



# From Semantics to Execution: Integrating Action Planning With Reinforcement Learning for Robotic Causal Problem-Solving

Manfred Eppe\*, Phuong D. H. Nguyen and Stefan Wermter

Department of Informatics, Knowledge Technology Institute, Universität Hamburg, Hamburg, Germany

## OPEN ACCESS

### Edited by:

Georg Martius,  
Max Planck Institute for Intelligent  
Systems, Germany

### Reviewed by:

Andrea Soltoggio,  
Loughborough University,  
United Kingdom  
Natalia Díaz-Rodríguez,  
ENSTA ParisTech École Nationale  
Supérieure de Techniques Avancées,  
France

### \*Correspondence:

Manfred Eppe  
eppe@informatik.uni-hamburg.de

### Specialty section:

This article was submitted to  
Computational Intelligence in  
Robotics,  
a section of the journal  
Frontiers in Robotics and AI

**Received:** 22 May 2019

**Accepted:** 04 November 2019

**Published:** 26 November 2019

### Citation:

Eppe M, Nguyen PDH and Wermter S  
(2019) From Semantics to Execution:  
Integrating Action Planning With  
Reinforcement Learning for Robotic  
Causal Problem-Solving.  
Front. Robot. AI 6:123.  
doi: 10.3389/frobt.2019.00123

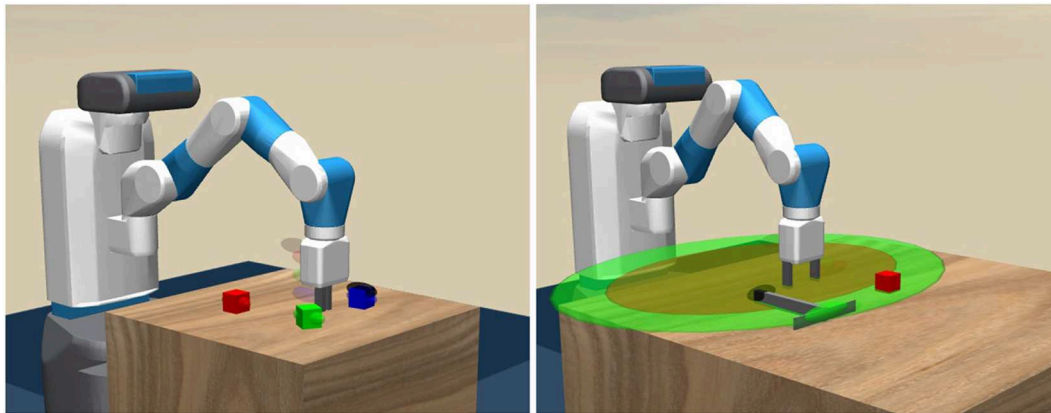
Reinforcement learning is generally accepted to be an appropriate and successful method to learn robot control. Symbolic action planning is useful to resolve causal dependencies and to break a causally complex problem down into a sequence of simpler high-level actions. A problem with the integration of both approaches is that action planning is based on *discrete high-level action- and state spaces*, whereas reinforcement learning is usually driven by a *continuous reward* function. Recent advances in model-free reinforcement learning, specifically, universal value function approximators and hindsight experience replay, have focused on goal-independent methods based on *sparse rewards* that are only given at the end of a rollout, and only if the goal has been fully achieved. In this article, we build on these novel methods to facilitate the integration of action planning with model-free reinforcement learning. Specifically, the paper demonstrates how the reward-sparsity can serve as a bridge between the high-level and low-level state- and action spaces. As a result, we demonstrate that the integrated method is able to solve robotic tasks that involve non-trivial causal dependencies under noisy conditions, exploiting both data and knowledge.

**Keywords:** reinforcement learning, hierarchical architecture, planning, robotics, neural networks, causal puzzles

## 1. INTRODUCTION

*How can one realize robots that reason about complex physical object manipulation problems, and how can we integrate this reasoning with the noisy sensorimotor machinery that executes the required actions in a continuous low-level action space?* To address these research questions, we consider reinforcement learning (RL) as it is a successful method to facilitate low-level robot control (Deisenroth and Rasmussen, 2011). It is well known that non-hierarchical reinforcement-learning architectures fail in situations involving non-trivial causal dependencies that require the reasoning over an extended time horizon (Mnih et al., 2015). For example, the robot in **Figure 1** (right) needs to first grasp the rake before it can be used to drag the block to a target location. Such a problem is hard to solve by RL-based low-level motion planning without any high-level method that subdivides the problem into smaller sub-tasks.

To this end, recent research has developed hierarchical and model-based reinforcement learning methods to tackle problems that require reasoning over a long time horizon, as demanded in domains like robotic tool use, block-stacking (Deisenroth and Rasmussen, 2011), and computer games (Aytar et al., 2018; Pohlen et al., 2018). Yet, the problem of realizing an agent that



**FIGURE 1** | A robot performing two object manipulation tasks. 1. *Block-stacking (Left)*: The gripper must stack three blocks at a random location within the robot's range on the table (indicated by the transparent goal markers behind the gripper). Herein, the robot needs to subdivide the course of actions into separate actions for grasping and placing the individual blocks. 2. *Tool use (Right)*: The red block is out of the gripper's range (indicated by the dark brown ellipsoid), so that solving the task of moving the block to a target location requires the robot to break the problem down into a sequence of high-level actions that involve grasping the rake, moving the rake toward the block and pulling the rake.

learns to solve open-domain continuous space causal puzzles from scratch, without learning from demonstration or other data sources, remains unsolved. The existing learning-based approaches are either very constrained (e.g., Deisenroth and Rasmussen, 2011), or they have been applied only to low-dimensional non-noisy control problems that do not involve complex causal dependencies (e.g., Bacon et al., 2017; Levy et al., 2019), or they build on learning from demonstration (Aytar et al., 2018).

A complementary method to address complex causal dependencies is to use pre-specified semantic domain knowledge, e.g., in the form of an action planning domain description (McDermott et al., 1998). With an appropriate problem description, a planner can provide a sequence of solvable sub-tasks in a discrete high-level action space. However, the problem with this semantic and symbolic task planning approach is that the high-level actions generated by the planner require the grounding in a low-level motion execution layer to consider the context of the current low-level state. For example, executing a high-level robotic action `move_object_to_target` requires precise information (e.g., the location and the shape) of the object to move it to the target location along a context-specific path. This grounding problem consists of two sub-problems **P.1** and **P.2** that we tackle in this article.

**P.1** *The first grounding sub-problem is to map the discrete symbolic action space to context-specific subgoals in the continuous state space.* For instance, `move_object_to_target` needs to be associated with a continuous sub-goal that specifies the desired metric target coordinates of the object.

**P.2** *The second grounding sub-problem is to map the subgoals to low-level context-specific action trajectories.* For instance, the low-level trajectory for `move_object_to_target` is specific to the continuous start- and target location of the object, and to potential obstacles between start and target.

Both problems are currently being recognized by the state of the art in combined task and motion planning (e.g., Toussaint et al., 2018), and, from a broader perspective, also in the field of state representation learning (e.g., Doncieux et al., 2018; Lesort et al., 2018). However, to the best of our knowledge, there exist currently no satisfying and scalable solutions to these problems that have been demonstrated in the robotic application domain with continuous state- and action representations (see section 2.4). In this research, we address **P.1** by providing a simple, yet principled, formalism to map the propositional high-level state space to continuous subgoals (section 3.1). We address **P.2** by integrating this formalism with goal-independent reinforcement learning based on sparse rewards (section 3.2).

Existing approaches that integrate action planning with reinforcement learning have not been able to map subgoals to low-level motion trajectories for realistic continuous-space robotic applications (Grounds and Kudenko, 2005; Ma and Cameron, 2009) because they rely on a continuous dense reward signal that is proportional to manually defined metrics that estimate how well a problem has been solved (Ng et al., 1999). The manual definition of such metrics, also known as *reward shaping*, is a non-trivial problem itself because the semantic distance to a continuous goal is often not proportional to the metric distance.

Recently, so-called *universal value function approximators* (UVFAs) (Schaul et al., 2015) in combination with *hindsight experience replay* (HER) (Andrychowicz et al., 2017) and neural actor-critic reinforcement learning methods (Lillicrap et al., 2016) have been proposed to alleviate this issue. HER realizes an efficient off-policy algorithm that allows for non-continuous sparse rewards without relying on reward shaping. Specifically, HER treats action trajectories as successful that do not achieve the desired specified goal, by pretending in hindsight that the achieved state was the desired goal state. Our research builds on this method because the sparse subgoal-specific rewards allow us to decouple the reward mechanism from the high-level action planning.

This approach enables us to address the following **central hypotheses**:

- H.1** *We hypothesize that model-free reinforcement learning with universal value function approximators (UVFAs) and hindsight experience replay (HER) is appropriate to learn the grounding of a discrete symbolic action space in continuous action trajectories. We measure the appropriateness by comparing the resulting hybrid discrete/continuous architecture with continuous hierarchical reinforcement learning (HRL). We consider our approach to be appropriate if it is better capable of learning to solve causal object-manipulation puzzles that involve tool use and causal chains of non-trivial length than HRL.*
- H.2** *We hypothesize that the approach is robust enough to handle a realistic amount of perceptual noise. We consider the approach to be robust to noise if there is no significant performance drop when moderate noise, e.g., 1–2% of the observational range, is added to the agent's state representation.*

We address these hypotheses by applying our method to three simulated robotic environments that are based on the OpenAI Gym framework. For these environments, we provide manually defined action planning domain descriptions and combine a planner with a model-free reinforcement learner to learn the grounding of high-level action descriptions in low-level trajectories.

**Our research contribution** is a principled method and proof-of-concept to ground high-level semantic actions in low-level sensorimotor motion trajectories and to integrate model-free reinforcement learning with symbolic action planning. The novelty of this research is to use UVFAs and HER to decouple the reward mechanism from the high-level propositional subgoal representations provided by the action planner: Instead of defining an individual reward function for each predicate, our approach allows for a single simple threshold-based sparse reward function that is the same for all predicates.

**Our research goal** is to provide a proof-of-concept and a baseline for the integration of action planning with reinforcement learning in continuous domains that involve complex causal dependencies.

The remainder of the article is organized as follows. In section 2 we investigate the state of the art in task and motion planning, hierarchical learning and the integration of planning with learning. We identify the problem of grounding high-level actions in low-level trajectories as a critical issue for robots to solve causal puzzles. We present our method and the underlying background in section 3. We describe the realization of our experiments in section 4 and show the experimental results in section 5 before we discuss and align our findings with the hypotheses in section 6. We conclude in section 7.

## 2. STATE OF THE ART

Our work is related to robotic task and motion planning, but it also addresses plan execution. Therefore, it is also related to

hierarchical learning algorithms and the integration of learning with planning.

### 2.1. Combined Task and Motion Planning

The field of combined task and motion planning (TAMP) investigates methods to integrate low-level motion planning with high-level task planning. The field aims at solving physical puzzles and problems that are too complex to solve with motion planning alone, often inspired by smart animal behavior (Toussaint et al., 2018). For example, crows are able to perform a sequence of high-level actions, using tools like sticks, hooks or strings, to solve a puzzle that eventually leads to a reward (Taylor et al., 2009). A set of related benchmark problems has recently been proposed by Lagriffoul et al. (2018). However, since TAMP focuses primarily on the planning aspects and not necessarily on the online action execution, the benchmark environments do not consider a physical action execution layer.

Toussaint et al. (2018) formulate the TAMP problem as an inverted differentiable physics simulator. The authors consider the local optima of the possible physical interactions by extending mixed-integer programs (MIP) (Deits and Tedrake, 2014) to first-order logic. The authors define physical interactions as action primitives that are grounded in contact switches. The space of possible interactions is restricted to consider only those interactions that are useful for the specific problem to solve. These interactions are formulated based on a fixed set of predicates and action primitives in the domain of robotic tool use and object manipulation. However, the authors provide only a theoretical framework for planning, and they do not consider the physical execution of actions. Therefore, an empirical evaluation to measure the actual performance of their framework, considering also real-world issues like sensorimotor noise, is not possible.

Other TAMP approaches include the work by Alili et al. (2010) and de Silva et al. (2013) who both combine a hierarchical symbolic reasoner with a geometrical reasoner to plan human-robot handovers of objects. Both approaches consider only the planning, not the actual execution of the actions. The authors do not provide an empirical evaluation in a simulated or real environment. Srivastava et al. (2014) also consider action execution and address the problem of grounding high-level tasks in low-level motion trajectories by proposing a planner-independent interface layer for TAMP that builds on symbolic references to continuous values. Specifically, they propose to define symbolic actions and predicates such that they refer to certain objects and their poses. They leave it to the low-level motion planner to resolve the references. Their approach scales well on the planning level in very cluttered scenes, i.e., the authors demonstrate that the planning approach can solve problems with 40 objects. The authors also present a physical demonstrator using a PR2 robot, but they do not provide a principled empirical evaluation to measure the success of the action execution under realistic or simulated physical conditions. Wang et al. (2018) also consider action execution, though only in a simple 2D environment without realistic physics. Their focus is on solving long-horizon task planning problems that involve sequences of 10 or more action primitives. To this end, the authors present a

method that learns the conditions and effects of high-level action operators in a kitchen environment.

A realistic model that also considers physical execution has been proposed by Leidner et al. (2018). The authors build on geometric models and a particle distribution model to plan goal-oriented wiping motions. Their architecture involves low-level and high-level inference and a physical robotic demonstrator. However, the authors build on geometric modeling and knowledge only, without providing a learning component. Noisy sensing is also not addressed in their work.

## 2.2. Hierarchical Learning-Based Approaches

Most TAMP approaches consider the planning as an offline process given geometrical, algebraic or logical domain models. This, however, does not necessarily imply the consideration of action execution. The consideration of only the planning problem under idealized conditions is not appropriate in practical robotic applications that often suffer from sensorimotor noise. To this end, researchers have investigated hierarchical learning-based approaches that differ conceptually from our work because they build on data instead of domain-knowledge to realize the high-level control framework.

For example, Nachum et al. (2018) and Levy et al. (2019) both consider ant-maze problems in a continuous state- and action space. The challenge of these problems lies in coordinating the low-level walking behavior of the ant-like agent with high-level navigation. However, these approaches do not appreciate that different levels of the problem-solving process require different representational abstractions of states and actions. For example, in our approach, the planner operates on propositional state descriptions like “object 1 on top of object 2” and generates high-level conceptual actions like “move gripper to object.” In those HRL approaches, the high-level state- and action representations are within the same state-and action space as the low-level representations. This leads to larger continuous problem spaces.

Other existing hierarchical learning-based approaches are limited to discrete action- or state spaces on all hierarchical layers. For example, Kulkarni et al. (2016) present the h-DQN framework to integrate hierarchical action-value functions with goal-driven intrinsically motivated deep RL. Here, the bottom-level reward needs to be hand-crafted using prior knowledge of applications. Vezhnevets et al. (2017) introduce the FeUdal Networks (FuNs), a two-layer hierarchical agent. The authors train subgoal embeddings to achieve a significant performance in the context of Atari games with a discrete action space. Another example of an approach that builds on discrete actions is the option-critic architecture by Bacon et al. (2017). Their method extends gradient computations of intra-option policies and termination functions to enable learning options that maximize the expected return within the options framework, proposed by Sutton et al. (1999). The authors apply and evaluate their framework in the Atari gaming domain.

## 2.3. Integrating Learning and Planning

There exist several robot deliberation approaches that exploit domain knowledge to deliberate robotic behavior and to perform

reasoning (e.g., Eppe et al., 2013; Rockel et al., 2013). The following examples from contemporary research extend the knowledge-based robotic control approach and combine it with reinforcement learning:

The Dyna Architecture (Sutton, 1991) and its derived methods, e.g., Dyna-Q (Sutton, 1990), queue-Dyna (Peng and Williams, 1993), RTP-Q (Zhao et al., 1999), aim to speed up the learning procedure of the agent by unifying reinforcement learning and incremental planning within a single framework. While the RL component aims to construct the action model as well as to improve the value function and policy directly through real experiences from environment interaction, the planning component updates the value function with simulated experiences collected from the action model. The authors show that instead of selecting uniformly experienced state-action pairs during the planning, it is much more efficient to focus on pairs leading to the goal state (or nearby states) because these cause larger changes in value function. This is the main idea of the prioritized sweeping method (Moore and Atkeson, 1993) and derivatives (Andre et al., 1998). The methods based on Dyna and prioritized sweeping have neither been demonstrated to address sparse rewards nor do they consider mappings between discrete high-level actions and states and their low-level counter parts.

Ma and Cameron (2009) present the policy search planning method, in which they extend the policy search GPOMDP (Baxter and Bartlett, 2001) toward the multi-agent domain of robotic soccer. Herein, they map symbolic plans to policies using an expert knowledge database. The approach does not consider tool use or similar causally complex problems. A similar restriction pertains to PLANQ-learning framework (Grounds and Kudenko, 2005): The authors combine a Q-learner with a high-level STRIPS planner (Fikes and Nilsson, 1972), where the symbolic planner shapes the reward function to guide the learners to the desired policy. First, the planner generates a sequence of operators to solve the problem from the problem description. Then, each of these operators is learned successively by the corresponding Q-learner. This discrete learning approach, however, has not been demonstrated to be applicable beyond toy problems, such as the grid world domain that the authors utilize for demonstrations in their paper. Yamamoto et al. (2018) propose a hierarchical architecture that uses a high-level abduction-based planner to generate subgoals for the low-level on-policy reinforcement learning component, which employed the proximal policy optimization (PPO) algorithm (Schulman et al., 2017). This approach requires the introduction of an additional task-specific evaluation function, alongside the basic evaluation function of the abduction model to allow the planner to provide the learner with the intrinsic rewards, similar to Kulkarni et al. (2016). The evaluation is only conducted in a grid-based virtual world where an agent has to pick up materials, craft objects and reach a goal position.

A very interesting approach has been presented by Ugur and Piater (2015). The authors integrate learning with planning, but instead of manually defining the planning domain description they learn it from observations. To this end, they perform clustering mechanisms to categorize object affordances and high-level effects of actions, which are immediately employed



in a planning domain description. The authors demonstrate their work on a physical robot. In contrast to our work, however, the authors focus mostly on the high-level inference and not on the robustness that low-level reinforcement-based architectures provide.

### 2.4. Summary of the State of the Art

The main weaknesses of the related state of the art that we address in this research are the following: TAMP approaches (section 2.1) mainly focus on the planning aspect. Whereas, they do not consider the physical execution of the planned actions or only evaluate the plan execution utilizing manually defined mapping between high-level (symbolic) and low-level (continuous value). These approaches require domain knowledge and model of robots and the environment to specify and execute the task and motion plans, which may suffer from noisy sensing conditions. On the contrary, hierarchical learning-based approaches (section 2.2) propose to learn both high-level and low-level from data, but mostly focus on solving problems with discrete action space, and they require internal hand-crafted reward functions. Methods with continuous action space like (Levy et al., 2019; Nachum et al., 2019) only consider setups without representational abstractions between the different hierarchical layers. Mixed approaches (section 2.3) that integrate learning and planning have similar disadvantages as the two other groups. In particular the lack of principled approaches to realize the mapping between discrete and continual spaces, the manual shaping of reward functions, and the lack of approaches that have been demonstrated and applied in a continuous-space realistic environment.

## 3. INTEGRATING REINFORCEMENT LEARNING WITH ACTION PLANNING

To tackle the problems P.1 and P.2, and to address the research goal of grounding high-level actions in low-level control trajectories, we propose the architecture depicted in Figure 2.

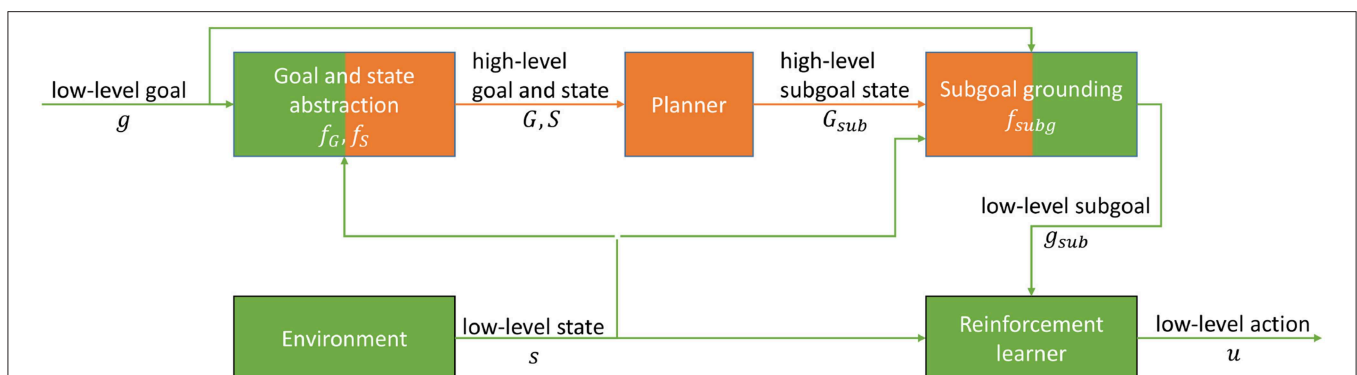
The novelty of the architecture with respect to the state of the art is its ability to learn to achieve subgoals that are provided in an abstract symbolic high-level representation by using a single universal sparse reward function that is appropriate for all discrete high-level goal definitions. This involves (i) the grounding of the high-level representations to low-level subgoals (section 3.1, Algorithm 1), and (ii) the formalization of the abstraction of the low-level space to the high-level space (section 3.1, Equation 4), (iii) a UVFA- and HER-based reinforcement learner to achieve the low-level subgoals (section 3.2), and (iv) the integration of an action planner with the reinforcement learning using the abstraction- and grounding mechanisms (section 3.3).

The resulting architecture is able to acquire a large repertoire of skills, similar to a multiple-goal exploration process (Forestier et al., 2017). However, instead of sampling and evaluating (sub-)goals through intrinsic rewards, the subgoals within our architecture are generated by the planner.

### 3.1. Abstraction and Grounding of States and Goals

Our abstraction and grounding mechanism tackles the research problem P.1, i.e., the mapping from high-level actions to low-level subgoals. STRIPS-based action descriptions are defined in terms of state changes based on predicates for pre- and postconditions. To appreciate that the state change is determined by the postcondition predicates, and not by the actions themselves, it is more succinct to define subgoals in terms of postcondition predicates because multiple actions may involve the same postconditions. Therefore, we define a grounding function for subgoals  $f_{subg}$ . The function is based on predicates instead of actions to avoid redundancy and to minimize the hand-engineering of domain models and background knowledge.

To abstract from low-level perception to high-level world states, we define abstraction functions  $f_S, f_G$ . These functions do not require any additional background knowledge because they fully exploit the definition of  $f_{subg}$ . In our formalization



**FIGURE 2 |** Our proposed integration model. Low-level motion planning elements are indicated in green and high-level elements in orange color. The abstraction functions  $f_S, f_G$  map the low-level state and goal representations  $s, g$  to high-level state and goal representations  $S, G$ . These are given as input to the planner to compute a high-level subgoal state  $G_{sub}$ . The subgoal grounding function  $f_{subg}$  maps  $G_{sub}$  to a low-level subgoal  $g_{sub}$  under consideration of the context provided by the current low-level state  $s$  and the low-level goal  $g$ . The reinforcement learner learns to produce a low-level motion plan that consists of actions  $u$  based on the low-level subgoal  $g_{sub}$  and the low-level state  $s$ .

of the abstraction and grounding, we consider the following conventions and assumptions C.1–C.7:

- C.1 The low-level environment state space is fully observable, but observations may be noisy. We represent low-level environment states with finite-dimensional vectors  $s$ . To abstract away from visual preprocessing issues and to focus on the main research questions, we adapt the state representations commonly used in deep reinforcement learning literature (Andrychowicz et al., 2017; Levy et al., 2019), i.e., states are constituted by the locations, velocities, and rotations of objects (including the robot itself) in the environment.
- C.2 The low-level action space is determined by continuous finite-dimensional vectors  $u$ . For example, the robotic manipulation experiments described in this paper consider a four-dimensional action space that consists of the normalized relative three-dimensional movement of the robot’s gripper in Cartesian coordinates plus a scalar value to represent the opening angle of the gripper’s fingers.
- C.3 Each predicate of the planning domain description determines a Boolean property of one object in the environment. The set of all predicates is denoted as  $\mathcal{P} = \{p_1, \dots, p_n\}$ . The high-level world state  $S$  is defined as the conjunction of all positive or negated predicates.
- C.4 The environment configuration is fully determined by a set of objects whose properties can be described by the set of high-level predicates  $\mathcal{P}$ . Each predicate  $p \in \mathcal{P}$  can be mapped to a continuous finite-dimensional low-level substate vector  $s_p$ . For example, the location property and the velocity property of an object in Cartesian space are both fully determined by a three-dimensional continuous vector.
- C.5 For each predicate  $p$ , there exists a sequence of indices  $p^{idx}$  that determines the indices of the low-level environment state vector  $s$  that determines the property described by  $p$ . For example, given that  $p$  refers to the object being at a specific location, and given that the first three values of  $s$  determine the Cartesian location of an object, we have that  $p^{idx} = [0, 1, 2]$ . A requirement is that the indices of the individual predicates must not overlap, i.e., abusing set notation:  $p_1^{idx} \cap p_2^{idx} = \emptyset$  (see Algorithm 1 for details).
- C.6 The high-level action space consists of a set of grounded STRIPS operators (Fikes and Nilsson, 1972)  $a$  that are determined by a conjunction of precondition literals and a conjunction of effect literals.
- C.7 A low-level goal  $g$  is the subset of the low-level state  $s$ , indicated by the indices  $g^{idx}$ , i.e.,  $g = s[g^{idx}]$ . For example, consider that the low-level state  $s$  is a six-dimensional vector where the first three elements represent the location and the last three elements represent the velocity of an object in Cartesian space. Then, given that  $g^{idx} = [0, 1, 2]$ , we have that  $g = s[g^{idx}]$  refers to the location of the object.

### 3.1.1. Mapping Predicates to Low-Level Subgoals

Abstracting from observations to high-level predicate specifications is achieved by mapping the low-level state space to a high-level conceptual space. This is realized with a set

of functions  $f_{subg}^p$  that we define manually for each predicate  $p$ . For a given predicate  $p$ , the function  $f_{subg}^p$  generates the low-level substate  $s_p$  that determines  $p$ , based on the current state  $s$  and the goal  $g$ :

$$f_{subg}^p(s, g) = s_p \tag{1}$$

To illustrate how  $f_{subg}^p$  can be implemented, consider the following two examples from a block-stacking task:

- 1 Consider a predicate (`at_target o1`) which indicates whether an object is at a given goal location on the surface of a table. Then the respective function  $f_{subg}^{(at\_target\ o1)}(s, g)$  can be implemented as follows:

$$f_{subg}^{(at\_target\ o1)}(s, g) = [g[0], g[1], g[2]] \tag{2}$$

In this case, the function extracts the respective target coordinates for the object `o1` directly from  $g$  and does not require any information from  $s$ .

- 2 Consider further a predicate (`on o2 o1`), which is true if an object `o2` is placed on top of another object `o1`. Given that the Cartesian location of `o1` is defined by the first three elements of the state vector  $s$ , one can define the following subgoal function:

$$f_{subg}^{(on\ o2\ o1)}(s, g) = [s[0], s[1], s[2] + h_{obj}], \text{ where } h_{obj} \text{ denotes the height of an object} \tag{3}$$

Here, the target coordinates of the object `o2` are computed by considering the current coordinates of `o1`, i.e., the first three values of  $s$ , and by adding a constant for the object height to the third (vertical axis) value.

### 3.1.2. Grounding High-Level Representations in Low-Level Subgoals

The function  $f_{subg}$  that maps the high-level subgoal state  $G_{subg}$  in the context of  $s, g$  to a low-level subgoal  $g_{subg}$  builds on Equation (1), as described with the following Algorithm 1:

---

**Algorithm 1:** Mapping propositional state representations to continuous state representations

---

```

1: function  $f_{subg}(G_{subg}, s, g)$ 
2:    $s_{subg} \leftarrow s$ 
3:   subgoal_changed  $\leftarrow$  True
4:   while subgoal_changed do
5:      $s_{subg}^{last} \leftarrow s_{subg}$ 
6:     for  $p \in \mathcal{P}$  do  $\triangleright$  For each predicate in high-level goal
7:       state, set low-level subgoal indices  $p^{idx}$ 
8:        $s_{subg}[p^{idx}] \leftarrow s_p$ , where  $s_p = f_{subg}^p(s_{subg}, g)$ 
8:     subgoal_changed  $\leftarrow$   $s_{subg}^{last} \neq s_{subg}$ 

```

---

The while loop is necessary to prevent the situation where the application of  $f_{subg}^p$  (in line 7, Algorithm 1) changes  $s_{subg}$

in a manner that affects a previous predicate subgoal function. For example, consider the two predicates  $(at\_target\ o1)$  and  $(on\ o2\ o1)$ . The predicate  $(at\_target\ o1)$  determines the Cartesian location of  $o1$ , and  $(on\ o2\ o1)$  depends on these coordinates. Therefore, it may happen that  $f_{subg}^{(at\_target\ o1)} = [x, y, z]$  causes the first three elements of  $s_{subg}$  to be  $x, y, z$ . However,  $f_{subg}^{(on\ o2\ o1)} = [x', y', z']$  depends on these  $x, y, z$  to determine the  $x', y', z'$  that encode the location of  $o2$  in  $s_{subg}$ . The while loop assures that  $f_{subg}^{(on\ o2\ o1)}$  is applied at least once after  $f_{subg}^{(at\_target\ o1)}$  to consider this dependency. This assures that all dependencies between the elements of  $s_{subg}$  are resolved.

To guarantee the termination of the Algorithm 1, i.e., to avoid that the alternating changes of  $s_{subg}$  cause an infinite loop, the indices  $p^{idx}$  must be constrained in such a way that they do not overlap (see assumption C.5).

### 3.1.3. Abstracting From Low-Level State- and Goal Representations to Propositional Statements

To realize the abstraction from low-level to high-level representations, we define a set of functions in the form of Equation 4. Specifically, for each predicate  $p$ , we define the following function  $f_p$  maps the current low-level state and the low-level goal to the predicates' truth values.

$$f_p(p, s) = \text{diff}(s[p^{idx}], s_p) < \epsilon, \text{ with } s_p = f_{subg}^p(s, g) \quad (4)$$

Equation 4 examines whether the subgoal that corresponds to a specific predicate is true, given the current observed state  $s$  and a threshold value for the coordinates  $\epsilon$ . In this article, but without any loss of generality, we assume that each predicate is determined by three coordinates. Equation 4 computes the difference between these coordinates given the current state, and the coordinates determined by  $f_{subg}^p$  as their Euclidean distance. For example,  $f_{subg}^{(at\_target\ o1)}$  may define target coordinates for  $o1$ , and Equation 4 considers the distance between the target coordinates and the current coordinates.

## 3.2. Generation of Adaptive Low-Level Action Trajectories

To address the research problem P.2, i.e., the grounding of actions and subgoals in low-level action trajectories, we consider continuous goal-independent reinforcement learning approaches (Schaul et al., 2015; Lillicrap et al., 2016). Most reinforcement learning approaches build on manually defined reward functions based on a metric that is specific to a single global goal, such as the body height, posture and forward speed of a robot that learns to walk (Schulman et al., 2015). Goal-independent reinforcement learning settings do not require such reward shaping (c.f. Ng et al., 1999), as they allow one to parameterize learned policies and value functions with goals. We employ the actor-critic deep deterministic policy gradient (DDPG) (Lillicrap et al., 2016) approach in combination with hindsight experience replay (HER) (Andrychowicz et al., 2017) to realize the goal-independent RL for the continuous control

part in our framework. Using the HER technique with off-policy reinforcement learning algorithms (like DDPG) increases the efficiency of sampling for our approach since HER stores not only the experienced episode with the original goal  $[episode \leftarrow (s_t, u_t, s_{t+1}, g)]$  in the replay buffer. It also stores modified versions of an episode where the goal is retrospectively set to a state that has been achieved during the episode, i.e.,  $episode' \leftarrow (s_t, u_t, s_{t+1}, g')$  with  $g' = s_{t'}[g^{idx}]$  for some  $t' > t$ .

To realize this actor-critic architecture, we provide one fully connected neural network for the actor  $\pi$ , determined by the parameters  $\theta^\pi$  and one fully connected neural network for the critic  $Q$ , determined by the parameters  $\theta^Q$ . The input to both networks is the concatenation of the low-level state  $s$  and the low-level subgoal  $g_{sub}$ . The optimization criterion for the actor  $\pi$  is to minimize the  $q$ -value provided by the critic. The optimization criterion for the critic is to minimize the mean squared error between the critic's output  $q$  and the discounted expected reward according to the Bellmann equation for deterministic policies, as described by Equation (5).

$$\arg \min_{\theta^Q} (q - Q(s_t, u_t, g_{sub} | \theta^Q))^2,$$

$$\text{where } Q(s_t, u_t, g_{sub} | \theta^Q) = r_t + \gamma \cdot Q(s_{t+1}, \pi(s_{t+1}, g_{sub} | \theta^\pi), g_{sub} | \theta^Q) \quad (5)$$

Given that the action space is continuous  $n$ -dimensional, the observation space is continuous  $m$ -dimensional, and the goal space is continuous  $k$ -dimensional with  $k \leq m$ , the following holds for our theoretical framework: At each step  $t$ , the agent executes an action  $u_t \in \mathbb{R}^n$  given a state  $s_t \in \mathbb{R}^m$  and a goal  $g \subseteq \mathbb{R}^k$ , according to a *behavioral policy*, a noisy version of the *target policy*  $\pi$  that deterministically maps the observation and goal to the action<sup>1</sup>. The action generates a reward  $r_t = 0$  if the goal is achieved at time  $t$ . Otherwise, the reward is  $r_t = -1$ . To decide whether a goal has been achieved, a function  $f(s_t)$  is defined that maps the observation space to the goal space, and the goal is considered to be achieved if  $|f(s_t) - g| < \epsilon$  for a small threshold  $\epsilon$ . This sharp distinction of whether or not a goal has been achieved based on a distance threshold causes the reward to be sparse and renders shaping the reward with a hand-coded reward function unnecessary.

## 3.3. Integration of High-Level Planning With Reinforcement Learning

Our architecture integrates the high-level planner and the low-level reinforcement learner as depicted in **Figure 2**. The input to our framework is a low-level goal  $g$ . The sensor data that represents the environment state  $s$  is abstracted together with  $g$  to a propositional high-level description of the world state  $S$  and goal  $G$ . An action planner based on the planning domain definition language (PDDL) (McDermott et al., 1998) takes these high-level representations as input and computes a high-level plan based on manually defined action definitions (c.f. the **Supplementary Material** for examples). We have implemented

<sup>1</sup>Applying action noise for exploration purposes is a common practice for off-policy reinforcement learning (e.g.,  $\epsilon$ -greedy). The kind of noise that we investigate in this article (e.g., in the experiments section 3.4) is supposed to simulate perceptual errors and not to be confused with this action noise for exploration.

the caching of plans to accelerate the runtime performance. The high-level subgoal state  $G_{sub}$  is the expected successor state of the current state given that the first action of the plan is executed. This successor state is used as a basis to compute the next subgoal. To this end,  $G_{sub}$  is processed by the subgoal grounding function  $f_{subg}$  (Algorithm 1) that generates a subgoal  $g_{sub}$  as input to the low-level reinforcement learner.

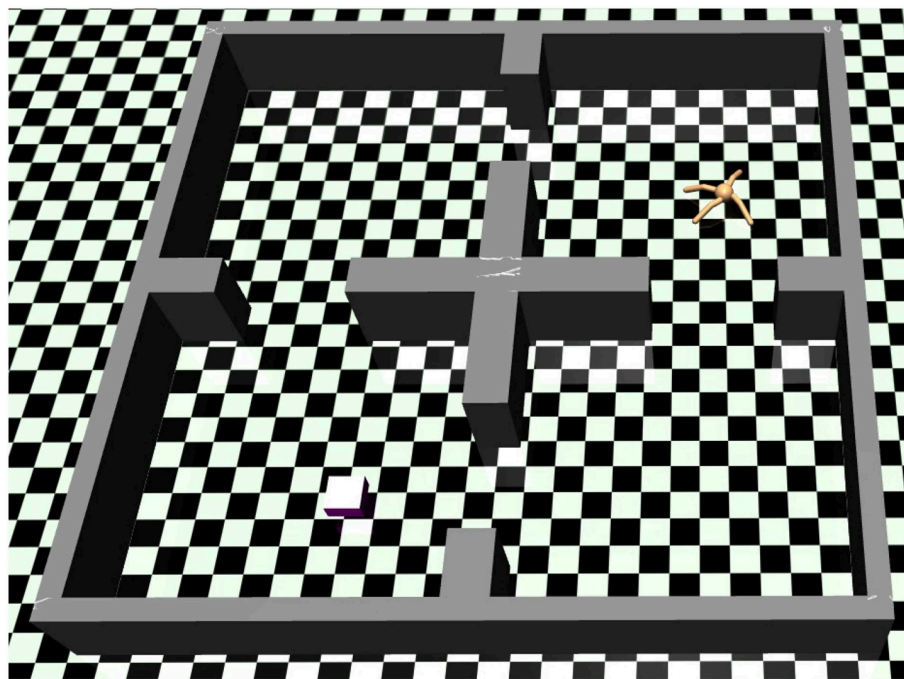
## 4. EXPERIMENTS

This section describes three experiments, designed for the evaluation of the proposed approach. We refer to the experiments as *block-stacking* (Figure 1, left), *tool use* (Figure 1, right), and *ant navigation* (Figure 3). The first two experiments are conducted with a Fetch robot arm, and the latter is adapted from research on continuous reinforcement learning for legged locomotion (Levy et al., 2019). All experiments are executed in the Mujoco simulation environment (Todorov et al., 2012). For all experiments, we use a three-layer fully connected neural network with the rectified linear unit (ReLU) activation function to represent the actor-critic network of the reinforcement learner in both experiments. We choose a learning rate of 0.01, and the networks' weights are updated using a parallelized version of the Adam optimizer (Kingma and Ba, 2015). We use a reward discount of  $\gamma = 1 - 1/T$ , where  $T$  is the number of low-level actions per rollout. For the block-stacking, we use 50, 100, and 150 low-level actions for the case of one, two and three blocks, respectively. For the tool use experiment, we use 100 low-level actions, and for the ant navigation, we used 900 low-level actions.

Preliminary hyperoptimization experiments showed that the optimal number of units for each layer of the neural networks for actor and critic of the reinforcement learning depends on the observation space. Therefore, we implement the network architecture such that the number of units in each layer scales with the size of the observation vector. Specifically, the layers in the actor and critic consist of 12 units per element in the observation vector. For example, for the case of the block-stacking experiment with two blocks, this results in 336 neural units per layer (see section 4.1). We apply the same training strategy of HER (Andrychowicz et al., 2017), evaluate periodically learned policies during training without action noise. We use a fixed maximum number of epochs and early stopping at between 80 and 95% success rate, depending on the task.

In all experiments, we evaluate the robustness of our approach to perceptual noise. That is, in the following we refer to perceptual noise, and not to the action noise applied during the exploration phase of the RL agent, if not explicitly stated otherwise. To evaluate the robustness to perceptual noise, we consider the amount of noise relative to the value range of the state vector. To this end, we approximate the continuous-valued state range, denoted  $rng$ , as the difference between the upper and lower quartile of the elements in the history of the last 5,000 continuous-valued state vectors that were generated during the rollouts<sup>2</sup>. For each action step in the rollout we randomly sample

<sup>2</sup>We use the upper and lower quartile and not the minimum and maximum of all elements to eliminate outliers.



**FIGURE 3** | An ant agent performing a navigation and locomotion task in a four-room environment. Herein, the agent needs to learn how to walk and find a way to reach the desired position. In this case, the agent needs to walk from the upper right room to the target location in the lower left room.



noise, denoted  $s_\gamma$ , according to a normal distribution with  $rng$  being the standard deviation. We add this noise to the state vector. To parameterize the amount of noise added to the state, we define a noise-to-signal-ratio  $\kappa$  such that the noise added to the state vector is computed as  $s_{noisy} = s + \kappa \cdot s_\gamma$ . We refer to the noise level as the percentage corresponding to  $\kappa$ . That is, e.g.,  $\kappa = 0.01$  is equivalent to a noise level of 1%.

For all experiments, i.e., block-stacking, tool use and ant navigation, we trained the agent on multiple CPUs in parallel and averaged the neural network weights of all CPU instances after each epoch, as described by Andrychowicz et al. (2017). Specifically, we used 15 CPUs for the tool use and the block-stacking experiments with one and two blocks; we used 25 CPUs for the block-stacking with 3 blocks. For the ant navigation, we used 15 CPUs when investigating the robustness to noise (Figure 7) and 1 CPU when comparing our approach to the framework by Levy et al. (2019) (Figure 8). The latter was necessary to enable a fair comparison between the approaches because the implementation of the architecture of Levy et al. (2019) supports only a single CPU. For all experiments, an epoch consists of 100 training rollouts per CPU, followed by training the neural networks for actor and critic with 15 batches after each epoch, using a batch size of 256. The results in Figures 4, 6 illustrate the median and the upper and lower quartile over multiple ( $n \geq 5$ ) repetitions of each experiment.

## 4.1. Block-Stacking

Figure 1 (left) presents the simulated environment for this experiment, where a number of blocks (i.e., up to three) are placed randomly on the table. The task of the robot is to learn how to reach, grasp and stack those blocks one-by-one to their corresponding random target location. The task is considered completed when the robot successfully places the last block on top of the others in the right order, and moves its gripper to another random target location. The difficulty of this task increases with the number of blocks to stack. The order in which the blocks need to be stacked is randomized for each rollout. The causal dependencies involved here are, that a block can only be grasped if the gripper is empty, a block (e.g., A) can only be placed on top of another block (e.g., B) if there is no other block (e.g., C) already on top of either A or B, etc.

The size of the goal space depends on the number of blocks. For this experiment, the goal space is a subset of the state-space that is constituted by the three Cartesian coordinates of the robot's gripper and three coordinates for each block. That is, the dimension of the goal- and subgoal space is  $k = (1 + n_o) \cdot 3$ , where  $n_o \in \{1, 2, 3\}$  is the number of objects.

The state-space of the reinforcement learning agent consists of the Cartesian location and velocity of the robot's gripper, the gripper's opening angle, and the Cartesian location, rotation, velocity, and rotational velocity of each object. That is, the size of the state vector is  $|s| = 4 + n_o \cdot 12$ , where  $n_o$  is the number of blocks.

The planning domain descriptions for all environments are implemented with the PDDL actions and predicates provided in the **Supplementary Material**, section A.

## 4.2. Tool Use

The environment utilized for this experiment is shown in Figure 1 (right). A single block is placed randomly on the table, such that it is outside the reachable region of the Fetch robot. The robot has to move the block to a target position (which is randomized for every rollout) within the reachable (dark brown) region on the table surface. In order to do so, the robot has to learn how to use the provided rake. The robot can drag the block either with the left or the right edge of the rake. The observation space consists of the Cartesian velocities, rotations, and locations of the robot's gripper, the tip of the rake, and the object. An additional approximation of the end of the rake is added in this task. The goal space only contains the Cartesian coordinates of the robot's gripper, the tip of the rake, and the object. The planning domain description for this tool use environment can be found in the **Supplementary Material**, section B.

## 4.3. Ant Navigation

The environment of navigation and locomotion in a four-connected-room scenario is shown in Figure 3, where the ant has to find a way to the randomly allocated goal location inside one of the four-rooms. The state-space consists of the Cartesian location and transitional velocity of the ant's torso, along with the joint position and velocity of the eight joints of the ant (i.e., each leg has one ankle joint and one hip joint). The goal space contains the Cartesian coordinate of the ant's torso. There are no other objects involved in the task. The planning domain description and the high-level action specifications for this navigation environment can be found in the **Supplementary Material**, section C.

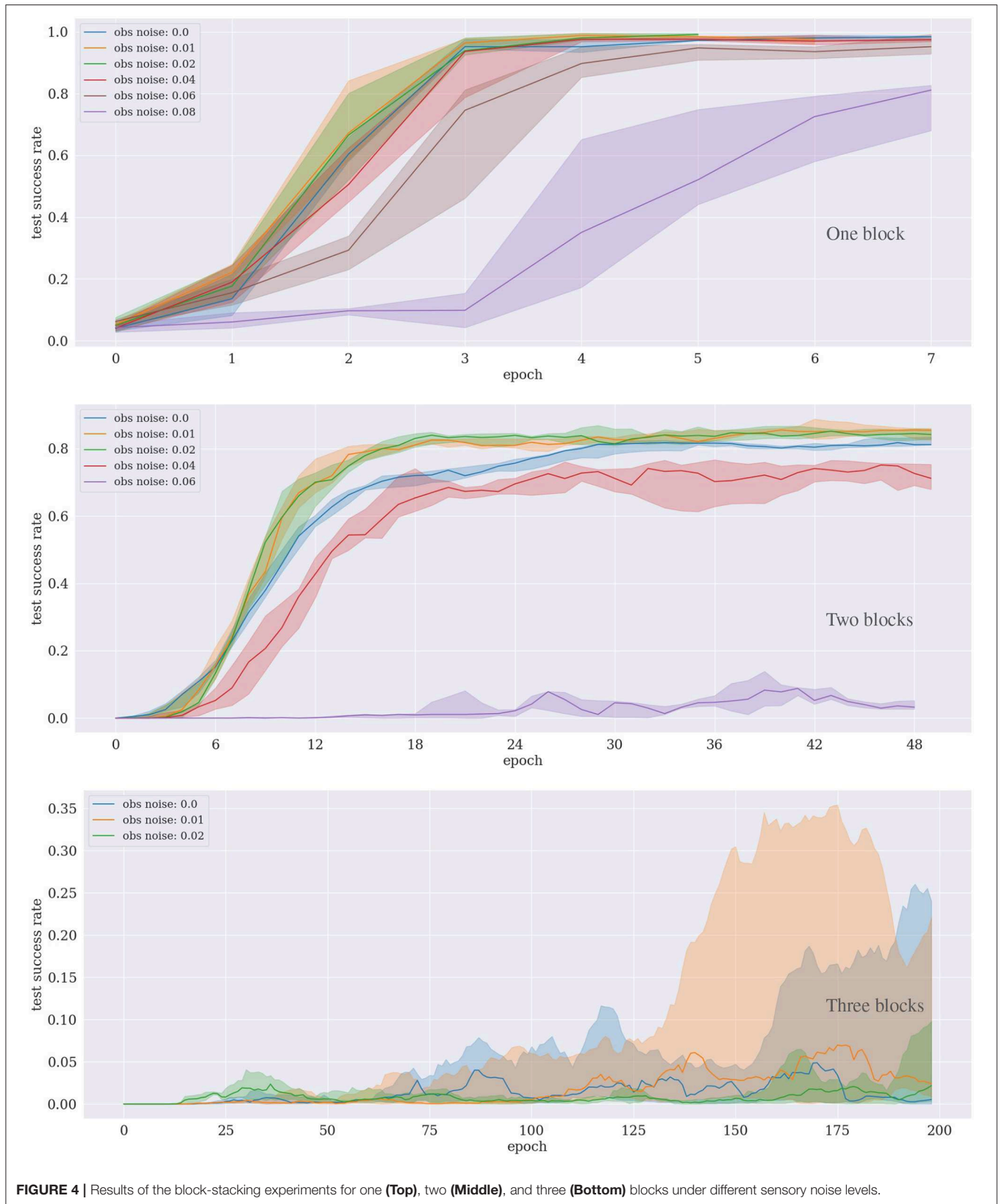
## 5. RESULTS

To evaluate our approach, we investigate the success rate of the testing phase over time for all experiments, given varying noise levels  $\kappa$  (see section 4). The success rate is computed per epoch, by averaging over the number of successful rollouts per total rollouts over ten problem instances per CPU.

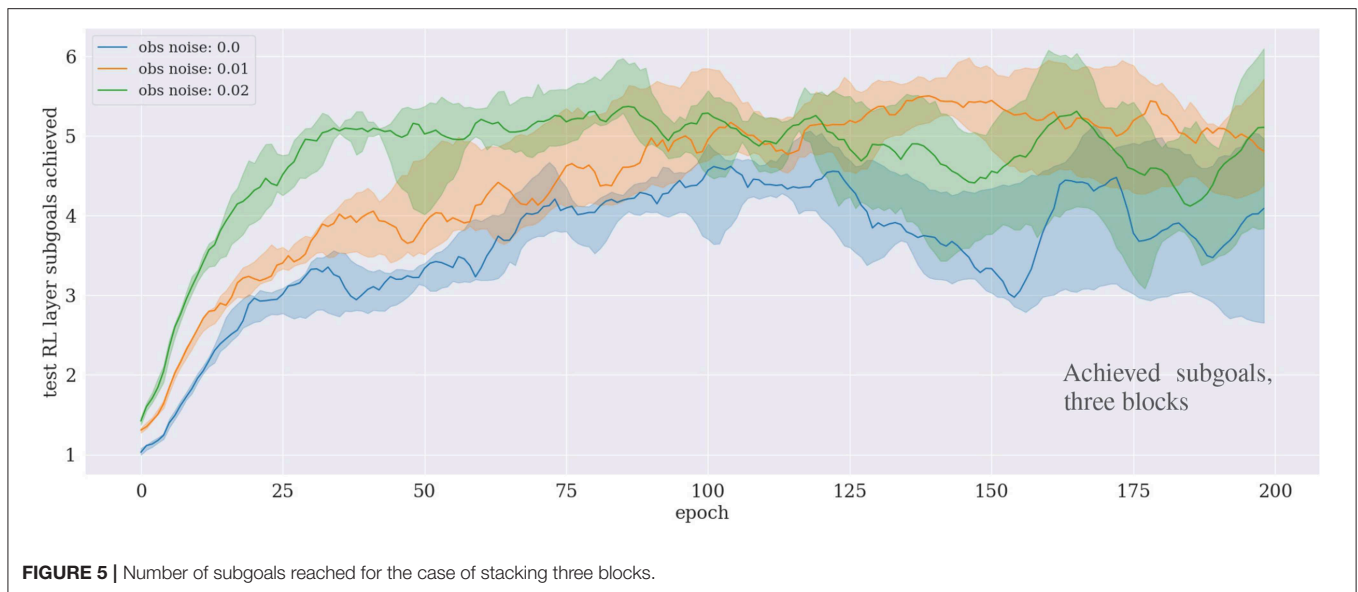
### 5.1. Block-Stacking

For the experiment with one block, the approach converges after around ten epochs, even with  $\kappa = 0.06$ , i.e., if up to 6% noise is added to the observations. The results are significantly worse for 8% and more noise. For the experiment with two blocks, the performance drops already for 6% noise. Interestingly, for both one and two blocks, the convergence is slightly faster if a small amount (1–2%) of noise is added, compared to no noise at all. The same seems to hold for three blocks, although no clear statement can be made because the variance is significantly higher for this task.

For the case of learning to stack three blocks consider also Figure 5, which shows how many subgoals have been achieved on average during each epoch. For our particular PDDL implementation, six high-level actions, and hence six subgoals, are at least required to solve the task: `[move_gripper_to(o1), move_to_target(o1), move_gripper_to(o2), move_o_on_o(o2,o1), move_gripper_to(o3), move_o_on_o(o3,o2)].`



**FIGURE 4** | Results of the block-stacking experiments for one (Top), two (Middle), and three (Bottom) blocks under different sensory noise levels.



**FIGURE 5** | Number of subgoals reached for the case of stacking three blocks.

First, the gripper needs to move to the randomly located object  $\circ 1$ , then, since the target location of the stacked tower is also randomly selected, the gripper needs to transport  $\circ 1$  to the target location. Then the gripper moves to  $\circ 2$  to place it on top of  $\circ 1$ , and repeats these steps for  $\circ 3$ . The result shows that the agent can consistently learn to achieve the first five subgoals on average, but is not able to proceed further. This demonstrates that the agent robustly learns to stack the first two objects, but fails to stack also the third one.

### 5.2. Tool Use

The results in **Figure 6** reveal that our proposed approach allows the agent to learn and complete the task in under 100 training epochs (corresponding to  $\sim 8$  h with 15 CPUs) even with a noise level increased up to 4% of the state range. We observe that it becomes harder for the agent to learn when the noise level exceeds 4%. In the case of 8% noise, the learning fails to achieve a reasonable performance in the considered time—first 100 training epochs (i.e., it only obtains  $<1\%$  success rate). Interestingly, in cases with very low noise levels (1–2%), the learning performance is better or at least as good as the case with no noise added at all.

### 5.3. Ant Navigation

**Figure 7** presents the performance of trained agents following our proposed approach in the ant navigation scenario. The results show that the agent can learn to achieve the task in  $<30$  training epochs under the low noise level conditions (up to 1%). The performance decreases slightly in the case of 1.5% but the agent still can learn the task after around 70 training epochs. With a higher noise level (i.e., 2%), the agent requires longer training time to cope with the environment.

### 5.4. Comparison With Hierarchical Reinforcement Learning

Results in **Figure 8** depict the benchmark experiment of our proposed approach with the HRL approach by Levy et al.

(2019). Though the HRL approach quickly learns the task at the beginning, it does not exceed a success rate of 70%. In comparison, our approach learns to solve the task more reliably, eventually reaching 100%, but the success rate grows significantly later, at around 50 epochs.

## 6. DISCUSSION

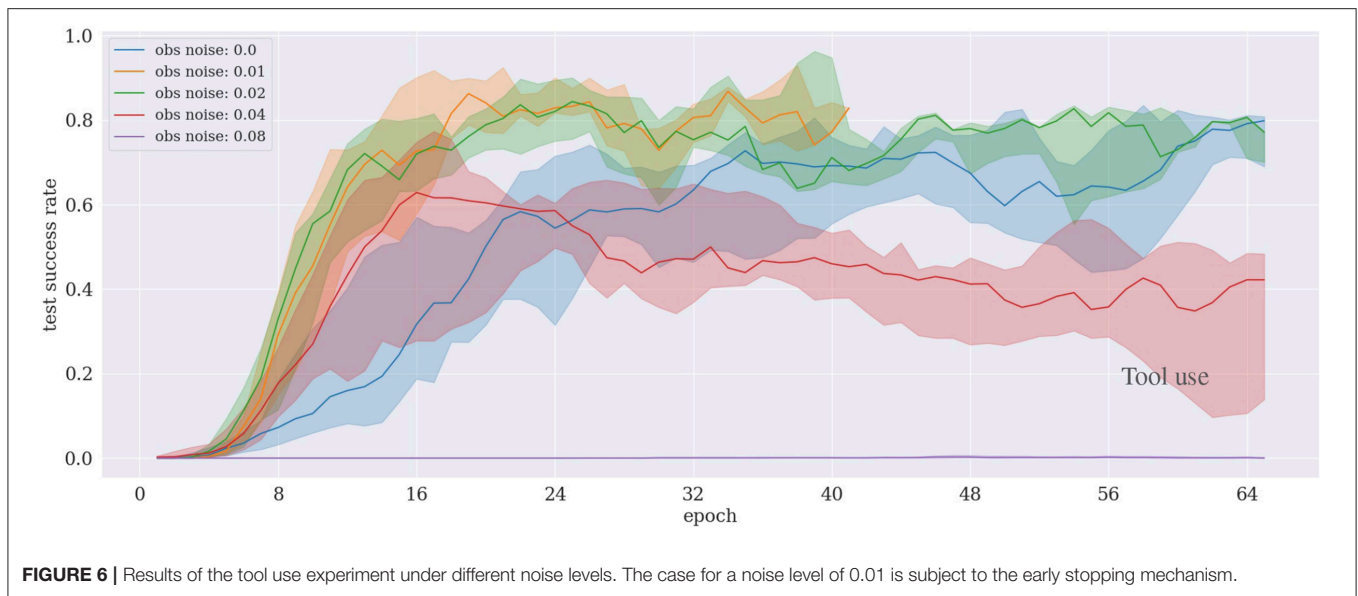
The results indicate that our proof-of-concept addresses the hypotheses **H.1** and **H.2** as follows:

### 6.1. Hypothesis H.1: Ability to Ground High-Level Actions in Low-Level Trajectories

Our experiments indicate that the grounding of high-level actions in low-level RL-based robot control using the HER approach performs well for small to medium-sized subgoal spaces. However, learning is not completely automated, as the approach requires the manual definition of the planning domain and of the functions  $f_{subg}^p$  that maps planning domain predicates to subgoals.

For the tasks of stacking two blocks and the tool use, the subgoal space involved nine values, and both tasks could be learned successfully. The qualitative evaluation and visual inspection of the agent in the rendered simulation revealed that the grasping of the first and second block failed more often for the experiment of stacking three blocks than for the experiment of stacking two blocks. Therefore, we conclude that the subgoal space for stacking three blocks, which involves twelve values, is too large.

However, the performance on the control-level was very robust. For example, it happened frequently during the training and exploration phase that the random noise in the actions caused a block to slip out of the robot’s grip. In these cases, the agent was able to catch the blocks immediately while they were



**FIGURE 6** | Results of the tool use experiment under different noise levels. The case for a noise level of 0.01 is subject to the early stopping mechanism.



**FIGURE 7** | Results of the ant navigation experiment under different noise levels. The curves are subject to early stopping.

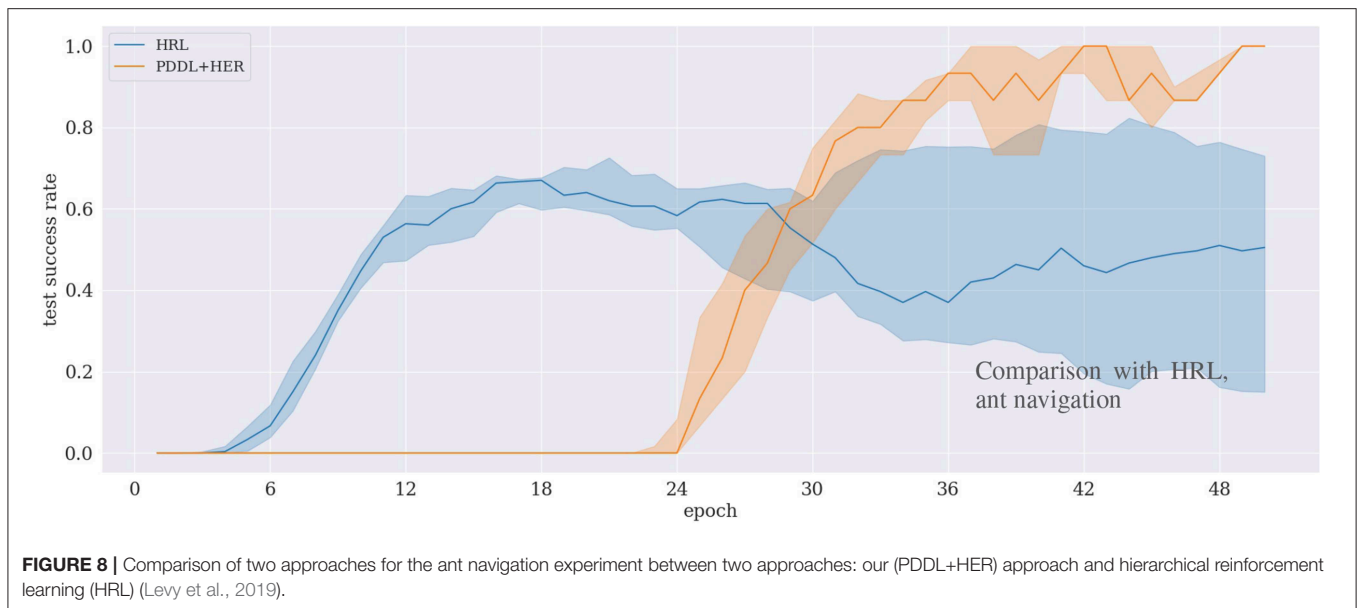
falling down. During the tool use experiment, the agent was also able to consider the rotation of the rake, to grasp the rake at different positions, and to adapt its grip it when it was slipping.

The results indicate that the approach is able to solve causal puzzles if the subgoal space is not too large. The architecture depends strongly on the planning domain representation that needs to be implemented manually. In practice, the manual domain-engineering that is required for the planning is appropriate for tasks that are executed frequently, such as adaptive robotic co-working at a production line or in a drone delivery domain. Due to the caching of plans (see section 3.3), we have not encountered issues with the computational complexity problem and run-time issues of the planning approach.

Our measure of appropriateness that we state in H.1 is to evaluate whether our method outperforms a state-of-the-art HRL

approach. **Figure 8** depicts that this is the case in terms of the final success rate. Specifically, the figure shows that the HRL approach learns faster initially, but never reaches a success rate of more than 70%, while our approach is slower but reaches 100%. A possible explanation for this behavior is that HRL implements a “curriculum effect” (cf. Eppe et al., 2019), in the sense that it first learns to solve simple subgoals due to its built-in penalization of difficult subgoals. However, as the success rate increases, there are fewer unsuccessful rollouts to be penalized which potentially leads to more difficult subgoals and, consequently, a lower overall success rate. This curriculum effect is not present in our approach because the planning mechanism does not select subgoals according to their difficulty. Investigating and exploiting this issue in detail is potentially subject to further research.





## 6.2. Hypothesis H.2: Robustness to Noise

For the block-stacking with up to two blocks and the tool-use experiments, the approach converged with a reasonable noise-to-signal ratio of four to six percent. For the block-stacking with three blocks and for the ant environment, a smaller amount of noise was required. An interesting observation is that a very low amount of random noise, i.e.,  $\kappa = 0.01$ , improves the learning performance for some cases. Adding random noise, e.g., in the form of dropout, is a common technique to improve neural network-based machine learning because it helps neural networks to generalize better from datasets. One possible explanation for the phenomenon is, therefore, that the noise has the effect of generalizing the input data for the neural network training, such that the parameters become more robust.

Noise is an important issue for real physical robots. Our results indicate that the algorithm is potentially appropriate for physical robots, at least for the case of grasping and moving a single block to a target location. For this case, with a realistic level of noise, the algorithm converged after  $\sim 10$  epochs (see **Figure 4**). Per epoch and CPU, the agent conducts 100 training rollouts. A single rollout would take around 20 seconds on a physical robot. Considering that we used 15 CPUs, the equivalent robot training time required is  $\sim 83$  h. For the real application, the physical training time can potentially further be lowered by applying more neural network training batches per rollout, and by performing pre-training using the simulation along with continual learning deployment techniques, such as the method proposed by Traoré et al. (2019).

## 7. CONCLUSION

We have developed a hierarchical architecture for robotic applications in which agents must perform reasoning over a

non-trivial causal chain of actions. We have employed a PDDL planner for the high-level planning and we have integrated it with an off-policy reinforcement learner to enable robust low-level control.

The innovative novelty of our approach is the combination of action planning with goal-independent reinforcement learning and sparse rewards (Lillicrap et al., 2016; Andrychowicz et al., 2017). This integration allowed us to address two research problems that involve the grounding of the discrete high-level state and action space in sparse rewards for low-level reinforcement learning. We addressed the problem of grounding of symbolic state spaces in continuous-state subgoals (P.1), by proposing a principled predicate-subgoal mapping, which involves the manual definitions of functions  $f_{subg}^p$  for each predicate  $p$ .

We assume that the manual definition of functions  $f_{subg}^p$  generally involves less engineering effort than designing a separate reward function for each predicate. Although this assumption heavily depends on the problem domain and may be subject to further discussion, the manual definition of functions  $f_{subg}^p$  is at least a useful scaffold for further research that investigates the automated learning of functions  $f_{subg}^p$ , possibly building on the research by Ugur and Piater (2015).

The predicate-subgoal mapping is also required to address the problem of mapping subgoals to low-level action trajectories (P.2) by means of reinforcement learning with sparse rewards using hindsight experience replay (Andrychowicz et al., 2017). Our resulting approach has two advantages over other methods that combine action planning and reinforcement learning, e.g., Grounds and Kudenko (2005), Yamamoto et al. (2018): The low-level action space for our robotic application is continuous and it supports a higher dimensionality.

We have realized and evaluated our architecture in simulation, and we addressed two hypotheses (H.1 and H.2): First, we

demonstrate that the approach can successfully integrate high-level planning with reinforcement learning and this makes it possible to solve simple causal puzzles (H.1); second, we demonstrate robustness to a realistic level of sensory noise (H.2). The latter demonstrates that our approach is potentially applicable to real-world robotic applications. The synthetic noise used in our experiments does not yet fully guarantee that our approach is capable of bridging the reality gap, but we consider it a first step toward real robotic applications (see also Andrychowicz et al., 2018; Nguyen et al., 2018; Traoré et al., 2019).

The causal puzzles that we investigate in this paper are also relevant for hierarchical reinforcement learning (HRL) (e.g., Levy et al., 2019), but we have not been able to identify an article that presents good results in problem-solving tasks that have a causal complexity comparable to our experiments. An empirical comparison was, therefore, not directly possible. Our approach has the advantage over HRL that it exploits domain knowledge in the form of planning domain representations. The disadvantage compared to HRL is that the domain knowledge must be hand-engineered. In future work, we plan to complement both approaches, e.g., by building on vector-embeddings to learn symbolic planning domain descriptions from scratch by means of the reward signal of the reinforcement learning. A similar approach, based on the clustering of affordances, has been presented by Ugur and Piater (2015), and complementing their method with reinforcement learning suggests significant potential. An overview of this topic and potential approaches is provided by Lesort et al. (2018). We also plan to apply the approach to a physical robot and to reduce the amount of physical training time by pre-training the agent in our simulation and by applying domain-randomization techniques (Andrychowicz et al., 2018).

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## REFERENCES

- Alili, S., Pandey, A. K., Sisbot, A., and Alami, R. (2010). "Interleaving symbolic and geometric reasoning for a robotic assistant," in *ICAPS Workshop on Combining Action and Motion Planning* (Toronto, ON).
- Andre, D., Friedman, N., and Parr, R. (1998). "Generalized prioritized sweeping," in *Advances in Neural Information Processing Systems 10*, eds M. I. J. Solla, M. J. Kearns, and S. A. Solla (Denver, CO: MIT Press), 1001–1007.
- Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., et al. (2018). Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*.
- Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., et al. (2017). "Hindsight experience replay," in *Conference on Neural Information Processing Systems (NIPS)* (Long Beach, CA), 5048–5058.
- Aytar, Y., Pfaff, T., Budden, D., Paine, T. L., Wang, Z., and de Freitas, N. (2018). "Playing hard exploration games by watching YouTube," in *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, QC), 2930–2941.
- Bacon, P.-L., Harb, J., and Precup, D. (2017). "The option-critic architecture," in *AAAI Conference on Artificial Intelligence* (San Francisco, CA), 1726–1734.

## AUTHOR CONTRIBUTIONS

ME has developed the overall concept and theoretical foundation of the method. He also implemented, developed and evaluated the computational framework together with PN. ME authored the core parts of sections 1 and 2, and he provided foundational contributions to all other sections. PN has researched and investigated significant parts of the related work in section 2, and he also worked on further refinements and adaptations of the theoretical background provided in section 3 of this article. He has revised and contributed also to all other sections of this article. Furthermore, PN has developed the tool use experiment. SW has actively contributed and revised the article.

## FUNDING

ME and SW acknowledge funding by the Experiment! Programme of the Volkswagen Stiftung. ME, SW, and PN acknowledge support via the German Research Foundation (DFG) within the scope of the IDEAS project of the DFG priority programme The Active Self. Furthermore, we acknowledge supports from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No. 642667 (SECURE) and No. 721385 (SOCRATES).

## ACKNOWLEDGMENTS

We thank Erik Strahl for the technical supports with the computing servers and Fares Abawi for the initial version of the Mujoco simulation for the experiments. We also thank Andrew Levy for providing the code of his hierarchical actor-critic reinforcement learning approach (Levy et al., 2019).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frobt.2019.00123/full#supplementary-material>

- Baxter, J., and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *J. Artif. Intel. Res.* 15, 319–350. doi: 10.1613/jair.806
- de Silva, L., Pandey, A. K., Gharbi, M., and Alami, R. (2013). "Towards combining HTN planning and geometric task planning. in *RSS Workshop on Combined Robot Motion Planning and AI Planning for Practical Applications* (Berlin).
- Deisenroth, M. P., and Rasmussen, C. E. (2011). "PILCO: a model-based and data-efficient approach to policy search," in *International Conference on Machine Learning (ICML)* (Bellevue, WA), 465–472.
- Deits, R., and Tedrake, R. (2014). "Footstep planning on uneven terrain with mixed-integer convex optimization," in *2014 IEEE-RAS International Conference on Humanoid Robots* (Madrid: IEEE), 279–286.
- Doncieux, S., Filliat, D., Díaz-Rodríguez, N., Hospedales, T., Duro, R., Coninx, A., et al. (2018). Open-ended learning: a conceptual framework based on representational redescription. *Front. Neurobot.* 12:59. doi: 10.3389/fnbot.2018.00059
- Eppe, M., Bhatt, M., and Dylla, F. (2013). "Approximate epistemic planning with postdiction as answer-set programming," in *International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)* (Corunna), 290–303.

- Eppe, M., Magg, S., and Wermter, S. (2019). "Curriculum goal masking for continuous deep reinforcement learning," in *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Berlin, Heidelberg), 183–188.
- Fikes, R., and Nilsson, N. (1972). STRIPS: a new approach to the application of theorem proving to problem solving. *Art. Intel.* 2, 189–208.
- Forestier, S., Mollard, Y., and Oudeyer, P.-Y. (2017). Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*.
- Grounds, M., and Kudenko, D. (2005). "Combining reinforcement learning with symbolic planning," in *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning* (Berlin; Heidelberg: Springer Berlin Heidelberg), 75–86.
- Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)* (San Diego, CA).
- Kulkarni, T. D., Narasimhan, K. K. R., Saeedi, A., and Tenenbaum, J. B. (2016). "Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation," in *Conference on Neural Information Processing Systems (NIPS)* (Barcelona), 3675–3683.
- Lagriffoul, F., Dantam, N. T., Garrett, C., Akbari, A., Srivastava, S., and Kavraki, L. E. (2018). Platform-independent benchmarks for task and motion planning. *IEEE Robot. Automat. Lett.* 3, 3765–3772. doi: 10.1109/LRA.2018.2856701
- Leidner, D., Bartels, G., Bejjani, W., Albu-Schäffer, A., and Beetz, M. (2018). Cognition-enabled robotic wiping: Representation, planning, execution, and interpretation. *Robot. Autonom. Syst.* 114, 199–216. doi: 10.1016/j.robot.2018.11.018
- Lesort, T., Díaz-Rodríguez, N., Goudou, J. I., and Filliat, D. (2018). State representation learning for control: an overview. *Neural Netw.* 108, 379–392. doi: 10.1016/j.neunet.2018.07.006
- Levy, A., Konidaris, G., Platt, R., and Saenko, K. (2019). "Learning multi-level hierarchies with hindsight," in *International Conference on Learning Representations (ICLR)* (New Orleans, LA).
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2016). "Continuous control with deep reinforcement learning," in *International Conference on Learning Representations (ICLR)* (San Juan).
- Ma, J., and Cameron, S. (2009). "Combining policy search with planning in multi-agent cooperation," in *RoboCup 2008: Robot Soccer World Cup XII. Lecture Notes in Computer Science*, vol 5399 LNAI (Berlin; Heidelberg: Springer), 532–543.
- McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., and Wilkins, D. (1998). *PDDL-The Planning Domain Definition Language*. Technical report, Yale Center for Computational Vision and Control.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). "Human-level control through deep reinforcement learning," in *Nature*, vol 518 (London: Nature Publishing Group), 529–533.
- Moore, A. W., and Atkeson, C. G. (1993). Prioritized sweeping: reinforcement learning with less data and less time. *Mach. Learn.* 13, 103–130.
- Nachum, O., Brain, G., Gu, S., Lee, H., and Levine, S. (2018). "Data-efficient hierarchical reinforcement learning," in *Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, QC), 3303–3313.
- Nachum, O., Gu, S., Lee, H., and Levine, S. (2019). "Near-optimal representation learning for hierarchical reinforcement learning," in *International Conference on Learning Representations (ICLR)* (New Orleans, LA).
- Ng, A. Y., Harada, D., and Russell, S. J. (1999). Policy invariance under reward transformations: theory and application to reward shaping. in *International Conference on Machine Learning (ICML)* (Bled: Morgan Kaufmann), 278–287.
- Nguyen, P. D., Fischer, T., Chang, H. J., Pattacini, U., Metta, G., and Demiris, Y. (2018). "Transferring visuomotor learning from simulation to the real world for robotics manipulation tasks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 6667–6674.
- Peng, J., and Williams, R. J. (1993). Efficient learning and planning within the Dyna framework. in *IEEE International Conference on Neural Networks* (San Francisco, CA), 168–174.
- Pohlen, T., Piot, B., Hester, T., Azar, M. G., Horgan, D., Budden, D., et al. (2018). Observe and look further: achieving consistent performance on atari. *arXiv preprint arXiv:1805.11593*.
- Rockel, S., Neumann, B., Zhang, J., Konecny, S., Mansouri, M., Pecora, F., et al. (2013). An ontology-based multi-level robot architecture for learning from experiences. In *AAAI Spring Symposium 2013* (Palo Alto, CA: AAAI Press), 52–57.
- Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). "Universal value function approximators," in *International Conference on Machine Learning (ICML)* (Lille), 13121320.
- Schulman, J., Levine, S., Jordan, M., and Abbeel, P. (2015). "Trust region policy optimization," in *International Conference on Machine Learning (ICML)* (Lille), 1889–1897.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.
- Srivastava, S., Fang, E., Riano, L., Chitnis, R., Russell, S., and Abbeel, P. (2014). "Combined task and motion planning through an extensible planner-independent interface layer," in *International Conference on Robotics and Automation (ICRA)* (Hong Kong), 639–646.
- Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Machine Learning Proceedings 1990* (Austin, TX: Elsevier), 216–224.
- Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull.* 2, 160–163.
- Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Art. Intel.* 112, 181–211.
- Taylor, A. H., Hunt, G. R., Medina, F. S., and Gray, R. D. (2009). "Do new caledonian crows solve physical problems through causal reasoning?" in *Proceedings of the Royal Society B: Biological Sciences* (London), 247–254.
- Todorov, E., Erez, T., and Tassa, Y. (2012). "MuJoCo: a physics engine for model-based control," in *IEEE International Conference on Intelligent Robots and Systems* (Vilamoura), 5026–5033.
- Toussaint, M., Allen, K. R., Smith, K. A., and Tenenbaum, J. B. (2018). "Differentiable physics and stable modes for tool-use and manipulation planning," in *Robotics: Science and Systems (RSS)* (Pittsburgh, PA).
- Traoré, R., Caselles-Dupré, H., Lesort, T., Sun, T., Rodríguez, N. D., and Filliat, D. (2019). Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer. *arXiv preprint arXiv:1906.04452*.
- Ugur, E., and Piater, J. (2015). "Bottom-up learning of object categories, action effects and logical rules: from continuous manipulative exploration to symbolic planning," in *International Conference on Robots and Automation (ICRA)* (Seattle, WA), 2627–2633.
- Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). "FeUdal networks for hierarchical reinforcement learning," in *International Conference on Machine Learning (ICML)* (Sydney, NSW), 3540–3549.
- Wang, Z., Garrett, C. R., Kaelbling, L. P., and Lozano-Perez, T. (2018). "Active model learning and diverse action sampling for task and motion planning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 4107–4114.
- Yamamoto, K., Onishi, T., and Tsuruoka, Y. (2018). "Hierarchical reinforcement learning with abductive planning," in *ICML / IJCAI / AAMAS Workshop on Planning and Learning* (Stockholm).
- Zhao, G., Tatsumi, S., and Sun, R. (1999). "RTP-Q: a reinforcement learning system with time constraints exploration planning for accelerating the learning rate," in *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* (Tokyo), 2266–2273.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Eppe, Nguyen and Wermter. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.