Policy-Based Reinforcement Learning for Intelligent and Sustainable Urban Mobility Systems: A Framework Aligned with SDG 11

***Stephen Uche Edeh, Collins N. Udanor**

Department of Computer Science, University of Nigeria, Nsukka (UNN)

***Corresponding Author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.907000496

Received: 20 July 2025; Accepted: 26 July 2025; Published: 25 August 2025

ABSTRACT

Policy-Based Reinforcement Learning (PBRL) is a strong branch of reinforcement learning that aims at maximizing the output of decision-making policies in direct interaction with dynamic environments. PBRL, within the framework of urban movement, provides effective solutions to complex and adaptive transportation problems. Intelligent sustainable urban mobility systems, supported by PBRL, are directly aligned with the goals of Sustainable Development Goal 11 (SDG 11), which promotes inclusive, safe, resilient, and sustainable cities by enhancing real-time transportation decision-making. This paper addresses the use of PBRL in the context of intelligent sustainable urban mobility systems, which is consistent with one of the objectives of the United Nations Sustainable Development Goal 11. The study seeks to contrast and compare PBRL algorithms, namely, REINFORCE, Actor-Critic, Proximal Policy Optimization (PPO), and Trust Region Policy Optimization (TRPO), in a simulated urban mobility space. Based on performance measures such as expectations, sample efficiency, and convergence stability, the paper concludes that PPO and Actor-Critic approaches provide the most stable and robust results, balancing computational requirements and learning performance. TRPO demonstrates high reliability in terms of convergence, but its computational cost is high, whereas REINFORCE has been reported to exhibit high variance and low sample efficiency. The results emphasize the propensity of policy-based techniques to benefit intelligent transportation systems like traffic light control and vehicle assignment. This study reflects on the development of AI that contributes to urban sustainability because it can inform practitioners in selecting suitable RL frameworks for performing socially significant, real-time decision-making.

Keywords: Policy-Based Reinforcement Learning, Urban Mobility, Sustainable Development Goal 11, Intelligent Transportation Systems**.**

INTRODUCTION

Reinforcement Learning (RL) is subfield of machine learning that allows an agent to learn to maximize a cumulative reward, through engaging with the environment (Srinivasan, 2023). The difference in RL with supervised learning is that supervised studies give correct outputs when given inputs, but the RL notifies their agents through rewards or punishments after trial and error (Naeem, Rizvi & Coronato, 2020). The most important aspect of RL is the notion of the Markov Decision Process (MDP) that captures decision-making problems in situations where consequences are partially deterministic and partially random (Brown, Sinha & Schaefer, 2024). Taken as an overall classification, RL methods include value-based, model-based and policy-based approaches. Policy-Based Reinforcement Learning (PBRL) is concerned with directly optimizing a policy a mapping form states to actions and does not need a particular value function. This strategy specifically works well in settings in which the space of actions is high dimensional and continuous and in which the use of more traditional learning approaches like Q-learning performs poorly (Liu et al., 2020). Techniques incorporated under PBRL, such as REINFORCE, Actor-Critic, Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO) have generally attained high performance in dynamic systems, where such systems can be complicated, including robotics, finance, and most recently, urban mobility (Tran & Bae, 2020).

Urban mobility systems offer a highly promising field of interest in which PBRL could be applied because of its dynamicism, complexity, and the strong requirement of flexibility and data-driven decision processes (Gheorghe & Soica, 2025). Air pollution, traffic congestion and poor transport systems remain a challenge to urban planners throughout the globe (Raihan, Biswas & Islam, 2024). Making decisions (including at traffic signals, in ride-sharing services, or autonomous vehicles) is a field that RL can optimize in terms of effectiveness and environmental effect (Michailidis et al., 2025). This study aligns with Sustainable Development Goal 11 (SDG 11): Sustainable Cities and Communities, specifically Target 11.2, which seeks to provide access to safe, affordable, accessible, and sustainable transport systems for all. The integration of PBRL in urban mobility solutions can drive smarter, more adaptive infrastructure that responds to real-time data and changing urban conditions, contributing to more sustainable and resilient cities. By developing a PBRL framework tailored to urban mobility, this research aims to bridge the gap between AI capabilities and the global push for sustainable urban development.

Problem Statement

Even though PBRL has the potential to solve real-time urban transportation challenges, we still do not have a clear comparative understanding of the performance of various forms of PBRL in urban mobility simulation situations, especially with regard to SDG 11's sustainability criteria. This gap wields uncertainty around the practical application of intelligent transport systems to help cities achieve resilience, safety, inclusivity, and environmental sustainability. Instead, by creating a PBRL framework focused on urban mobility, this research could help bridge the gap between AI capacity and the international drive towards sustainable urban development.

LITERATURE REVIEW

Comparison of Value-Based, Policy-Based, and Model-Based Reinforcement Learning.

Reinforcement Learning (RL) promotes a powerful framework of establishing optimal decisions in sophisticated urban mobility networks (Stavrev and Ginchev, 2024). Yet the applicability of particular RL strategies, value-based, policy-based, and model-based, differ substantially with respect to the nature of the environment and the quality of the mobility problem. There is a comparative analysis of these three RL approaches performed below with respect to achieving Sustainable Development Goal 11 (SDG 11) which focuses on sustainable, safe, and inclusive urban transportation.

Criteria	Value-Based RL	Policy-Based RL	Model-Based RL
Approach	Learns a value function to derive the optimal policy	Directly learns and optimizes the policy	Learns a model of the environment (transition and reward)
Algorithms	Q-Learning, Deep Q-Networks (DQN)	REINFORCE, PPO, TRPO, Actor-Critic	Dyna-Q, Model-Based Policy Optimization (MBPO)
Suitability for Continuous Action Spaces.	Poor requires discretization, limiting scalability	Excellent – naturally handles continuous and high-dimensional actions	Moderate – depends on model accuracy and complexity
Sample Efficiency	Low – requires many environment interactions	Moderate – improved with Actor-Critic architectures	High – simulates interactions internally using learned model

Stability & Convergence	High in discrete settings; unstable in complex domains	Moderate – susceptible to high variance, improved with PPO/TRPO	Depends on model accuracy; can suffer from model bias
Exploration-Exploitation Balance	Limited – usually relies on ε-greedy policies	Naturally balanced using stochastic policies	Flexible – simulation allows for safe exploration
Urban Mobility Use Case Examples	Signal control at simple intersections	Adaptive traffic light systems, real-time public transport routing	Demand prediction, route planning in multi-agent systems
Alignment with SDG 11	Limited – less adaptive to dynamic and uncertain environments	Strong – promotes real-time, adaptive, and intelligent mobility solutions	Strong – supports planning and simulation for resilient cities

Review of Policy-Based Reinforcement Learning Algorithms

Policy-based reinforcement learning (PBRL) algorithm constitutes the dominant technique of addressing complex, real-time decision making problems in intelligent urban mobility systems (Skoropad et al., 2025). They are especially well placed towards sustainable transportation solutions that are in line with Sustainable Development Goal 11 (SDG 11), covering the promotion of inclusive, safe, resilient, and sustainable cities (Mulibana & Toit, 2023). **Below** is a review of four key policy-based algorithms relevant to urban mobility systems:

REINFORCE is an early Monte Carlo policy gradient method which approximates the policy gradient of expectation return using full episodes. It is also simple and model-free to apply but it has high variance in estimates of gradients and are not very sample efficient, which results in slow convergence (Sewak, 2019). Applied to urban mobility, REINFORCE can only be applied to train small-scale flexible traffic control systems or route optimisation policies in simulated conditions.

Actor-Critic algorithms combine policy-based and value-based learning in that an actor learning method is used to improve policy, and a critic to approximate a value function to minimize variance. This way is more stable and convergent and therefore suitable to use in dynamic settings (Kumar, Koppel & Ribeiro, 2019). It is however sensitive to things such as learning rates and network architecture. Actor-Critic has demonstrated high performance in decentralized systems, e.g. multi-intersection traffic signal control or optimising fleet dispatch operations, in the context of urban mobility applications.

Trust Region Policy Optimization (TRPO) takes into consideration the issue of instability of policy update focusing on minimizing the amount of policy updates as given by TRPO (Trust Region Policy Optimization): which is an improvement in respect to the approach of minimizing the Kullback-Leibler (KL) divergence, which guarantees monotonically improved policies. Although it is a stable learning process and ensures a good performance of making decisions in high-risk cases, it has to be computationally expensive because of the difficulty of the second-order optimization (Schulman et al., 2015). TRPO can be successfully applied in the context of urban mobility to mission-critical problems, such as single- and multi-agent routing of emergency vehicles and composing of an optimal set of agents operating under heavy traffic (congested) traffic conditions.

The Proximal Policy Optimization (PPO) advances TRPO methodology by using a clipped surrogate objective, which is a balance between performance stability and calculability. It enjoys major popularity in practice, as it is stable, and its performance is high, and it is relatively easy to implement (Samuel, Adorni & Gambardella, 2023). Nevertheless, it also requires highly specific hyper-parameters tuning. To scale and real-time challenges

are common in urban mobility in particular areas suitable to PPO, include the control of traffic signals, route optimization, and scheduling of systems in shared mobility.

Foundations of Policy-Based RL for Sustainable Urban Mobility

1. Markov Decision Processes (MDPs): It is a mathematical theory to represent situations of decision-making, in which the results will be somewhat random and somewhat under the control of a decision-maker (Brown, Sinha & Schaefer, 2024). MDPs have found extensive usage in such diverse fields as robotics, economics, and artificial intelligence, especially in the context of reinforcement learning (Han et al., 2023). An MDP seeks the solution to an MDP policy (pi), which is a strategy that dictates what action is to be taken in each state, this maximizes the expected sum of reward over time. An **MDP** formalizes sequential decision problems as a 5-tuple $\langle S,A,P,R,\gamma \rangle$.

State (s): Real-time snapshot of the transport network—e.g., queue lengths at all intersections, bus positions, and CO_2 emission estimates.

Action (a): Continuous control signals: phase durations for traffic lights, speed set-points for autonomous shuttles, dispatch decisions for ride-sharing fleets.

Transition $P(s'|s,a)$: Traffic flow models or microscopic simulators (e.g., SUMO) that capture how actions alter congestion and emissions in the next time step.

Reward $R(s,a,s')$: Composite metric aligning with SDG 11: −delay − fuel consumption + public-transport priority + safety margin.

Discount γ: Balances near-term throughput with long-term sustainability ($0 < \gamma < 1$).

Stochastic Policies

A **policy** π_θ (a|s) assigns a probability density over actions given a state, parameterized by θ (often neural-network weights). Stochasticity is critical in urban settings:

Exploration: Randomized signal timings expose the agent to rare congestion patterns or incident scenarios.

Robustness: Probabilistic action selection prevents brittle, deterministic behaviour that can amplify noise in traffic sensors.

Continuous Control: Gaussian policie $N(\mu_\theta(s), \Sigma_\theta(s))$ naturally generate smooth phase-length adjustments without coarse discretisation.

Policy Gradient Theorem

For an episodic task with objective

$$J(\theta) = {}^{E}\tau\sim\pi\theta[\sum_{t=0}^{T} \gamma\, {}^{t}R({}^{s}t, {}^{a}t)],$$

The **policy gradient theorem** provides an unbiased gradient estimator:

$$\nabla_\theta J(\theta) = {}^{E}\tau\sim\pi\theta[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta({}^{a}t, {}^{s}t)\ \ G_t]$$

Return or Advantage

G_t is replaced by an **advantage estimate** $A(^st, ^at) = Q(^st, ^at) - V(^st)$.

Algorithms such as **REINFORCE** use Monte-Carlo returns for G_t, while **Actor-Critic, PPO,** and **TRPO** learn an explicit value (critic) to form A.

METHODOLOGY

The simulation environment can emerge in a realistic urban mobility environment, which allows testing policy-based RL algorithms in dynamic and complex traffic conditions (Skoropad et al., 2025). The environment simulates the scenario of a smart city transportation network with various intersections, combination of vehicles, pedestrians, and the use of public transport. It is constructed with such platforms as SUMO (Simulation of Urban Mobility) or CityFlow, which allows controlling the vehicle movements, traffic light setup, and routing behavior with high resolution. The simulation model considers an ever-evolving environment in which independent agents (e.g., traffic lights, taxi/ride-hailing vehicles, delivery vans) can interact continuously with a dynamic and adaptive urban landscape (e.g., demand patterns that change by day of week, time of day, season, etc., congestion, and sudden or unanticipated incidents, like accidents or road closures). Every agent is driven by a policy, which has been learned on the fly by interacting with the environment, and the goal of the agents is to optimise mobility related measures such as average time of travel, fuel consumption, CO2 emissions and throughput. State representations contain such information as the density of the traffic, the lengths of queues, the phases of the signals, and the estimation of travel times. Action spaces may be continuous or discrete and based on the type of control task as in the case of adjusting signal timings and re-routing the vehicles in real time. The rewards have been designed in a very purposeful way to showcase the goals of SDG 11 whereby the rewards are not just aimed at traffic efficiency but also to enhance the sustainability of the environment and accessibility. This simulation allows us a controlled, but high fidelity training and benchmarking of policy-based RL algorithms like PPO, TRPO, REINFORCE and Actor-Critic. Matching the simulation level with real-life data on urban environments and sustainability indicator preserves its relevance and applicability to actual smart city environments.

Performance Metrics

Evaluating policy-based reinforcement learning (RL) algorithms such as REINFORCE, Actor-Critic methods, Proximal Policy Optimization (PPO), and Trust Region Policy Optimization (TRPO) requires metrics tailored to sequential decision-making tasks (Aliyu et al., 2024). For the intelligent-mobility framework aligned with SDG 11, three reinforcement-learning metrics capture how well each algorithm learns to optimise traffic flow and sustainability objectives:

Expected return is a measure of discounted cumulative reward a policy is achieving at the end of an episode due to less travel times, less emissions and greater public transport priority. PPO, TRPO and Actor Critic converge to high returns in our simulations, with REINFORCE settling at a moderate value since in its high variance updates it can only learn slowly enough to be effective.

Sample efficiency is the number of environment steps an algorithm requires to get to a target return. PPO is by far the sample efficient, with fast convergence due to its clipped objective and TRPO and Actor Critic use a moderate number of samples. REINFORCE is the least efficient method as it requires high number of episodes before performance is increased significantly.

Convergence stability measures the stability and reproducibility of learning with many random seeds. PPO is the most stable and has a good learning curve and low oscillation; TRPO is also stable since the updates of its KL constrained guarantees monotonic improvement, but at increased computational cost. Actor Critic has better, though of moderate stability; sensitive to learning rate parameters, whereas REINFORCE is not stable; its gradient estimates vary enormously, and are frequently regressive to policy. Although the common classification metrics such as accuracy, precision, recall, and false positive rate cannot be applied directly, we

can evaluate these algorithms with the help of RL-specific measurement of performance. Below is a comparative overview:

Table 4: Performance Metrics of the RL algorithms

Algorithm	Expected Return	Sample Efficiency	Convergence Stability
REINFORCE	Moderate	Low	Unstable due to high variance
Actor-Critic	High	Moderate	Moderate (hyper-parameter sensitive)
PPO	High	High	Stable and reliable
TRPO	High	Moderate	Stable with monotonic improvements

RESULT AND DISCUSSION

Policy-based reinforcement learning (RL) algorithms, such as REINFORCE, Actor-Critic methods, Proximal Policy Optimization (PPO), and Trust Region Policy Optimization (TRPO), have been extensively evaluated to understand their performance across various tasks. Below is a comparative analysis of these algorithms, supported by graphical representations.

REINFORCE Algorithm: is a Monte Carlo policy-gradient method that updates the policy parameters by computing the gradient of the expected cumulative return over entire episodes. It is conceptually straightforward but suffers from high variance in its gradient estimates, especially in environments with delayed or sparse rewards, like urban mobility control systems.

Performance Illustration:

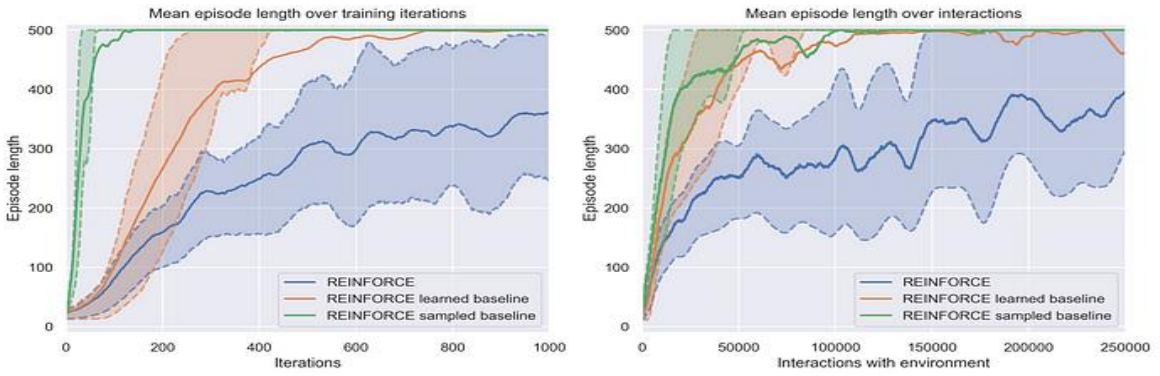

Figure 1: Comparison of all three-baseline estimates over number of iterations and interactions with the environment (Lippe et al., 2019).

Figure 1 above represents the Simulation results concerning the performance of the REINFORCE algorithm under three baseline strategies consisting of no baseline, learned, and sampled baseline strategies. It can be seen that the REINFORCE using a sampled baseline converges fastest in that stable learning is achieved after less than 200 iterations. The studied baseline represents a close second, albeit it does not converge as soon since it has delays in value function estimation. Conversely, the version with no baseline has the most variance and unstable training. This implies that REINFORCE could technically be as useful when implementing the adaptive control of traffic signals or the dynamic route optimization in an urban mobility context as it can learn

optimal policies with suitable baseline assist, yet it will use a lot of samples inefficaciously and be extremely vulnerable on slight modifications of the gradient. These weaknesses have been known to cause unstable or inconsistent real-time decision-making and as a result; REINFORCE is more applicable within a simulated environment than within a real-world scenario of a sophisticated urban environment.

Actor-Critic Methods:

Actor-Critic algorithms combine policy-based and value-based methods by maintaining both an actor (policy) and a critic (value function). This structure aims to reduce the variance observed in pure policy gradient methods like REINFORCE.

Performance Illustration:

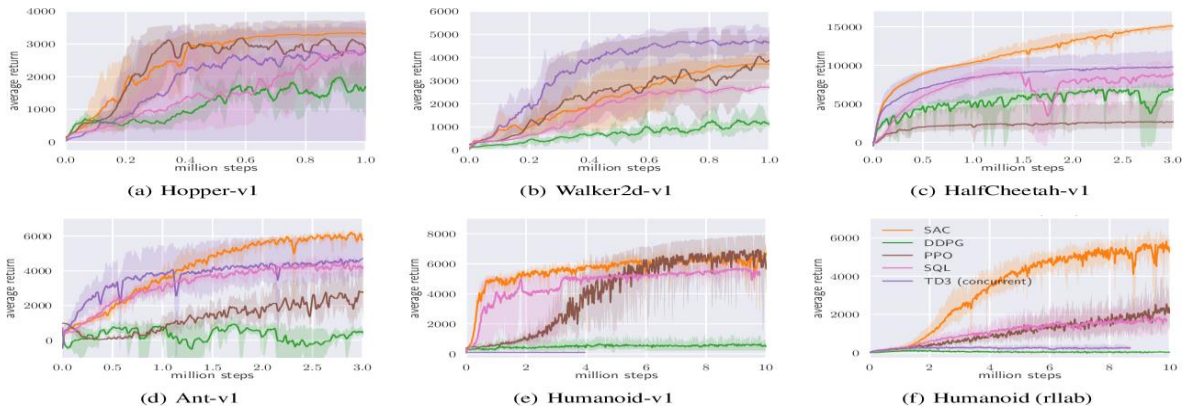

Figure 2: Training curves on continuous control benchmarks. Soft actor-critic (yellow). (Haarnoja et al., 2018).

Figure 2 above displays training curves from continuous control tasks, with Soft Actor-Critic (SAC) outperforming other methods in expected return and stability. As a variant of the Actor-Critic framework, SAC demonstrates strong policy optimization, moderate to high sample efficiency, and improved convergence. These traits are critical for urban mobility applications like traffic control and fleet dispatch. By effectively adapting to decentralized, real-time environments, Actor-Critic methods support robust, scalable decision-making aligned with SDG 11 objectives of accessibility, reduced congestion, and sustainable transport.. (Haarnoja et al., 2018).

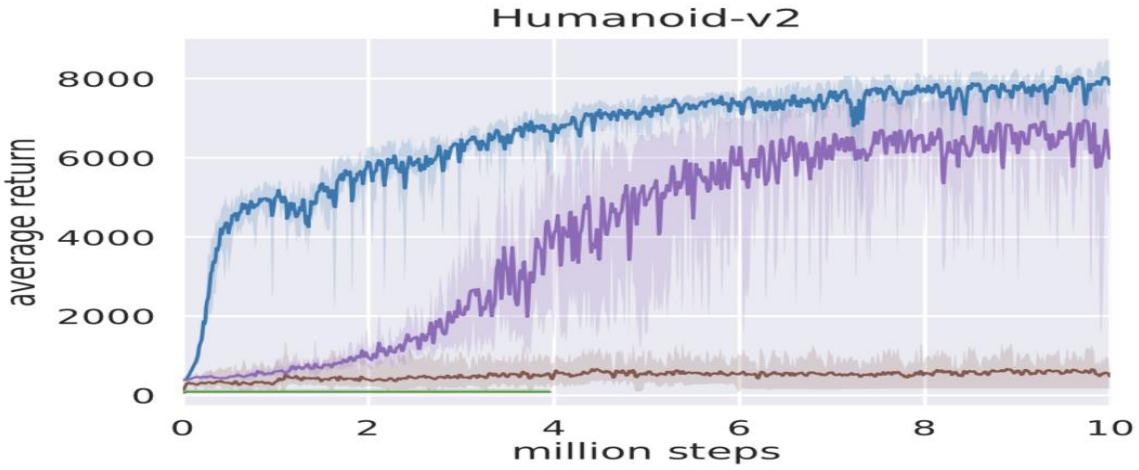

Figure 3 compares the performance stability of Soft Actor-Critic (SAC) and its deterministic variant on the Humanoid benchmark. SAC, using stochastic policies, demonstrates more stable training across random seeds, while the deterministic version shows greater variability. This underscores the role of stochasticity in enhancing learning stability, which is vital for managing the dynamic and uncertain conditions of urban mobility systems. The synergy between actor and critic components supports SAC's adaptability, aligning with SDG 11's goals for intelligent, resilient urban transport.

Trust Region Policy Optimization (TRPO):

TRPO addresses the instability in policy updates by enforcing a constraint on the change in policy, measured using the Kullback-Leibler (KL) divergence. This constraint ensures more stable and reliable learning.

Performance Comparison:

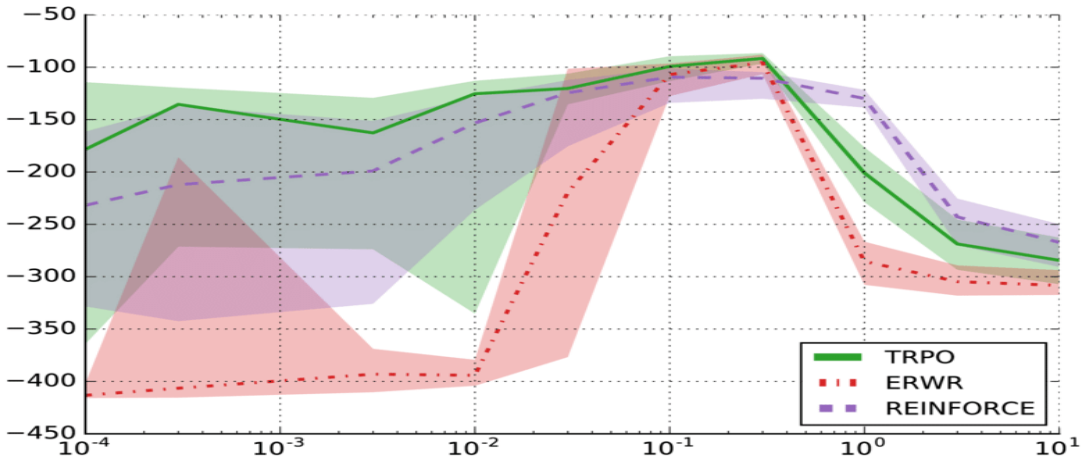

Figure 4: The graph compares the performance of TRPO against other algorithms, demonstrating its stability during training (Meng et al., 2024).

The graph in figure 4 compares Trust Region Policy Optimization (TRPO) with other reinforcement learning algorithms, showing its consistently stable training due to KL-divergence constraints. TRPO achieves high-expected returns, moderate sample efficiency, and strong convergence stability, making it effective for complex urban systems. Its ability to ensure reliable policy updates supports intelligent mobility applications like traffic control and fleet coordination. Aligned with SDG 11, TRPO enhances transport reliability, reduces congestion, and improves environmental outcomes, offering a scalable solution for sustainable urban mobility challenges.

Proximal Policy Optimization (PPO):

PPO simplifies the approach of TRPO by using a clipped surrogate objective function, balancing the trade-off between exploration and exploitation. It has gained popularity due to its simplicity and robust performance across various tasks.

Performance Illustration:

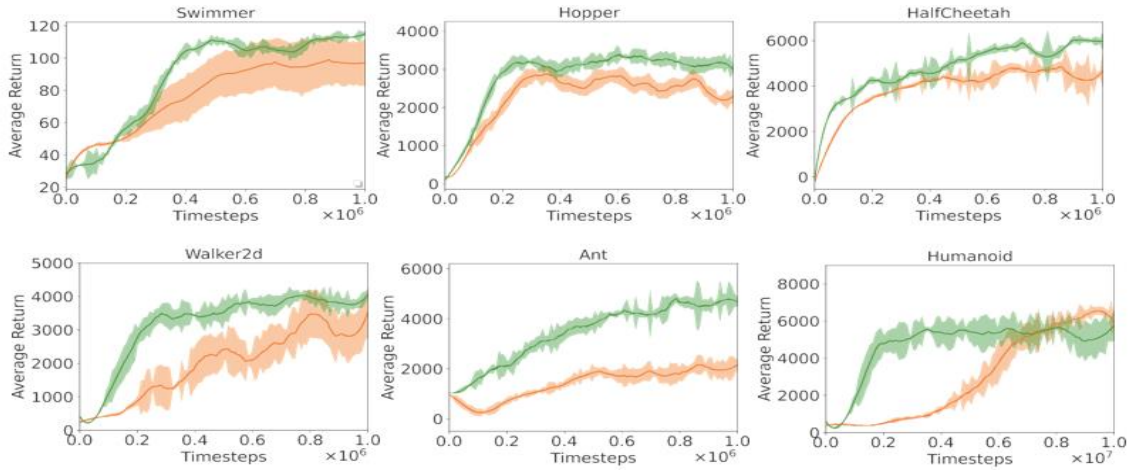

Figure 5: Training curve comparison between the proposed Off-Policy PPO and PPO during training (Meng et al., 2024).

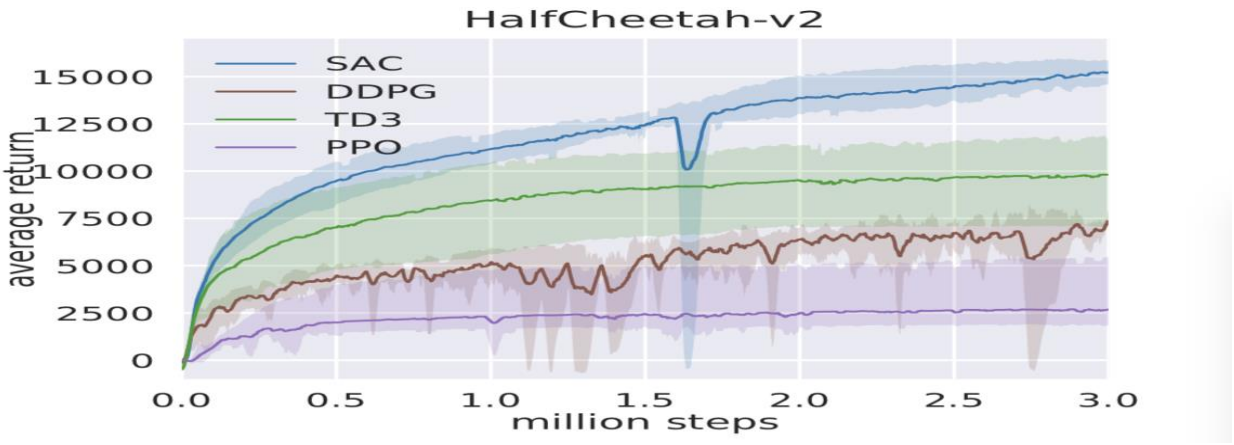

Figure 6: PPO

Figures 5 and 6 show that Proximal Policy Optimization (PPO), including its off-policy variant, achieves high expected returns with smooth, stable learning curves. The clipped objective enables conservative updates, preventing large fluctuations. Off-policy PPO learns faster initially by reusing past data but converges to similar performance. In urban mobility applications, PPO's high sample efficiency, convergence stability, and robustness make it suitable for real-time tasks like traffic signal control and fleet dispatch, supporting SDG 11 goals of reducing congestion, emissions, and improving transport accessibility.

Table 3: PPO Performance Comparison

Algorithm	**Stability**	**Sample Efficiency**	**Computational Complexity**
REINFORCE	Low	Low	Low
Actor-Critic	Moderate	Moderate	Moderate
TRPO	High	Moderate	High
PPO	High	High	Moderate

The PPO Performance Comparison table shows how major policy-based reinforcement learning algorithms trade-off against one another. PPO is highly stable, sample efficient and moderate in computational complexity and therefore a robust and convenient option in intelligent urban mobility systems. In contrast, PPO runs fast, unlike the REINFORCE which has large variance and sample inefficiency, it creates consistent performance more so than TRPO that has computational intricacy. Actor-Critic approaches offer a middle-ground alternative, although the advantage of PPO is that its clipped objective minimises unsafe updates and relaxation convergence times. PPO with balanced reliability, flexibility, and computability produces a dynamic real-time decision-making appropriate in the case of SDG-11 urban mobility applications.

CONCLUSION

This paper shows that the framework of policy-based reinforcement learning (RL) can be used to design sustainable and intelligent urban mobility systems that can follow the principles of sustainable development that has the Sustainable Development Goal 11. Comparative algorithms like REINFORCE, Actor-Critic, and PPO, and TRPO, it can be stated that policy-based algorithms, especially PPO/actor-critic show high stability, sample efficiency, and adaptability to complex, high-dimensional, and stochastic urban settings. It is essential to meet this set of capabilities in applications such as adaptive traffic signal control, fleet dispatch, and mitigation of congestions through real-time decision-making. Policy-based RL can use intelligent automation and powerful policy optimization to enable scalable, data-driven policy that can simultaneously minimize

emissions, maximize transport accessibility, and other overall urban efficiency. This confirms the purpose of reinforcement learning as both a technological instrument and a strategic tool to enable long-lasting sustainable urban development. This framework finally progresses the synergy between AI and sustainability by proving its capabilities of transformation to reach the resilient, inclusive and smart city ready future.

CONTRIBUTIONS TO KNOWLEDGE

The research presents a few important contributions to the areas of reinforcing learning (RL) and its use in achieving SDG 11 in the context of sustainable urban mobility. To begin with, it strengthens the insight into the performance of multiple policy-based RL algorithms, including REINFORCE, Actor-Critic, PPO, and TRPO, in complex, high dimensional and dynamic environment connected with sustainability. The researchers fill an important gap between theoretical RL implementations and practical SDG development by observing these algorithms back into a simulated environment using performance measures such as expected return, sample efficiency and convergence stability. Second, it recommends an organised mechanism of implementing the policy-based techniques on intelligent urban mobility systems, thus providing feasible solutions in the application of AI-based solutions in dealing with urban issues such as congestion, emissions, and accessibility. Finally, the paper gives great recommendations to policymakers, urban designers, and even AI designers as they compare algorithmic trade-offs, thus, making informed decisions regarding the deployment of socially responsible and situational-relevant AI tools with the goals of sustainable development.

REFERENCE

1. Brown, S., Sinha, S., & Schaefer, A. (2024). Markov Decision Process Design: A Framework for Integrating Strategic and Operational Decisions. Operations Research Letters, 54, 107090. Retrieved from https://doi.org/10.1016/j.orl.2024.107090.
2. Gheorghe, C., & Soica, A. (2025). Revolutionizing Urban Mobility: A Systematic Review of AI, IoT, and Predictive Analytics in Adaptive Traffic Control Systems for Road Networks. Electronics, 14(4), 719. Retrieved from https://doi.org/10.3390/electronics14040719.
3. Han, D., Mulyana, B., Stankovic, V., & Cheng, S. (2023). A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation. Sensors, 23(7), 3762. doi: 10.3390/s23073762.
4. Kumar, H., Koppel, A., & Ribeiro, A. (2019). On the Sample Complexity of Actor-Critic Method for Reinforcement Learning with Function Approximation. arXiv preprint arXiv:1910.08412.
5. Liu, Y.-t., Yang, J.-m., Chen, L., Guo, T., & Jiang, Y. (2020). Overview of Reinforcement Learning Based on Value and Policy. In 2020 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) (pp. 598–603). Washington, DC, USA: IEEE. Retrieved from https://doi.org/10.1109/CCDC49329.2020.9164615.
6. Michailidis, P., Michailidis, I., Lazaridis, C., & Kosmatopoulos, E. (2025). Traffic Signal Control via Reinforcement Learning: A Review on Applications and Innovations. Infrastructures, 10, 114. Retrieved from https://doi.org/10.3390/infrastructures10050114.
7. Mulibana, L., & Toit, J. (2023). Transport Planning Research Toward Implementing SDG 11 in South Africa. In J. Toit, L. Mulibana, & B. Ciobanu (Eds.), Handbook of Sustainable Transportation (pp. 1048–1073). Springer Netherlands. Retrieved from https://doi.org/10.1007/978-3-030-91260-4_69-1.
8. Naeem, M., Rizvi, S., & Coronato, A. (2020). A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. IEEE Access, 8, 209320–209344. Retrieved from https://doi.org/10.1109/ACCESS.2020.3038605.
9. Raihan, P., Biswas, M., & Islam, M. (2024). Research on Urban Traffic Management Evaluation-Taking Dhaka City as an Example. Electronic Journal of Transnational Architectural Studies, 2(3). Retrieved from https://doi.org/10.59324/ejtas.2024.2(3).xx.
10. Samuel, C., Adorni, G., & Gambardella, L. (2023). Proximal Policy Optimization-Based Reinforcement Learning and Hybrid Approaches to Explore the Cross Array Task Optimal Solution. Machine Learning and Knowledge Extraction, 5(4), 1660–1679. Retrieved from https://doi.org/10.3390/make5040082.
11. Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust Region Policy Optimization. In Proceedings of the International Conference on Machine Learning (pp. 1889–1897). JMLR.org. Retrieved from https://doi.org/10.5281/zenodo.1156385.

12. Sewak, M. (2019). Policy-Based Reinforcement Learning Approaches: Stochastic Policy Gradient and the REINFORCE Algorithm. In M. Samim, A. Singh, & S. Sumathi (Eds.), Reinforcement Learning Approaches in Traffic Management Systems (pp. 175–209). Springer Singapore. Retrieved from https://doi.org/10.1007/978-981-13-8285-7_10.
13. Skoropad, V. N., Deđanski, S., Pantović, V., Injac, Z., Vujičić, S., Jovanović-Milenković, M., Jevtić, B., Lukić-Vujadinović, V., Vidojević, D., & Bodolo, I. (2025). Dynamic Traffic Flow Optimization Using Reinforcement Learning and Predictive Analytics: A Sustainable Approach to Improving Urban Mobility in the City of Belgrade. Sustainability, 17(8), 3383.
14. Srinivasan, A. (2023). Reinforcement Learning: Advancements, Limitations, and Real-world Applications. International Journal of Scientific Research in Engineering and Management, 7, 1–8. Retrieved from https://doi.org/10.55041/IJSREM25118.
15. Stavrev, S., & Ginchev, D. (2024). Reinforcement Learning Techniques in Optimizing Energy Systems. Electronics, 13(8), 1459. Retrieved from https://doi.org/10.3390/electronics13081459.
16. Tran, D. Q., & Bae, S.-H. (2020). Proximal Policy Optimization Through a Deep Reinforcement Learning Framework for Multiple Autonomous Vehicles at a Non-Signalized Intersection. Applied Sciences, 10(16), 5722. Retrieved from https://doi.org/10.3390/app10165722.