



Multi-agent Path Finding for Timed Tasks Using Evolutionary Games

Sheryl Paul^(✉), Anand Balakrishnan, Xin Qin, and Jyotirmoy V. Deshmukh

University of Southern California, Los Angeles, CA 90007, USA
{sherylpa,anandbal,xinqin,jdeshmuk}@usc.edu

Abstract. Autonomous multi-agent systems such as hospital robots and package delivery drones often operate in highly uncertain environments and are expected to achieve complex temporal task objectives while ensuring safety. While learning-based methods such as reinforcement learning are popular methods to train single and multi-agent autonomous systems under user-specified and state-based reward functions, applying these methods to satisfy trajectory-level task objectives is a challenging problem. Our first contribution is the use of weighted automata to specify trajectory-level objectives, such that, maximal paths induced in the weighted automaton correspond to desired trajectory-level behaviors. We show how weighted automata-based specifications go beyond time-liness properties focused on deadlines to performance properties such as expeditiousness. Our second contribution is the use of evolutionary game theory (EGT) principles to train homogeneous multi-agent teams targeting homogeneous task objectives. We show how shared experiences of agents and EGT-based policy updates allow us to outperform state-of-the-art reinforcement learning (RL) methods in minimizing path length by nearly 30% in large spaces. We also show that our algorithm is computationally faster than deep RL methods by at least an order of magnitude. Additionally our results indicate that it scales better with an increase in the number of agents as compared to other methods.

1 Introduction

Large-scale deployment of multi-agent autonomous mobile systems is becoming a reality in many sectors such as automated warehouses [9, 57], surveillance and patrolling [1, 19], package delivery [50], and logistics support for cargo in aviation, and railways [6, 40]. On the one hand, these autonomous systems are safety-critical and require careful planning to avoid unsafe events such as collisions with other agents or the environment, transgressing zoning restrictions, etc. On the other hand, the deployed systems are expected to be high-performing; for instance, where performance is measured in the number of tasks performed, task completion times, and load balancing across agents. This is an especially difficult challenge when the agents do not know the location of the goal (making it difficult to use heuristic search methods). The core technical challenge for such systems is the *planning* problem: how do we synthesize plans for the agents to

move in their environment while satisfying safety specifications and task objectives with unknown goal locations, while simultaneously ensuring performance or throughput of the overall system?

Traditional AI approaches such as the *multi-agent pathfinding* (MAPF) problem use various heuristic methods in combinatorial optimization and graph search to plan paths for many agents to reach their respective goals while avoiding collisions [55]. However, these approaches involve a centralized planner with full observability of the system, as well as full knowledge of the goal locations. These techniques usually suffer from poor scalability with increasing numbers of agents and world sizes. For highly uncertain environments, an alternative is to endow agents with autonomy that allows them to navigate an environment with the help of sensors. In this setting, each individual agent has a control policy to determine the optimal action to perform based on its sensor observations. Reinforcement learning (RL) has emerged as a promising approach to learning such control policies in uncertain environments modeled as Markov decision processes. The main idea in RL is to use hand-crafted and state-based rewards, and learn policies that optimize the cumulative long-term rewards. An important assumption is that additive long-term rewards capture desired global behavior. There are many instances of *reward hacking*, where myopic rewards for contrasting objectives lead to undesired long-term behaviors [5].

An approach to tackle this problem of specifying time-sensitive and sequential tasks with complex dependencies is to use logic-based specifications like in Linear Temporal Logic (LTL) [45] or its extensions, including Signal Temporal Logic (STL) [36] and Mission-time Linear Temporal Logic (MLTL) [48]. Moreover, such specification languages can be used to define complex tasks for multi-agent systems [32]. In recent years, *reward shaping* using temporal logic-based objectives [17, 24, 35, 46] has emerged as a promising approach to address the problem of defining state-based rewards to satisfy complex spatiotemporal tasks. A prevailing theme in many reward shaping papers is to utilize either the Boolean satisfaction semantics for LTL [6] or quantitative satisfaction semantics (i.e., robustness values for STL) [7, 16] as reward signals.

However, while these logical specifications are well-suited for specifying time-liness properties such as deadlines, they are not as effective at specifying performance properties such as *expeditiousness*. For example, consider the environment shown in Fig. 1 and the trajectories shown in green and red, respectively. Let a_i

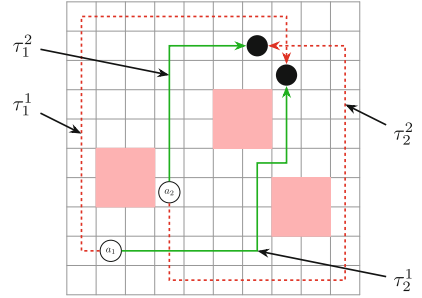


Fig. 1. Shortest paths in the grid environment: agents a_1 and a_2 must get to designated final states (black circles) while avoiding red obstacle regions. Red dashed trajectories τ_1^1 of agent a_1 and τ_2^1 of a_2 satisfy the task but are not the shortest paths, green solid trajectories τ_1^2 and τ_2^2 are the shortest paths. Arrows point towards goals.

denote the position of the i^{th} agent, then the formula $F_{[0,25]}(a_i \in \text{Goal})$ specifies that the agent should reach the goal within 25 time-steps.

While all the trajectories satisfy the formula, it is clear that the green trajectories are more expeditious as they reach the goal faster. Note that we do not want to replace this specification with one that indicates a shorter deadline. Rather, we would like to prefer the ones that reach the goal faster. Such requirements are crucial in multi-agent systems where there may be hard deadlines on agents reaching their goals, with agents reaching goals at varying times, and we want to prefer behaviors where average expeditiousness across agents is preferred.

To address this problem, we use *weighted automata* specifications as they allow us to capture performance metrics like expeditiousness [18, 39]. Weighted automata are a class of automata with transition *weights* interpreted over algebraic structures and generalize the various qualitative and quantitative semantics of discrete-time temporal logics [28, 29].

In this paper, we present an innovative application of an *evolutionary algorithm* rooted in evolutionary game theory (EGT) [53] to address the Multi-Agent Pathfinding (MAPF) problem. This algorithm is adept at deriving optimal policies for multiple agents in complex environments, advancing beyond existing optimization-focused techniques that rely on preset goal knowledge and environmental models [25, 42, 43, 59]. We examine a scenario involving numerous homogeneous agents without predefined goal locations, where a single, unified policy learned from collective experiences supersedes individual training. This shifts the problem to a single-agent learning paradigm. Our method is benchmarked against single-agent algorithms like Q-learning [56], PPO [52], and heuristic searches such as A^* [23]. Here, A^* benefits from its heuristic approach to goal navigation. Our results show that our evolutionary strategy not only achieves optimal policy faster but also competes with, and can surpass the aforementioned methods in efficiently reaching targets.

Contributions: To summarize, the main contributions of this paper are as follows: (1) we introduce weighted automata (WA)-based task objectives and show that these encourage agents to finish their tasks sooner; (2) we propose an algorithm ‘MAPF-EGT’ (Multi Agent Pathfinding using Evolutionary Game Theory) that uses evolutionary techniques to learn stochastic control policies for homogeneous agents trying to fulfill homogeneous task objectives (but with *a priori* unknown goals); and (3) we show empirical results that our method outperforms state-of-the-art RL and heuristic search methods in various grid-world environments.

2 Preliminaries

Policy Synthesis for Stochastic Games: To formalize the problem of synthesizing policies for agents such that they satisfy their assigned task specifications, we use the framework of stochastic games.

Definition 1 (Stochastic Game). A stochastic game over a finite set of agents \mathcal{N} , where each agent $i \in \mathcal{N}$ can be modeled as a tuple $(\mathcal{S}^i, \mathcal{A}^i, \Delta^i, \gamma, \mathcal{R}^i)$, where \mathcal{S}^i denotes the set of possible states of agent i , \mathcal{A}^i is the set of actions available to agent i , and $\Delta^i : \mathcal{S}^i \times \mathcal{A}^i \times \mathcal{S}^i \rightarrow [0, 1]$ is a joint probability distribution over states, actions and next states that defines the transition dynamics of the game. At each time-step t , agent i is in some state s_t^i and executes action a_t^i , transitioning to the next state $s_{t+1}^i \sim \Delta^i(s' \mid s = s_t^i, a = a_t^i)$, and receives reward $r_t^i = \mathcal{R}^i(s_t^i, a_t^i)$. The discount factor $\gamma \in [0, 1]$ prioritizes early rewards by discounting a reward r_t^i at time t , by a factor of γ^t . A joint action at time t is a tuple $\mathbf{a}_t = (a_t^1, a_t^2, \dots, a_t^n)$. The system transitions from a joint state $\mathbf{s}_t = (s_t^1, s_t^2, \dots, s_t^n)$ based on the joint action \mathbf{a}_t to a joint state sampled from the joint transition probability distribution $\mathbf{s}_t' \sim \Delta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$.

A policy π^i for agent i is defined as a joint probability distribution over the set of states and action, i.e., $\pi^i : \mathcal{S}^i \times \mathcal{A}^i \rightarrow [0, 1]$. A joint policy π can be viewed as the tuple of all individual agent policies: $\pi = (\pi^1, \pi^2, \dots, \pi^n)$. A trajectory τ^i of agent i induced by policy π^i is defined as a $(T + 1)$ -length sequence of state-action pairs: $\tau^i = \{(s_0^i, a_0^i), (s_1^i, a_1^i), \dots, (s_T^i, a_T^i)\}$, where, $\forall t < T : a_t^i \sim \pi^i(a^i \mid s = s_t^i), (s_{t+1}^i, r_{t+1}^i) \sim \mathcal{T}$. A trajectory τ for all agents in the game is a tuple of individual agent trajectories: $\tau = (\tau^1, \tau^2, \dots, \tau^n)$, where each τ^i is as defined previously for agent i . Let η^i represent the expected sum of rewards (return) for agent i under policy π^i , and η the total return across all agents under the joint policy π :

$$\eta^i = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_t^i \mid s_0^i, \pi^i \right], \quad \eta = \sum_{i \in \mathcal{N}} \eta^i$$

Here, s_0^i is the initial state of agent i . An optimal policy π^* is defined as the policy that maximizes the total expected reward η across all agents in the system. The optimal policy Π^* is composed of individual policies $\Pi^* = (\pi^{1*}, \pi^{2*}, \dots, \pi^{n*})$ for each agent, and $\eta(\Pi^*)$ is the optimal joint policy consisting of individual joint policies. We note that in our setting, we use homogeneous agents with identical tasks, so instead of using per-agent policies, we have n instances of the same policy π^* .

In our setting, we assume that at time $t = 0$, agents can start at some location defined by the set of initial states \mathcal{I} , must avoid obstacles defined by the set of locations in \mathcal{O} at all times, and must reach a goal in the set of goal states: \mathcal{F} within time T while avoiding obstacle states at all times. Formally, $s_0^i \in \mathcal{I}, s_t^i \notin \mathcal{O}$ and $s_T^i \in \mathcal{F}, \forall i \in \mathcal{N}$ and $0 \leq t \leq T$.

Weighted Automata (WA) Based Rewards: We now introduce *weighted automata* (WA), and show how they can be used to define rewards. In general, weighted automata are finite-state machines that have a notion of *accepting* conditions and, for each transition in the automaton, the machine outputs a *weight* for the transition [18]. When interpreted over algebraic *semirings*, the weights in the automaton can be used to define various different types of quantitative objectives along with the acceptance condition.

We eschew the general definition of weighted automata, and look at the specific instances of *quantitative languages* [11, 14, 15], where the weights of the transitions in the automaton are interpreted over *valuation functions*. This formalism is a generalization of the recently popular *reward machine* approach used in reinforcement learning frameworks with temporally dependent tasks [13, 26, 27, 60]. Weighted automata are more expressive than reward machines as they allow a notion of acceptance in the automaton. We will later empirically show how specific weighted automata can generalize discrete-time temporal logic objectives (such as Signal Temporal Logic).

Definition 2 (Weighted Automata). *A weighted automaton is defined as a tuple $\mathcal{A} = (Q, Q_I, Q_F, \Sigma, T, \lambda)$, where:*

- Q is a finite set of locations, and $Q_I \subseteq Q$ and $Q_F \subseteq Q$ are respectively the set of initial locations and final locations;
- Σ is an input alphabet;
- $T : Q \times \Sigma \rightarrow 2^Q$ is a (partial) labeled transition function, where 2^Q is the powerset of Q ;
- $\lambda : Q \times \Sigma \times Q \rightarrow \mathbb{R}$ is a weight function.

An automaton is *complete* if for all $q \in Q$ and $s \in \Sigma$, there is at least one successor location in the automaton, i.e., $|T(q, s)| > 0$. Likewise, the automaton is *deterministic* if for all $q \in Q$ and $s \in \Sigma$, there is exactly one successor location, i.e., $|T(q, s)| = 1$. As we intend to use WA-based reward functions, the set of input symbols for the WA is essentially the set of states \mathcal{S} of a stochastic game as defined before. Thus, given an input trajectory, $\tau = (s_0, s_1, \dots, s_l) \in \Sigma^*$, a *run* in the automaton \mathcal{A} is a sequence of locations $(q_0, q_1, \dots, q_{l+1})$ such that $q_{i+1} \in T(q_i, s_i)$, for $i \in 0, \dots, l$. We use $\text{run}_{\mathcal{A}}(\tau)$ to denote the set of runs induced in \mathcal{A} by $\tau \in \Sigma^*$. Note that if $|\text{run}_{\mathcal{A}}(\tau)| > 0$, the automaton is non-deterministic, and the run is *accepting* if the last location q_{l+1} in the run is in Q_F . Moreover, for a given input $\tau = (s_0, s_1, \dots, s_T) \in \Sigma^*$ and a corresponding run $\sigma = (q_0, q_1, \dots, q_l, q_{T+1})$, a sequence of weights (w_0, w_1, \dots, w_T) is produced such that $w_i = \lambda(q_i, s_i, q_{i+1})$. For some $\tau \in \Sigma^*$ and a corresponding run $r \in Q^*$ in the automaton, we let $\text{weights}_{\mathcal{A}}(\tau, r) \in \mathbb{R}^*$ denote the sequence of weights induced by the run.

Definition 3 (Valuation function). *Given the weights $w \in \text{weights}_{\mathcal{A}}(\tau, r)$ corresponding to a run $r \in Q^*$ induced by an input $\tau \in \Sigma^*$ on a weighted automaton \mathcal{A} , a valuation function $\text{Val}(w) \in \mathbb{R}$ is a scalar function that outputs the weight of the sequence. The weight of a trajectory $\tau \in \Sigma^*$ in an automaton \mathcal{A} can be defined as*

$$w_{\mathcal{A}}(\tau) = \max_{r \in \text{run}(\tau)} \text{Val}(\text{weights}(\tau, r)).$$

For a sequence of weights $w = (w_0, w_1, \dots, w_l)$, examples of valuation functions include:

$$\begin{aligned}
- \quad \text{Sum}(w) &= \sum_{t=0}^l w_t & - \quad \text{Avg}(w) &= \frac{1}{l+1} \sum_{t=0}^l w_t \\
- \quad \text{For a discount factor } \gamma \in [0, 1], \text{DiscountedSum}_\gamma(w) &= \sum_{t=0}^l \gamma^t w_t
\end{aligned}$$

In this paper, we are particularly interested in the $\text{DiscountedSum}_\gamma$ valuation function defined on deterministic weighted automata, as it captures the total discounted rewards semantics for the policy synthesis problem described above.

Remark 1. While we restrict ourselves to deterministic automata for brevity, one can extend the presented framework to non-deterministic automata by resolving the non-deterministic transitions using the max operation (or the general semiring addition) [11, 39]. Moreover, since our framework evaluates the weight of trajectories at the end of *episodes*, one can resolve the non-deterministic runs using the weighting function w_A as described above.

Problem Definition: A specific setting for multi-agent pathfinding that we wish to solve in this paper is *reach-avoid* specifications, i.e., where multiple agents are each randomly assigned start and goal locations in a shared environment. Their task is to navigate to their goals as quickly as possible while avoiding collisions with one another. We can express these conditions as:

1. **Expeditious goal achievement:** Each agent must reach the goal state within T time-steps, and as expeditiously as possible.
2. **Collision avoidance:** Every agent must avoid collisions with other agents and obstacles at all times.

Solution Outline: Our goal, is to define a reward function for each agent in the system such that by maximizing the expected total reward across all agents (denoted by η above), the resulting policy Π^* achieves the high-level reach-avoid task in a performant fashion. We do this by defining a weighted automaton that specifies the task for each agent in the system, i.e., each agent aims to maximize the weight of its trajectory in the system with respect to the same automaton specification. The weighted automaton describes a pattern such that maximizing the composed weights of all agent trajectories, or the *utility* of the system.

3 Reward Shaping for Timed Multi-agent Reach-Avoid Problems

To ensure a timely multi-agent reach-avoid system, we propose a reward structure that emphasizes the importance of efficient trajectories to the goal. Previously in literature, AvSTL [2] was developed to assess behavior over time by measuring the area under the signal curve within a given timeframe. Yet, for reach-avoid tasks where goals must be reached quickly and safely, AvSTL may inadvertently favor less direct routes due to its integral-based evaluation. Consider the robustness metric for reaching a goal within time T , denoted by

$F_{[0,T]}(reachgoal)$, as 1 for success and 0 otherwise. As shown in Fig. 2, an agent that swiftly reaches the goal and moves away may have a lower cumulative value than one following a longer route that lingers at the goal, highlighting a potential discrepancy in AvSTL's approach for such tasks.

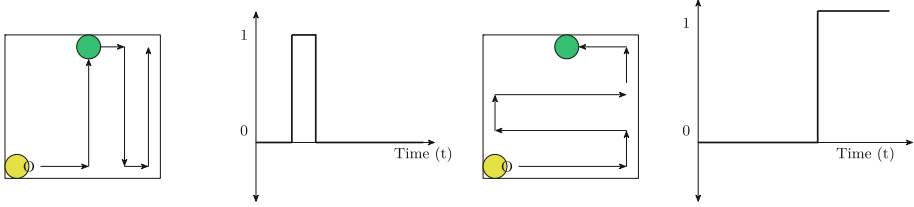


Fig. 2. Counter-example: An agent (O) tries to go from the initial location (yellow) to the goal location (green). Under general expeditious semantics (such as in AvSTL) the trajectory in the left figure where the agent reaches the goal quickly but wanders after would receive a lower reward (area under the curve) than the one in the right figure where the agent takes a longer path to reach the goal but stays in the goal for longer. (Color figure online)

In this section, we outline our reward-shaping mechanism which is designed to promote efficient goal attainment with collision avoidance. The reward function provides a large reward ($+b$) for reaching the goal as sufficient incentive for doing so, while penalizing each step *before* reaching the goal with a minor negative reward ($-a$) to hasten the process. It also generates a hefty penalty for collisions ($-c$) at any point which could be smaller than or equal to the reward for reaching the goal, depending on whether or not the system wants to permit trajectories with collisions or not. This structured reward shaping aims to balance the urgency of completing tasks with the imperative of maintaining safe

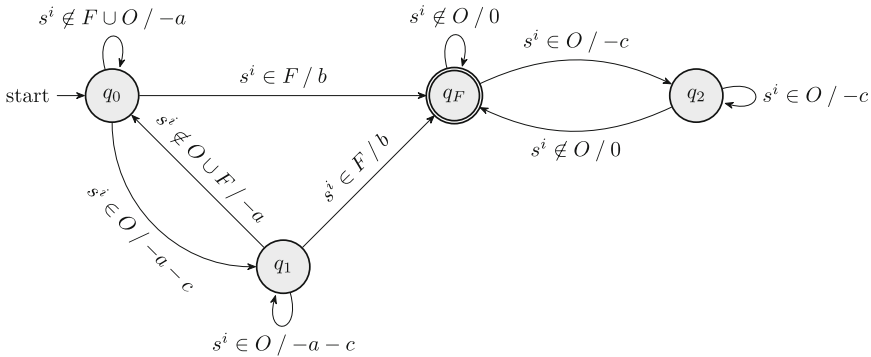


Fig. 3. A deterministic weighted automaton defining the reach-avoid task that needs to be completed by each agent in the system. In the figure, s^i refers to the current state of agent i in the multi-agent system.

operations within the multi-agent system. We abuse notation and denote $\tau^i(t)$ to be the state of the agent i at time t in the trajectory. The reward function R for an agent i at time t taking action a resulting in a state transition from s_t^i to s_{t+1}^i can be defined as:

$$R_t^i(s_t^i, a, s_{t+1}^i) = f_F(\tau^i(t)) + f_O(\tau^i(t)) \quad (1)$$

Now we can define:

$$f_O(\tau^i(t)) = \begin{cases} -c & \text{if } s_t^i \in O \\ 0 & \text{if } s_t^i \notin O \end{cases}, f_F(\tau^i(t)) = \begin{cases} 0 & \text{if } \exists t' < t : s_{t'}^i \in F \\ b & \text{if } s_t^i \in F \text{ \& } \nexists t' < t : s_{t'}^i \in F \\ -a & \text{if } s_t^i \notin F \text{ \& } \nexists t' < t : s_{t'}^i \in F \end{cases} \quad (2)$$

where $a, b, c \in \mathbb{R}^+$; $s_t^i \in F$, indicates agent i reaching the goal and $s_t^i \in O$ indicates a collision with an obstacle. We also define the constraint: $b \geq c > a \cdot T$, where T is the total time permitted to reach the goal (i.e., the length of an episode in terms of timesteps). This provides a hierarchical framework where reaching the goal is paramount, followed by the penalty for collisions, with a smaller penalty for taking steps before reaching the goal.

The above reward structure generally follows the weights from the automaton shown in Fig. 3. Moreover, for a discount factor $\gamma \in (0, 1)$ we can interpret the trace over the DiscountedSum $_\gamma$ valuation function.

Let us denote $toa(\tau)$ as the time of arrival of the trajectory into the goal, where $toa(\tau) = \min\{t \mid (s_t, a_t) \in \tau \text{ and } s_t \in F\}$.

Proposition 1. *If the sum of rewards over a trajectory is positive, the trajectory satisfies the condition of reaching the goal within T timesteps, and not colliding with an obstacle. $w_A(\tau) > 0 \implies F_{[0,T]}(s_{toa} \in F) \wedge G_{[0,T]}(\neg s \in O)$ ¹*

All timesteps after $toa(\tau)$ are weighted 0 according to our reward function, as we assume the absence of collisions, so we only consider the part upto $toa(\tau)$.

Lemma 1. *Trajectories that are more expeditious i.e. that reach the goal sooner (assuming they do not have collisions) have higher rewards returned by the weighted automata.*

$$toa(\tau) < toa(\tau') \implies w_A(\tau) > w_A(\tau')$$

Proof. The proof follows from the definitions of the reward function and $w_A(\tau)$.

4 Policy Optimization

In our multi-agent system, we employ homogeneous agents that are centrally trained under a unified framework, allowing for the sharing of a joint policy. This homogeneity simplifies the training process and enables the application of

¹ $F_{[a,b]}(x > 0)$ denotes $\exists t \in [a, b]$ s.t. $x(t) > 0$ and $G_{[a,b]}(x > 0)$ denotes $\forall t \in [a, b]$ s.t. $x(t) > 0$.

single-agent methodologies, such as A^* and single-agent reinforcement learning algorithms to manage the collective behavior.

Search-Based Methods for Policy Learning: Heuristic-based approaches to pathfinding, exemplified by algorithms such as A^* [23], D-star [54], and D-star Lite [33], utilize heuristics to navigate efficiently from one point to another within an environment. These algorithms operate by employing a heuristic to estimate the cost from any node in the search space to the goal, effectively guiding the search process toward the most promising paths while minimizing unnecessary exploration. Common heuristics include the Manhattan distance, which provides a direct estimation of the minimal possible distance to the goal, assuming a grid-like path with no obstructions. This heuristic foreknowledge is crucial as it significantly influences the efficiency and effectiveness of the search.

Learning-Based Methods: Data-driven policy optimization methods are increasingly popular in environments where the system is stochastic, the model of the system is not available, and when the state-space is large. Model-free reinforcement learning methods such as Q-learning [58], Deep Q-Networks (DQN) [38], Advantage Actor Critic (A2C) [37], and Proximal Policy Optimization (PPO) [52] refine their strategies through extensive interaction with the environment.

Q-learning is a foundational off-policy algorithm where each pair of state s and action a is associated with a $Q(s, a)$ value representing the expected future reward. This value is iteratively updated using the Bellman equation [58]. In multi-agent systems, Q-learning can be adapted by extending the state-action space to include all possible combinations of states and actions for every agent, effectively using a shared Q-table. This approach allows all agents to follow a centrally managed, uniform policy that operates based on the joint action space. Deep Q-Networks (DQN) extends Q-learning by employing deep neural networks to approximate the $Q(s, a)$ function, allowing it to handle continuous and high-dimensional state spaces efficiently.

A2C is an actor-critic method that uses multiple parallel environments to update its policy and value networks reducing variance and improving learning speed. PPO is a more stable and efficient actor-critic method that optimizes a clipped version of the objective function instead of directly optimizing the objective function to provide smoother updates. Actor-critic methods can be implemented with each agent operating simultaneously in a shared environment, all contributing to a central policy update mechanism. This ensures that learning is synchronized.

Monte Carlo methods, another branch of model-free RL, do not assume knowledge of the environment's dynamics and instead rely on sampling full trajectories to estimate expected returns. These methods can be adapted for multi-agent use by sampling and averaging returns across all agents, using these aggregated insights to update a central policy that guides all agents.

We explored a spectrum of strategies for addressing the multi-agent pathfinding challenge. These included the traditional heuristic-driven approach such as A^* , tabular temporal difference methods like Q-learning, neural network-driven

techniques exemplified by Proximal Policy Optimization (PPO), and stochastic sampling-based approaches such as those used in Monte Carlo methods.

Limitations of These Approaches: Limitations of search-based algorithms include scalability as the frontier grows exponentially with an increase in the number of agents, lack of applicability in stochastic and dynamic environments, and a priori knowledge of the environment in terms of an admissible heuristic which we *do not provide* to any of the other algorithms including our own. The primary shortcoming of learning-based algorithms such as the ones we described above that we aim to overcome is sample-inefficiency. These algorithms can require a large number of interactions with the environment to learn effective policies, which is impractical in complex or time-sensitive applications. These algorithms also treat each step or trajectory as equally important for a policy update. However, we aim to replace that with a weighted update that ensures that steps in trajectories that overperform or underperform by a significant magnitude, also make a more significant impact to the policy.

4.1 Algorithmic Framework Using Evolutionary Game Theory

Evolutionary Game Theory (EGT) presents a dynamic alternative to classical game theory and is better suited to the realities of multi-agent systems. [4, 51, 53] Unlike traditional game theory, it does not presume rationality among agents, making it adaptable to any scenario. It focuses on the dynamics of strategy changes driven by the success of current strategies in the population, reflecting a more naturalistic approach to agent learning. Evolutionary strategies also offer quicker and more robust convergence guarantees, making them particularly appealing in environments where reinforcement learning is an intuitive fit. Given these advantages, it is a promising direction for advancing research and practical applications in multi-agent settings.

Replicator Equation: The key concept within EGT pertinent to us is that of *replicator dynamics*, which describes how the frequency of strategies (or policies in RL) changes over time based on their relative performance. The classic replicator equation in evolutionary game theory describes how the proportion of a population adopting a certain strategy evolves over time. Mathematically, it is expressed as [41]:

$$x_j(i+1) = x_j(i) \cdot \frac{f_j(i)}{\bar{f}(i)} \quad (3)$$

where $x_j(i)$ represents the proportion of the population using strategy j at time i , $f_j(i)$ is the fitness of strategy j , and $\bar{f}(i)$ is the average fitness of all strategies at time i . The equation indicates that the growth rate of a strategy's proportion is proportional to how much its fitness exceeds the average fitness, leading to an increase in the frequency of strategies that perform better than average.

Representation of Populations and the Fitness Equivalent: We represent the probability distribution over actions in a given state as a population. Each individual in the population corresponds to an action, and the proportion of (individuals corresponding to a specific action) in the population represents the probability of taking the action in that state. Thus, the population represents a stochastic policy. The fitness function measures the reproductive success of strategies based on payoffs from interactions, similar to utility in classical game theory, or how in RL, the expected return measures the long-term benefits of actions based on received rewards. Both serve as optimization criteria: strategies or policies are chosen to maximize these cumulative success measures to guide them towards optimal behavior. Therefore, in our model, the fitness for a state $f(s)$ corresponds to the expected return from that state, equivalent to the value function $v(s)$, and the fitness for a state-action pair $f(s, a)$ corresponds to the expected return from taking action a in state s , equivalent to the action-value function $q(s, a)$.

Under the assumption of sparse rewards- where significant rewards are received only upon reaching specific states or goals, $f(s)$ is defined as $\mathbb{E}[f(\tau_s)]$, the expected return across all trajectories through state s . Likewise, $f(s, a)$ is defined as $\mathbb{E}[f(\tau_{(s,a)})]$, the expected return across trajectories involving the state-action pair (s, a) .

Policy Update Mechanism: The replicator equation can be adapted to update the probability of selecting certain actions based on their relative performance compared to the average. The adaptation of the replicator equation is as follows:

$$\pi^{i+1}(s, a) = \frac{\pi^i(s, a)f(s, a)}{\sum_{a' \in A} \pi^i(s, a')f(s, a')} \quad (4)$$

where $\pi^i(s, a)$ is the probability of action a in state s , in the i^{th} iteration, and π^{i+1} represents the policy in the $i + 1^{th}$ iteration.

Heterogeneous Agents: We note that in our approach agents are assumed to be homogeneous, i.e., the learned policy is shared across all the agents. Our approach can theoretically work for heterogeneous agents as well. However, each agent is then required to maintain its own policy. This means that for each agent, per agent state, a population of actions will require to be updated.

Prioritizing Expeditious and Safe Trajectories: The adaptation of our policy update equation focuses on the ratio of the expected fitness of a particular action to the average expected fitness, allowing for a more nuanced update. It scales the probability of each action relative to how much better (or worse) it performs compared to the average, rather than simply whether it is better or worse. Actions that lead to higher returns relative to the average are thus promoted in subsequent iterations of the policy. This selective pressure inherently favors actions that contribute to reaching goals faster and avoiding collisions, as they have a greater impact on the expected return due to the structured rewards. As a result, the policy evolves towards a strategy that seeks to maximize rewards

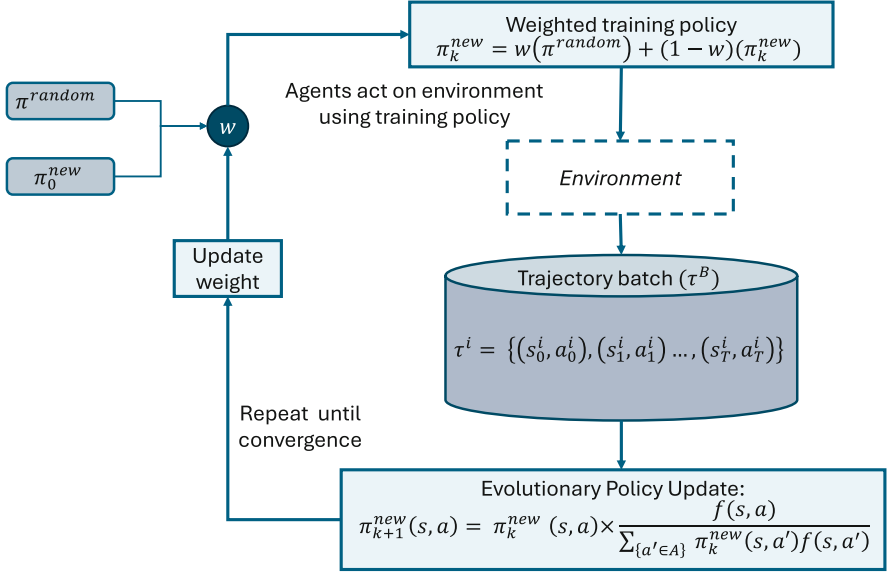


Fig. 4. Overview of the evolutionary based learning approach used in our algorithms.

by combining efficiency with safety. The learning rate α ensures that the update remains bounded and allows for fine-tuning of the learning process (Fig. 4).

Proposed Algorithm: Our algorithm uses batch-based updates: We initialize the policy π^{new} to be initially random. We sample trajectories as part of a batch, and the state-action pairs in these trajectories are updated according to the update rule. The return for the $(k+1)^{th}$ iteration is set as the return for the current batch of trajectories. We maintain a weighted discounted policy while training, with the weight decreasing with each iteration. This process is repeated until our termination condition has been met i.e. $\eta_{k+1} - \eta_k > \delta$: This condition checks if our policy is improving with each update. If it does not, we say it has converged.

Termination and Convergence: We can show that the algorithm terminates after a finite number of iterations, based on the observation that policy updates ensure that the utility/value of each state monotonically increases. The evolutionary update guarantees that the probability $\pi^{new}(s, a)$ increases only when $f(s, a)$ is greater than $\mathbb{E}_a(f(s, a))$, while $\pi^{new}(s, a)$ decreases whenever $f(s, a) < \mathbb{E}_a(f(s, a))$. This in turn guarantees that under π^{new} , the value of the state s increases. For states not sampled in batch, $\pi^{new}(s, a)$ is unchanged. Now there are two cases: (1) The maximum improvement across all states between two consecutive iterations is less than δ , in which case the algorithm goes to Line 15 and returns the policy π^{new} . (2) The maximum improvement is greater than δ . In the second case, the algorithm cannot forever increase the expected

Algorithm 1: Multi-agent Pathfinding using EGT**Input:** Hyper-parameters

- To alter weight: $\nu \in (0, 1)$
- To ensure discounted policies: $\epsilon > 0$
- To check convergence: $\delta > 0$
- Scaling factor/learning rate: $\alpha > 0$

Output: Trained policy

```

1  $k \leftarrow 1$  //  $k$  is the iteration number
2  $\eta_0 \leftarrow 0$  //  $\eta_k$  contains the expected at the end of the  $k^{th}$  iteration
3  $\pi_0^{new}(s, a) \leftarrow \frac{1}{|A|}, \forall s \in S, a \in A$  // Initialize policies as random
4 while True do
5   for each episode in batch  $b = 1$  to  $B$  do
6     Generate trajectory  $\tau_b^i$  for each agent using  $\pi_k^{new}$ 
7     Append the collective generated trajectory  $\tau_b$  to batch
8   Compute expected return  $\eta_k$  over the batch
9   if  $\eta_k - \eta_{k-1} \geq \delta$  then
10     for all trajectories  $\tau_b$  in batch do
11       for each state-action pair  $(s, a)$  in trajectory  $\tau_b$  do
12         Evolutionary Policy Update:
13          $\pi_{k+1}^{new}(s, a) \leftarrow \pi_k^{new}(s, a) \cdot \alpha \cdot \frac{\pi_k^{new}(s, a) f(s, a)}{\sum_{a' \in A} \pi_k^{new}(s, a') f(s, a')}$ 
14        $w \leftarrow \max(\epsilon, w - \nu)$  // Decrement weight on exploratory policy
15        $\pi_{k+1}^{new} \leftarrow w \cdot \frac{1}{|A|} + (1 - w) \cdot \pi_{k+1}^{new}$  // Weighted policy update  $k \leftarrow k + 1$ 
16   else return  $\pi_k^{new}$  // Return the final optimized policy

```

return η_k , as the maximum expected return is guaranteed to be finitely upper bounded (as trajectories are finite in length).

We note that our algorithm may converge to a policy that is sub-optimal; however, as long as the return η_k is positive, the trained policy satisfies the weighted automaton objectives. To prove that the algorithm terminates to an optimal policy would require the policy update operator to be a contraction mapping, similar to learning algorithms like Q-learning [58] or value iteration [10]; we defer this extension to future work.

To analyze the complexity of our algorithm, we note that each loop (i.e. from Lines 5 to 13) runs until $\eta_k - \eta_{k-1} < \delta$. If we assume the existence of an optimal policy that maximizes returns, and let the return under the optimal policy be denoted as η^* , then this loop can run for a maximum of $\frac{\eta^* - \eta_0}{\delta}$ times. Each loop itself has a complexity of $T \cdot B$ (with T being the length of an episode and B being the number of episodes in a batch). Then, the complexity of our algorithm is $O\left(\frac{\eta^* - \eta_0}{\delta} \cdot T \cdot B\right)$.

5 Experiments and Results

For our experiments, we define the world as a two dimensional $n \times n$ grid, where each state $s \in S$, is defined as (x, y) where x and y are coordinates. The agents are homogeneous and the set of actions available to each agent $A^i = \{\text{up, down, left, right, stay}\}$. At the beginning of each episode the agent is assigned a random start location and must reach one of the goal locations by the end of it. We vary the grid size from 20×20 to 200×200 , and vary the number of agents from 2 to 50. We benchmark our algorithm against a planning algorithm A^* [23], a Monte-Carlo approach [56], tabular Q-learning [56] and a Deep RL algorithm PPO [52]. We use Manhattan-distance as the heuristic for A^* and the same reward function as our algorithm for the other approaches. Our experiments were carried out on a laptop with 2.0 GHz dual-core Intel Core i5 processor and 16GB RAM. We use the ‘stablebaselines3’ [47] implementation of PPO with standard values for the hyperparameters. Our algorithms have been implemented on OpenAI gym [12].

5.1 Results and Discussion

Comparison with A^* : We include A^* as a baseline heuristic search algorithm. A^* uses a heuristic based on Manhattan distance and does not account for proximity to obstacles, while the other algorithms incorporate this information into their reward structures. We note that each agent invokes A^* separately to plan its path, while the other algorithms learn a common policy. The paths identified by A^* are close to optimal as the Manhattan distance is an admissible heuristic, and thus a good baseline for comparison.

Timesteps/Path Length: MAPF-EGT excels in path length (i.e. timesteps to reach the goal) for large grids. PPO and Monte Carlo rival it on smaller grids but lag behind on larger ones.

Distance from Obstacles: Monte Carlo keeps a relatively larger distance from obstacles, showing a conservative approach. Q-Learning and PPO have similar distances, indicating a balanced approach in avoiding obstacles. MAPF-EGT shows a more adaptive strategy, maintaining sizeable distances from obstacles even as grid size increases.

Computation Time: Q learning struggles with scalability and training, affecting its large-grid success rate. PPO takes longer time on smaller grids but scales well with increase in size. MAPF-EGT has a comparable computation time to the Monte-Carlo method (Fig. 5).

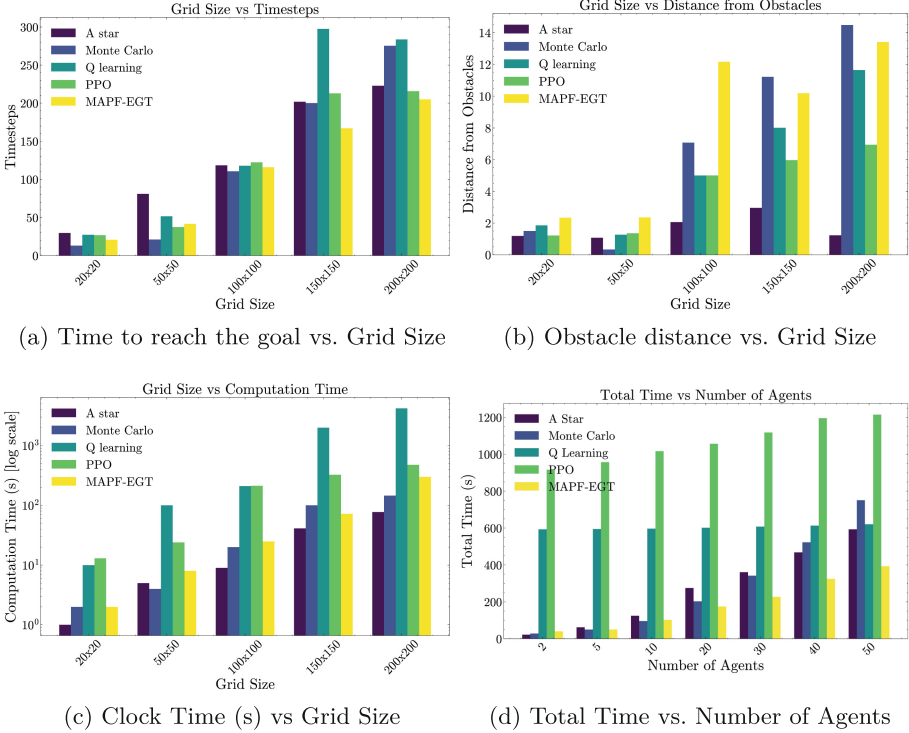


Fig. 5. MAPF-EGT benchmarked against the algorithms: A^* , Monte-Carlo search, PPO, and Q learning. Timesteps to reach the goal (Fig. a), Expected minimum distance from obstacles (greater distance indicates safer paths) (Fig. b), and clock time (seconds) required for computation (Fig. c) compared across the grid sizes: 20×20 , 50×50 , 100×100 , 150×150 and 200×200 . Fig. d shows scaling in total time taken, with number of agents varied from 2 to 50 on a 100×100 grid. We also note that the A^* algorithm is given a heuristic i.e. the Manhattan distance to the closest goal, information that the other algorithms are not given.

Number of Agents: Testing MAPF-EGT with 2 to 100 agents on a 100×100 grid, we see that PPO scales rather poorly, while Q-learning remains steady. Although Q-learning had long training times as indicated previously, its short run times make it scale well with increase in the number of agents. MAPF-EGT scales the best with increase in the number of agents. Our intuition for this is that more agents represent more trajectories generated, and more diverse and richer data to learn from.

6 Related Work

Early work in modeling environment dynamics used Markov decision processes (MDPs) [22, 49] or differential equations [20, 44]. Recently, the focus has shifted

to data-driven and automata-based methods for control synthesis and achieving temporal logic-based objectives [8]. Modern techniques incorporate robustness metrics into deep reinforcement learning (RL) frameworks, replacing traditional reward functions with back-propagation to train controllers for complex temporal tasks [7]. In RL, reward engineering is crucial, with methods ranging from Q-learning for robust controller learning [3], to using deterministic finite automata (DFA) for task specifications and potential functions. Automata are preferred for their robustness in symbolic weighted automata frameworks [29] and the ability to translate various specification formalisms into automata [30]. Addressing Non-Markovian Rewards (NMRs) now involves Non-Markovian Reward Decision Processes (NMRDPs) with history-sensitive reward functions [13]. Model-based MDP synthesis under Linear Temporal Logic (LTL) has been streamlined to mixed integer linear programming, using task progression concepts to develop policies [31]. Advanced techniques for LTL controller synthesis handle unsatisfiable tasks [21], and Temporal Logic Policy Search introduces a robustness-oriented approach to model-free RL [34].

7 Conclusion

In this paper, we have addressed the complex challenge of trajectory-level task objectives for autonomous multi-agent systems in uncertain environments. Our contributions include the introduction of weighted automata-based task objectives, which enhance the agents' ability to complete tasks more expeditiously, and the development of the MAPF-EGT algorithm, which leverages evolutionary game theory to train homogeneous agent teams more effectively. Our empirical results demonstrate that our approach outperforms state-of-the-art reinforcement learning and heuristic search methods, achieving a reduction in path length and faster computation times. These findings highlight the potential of our methods to improve the efficiency and scalability of multi-agent systems in various practical applications.

Limitations and Future Work: Our current setup is confined to discrete state-action spaces, but we are actively developing an extension to accommodate continuous spaces. This expansion will utilize function approximation through radial basis functions, enabling policy updates for states within proximity to the updated state. Furthermore, given our methodology of normalizing the probability distribution across a state's actions, we plan to implement a continuous model, which will be adjusted using probability density functions and updated via Dirac delta functions. Our research currently focuses on homogeneous agents and tasks. Future developments will aim to include diverse forms of multi-agent learning, providing convergence guarantees. Additionally, we intend to explore game-theoretic guarantees within the realm of multi-agent learning to ensure robust and strategic interactions among agents.

References

1. Agmon, N., Urieli, D., Stone, P.: Multiagent patrol generalized to complex environmental conditions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, pp. 1090–1095 (2011)
2. Akazaki, T., Hasuo, I.: Time robustness in MTL and expressivity in hybrid system falsification. In: Kroening, D., Păsăreanu, C.S. (eds.) *CAV 2015*. LNCS, vol. 9207, pp. 356–374. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21668-3_21
3. Aksaray, D., Jones, A., Kong, Z., Schwager, M., Belta, C.: Q-learning for robust satisfaction of signal temporal logic specifications. In: *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 6565–6570 (2016). <https://doi.org/10.1109/cdc.2016.7799279>
4. Alexander, J.M.: Evolutionary game theory. In: Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2021 edn., Metaphysics Research Lab, Stanford University (2021)
5. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety (2016)
6. Badings, T., Romao, L., Abate, A., Jansen, N.: A stability-based abstraction framework for reach-avoid control of stochastic dynamical systems with unknown noise distributions (2024)
7. Balakrishnan, A., Deshmukh, J.V.: Structured reward shaping using signal temporal logic specifications. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3481–3486 (2019). <https://doi.org/10.1109/IROS40897.2019.8968254>
8. Balakrishnan, A., Jakšić, S., Aguilar, E.A., Ničković, D., Deshmukh, J.V.: Model-free reinforcement learning for spatiotemporal tasks using symbolic automata. In: *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 6834–6840 (2023). <https://doi.org/10.1109/CDC49753.2023.10383559>
9. Bellusci, M., Basilico, N., Amigoni, F.: Multi-agent path finding in configurable environments. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 159–167 (2020)
10. Bertsekas, D.P.: *Dynamic Programming and Optimal Control*, vol. I, 3rd edn. Athena Scientific, Belmont (2005)
11. Boker, U.: Quantitative vs. weighted automata. In: Bell, P.C., Totzke, P., Potapov, I. (eds.) *RP 2021*. LNCS, vol. 13035, pp. 3–18. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89716-1_1
12. Brockman, G., et al.: OpenAI gym. arXiv preprint [arXiv:1606.01540](https://arxiv.org/abs/1606.01540) (2016)
13. Camacho, A., Chen, O., Sanner, S., McIlraith, S.A.: Non-Markovian rewards expressed in LTL: guiding search via reward shaping (extended version). In: *Goal-sRL, a Workshop Collocated with ICML/IJCAI/AAMAS* (2018)
14. Chatterjee, K., Doyen, L., Henzinger, T.A.: Quantitative languages. In: Kaminski, M., Martini, S. (eds.) *CSL 2008*. LNCS, vol. 5213, pp. 385–400. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87531-4_28
15. Chatterjee, K., Henzinger, T.A., Otop, J.: Quantitative monitor automata. In: Rival, X. (ed.) *SAS 2016*. LNCS, vol. 9837, pp. 23–38. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53413-7_2
16. Cohen, M., Belta, C.: Temporal logic guided safe model-based reinforcement learning. In: Cohen, M., Belta, C. (eds.) *Adaptive and Learning-Based Control of Safety-Critical Systems*. SLCS, pp. 165–192. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-29310-8_9

17. Cohen, M.H., Belta, C.: Model-based reinforcement learning for approximate optimal control with temporal logic specifications. In: Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control, pp. 1–11 (2021)
18. Droste, M., Gastin, P.: Weighted automata and weighted logics. In: Droste, M., Kuich, W., Vogler, H. (eds.) Handbook of Weighted Automata. Monographs in Theoretical Computer Science. An EATCS Series, pp. 175–211. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01492-5_5
19. Elmaliach, Y., Agmon, N., Kaminka, G.A.: Multi-robot area patrol under frequency constraints. *Ann. Math. Artif. Intell.* **57**, 293–320 (2009)
20. Gilpin, Y., Kurtz, V., Lin, H.: A smooth robustness measure of signal temporal logic for symbolic control. *IEEE Control Syst. Lett.* **5**(1), 241–246 (2020)
21. Guo, M., Zavlanos, M.M.: Probabilistic motion planning under temporal tasks and soft constraints. *IEEE Trans. Autom. Control* **63**(12), 4051–4066 (2018)
22. Haesaert, S., Soudjani, S., Abate, A.: Temporal logic control of general Markov decision processes by approximate policy refinement. *IFAC-PapersOnLine* **51**(16), 73–78 (2018)
23. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* **4**(2), 100–107 (1968). <https://doi.org/10.1109/TSSC.1968.300136>
24. Hasanbeig, M., Kantaros, Y., Abate, A., Kroening, D., Pappas, G.J., Lee, I.: Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 5338–5343. IEEE (2019)
25. Hashemi, N., Hoxha, B., Prokhorov, D., Fainekos, G., Deshmukh, J.: Scaling learning based policy optimization for temporal tasks via dropout (2024)
26. Icarte, R.T., Klassen, T., Valenzano, R., McIlraith, S.: Using reward machines for high-level task specification and decomposition in reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning, pp. 2107–2116. PMLR (2018)
27. Icarte, R.T., Klassen, T.Q., Valenzano, R., McIlraith, S.A.: Reward machines: exploiting reward function structure in reinforcement learning. *J. Artif. Intell. Res.* **73**, 173–208 (2022). <https://doi.org/10.1613/jair.1.12440>
28. Jakšić, S., Bartocci, E., Grosu, R., Nguyen, T., Ničković, D.: Quantitative monitoring of STL with edit distance. *Formal Methods Syst. Design* **53**(1), 83–112 (2018). <https://doi.org/10.1007/s10703-018-0319-x>
29. Jaksic, S., Bartocci, E., Grosu, R., Nickovic, D.: An algebraic framework for runtime verification (2018)
30. Jothimurugan, K., Bansal, S., Bastani, O., Alur, R.: Compositional reinforcement learning from logical specifications. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
31. Kalagarla, K.C., Jain, R., Nuzzo, P.: Synthesis of discounted-reward optimal policies for Markov decision processes under linear temporal logic specifications. arXiv preprint [arXiv:2011.00632](https://arxiv.org/abs/2011.00632) (2020)
32. Kempa, B., Cramer, N.B., Frank, J.D.: Swarm mentality: toward automatic swarm state awareness with runtime verification. In: AAAI 2022 Spring Symposium Series (2022)
33. Koenig, S., Likhachev, M.: Fast replanning for navigation in unknown terrain. *IEEE Trans. Rob.* **21**(3), 354–363 (2005)
34. Li, X., Ma, Y., Belta, C.: A policy search method for temporal logic specified reinforcement learning tasks. In: 2018 Annual American Control Conference (ACC), pp. 240–245. IEEE (2018)

35. Li, X., Vasile, C.I., Belta, C.: Reinforcement learning with temporal logic rewards. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3834–3839. IEEE (2017)
36. Maler, O., Nickovic, D., Pnueli, A.: Real time temporal logic: past, present, future. In: Pettersson, P., Yi, W. (eds.) FORMATS 2005. LNCS, vol. 3829, pp. 2–16. Springer, Heidelberg (2005). https://doi.org/10.1007/11603009_2
37. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning (2016)
38. Mnih, V., et al.: Playing Atari with deep reinforcement learning (2013)
39. Mohri, M.: Weighted automata algorithms. In: Droste, M., Kuich, W., Vogler, H. (eds.) Handbook of Weighted Automata. Monographs in Theoretical Computer Science. An EATCS Series, pp. 213–254. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01492-5_6
40. Morris, R., et al.: Planning, scheduling and monitoring for airport surface operations. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)
41. Mukhopadhyay, A., Chakraborty, S.: Replicator equations induced by microscopic processes in nonoverlapping population playing bimatrix games. *Chaos: Int. J. Nonlinear Sci.* **31**(2) (2021). <https://doi.org/10.1063/5.0032311>
42. Pant, Y.V., Abbas, H., Mangharam, R.: Smooth operator: control using the smooth robustness of temporal logic. In: 2017 IEEE Conference on Control Technology and Applications (CCTA), pp. 1235–1240 (2017). <https://doi.org/10.1109/CCTA.2017.8062628>
43. Pant, Y.V., Abbas, H., Quaye, R.A., Mangharam, R.: Fly-by-logic: control of multi-drone fleets with temporal logic objectives. In: 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS), pp. 186–197 (2018). <https://doi.org/10.1109/ICCPS.2018.00026>
44. Pant, Y.V., Abbas, H., Quaye, R.A., Mangharam, R.: Fly-by-logic: control of multi-drone fleets with temporal logic objectives. In: Proceedings of the International Conference on Cyber-Physical Systems (ICCPS), pp. 186–197 (2018)
45. Pnueli, A.: The temporal logic of programs. In: 18th Annual Symposium on Foundations of Computer Science (SFCS 1977), pp. 46–57. IEEE (1977)
46. Qin, X., Aréchiga, N., Deshmukh, J., Best, A.: Robust testing for cyber-physical systems using reinforcement learning. In: 2023 21st ACM-IEEE International Symposium on Formal Methods and Models for System Design (MEMOCODE), pp. 36–46 (2023)
47. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: reliable reinforcement learning implementations. *J. Mach. Learn. Res.* (2021)
48. Reinbacher, T., Rozier, K.Y., Schumann, J.: Temporal-logic based runtime observer pairs for system health management of real-time systems. In: Ábrahám, E., Havelund, K. (eds.) TACAS 2014. LNCS, vol. 8413, pp. 357–372. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54862-8_24
49. Sadigh, D., Kapoor, A.: Safe control under uncertainty with probabilistic signal temporal logic. In: Proceedings of Robotics: Science and Systems XII (2016)
50. Salzman, O., Stern, R.: Research challenges and opportunities in multi-agent path finding and multi-agent pickup and delivery problems. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1711–1715 (2020)
51. Sandholm, W.H.: Evolutionary game theory. In: Meyers, R. (ed.) Encyclopedia of Complexity and Systems Science, pp. 3176–3205. Springer, New York (2009). https://doi.org/10.1007/978-0-387-30440-3_188

52. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint [arXiv:1707.06347](https://arxiv.org/abs/1707.06347) (2017)
53. Smith, J.: *Evolution and the Theory of Games*. Cambridge University Press, Cambridge (1982)
54. Stentz, A.: Optimal and efficient path planning for partially known environments. In: Hebert, M.H., Thorpe, C., Stentz, A. (eds.) *Intelligent Unmanned Ground Vehicles*. The Springer International Series in Engineering and Computer Science, vol. 388, pp. 203–220. Springer, Boston (1997). https://doi.org/10.1007/978-1-4615-6325-9_11
55. Stern, R., et al.: Multi-agent pathfinding: definitions, variants, and benchmarks. In: *Twelfth Annual Symposium on Combinatorial Search* (2019)
56. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press (2018)
57. Varambally, S., Li, J., Koenig, S.: Which MAPF model works best for automated warehousing? In: *Proceedings of the International Symposium on Combinatorial Search*, vol. 15, pp. 190–198 (2022)
58. Watkins, C.J., Dayan, P.: Q-learning. *Mach. Learn.* **8**(3), 279–292 (1992)
59. Williams, S., Deshmukh, J.: Potential games on cubic splines for multi-agent motion planning of autonomous agents. In: *2024 International Conference on Autonomous Agents and Multiagent Systems*. University of Southern California, Los Angeles (2024)
60. Zhou, W., Li, W.: A hierarchical Bayesian approach to inverse reinforcement learning with symbolic reward machines (2022). <https://doi.org/10.48550/arXiv.2204.09772>