Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30958

Research on the Application of Q-learning in Braitenberg Car

Jinyu Liu

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
859542454@qq.com

Abstract. Addressing the issues of insufficient behavioral stability in Braitenberg vehicles within complex light fields and obstacle-laden environments, this paper adopts tabular Q-learning to learn phototaxis and obstacle avoidance strategies. The vehicle constructs discrete states based on the intensity of left and right light sensors and collision flags, utilizing actions including moving forward, turning left, turning right, and stopping. Reward shaping is employed to encourage approaching the light source while penalizing collisions and ineffective idling. The strategy employs an ϵ-greedy approach, with decay applied to both the learning rate and exploration rate during training, while a discount factor balances long-term returns. Moderate domain randomization is used during training to close the gap between simulation and reality, and safety shielding is used in the early stages to keep high-risk actions to a minimum. Experiments conducted in a two-dimensional grid environment with random obstacles demonstrate that the algorithm converges within thousands of steps, significantly reducing the average number of collisions per episode and markedly improving the success rate of phototaxis. This paper provides a minimal reproducible implementation and key hyperparameter settings, offering a concise and effective baseline for low-cost mobile robot teaching and Braitenberg behavioral research.

Keywords: Q-Learning, Braitenberg Vehicles, Phototaxis and Obstacle Avoidance, Reinforcement Learning, Simulation-to-Reality Transfer

1. Introduction

Braitenberg vehicles demonstrate diverse behaviors through minimalist perception-action mechanisms. However, in scenarios involving complex light fields and obstacles, hand-crafted rules struggle to simultaneously guarantee stability and transferability. Their overall performance is sensitive to parameters and environmental perturbations, leading to the easy degradation of phototaxis and obstacle avoidance capabilities in real-world applications. Reinforcement learning offers a lightweight alternative path. Compared to deep learning techniques, tabular Q-learning is less reliant on processing power and data since it is easy to implement, interpretable, and computationally inexpensive in low-dimensional discrete states. However, the majority of current work concentrates on deep networks or broad mobile platforms, lacking a systematic baseline that is replicable, focused on simulation-to-reality transfer, and targeted towards Braitenberg cars.

DOI: 10.54254/2755-2721/2026.TJ30958

To this end, this paper proposes a minimal reproducible experimental scheme centered on the task of "phototaxis with obstacle avoidance." It clarifies states/actions and reward shaping, adopts an ϵ-greedy strategy with decaying exploration and learning rates, and introduces moderate domain randomization and safety shielding during the training phase to mitigate early-stage risks and transfer bias. Convergence within limited samples and a reduction in collisions are verified in a 2D environment with random obstacles. Furthermore, benchmark performance curves and practical points are provided, serving as a concise baseline for subsequent method comparisons and teaching experiments.

2. Theoretical overview

Q-learning belongs to the class of model-free temporal difference (TD) methods. Its objective is to learn the optimal state-action value function within a Markov Decision Process (MDP) and subsequently adopt an ϵ-greedy strategy to maximize cumulative returns [1]. The algorithm performs a bootstrap update on the current $Q(s,a)$ after interacting with the environment:

$$Q_{t+1}\left(s_t,a_t\right)=\left(1-\alpha\right)Q_t\left(s_t,a_t\right)+\alpha\left[r_t+\gamma\underset{a'}{\max}Q_t\left(s_{t+1},a'\right)\right] \tag{1}$$

Since the update target uses an estimate of the optimal value rather than the behavior policy itself, this method is classified as off-policy. Under classical conditions, it has been proven to converge to the optimal $Q^*(s,a)$.

During the learning process, the trade-off between exploration and exploitation determines convergence efficiency and final performance. The commonly used ϵ-greedy strategy explores randomly with probability ϵ and selects the action with the current maximum value with probability $1-\epsilon$. The value of ϵ decays gradually with the training process to achieve a transition from broad exploration to stable exploitation. As long as each state-action pair is visited sufficiently and the learning rate satisfies standard decay conditions, the algorithm is almost guaranteed to converge [1,2].

The setting of the reward function directly shapes the learning objective and affects sample efficiency. For phototaxis and obstacle avoidance, in addition to sparse signals such as terminal arrival and collision, incremental rewards/penalties related to changes in distance to the light source can be introduced to encourage "approaching the target and staying away from obstacles." This is supplemented by slight time penalties to suppress ineffective wandering. At the same time, it is necessary to avoid shaping methods that are inconsistent with the task goal, as improper rewards may induce sub-optimal behaviors.

Tabular Q-learning is concise and easy to reproduce in discrete, low-dimensional scenarios. However, as local occupancy information and multiple sensor quantities are incorporated, the state-action space expands rapidly. Coverage and storage overhead become major bottlenecks, leading to problems such as slow convergence and sensitivity to hyperparameters. When the perception dimension is high or states are continuous, function approximation is often used to replace the Q-table. Approximate methods represented by Deep Q-Networks (DQN) demonstrate stronger scalability and generalization capabilities on complex tasks, providing a path for subsequent moves from simulation to more realistic sensing and dynamics modeling [3]. The theoretical points above are directly relevant to the Braitenberg phototaxis and obstacle avoidance tasks focused on in this paper and lay the foundation for subsequent modeling and experimental design.

DOI: 10.54254/2755-2721/2026.TJ30958

3. Application research of Q-learning in braitenberg vehicles

3.1. Q-learning modeling and training for phototaxis behavior

Phototaxis is regarded as an instinctual response to the gradient of light source intensity. Braitenberg vehicles can exhibit such behavior under a minimal perception-execution architecture, making them suitable carriers for learning-based control [4,5]. In a discrete two-dimensional grid, the state can be composed of the vehicle's coordinates, the relative bearing of the light source (quadrant or discrete angular sector), and the segmented distance to the light source. Actions can be limited to moving forward, turning left, and turning right to stably update values and reduce the search space. Termination conditions are typically defined as reaching the light source, colliding, or reaching the maximum number of steps.

In order to steer the value update toward a strategy of "approaching the target and avoiding ineffective wandering," the reward function incorporates incremental rewards/penalties associated with changes in distance to the light source, as well as small time penalties, in addition to the terminal reward for reaching the target and collision penalties. This provides a smooth learning gradient beyond sparse signals. The policy selection adopts an ϵ -greedy mechanism that gradually decays as training progresses, ensuring sufficient exploration in the early stages and tending towards greedy exploitation in the later stages. When state-action pairs are sufficiently visited and the learning rate meets standard decay conditions, value iteration almost inevitably converges to the optimal Q^* . In contrast to the early "hard-wired" paradigm of Braitenberg vehicles, the phototaxis task can generate a stable and repeatable approximate shortest-path strategy through bootstrap updates without depending on an explicit environment model [4].

3.2. State space construction and Q-table update for obstacle avoidance

Obstacle avoidance requires explicitly incorporating "safety" information into decision-making. Therefore, the state needs to be extended to reflect local obstacle distribution or rangefinding readings. A common practice is to add discrete rangefinding segments in the front, left, and right directions, or to encode passable/impassable units in a 3x3 or 5x5 occupancy subgraph centered on the vehicle. This provides learnable local geometric cues without introducing continuous high-dimensional perception. While maintaining the phototaxis goal, the reward function imposes strong penalties for collisions, small penalties for approach actions entering a safety threshold, and micro-rewards for successfully bypassing obstacles while approaching the target, thereby reinforcing "safe and efficient" sequential decision-making [1].

After the aforementioned expansion, the search space of tabular Q-learning increases significantly, making the learning process more sensitive to hyperparameters and reducing sample efficiency. This is consistent with empirical observations of the "curse of dimensionality" in existing mobile robot research [1,6]. To reduce deployment risks, actual systems can superimpose policy shielding or equivalent safety filtering at the execution end. This replaces high-risk actions with safe alternatives that meet constraints before issuance, thereby significantly reducing collision events during training and evaluation without altering the learning algorithm [7].

3.3. Simulation environment and training framework design

The simulation environment is implemented as an episodic interaction system with reset and step interfaces. In Figure 1, the agent moves from the top-left starting point towards the bottom-right

DOI: 10.54254/2755-2721/2026.TJ30958

light source in an $N \times N$ grid, interacting under fixed or random obstacle configurations. At the beginning of each episode, the environment is reset and returns the initial observation. At each step, the policy selects an action according to the ϵ -greedy rule (which decays with training progress), and the environment returns the next observation, immediate reward, and termination flag. The tabular Q-values are bootstrap-updated based on this information. The episode ends upon arrival, collision, or exceeding the maximum number of steps.

During training, metrics such as arrival rate, collision count, cumulative return, and path sub-optimality are statistically rolled to monitor learning progress and set early stopping conditions. To improve robustness against uncertainty, randomization within a reasonable range is implemented for critical parameters such as friction, sensor noise, lighting, and action delay during the training phase. This reduces the overfitting of the strategy to environmental details and provides a basis for subsequent transfer from simulation to physical platforms [8].

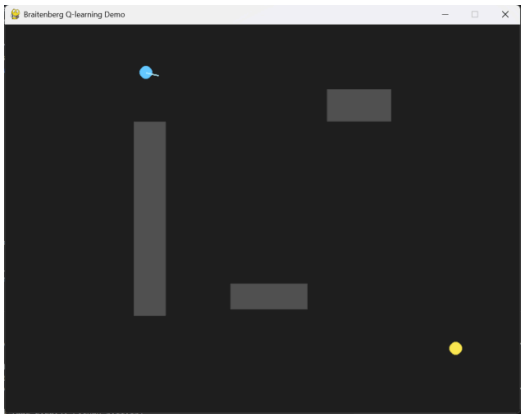

Figure 1. Map used in the experiment

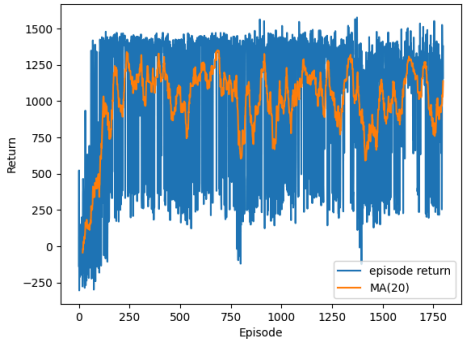

Figure 2. Learning curve

3.4. Comparison with related research cases

Under low-dimensional discrete modeling, tabular Q-learning has the advantages of simple implementation, no need for an environment model, and ease of reproduction. However, it faces bottlenecks in scalability and sample efficiency when perception dimensions and scene complexity increase [1]. Addressing high-dimensional or continuous observations, Deep Q-Networks approximate $Q(s, a)$ using neural networks combined with target networks and experience replay,

DOI: 10.54254/2755-2721/2026.TJ30958

significantly improving representation capability and performance ceilings on complex tasks. This approach is widely adopted in the fields of reinforcement learning and robotics [3].

In engineering practice, navigation and obstacle avoidance often appear simultaneously. Domestic work on search and path planning indicates that, under reasonable state abstraction and reward shaping, Q-learning-based vehicles can achieve high arrival rates and maintain low collision frequencies in obstacle-laden environments. This provides an actionable baseline scheme for small-scale platforms and teaching prototypes. In summary, the above design and comparison directly serve the modeling details and experimental reproduction of this paper, and align with the biologically inspired localization of Braitenberg vehicles [6,9-11].

3.5. Successful implementation cases and visualization effects

The training results of this simulation project demonstrate that the intelligent vehicle based on Q-learning can indeed learn the combined behavior of phototaxis and obstacle avoidance, ultimately exhibiting intelligence similar to Braitenberg vehicles. After training for a sufficient number of episodes, the vehicle can autonomously find a collision-free path from the starting point to the light source. The driving route planned by the vehicle in a complex obstacle environment is marked with a blue polyline. It can be seen that the vehicle changes direction multiple times to bypass obstacles and finally successfully arrives at the light source located in the lower right. This trajectory clearly showcases the obstacle avoidance and bypassing action sequence learned by the Q-learning strategy, as well as the phototactic tendency to continuously advance towards the target.

During the training process, this study monitored the evolution of agent performance through metrics such as cumulative reward per episode and collision counts. The cumulative reward exhibited an upward trend, indicating the vehicle's improvement in strategies for gaining positive rewards (e.g., approaching a light source) and minimizing negative rewards (collisions). Correspondingly, the collision count curve drops rapidly from an initially high level, approaching zero and remaining stable in the later stages of training, as shown in Figure 2. This implies that the vehicle no longer collides in most episodes, successfully achieving reliable obstacle avoidance movement. Furthermore, we observed that there may be significant fluctuations in the curve during the early stages of training. This is due to the unstable behavior of the vehicle under the exploration strategy, leading to fluctuating episode performance. However, as ϵ gradually decreases and the strategy approaches optimality, the curve fluctuations diminish and converge. These visualization results strongly illustrate the effectiveness of Q-learning training: through repeated interaction with the environment, the intelligent vehicle continuously improves its performance metrics, eventually achieving the target behavior of phototaxis and obstacle avoidance.

4. Conclusion

This paper systematically reviews the application of the Q-learning algorithm in the phototaxis and obstacle avoidance tasks of Braitenberg vehicles. Research indicates that Q-learning can endow Braitenberg vehicles, which originally operated on fixed reaction mechanisms, with certain learning capabilities, enabling them to learn goal-oriented navigation strategies through environmental interaction without manual programming. This provides a feasible pathway for intelligent vehicles and even broader autonomous robot control, reflecting the potential of reinforcement learning in the robotics field.

Traditional Q-learning algorithms experience challenges like slow convergence and poor scalability in complex scenarios. However, these issues can be addressed through emerging

DOI: 10.54254/2755-2721/2026.TJ30958

technologies such as deep learning and transfer learning, alongside better algorithm design. The Deep Q-Network, a significant evolution of Q-learning, has shown success in complex tasks, indicating the potential for broader applications. Fields such as autonomous driving and intelligent unmanned vehicles are poised to benefit from enhancements in Q-learning methods. Ongoing research aims to refine algorithms and reduce the simulation-to-reality gap, expanding the applications of Q-learning in intelligent mobile entities.

References

[1] Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., & Topcu, U. (2018). Safe reinforcement learning via shielding. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) (pp. 2669–2678). AAAI Press.

[2] Braitenberg, V. (1986). Vehicles: Experiments in synthetic psychology. MIT Press.

[3] Chu, J., Deng, X., & Yue, Q. (2024). Autonomous path planning for search-and-rescue robots based on Q-learning. Journal of Nanjing University of Aeronautics & Astronautics, 56(2), 364–374. https: //doi.org/10.16356/j.1005-2615.2024.02.020

[4] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https: //doi.org/10.1038/nature14236

[5] Shaikh, D., & Rañó, I. (2020). Braitenberg vehicles as computational tools for research in neuroscience. Frontiers in Bioengineering and Biotechnology, 8, 565963. https: //doi.org/10.3389/fbioe.2020.565963

[6] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[7] Tian, X., & Dong, X. (2024). Obstacle-avoidance path planning for mobile robots based on improved DQN. Journal of Chinese Inertial Technology, 32(4). https: //doi.org/10.13695/j.cnki.12-1222/O3.2024.04.012

[8] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. arXiv preprint arXiv: 1703.06907. Presented at the IROS 2017 Workshop.

[9] Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3), 279–292. https: //doi.org/10.1007/BF00992698

[10] Zhao, J., Pei, Z., Jiang, B., Lu, N., Zhao, F., & Chen, S. (2024). Visual obstacle avoidance for UAV virtual pipelines based on deep reinforcement learning. Acta Automatica Sinica, 50(11), 2245–2258. https: //doi.org/10.16383/j.aas.c230728

[11] Zhou, Z., Yu, S., Yu, J., Duan, J., Chen, L., & Chen, C. L. P. (2023). T-DQN intelligent obstacle-avoidance algorithm for unmanned surface vehicles. Acta Automatica Sinica, 49(8), 1645–1655. https: //doi.org/10.16383/j.aas.c210080