**Comments to author (Associate Editor)**

In this article, the authors propose an evaluation and comparison of reward functions for the RL applied to the urban traffic control problem.

In accordance with the reviewers’ comments, the topic is interesting, however, the structure and content of the work are weak under various aspects.

The main issues can be summarized as follows:

- the contribution of the work should be clearly highlighted and technically described;

- a detailed discussion of the state of the art and the contribution with respect to the article [27] should be provided;

- a clearer presentation of the problem statement and method should be ensured.

**Reviewer 1:**

This paper is concerned with assessing multiple reward functions for the Q-Learning solution of a traffic signal control problem.

The paper is well-written and the simulations are profound, however motivating the importance of the paper based on assessing different reward functions does not justify the work, specially the original work (refereed to [27]) is not published yet.

* What are the main differences between this work and the one proposed in Ref [27]?
  + This one includes pedestrians. Very few other papers do. The previous paper referred to single-objective optimisation. Here 2 different control tasks need to be balanced since they are mutually exclusive.
* Why the authors did not compare to the previous results obtained prior to this work as referred to in the abstract?
  + They cannot be compared, as the previous work only dealt with a vehicular intersection.
* The authors are encouraged to add an algorithm about implementing the Deep Q-Network.
  + Included.
* It is noted that the learning rate α is very small. How this would fit in the attempted high dimensional space?
  + The learning rate is adequate. Assuming one action every 9s on average, this provides 200 transitions/episode\*1500 episodes = 3x10E6 transitions. Given that the size of a batch for SGD is 32, this provides 9375 SGD steps, which is enough for model convergence under the current learning rate.
* The authors need to parameterize the different segments (variables) of each state of the Q-Learning process.
  + No. Q learning requires parametrisation and segmentation of variables, DQN does not, which is why it was selected.
* Is it possible to couple the traffic in each two-directions to simplify the structure of the states and hence customize the measurements and optimization process?
  + This would defeat the purpose of the experiment as this is a real intersection and hence the current limitations imposed by the transport authority are enforced. These limitations cannot be changed at will by the researchers without losing the real-world focus of the problem.
* The authors need to fix the citations notations mentioned in Section II (e.g. [9]-[18])
  + Fixed, providing more detail on their contributions and their relation to this paper.

Finally: I would prefer to structure the paper around the Q-Learning solution process and use the two best rewarding functions (If they were not actually used in the original work) to highlight the contribution of the work.

* + The Q-Learning solution process has been previously reviewed in the literature, this paper does not concern itself about the feasibility of providing a viable solution, but instead explores what state variables provide the best quality of solution. One of the major open points revolves around the performance of the different reward functions based on available data. This is the problem this paper addresses.

**Reviewer 2:**

This work aims at evaluating various reward functions for an off-policy model-free value-based Reinforcement Learning algorithm, particularly the Q-Learning, in the case of a

traffic signal control problem.

The topic of the article is actual, however, the work presents various major weaknesses.

The abstract only vaguely presents the contribution of the article that mainly consists in the comparison of reward functions of a reinforcement learning method. It is not clear if the authors are presenting a novel methodology for the computation of the rewards or if the article is only an

evaluation paper. However, if this is the second case, the authors should review various techniques and provide a comparison of the results.

* + Abstract, introduction and literature review reworked.

The state of the art discussion is particularly reduced and does not provide clear evidence of the contribution of this work with respect to the existing literature, also a detailed discussion of the advances of this work with respect to the previous work by some of the authors [27] is not provided.

* + Removed references to [27] from introduction. Framed as its own problem.

The problem definition only provides details on reinforcement learning without highlighting the issues related to the application of RL or Q-learning to the urban traffic control, the methodology proposed to solve the problem, and the possible advantages and outcomes.

The methods section can be improved by providing an overview of the algorithm that the authors are going to present.

* + Included

The simulation section is particularly consistent and sufficiently detailed.

**Changelist:**

* Abstract reworked, removed reference to previous work.
* Introduction reworked for stress on pedestrian side, removed citation of previous work.
* Reworked literature review section. More stress on what each bit of literature covers, different approaches, and lack of pedestrians and comparisons in literature.