Reviewer #1: This study proposes a deep reinforcement learning-based car-following model for autonomous vehicles. Overall, this study is straightforward, and the topic is interesting.  
Regarding traffic flow modeling and learning-based models in TR-B, **the reviewer would like to see more novel methodology parts (e.g., models, properties, and algorithms)**. For example, Yuan, et al. (2021) proposed a new macroscopic traffic flow model with physics models and Gaussian process (GP). New theorems are provided to show that the proposed shadow-GP is learnable and the feasibility of employing the physical knowledge regularization. Ahamed, et al. (2021) proposed a DQN-based approach for urban delivery. New heuristics are provided to guide specific actions. The computational complexities of the heuristics and the DQN training are analytically investigated. Zhou, et al. (2019) presented a robust car-following control strategy under uncertainty for connected and automated vehicles (CAVs). To analyze the stability of the proposed control system, they mathematically proved a sufficient and necessary condition for local stability and sufficient conditions for robust string  
stability.  
However, the **methodology part in this paper is the common knowledge** that has been used in existing studies. The authors revised the reward function and traffic environment of the free-driving policy based on a widely used actor-critic network framework, and then used the existing Deep Deterministic Policy Gradient (DDPG) Algorithm for training. There are no mathematical analyses to investigate the **feasibility**, **stability**, and **computational complexity** of the proposed model and algorithm. Therefore, the reviewer suggests the authors make **serious revisions to justify the methodological contributions** or submit this manuscript to a more application oriented journal .

Dear Reviewer,

the main critique addresses the missing methodology parts in this study. However, from our perspective we see two main methodological contributions:

1. To our knowledge, all existing car-following models that have been trained with Reinforcement Learning (RL) use (i) real driving data, e.g. NGSIM, (ii) standard driving cycles, e.g. the Standard European Driving Cycle or (iii) even constant speeds, to model the leading vehicle’s trajectories in the training phase. However, such ‘usual’ driving data lack in situations that are safety critical, such as full-braking of the leader vehicle. This leads to car-following models that are suitable for this special kind of training data, but they cannot be generalized for different kind of car-following scenarios that are not reflected in training. All analyzed RL based car-following models are validated on the same leading trajectory data that has been also used in training, so that the known problem of machine learning generalization cannot be evaluated. There is just one study that validates the performance of the trained model in scenarios that are not in the scope of the training data and found that in this case, there is a significant degradation of the DRL performance (Lin et al., 2020). The main problem of all analyzed RL based car-following models in the literature is, that through the choice of standardized or real driving data serving as leading trajectory in the training phase, the training itself covers just a small part of the distribution of possible car-following states. This leads to the known problem of robust out-of-distribution (OOD) generalization of the trained models (Träuble et al., 2021).

Based on these insights, our first methodological contribution is a RL based car-following model that is trained based on stochastic processes serving as leading trajectories with the aim to cover a bigger part of the distribution of possible car-following states, especially safety-critical states, such as the follower reacting on high decelerations of the leader vehicle. To test the trained model’s robustness of OOD generalization, we validate the car-following model in different scenarios that it has never been seen in training. This includes experimental car-following scenarios recorded in urban and peripheral arterials in Napoli. Moreover, we bring the trained model to its limits by synthetic scenarios where we manipulate the leader vehicle to accelerate and decelerate extremely, including a sudden full-braking maneuver to a standstill. Since there is no analytical way of showing the robustness of machine learning models in general, the best one can do is demonstrating string stability as well as a plausible and crash free behaviour in extreme situations that the model has never seen before.

1. Our second methodological contribution is a modularized RL architecture to combine different driving objectives. To our knowledge, there exist no complete RL-based model reflecting not only car-following situations but also free-driving and a transition between the two. In this study, we consider both states in order to define a model whose application is more general. Moreover, instead of combining the objective of free-driving (keeping to speed limit) and the objective of car-following (driving safely, efficiently and comfortably) in a single unified RL architecture, we show that, by decomposing the system into a modularized RL architecture, we avoid the known problem of multi-objective RL where too many different objectives have to be considered simultaneously (Liu et al., 2014). Still, instead of using an architecture that explicitly separates between free-driving and car-following (for example, by using a fixed gap threshold), we propose an approach that distinguishes between both states implicitly. Particularly, both policies work simultaneously and both compute an acceleration value each time step.

Based on this approach and by considering a comfortable approaching of the leader vehicle in the reward function of the car-following policy, we prove realistic and comfortable vehicle behavior for the transition from free-driving into approaching a standing vehicle in the validation scenarios.

In the following we refer to the other comments to improve this study.

1. The free-driving policy only considers the acceleration part. However, there is another scenario in free-driving traffic that vehicles need to decelerate when the speed limit drops. The authors should also **consider a deceleration part in the free-driving policy**.

TODO: insert additional reward function part [e.g., Treiber’s proposition S1] and train model again

1. In the car-following policy, the authors said "when gt exceeds gmax or there is no leader, g is set to gmax". If there is no leader, it seems the vehicle is in a free-driving scenario. Then, this should not be a car-following policy.

As stated in the beginning, our modularized RL system does not really distinguish between free- driving and car-following scenarios. Both policies always compute an acceleration value, and just the minimum of both accelerations is the control input for the vehicle. When the vehicle is far away from a leading vehicle (g\_t tends to infinity), the output of the car-following policy should be a\_max to ensure that the vehicle is accelerating until the free-driving policy reacts with lower accelerations to the nearing to the desired speed v\_des. But as we don’t want to train the car-following policy for gap values between zero and infinity, we replace the actual gap g\_t with g\_max for states where we are definitely not in a car-following state anymore, resulting in an constant car-following policy’s output of a\_t = a\_max in free-driving states.

1. The authors used several datasets in this study. One dataset is field data, and the others are simulated data. Why did the authors use these datasets? It seems the field dataset is enough. **Why did the authors use simulated datasets?** **How did the authors use these datasets for training and validation?**

We hope that we could clarify the choice of different datasets for training and validation in the first part of our answer. [please make the numbering unique, for example m1, m2, 1,2, ... (the m stands for methodological issue). In this system, we will refer to m1]

1. The authors compared the proposed model with the IDM. But the IDM is used for human-driven vehicles, not for autonomous vehicles. The authors should **compare the proposed model with existing RL-based models.**

The IDM [not “IDM model” since the M in IDM already stands for “model”] does not consider reaction times and human perception failures. Therefore the IDM can indeed be seen as an ACC model. Furthermore, our RL model uses the same input and output as the IDM. Therefore, from our perspective, the IDM is a suitable model to compare the performance of our RL model with.  
Furthermore we did not consider other RL based car-following models in the validation because our model is the only one that considers free-driving and the transition to car-following as well as a bigger range of possible accelerations in the conceptualization of the RL system. As, to our knowledge, all other RL based car-following models don’t consider safety-critical scenarios, a comparison of those models with ours is not suitable.

1. The **organization** is terrible. For example, table 1 and figure 2 show the results of the car-following policy before it is mentioned.

Werde ich noch umstrukturieren.

1. In the last paragraph on page 2, the authors mentioned some RL-based AV car-following models. However, one study **(Zhou, et al., 2017)** is a supervised learning-based model for human-driven vehicle trajectory prediction. **It should not be considered here**. The authors should carefully review these studies.

Aus meiner Sicht hat dieses Paper schon seine Daseinsberechtigung. Es geht in diesem Abschnitt der Introduction nicht um RL based car-following models, sondern vielmehr um data-driven approaches using Deep Learning methods -> supervised learning [the confusion arises because both model classes are called „DRL“. One could clarify this in the manuscript on p 3 top (not yet included): „To overcome this issue, \*other\* DRL methods do not try to emulate the human driver but directly optimize metrics such as safety, efficiency, and comfort.“]

1. There is a **wrong formulation in equation** (8). The reviewer believes that "d" is supposed to be "Δt". Please carefully proofread this paper.

We thank the reviewer for spotting this typo and have carefully proofread the revised paper

1. In section 5.5, the authors said, "**cross-validation with the IDM**". Cross-validation is a specific term. It would be better to use "comparison".

We thank the reviewer for pointing at the incorrect usage of the technical term "cross validation" and have revised the wording as suggested

**References**

Yuan, Y., Zhang, Z., Yang, X. T., & Zhe, S. (2021). Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation. Transportation Research Part B: Methodological, 146, 88-110.

Ahamed, T., Zou, B., Farazi, N. P., & Tulabandhula, T. (2021). Deep Reinforcement Learning for Crowdsourced Urban Delivery. Transportation Research Part B: Methodological, 152, 227-257.

Zhou, Y., & Ahn, S. (2019). Robust local and string stability for a decentralized car-following control strategy for connected automated vehicles. Transportation Research Part B: Methodological, 125, 175-196.

**Additional References**

Lin, Y., McPhee, J., Azad, N.L., 2020. Comparison of deep reinforcement learning and model predictive control for adaptive cruise control. IEEE Transactions on Intelligent Vehicles 6, 221-231.

Liu, C., Xu, X., & Hu, D. (2014). Multiobjective reinforcement learning: A comprehensive overview. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(3), 385-398.

Träuble, F., Dittadi, A., Wuthrich, M., Widmaier, F., Gehler, P. V., Winther, O., ... & Bauer, S. (2021). Representation Learning for Out-of-distribution Generalization in Reinforcement Learning. In ICML 2021 Workshop on Unsupervised Reinforcement Learning.