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4$ are free variables. With the subformula $p_0 \geq \bar{x}_1$ we say that the robot can move at high speed if its confidence of being in an easy to navigate (i.e., difficulty 0) segment is above a certain threshold. In the hard requirement, we also force this threshold to be at least 0.9. The robot can also move at high speed if it is confident of not being in a really hard segment (subformula \bar{x}_2), or if the combination of p_0 and p_1 (i.e., clear or slightly cluttered segment) are above other thresholds (subformula $p_0 \geq \bar{x}_3 \wedge p_1 \geq \bar{x}_4$). We do not force any requirement for these other thresholds, we simply train the values from the data. We use a trace with 100 runs to train and test the shield. The shield generation takes 50 seconds to generate velocity regulation shields. Table 1.b shows the result of the experiment. As in *Tiger*, a lower value of c produce a lower return. In this case, the best c is 103 (the difference between a collision when we move slowly in the shortest segment and a fast movement in the longest segment without a collision). The first row shows that, unlike *Tiger*, the usage of a shield can improve the performance even when c is correct. In this case, the shield intervenes only 7 times (over the 3500 analyzed steps), yielding a 5.38% increment in the return. This happens because the shield blocks the rare cases in which the POMCP simulations are not enough to properly assess the risk of moving at high speed. When c decreases, the shield intervenes more

often (see column *#SA*) since the error due to the limited number of simulations is combined with the errors generated by an incorrect value of c . Table 1.b also shows that a higher number of interventions leads to a bigger relative increase in the performance (column *RI*). The difference is statistically significant in the case of $c \in \{103, 90, 70\}$, and show that the introduction of the shield improves the performance up to the 81%, even in cases in which the shield is trained using traces generated by a POMCP process that makes some mistakes. In the case of $c = 50$ the return increase but the difference is not statistically significant. The shield intervenes 171 times by blocking risky high-speed moves, but unlike *Tiger*, in which we use a rule for every possible action, here POMCP made many wrong decisions when it moves at low or medium speed (for example, by moving slowly when the path is clear). The usage of the shield does not significantly increase the time required to perform the simulations.

6 Conclusions and Future Work

In this work, we present a methodology that generates a shielding mechanism for POMCP exploiting a high-level representation of expected policy behavior provided by human experts. The shielding mechanism preemptively blocks unexpected actions. The approach is proved to provide a statistically significant improvement to the performance of POMCP in a standard domain and a robotics-inspired domain. This work paves the way towards several interesting research directions. First, we aim at improving the expressiveness of logical formulas by employing temporal logic to verify safe reachability requirements. Second, we aim to further improve the integration between POMCP and the shielding mechanism (e.g., by considering the effect of shielding on other actions besides the first one of the simulation). Third, most important, we aim at developing an approach for synthesizing logical rules online, i.e., while the POMCP algorithm is running and not based only on previously generated traces.

References

- Acampora, G.; Vitiello, A.; Di Stefano, B.; van der Aalst, W.; Gunther, C.; and Verbeek, E. 2017. IEEE 1849: The XES Standard: The Second IEEE Standard Sponsored by IEEE Computational Intelligence Society [Society Briefs]. *IEEE Computational Intelligence Magazine*.
- Anjomshoae, S.; Najjar, A.; Calvaresi, D.; and Främling, K. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, Richland, SC.
- Bastani, O.; Pu, Y.; and Solar-Lezama, A. 2018. Verifiable Reinforcement Learning via Policy Extraction. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2499–2509. Red Hook, NY, USA.
- Bjørner, N.; Phan, A.-D.; and Fleckenstein, L. 2015. vZ - An Optimizing SMT Solver. In *Proceedings of the 21st International Conference on Tools and Algorithms for the Construction and Analysis of Systems - Volume 9035*, 194–199. Berlin, Heidelberg.
- Cashmore, M.; Collins, A.; Krarup, B.; Krivic, S.; Magazzeni, D.; and Smith, D. 2019. Towards Explainable AI Planning as a Service. 2nd ICAPS Workshop on Explainable Planning, XAIP 2019.
- Cashmore, M.; Fox, M.; Long, D.; and Magazzeni, D. 2016. A Compilation of the Full PDDL+ Language into SMT. In *Proceedings of the Twenty-Sixth International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'16, 79–87.
- Cassandra, A.; Littman, M. L.; and Zhang, N. L. 1997. Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 54–61.
- Castellini, A.; Chalkiadakis, G.; and Farinelli, A. 2019. Influence of State-Variable Constraints on Partially Observable Monte Carlo Planning. In *Proc. 28-th International Joint Conference on Artificial Intelligence, IJCAI-19*, 5540–5546.
- Castellini, A.; Marchesini, E.; and Farinelli, A. 2020. Online Monte Carlo Planning for Autonomous Robots: Exploiting Prior Knowledge on Task Similarities. In *Proceedings of the 6th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2019@AI*IA2019)*, volume 2594 of *CEUR Workshop Proceedings*, 25–32. CEUR-WS.org.
- Castellini, A.; Marchesini, E.; Mazzi, G.; and Farinelli, A. 2020. Explaining the influence of prior knowledge on POMCP policies. In *Proceedings of the 17th European Conference on Multi-Agents Systems*, volume 12520 of *Lecture Notes in Artificial Intelligence*.
- De Moura, L.; and Bjørner, N. 2008. Z3: An Efficient SMT Solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, TACAS'08/ETAPS'08, 337–340. Berlin, Heidelberg.
- Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable Planning. *CoRR* abs/1709.10256.
- Gunning, D. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. ii–ii.
- Hellinger, E. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* 136: 210–271.
- Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artif. Intell.* 101(1–2): 99–134.
- Kocsis, L.; and Szepesvári, C. 2006. Bandit Based Monte-Carlo Planning. In *Proc. ECML'06*, 282–293. Berlin, Heidelberg.
- Langley, P.; Meadows, B.; Sridharan, M.; and Choi, D. 2017. Explainable Agency for Intelligent Autonomous Systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4762–4763.
- Mazzi, G.; Castellini, A.; and Farinelli, A. 2021a. Identification of Unexpected Decisions in Partially Observable Monte Carlo Planning: A Rule-Based Approach. In *accepted at the 21th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21.
- Mazzi, G.; Castellini, A.; and Farinelli, A. 2021b. Policy Interpretation for Partially Observable Monte-Carlo Planning: a Rule-based Approach. In *Proceedings of the 7th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2020@AI*IA2020)*, volume 2806 of *CEUR Workshop Proceedings*, 44–48. CEUR-WS.org.
- Newaz, A. A. R.; Chaudhuri, S.; and Kavraki, L. E. 2019. Monte-Carlo Policy Synthesis in POMDPs with Quantitative and Qualitative Objectives. In *RSS 2019*.
- Norman, G.; Parker, D.; and Zou, X. 2017. Verification and control of partially observable probabilistic systems. *Real-Time Systems* 53(3): 354–402.
- Papadimitriou, C. H.; and Tsitsiklis, J. N. 1987. The Complexity of Markov Decision Processes. *Math. Oper. Res.* 12(3): 441–450.
- Silver, D.; and Veness, J. 2010. Monte-Carlo Planning in Large POMDPs. In Lafferty, J. D.; Williams, C. K. I.; Shawe Taylor, J.; Zemel, R. S.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 23*, 2164–2172. Curran Associates, Inc.
- Wang, Y.; Chaudhuri, S.; and Kavraki, L. E. 2018. Bounded Policy Synthesis for POMDPs with Safe-Reachability Objectives. *ArXiv* abs/1801.09780.
- Zhu, H.; Xiong, Z.; Magill, S.; and Jagannathan, S. 2019. An Inductive Synthesis Framework for Verifiable Reinforcement Learning. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, 686–701. New York, NY, USA.