Advanced Reinforcement Learning-Based Thermal Management Strategy For Battery Electric Vehicles

Tomislav Bukic´*[1], Amr Mousa[2], Milan Živadinovic´[3], Bernhard Peischl[3], Harisyam Manda[3], Wolfgang Koenig[3], Armin Traussnig[3], Charles F. Gaylard[4]

[1]*AVL-AST d.o.o., Croatia*
[2]*Virtual Vehicle Research GmbH, Austria*
[3]*AVL List GmbH, Austria*
[4]*University of Bath, Claverton, Down, UK*

DOI: https://doi.org/10.46720/FWC2023-SDV-001

ABSTRACT: The vehicle control systems domain encounters an increasing number of parameters and calibration targets considering the emerging technologies such as connected vehicles and automated driving. Accordingly, the calibration processes for such systems have become more complex and thus error prone and tedious. Moreover, the derived control policies are not easily transferable between different vehicle configurations, hence, the calibration effort is increasing dramatically with each configuration change. Therefore, the reduction of such efforts needed to setup the control policy is inevitable to further reduce cost of drivetrain development. The proposed methodology therefore is an important means to make BEVs (Battery Electric Vehicles) more attractive for car buyers.

The fast development of the Artificial Intelligence (AI) domain is opening the door to numerous opportunities and applications in the automotive industry. We utilize Reinforcement Learning (RL) techniques to design appropriate control strategies for different vehicle systems, thus improving the conventional approaches and reducing the development effort. Combining the expertise in simulation and big data, we propose a cloud-based solution that runs a high-fidelity simulation to train, test and deploy the thermal management control strategy for a fleet of BEVs. The benefits, among others, are that the RL-based control policies can be designed more rapidly and run more efficient than the traditional rule-based approaches. After deploying the initial model trained against the simulation, we have the capability of collecting data from a fleet of vehicles operating with the latest control strategy. Using collected data, we iteratively train and customize the strategy throughout the operation time.

We have tested the above-mentioned framework on the use case of cabin heating mode selection for BEVs. Our RL agents are trained and evaluated in a model-in-the-loop simulation environment. The policy evaluation is based on the agents' performance on representative vehicle test measurements (drive cycles). The metrics are selected to quantify the energy efficiency and comfort individually, as well as aggregated to enable a fair comparison. Notably the trained agents achieved better results than the original control policy on most of the individual metrics and significantly better results on the aggregated metric.

At this moment, our framework is tested against simulated vehicle fleet. The first reasonable research question is if the trained control system can be directly transferred to the real vehicle, or whether additional adjustments must be performed to achieve the needed flexibility. Another open question refers to the adequate combination of RL algorithms to achieve even better performance on telemetry data from a connected fleet. Time will tell if the idealized case with continuing on-policy training, or the more complex case using the policy-agnostic offline algorithms, will provide the stronger solution.

The main technical contribution of our work is the use case agnostic framework which iteratively improves a conventional rule-based control strategy. Following the automotive V-model, the design-, implementation-, and testing -phase is strictly separated from the in-use phase of a vehicle function. To leverage historic data from the in-use phase, our framework disrupts this classical model and embeds the DevOps and ML-Ops practices into the automotive engineering process.

In this article we suggest a framework which automates the complex and time-consuming creation process of control strategies. The trained control policies provide better results and allow for a continuous improvement after they are finally deployed on a fleet of vehicles.

KEY WORDS: reinforcement learning, adaptive control strategies, thermal management, artificial intelligence, battery electric vehicles

1 Introduction

1.1 Trends in industry

The automotive domain in general is a key industrial sector for Europe (1) by securing 13.3 million jobs, producing 20% of the vehicles worldwide (out of 99 million vehicles produced yearly worldwide), and generating a yearly trade balance over =C99 billion. At the same time, the automotive market is impacting different major societal challenges such as reduction of pollutant emissions (2) , reduction of traffic fatalities (3) , increased mobility for an ageing population, or reducing congestion. Parallel to that, the habits of the consumers are evolving, and new needs are emerging such as infotainment and connectivity, human-machine interaction, and customization, as well as mobility as a service. Nowadays, the automotive sector is confronted by four main trends:

- Electrification (4), with the introduction of e-mobility (hy-

brid, pure electric vehicle) to optimize or even completely remove the internal combustion engine, finally reducing the resulting local pollutant emissions during vehicle operation.

- ADAS and autonomous driving functions (5) , with the purpose of providing more comprehensive information to the driver for better context awareness, up to taking over specific driving maneuvers – finally reducing the demands on the driver and lowering number and impact of possible accidents
- Connected vehicles (6; 7) enabling optimization of vehicle's operation or the emergence of new services while relying on external information, e.g., from other vehicles or from the infrastructure
- Diverse mobility (8) targeting the efficient movement of people and goods with respect to different factors such that time, energy consumption, ecological footprint.

Reinforcement learning (RL) has emerged as a powerful tool in control engineering for optimizing the performance of battery electric vehicles (BEVs). As the automotive industry shifts towards electrification, there is a growing need for efficient and intelligent control strategies to enhance the range, charging behavior, and overall operation of BEVs. RL, with its ability to learn from interactions with the environment, offers a promising approach to address these challenges.

Battery electric vehicles rely on complex control systems to manage various components, such as the electric motor, battery pack, power electronics, and regenerative braking systems. Traditional control techniques often rely on fixed models or rule-based strategies, which may not fully adapt to the dynamic and uncertain nature of real-world driving conditions.

Reinforcement learning provides a solution to this limitation by enabling BEV control systems to learn and optimize their behavior through trial and error. By interacting with the environment, which includes factors such as traffic, road conditions, and energy consumption, RL algorithms can learn optimal control policies that maximize specific objectives, such as range, energy and thermal efficiency, or driver comfort.

However, there are challenges associated with the application of RL to control engineering for BEVs. RL algorithms typically require extensive computational resources and substantial amounts of training data, which can be challenging to obtain in real-world driving scenarios. Additionally, safety considerations and the need to validate and certify RL-based control systems pose significant hurdles.

2 Thermal management

In today's rapidly evolving automotive industry, one of the critical areas of focus is vehicle thermal management. As vehicles become more advanced and complex, with the integration of electric powertrains and sophisticated electronics, managing the thermal dynamics within a vehicle becomes increasingly challenging. The efficient orchestration of heat generation, dissipation, and distribution is crucial for optimal vehicle performance, passenger comfort and safe vehicle operation.

The thermal management system in a vehicle encompasses various components and processes that control the temperature of critical systems such as the e-motor, battery, and passenger cabin. However, several challenges arise in ensuring effective thermal management in modern vehicles, necessitating innovative solutions to address them.

Figure 1 shows the thermal system architecture of a battery electric vehicle. A heat pump system can transfer the heat from one location to the another by utilizing the refrigerant circuit. It can use the ambient air surrounding the vehicle as a heat source. In this mode, the heat pump extracts heat from the outside air and transfers it into the vehicle cabin for heating purposes. However, the efficiency of the heat pump may be affected in cold weather conditions when the outside air temperature is significantly lower. In that case the heat pump can recover waste heat generated by various vehicle powertrain elements. For example, heat can be extracted from the vehicle's e-motor or the battery. This approach optimizes the utilization of the vehicle's internal heat sources and enhances energy efficiency.

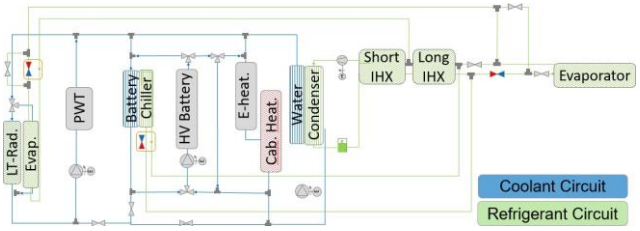

Figure 1: Thermal system architecture

Different modes in the vehicle thermal management system ensure the best performance, passenger comfort and safe vehicle operation by using the correct heat sources and sinks. These modes allow the system to adapt to various conditions and requirements, providing the right heat source under different scenarios. The challenge is to choose the optimal mode in terms of efficiency but also to guarantee the passenger comfort and safe operation for different boundary conditions. For this use case we have created the following modes:

Mode 1 self-heating PWT and battery to operating temperature. Cabin is heated using air heat pump or electrical heater

Mode 2 heating the cabin with PWT or ambient air heat pump.

Mode 3 heating cabin from battery

Mode 4 heating cabin and PWT from battery

Mode 5 heating cabin using combined heat of PWT and battery

Mode 6 heating battery and cabin from ambient air heat pump or electrical heater

Mode 7 heating cabin and battery from PWT or electrical heater

Mode 8 cooling cabin, PWT and battery

To determine the thermal mode that achieves the desired performance objectives a simplified plant model of the vehicle was generated in MATLAB/Simulink. The vehicle model includes the most important subsystems that are related to the cooling performance, cabin comfort and energy consumption. The model was built and calibrated to match the system behaviour of the reference vehicle.

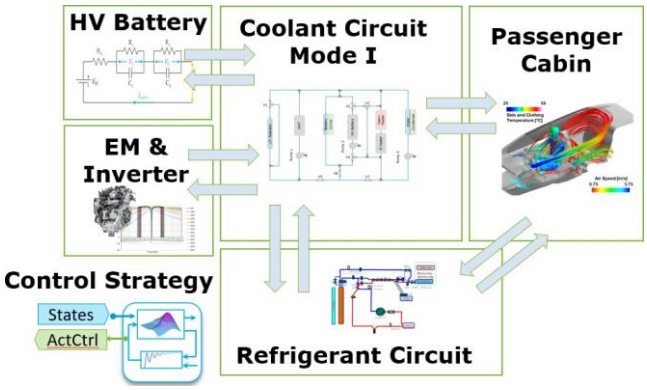

Figure 2: Vehicle model

The objective of this project is to evaluate the benefits of using a RL agent compared to a conventional control strategy in the context of electric vehicle thermal management. The goal is to optimize the thermal mode selection to ensure cabin comfort, securing battery safety limits and enhance energy efficiency. To compare both approaches, first a conventional control strategy for thermal management was implemented in the vehicle model. This strategy relies on predefined rules and fixed algorithms based on threshold values and predefined operating conditions.

3 Reinforcement learning in Automotive industry

Reinforcement Learning has emerged as a key component in the optimization of thermal management systems in Battery Electric Vehicles. The inherent value of RL lies in the adaptive nature and capacity to optimize processes in real-time. RL works by training an *agent* to make decisions within a specific *environment* to maximize a defined *reward function*. In the context of thermal management systems in BEVs, the *agent* represents the thermal control system, and the *environment* includes everything else except the agent. The *reward function* is designed to prioritize battery life, performance, and energy efficiency.

The integration of RL in thermal management systems of BEVs enables more adaptive, efficient, and intelligent control of battery temperature, thereby improving vehicle performance, enhancing battery lifespan, and maximizing energy efficiency. However, challenges such as defining optimal reward functions and managing the complex state-action space in real-world driving scenarios persist. As RL continues to evolve and as computational capabilities and data availability increase, it will unquestionably continue to influence the future of thermal management systems in BEVs.

3.1 Literature

Huang et al. (9) employed double deep Q network (double DQN) algorithm to optimize the battery thermal effects on energy efficiency. The proposed method, based on a gated recurrent unit (GRU) for state feature extraction, outperformed the existing fuzzy control, achieving over a 6.7% energy reduction during aggressive driving across various tested cycles.

Other scholars such as Chen (10) applied Policy Gradient Reinforcement Learning (PGRL) approaches to vehicle climate control, utilizing a multilayer perceptron-based neural network with a softmax output layer, to maximize occupant comfort while maintaining reasonable energy consumption. Compared to other SARSA-based RL approaches, the implementation of the PGRL approach, particularly Proximal Policy Optimization (PPO), resulted in a significantly improved occupant comfort (from 67% to 92.3%) and a faster learning time (reduced to 0.63 years), with potential for further enhancements through additional policy gradient techniques and a more realistic human thermal comfort model.

Brusey et al. (11) presented an energy-efficient vehicle climate control strategy, formulated as a Markov Decision Process (MDP) and solved using the Sarsa(λ) algorithm, providing a significant improvement over traditional climate control methods. Compared to their best performing controller, this approach reduces energy consumption by 13% while increasing the time passengers spend in thermal comfort by 23%, signifying its promise for substantial comfort and energy improvements in electric vehicles.

4 Methodology

4.1 Mathematical formulation

Reinforcement learning considers the world in which an agent acts every discrete time step. Agent's environment is modeled by the set of states S, actions A, and a reward function $r : S \times A \times S \rightarrow \mathrm{R}$. World is modeled stochastically with Markov decision process. This means that the transition to the next state s' conditionally depends only on the current state of the world s and the last chosen action a. Formally, in step k transition probability is

$$P(s_{k+1}, r_{k+1} | a_{k+1}, s_k) = P(s_{k+1}, r_{k+1} | a_{k+1}, s_k, a_k, \ldots, s_0) \quad (1)$$

where s_0 is a starting state, a_k action to which the agent has committed and r_{k+1}.

A sequence of all consequential states, actions, and rewards defines the episode. Our goal is to find the decision-making policy $\pi : S \rightarrow A$ which will maximize the expectation of the total discounted sum of rewards collected through

$$\mathrm{R} = \sum_{k=0}^{\infty} \gamma^k r(s_k, a_k, s_{k-1}) \quad (2)$$

In the formula 2 $\gamma \in [0, 1]$ is called a forgetting factor. It models how much the near future will be prioritized over the uncertain more distant future in the agent's decision-making.

The agent is not necessarily aware of the complete world's state s. The part of the world observable to the agent is called an observation space. The world of our agent is modeled by the dynamics of the car during its ride. In our case actions are discrete cabin heating modes, and states are all possible readouts from the current step of the simulation:

$$s_k = \begin{bmatrix} T_{Cabin\_Air}, & T_{Cabin\_Target}, & T_{Bat}, & T_{Amb}, \\ T_{PWT}, & T_{CH}, & T_{Chlr}, & T_{Co}, \\ T_{CoB}, & T_{HC}, & T_{Rad}, & T_{WCDS} \\ v_{Veh} & Pwr_{Bat} & Eff_{Bat} & Eff_{PWT} \\ & \cdots & & \end{bmatrix} \quad (3)$$

$$a_k \in \{1, \ldots, 8\}$$

The optimized function - agent's policy - is modeled by the neural network. To reduce the input space of the neural network, observation space is chosen as a subset of state space as follows:

$$\mathbf{Obs}_k = \begin{bmatrix} T_{Cabin\_Air}, & T_{Cabin\_Target}, & T_{Bat}, & T_{Amb}, \\ T_{PWT}, & T_{CH}, & T_{Chlr}, & T_{Co}, \\ T_{CoB}, & T_{HC}, & T_{Rad}, & T_{WCDS} \end{bmatrix} \quad (4)$$

The reward function was crafted to include multiple objectives. The first objective was minimizing the absolute difference between the target and the actual cabin temperature, T_{Dev}. The second was minimizing the signed change in temperature ΔSoC. Last two terms P_{CH}, power consumed by the chiller, and P_{Comp}, power consumed by the compressor. The rest of the reward function was imposing soft and hard constraints to T_{Bat}, T_{PWT}, and SoC as follows

$$r(s_k, a_k, s_{k-1}) = \frac{T_{Dev}}{31} + \frac{\Delta SoC + 0.00856}{0.016} + \frac{P_{CH}}{9000} + \frac{P_{Comp}}{5500} - Constr(T_{Bat}) - Constr(T_{PWT}) - Constr(SoC) \quad (5)$$

where constraints are defined with equations

$$Constr(T_{Bat}) = \begin{cases} \frac{-10 - T_{Bat}}{20}, & T_{Bat} \in <-30, -10> \\ 0, & T_{Bat} \in [-10, 50> \\ \frac{T_{Bat} - 50}{10}, & T_{Bat} \in [50, 60> \\ terminate, & T_{Bat} \notin <-30, 60> \end{cases} \quad (6)$$

$$Constr(T_{PWT}) = \begin{cases} \frac{-10 - T_{PWT}}{20}, & T_{PWT} \in <-30, -10> \\ 0, & T_{PWT} \in [-10, 50> \\ \frac{T_{PWT} - 50}{20}, & T_{PWT} \in [50, 70> \\ terminate, & T_{PWT} \notin <-30, 70> \end{cases} \quad (7)$$

$$Constr(SoC) = \begin{cases} 0, & SoC \in <0, 1> \\ terminate, & SoC \notin <0, 1> \end{cases} \quad (8)$$

Normalization and centralization constants in Equation 5 are computed from the empirical mean and standard deviation from multiple runs of the simulated environment. Constants in constraint definitions 6 and 7 are defined by the simulation team to keep the vehicle in the safe working conditions, while the constraint 8 is used to stop the episode when SoC is drained .

4.2 Environment

All experiments in this paper were executed in our framework which is built using Python language (12). Plant model provided by the simulation team was compiled to FMU and wrapped in Gymnasium (13) to provide a standardised Python interface for reinforcement learning training. Training was done using algorithms from RLlib (14) library. There are two main parts of our training process (Figure 3)

- Training in simulation
- Training using real world data

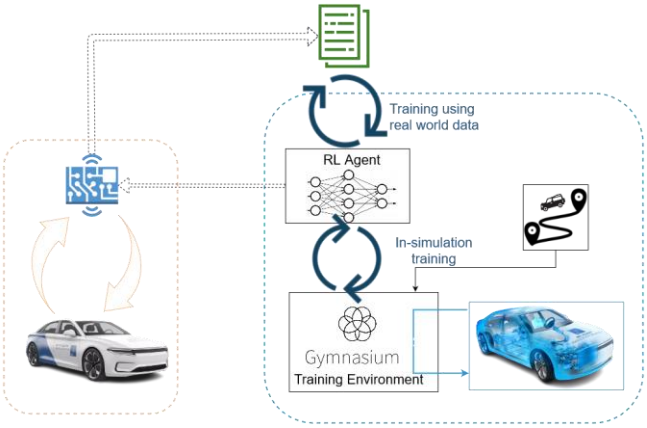

Figure 3: Training system architecture

which should show our capability to create automated initial strategy and improve on it using real world data. Currently, while we haven't deployed the algorithm to the vehicle, we are simulating real world data using new drive cycles and varying parameters.

Our framework enables us to automate of in-simulation training of thermal control strategies. Moreover, it allows periodic fine-tuning of control strategies after initial strategies are deployed to the vehicle. This solution introduces MLOps and DevOps practises in automotive engineering process enabling both the time saving during the initial creation process of control strategies. It also allows for a continuous improvement of the initial control strategies.

5 Experiments

In the in-simulation training with online algorithms approach, we investigate several state-of-the-art algorithms such as DQN, PPO and A3C. After careful hyperparameter tuning, the best performing agent was the PPO 0c6c9 with the hyperparameters in Table 1 and Table 2.

Parameter	Value	Parameter	Value
Agent stepsize	60s	$T_{Ambient}$	$-7°C$
SoC_0	90%	T_{Cabin_Target}	$23°C$

Table 1: Environmental parameters for in-simulation training with online algorithms

Parameter	Value	Parameter	Value
fcnet_hiddens	[256, 256]	Entropy_Coeff	0.01
lr	0.001	train_batch_size	5000
γ	0.99	sgd_minibatch_size	500
λ	0.95	num_sgd_iter	32
KL_coeff	0.03	vf_share_layers	True

Table 2: Training parameters for in-simulation training with online algorithms

On the other hand, the resultant PPO agent (0c6c9) was deployed to 985 different drive cycles with different initial ambient temperatures and SoC levels. A total of 3940 test cycles generated 76305 training instances for learning from data which replicates real world trajectories. The PPO agent's neural network was used to initialize the offline learning agents utilizing the CRR and MARWIL algorithms and the training was completed with hyperparameters in Table 3 and Table 4.

Parameter	Value	Parameter	Value
Agent stepsize	60s	$T_{Ambient}$	$[-7, 5]°C$
SoC_0	$[40, 90]\%$	T_{Cabin_Target}	$23°C$

Table 3: Environmental parameters for learning from data which replicates real world trajectories

Parameter	Value	Parameter	Value
fcnet_hiddens	[256, 256]	VF_coeff	1.0
lr	0.0001	train_batch_size	10
γ	0.999	num_workers	5
β	1.0	envs_per_worker	3

Table 4: Training parameters for learning from data which replicates real world trajectories

6 Results

The agents trained on a driving cycle called DC1-Return provided by the simulation team as shown in Figure 4. It has a total driving time of 860 seconds, a cumulative driving distance of 11.347 km, a max and minimum road slope of 7.65 and −6.86 respectively.

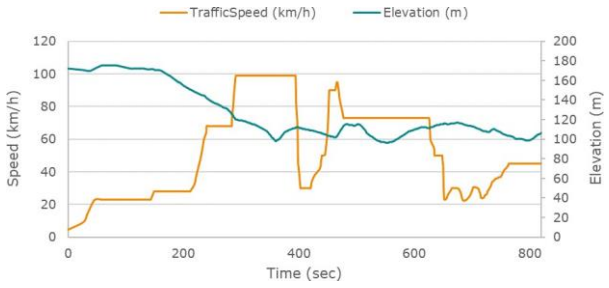

Figure 4: Performance of agents trained on DC1-Return

The best performing agent was the MARWIL 69081 where it outperformed the CRR agent. Accordingly, the focus of the comparison will be to benchmark the rule-based strategy against the PPO-0c6c9 agent and the MARWIL-69081 agent. The testing was conducted on a 24 test cases incorporating six drive cycles with two initial SoC levels (90% and 60%) and two ambient temperatures ($-7^{\circ}C$ and $10^{\circ}C$). The following three cases are in focus to discuss and highlight some key findings:

6.1 Short cycle testing with cold start

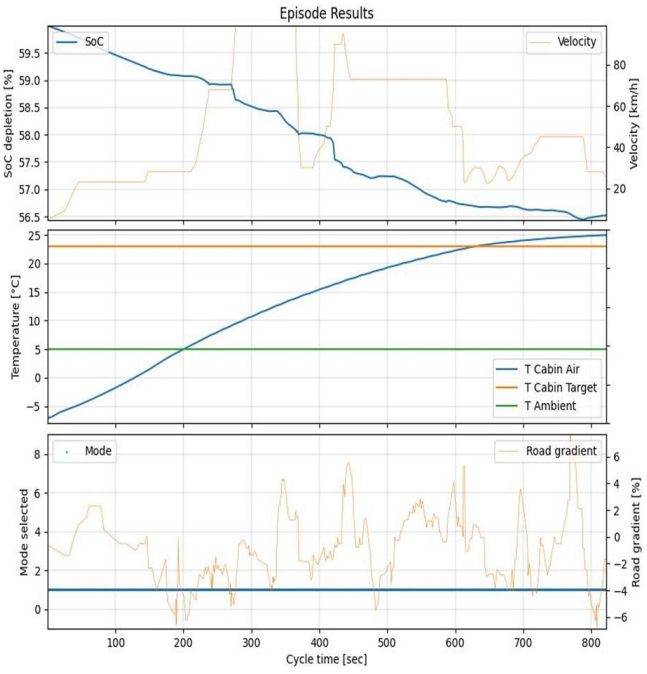

Figure 5: Results of **rule based agent** (the best in this test case!) on short cycle testing with cold start

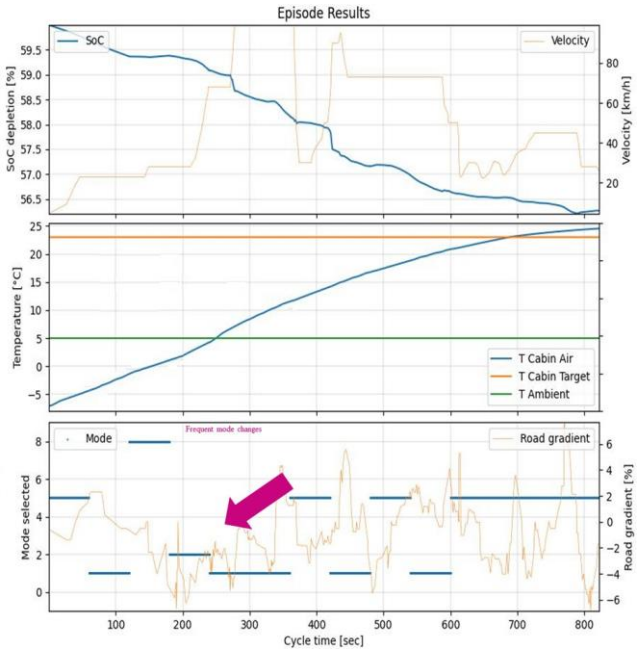

Figure 6: Results of PPO 0c6c9 on short cycle testing with cold start

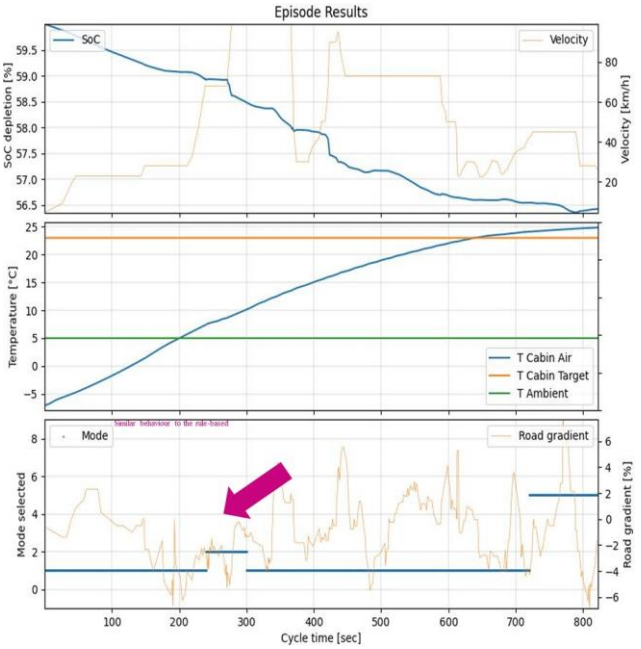

Figure 7: Results of MARWIL 69081 on short cycle testing with cold start

Figure 5, Figure 6 and Figure 7 compare the performance of the three agents while the results are summarized in the table Table 5.

Decription	**Rule based**	PPO	MARWILL
ΔSoC/100km	**30.66**	32.87	31.46
Rise time $/^{\circ}C$	**19.71**	21.54	19.96

Table 5: Test results for short cycle testing with cold start (drive cycle is DC-1-Return, $SoC_0 = 60\%$, $Temp_{init} = -7^{\circ}C$)

The rule-based strategy outperformed both PPO and MARWIL agents in the ΔSoC/100km and the Rise time $/^{\circ}C$. MARWIL agent

was the second closest to the rule-based strategy and showed similar behavior in the mode selections as shown in figure X while the PPO agent followed a different approach that incorporated frequent mode changes.

6.2 Long cycle testing with warm start

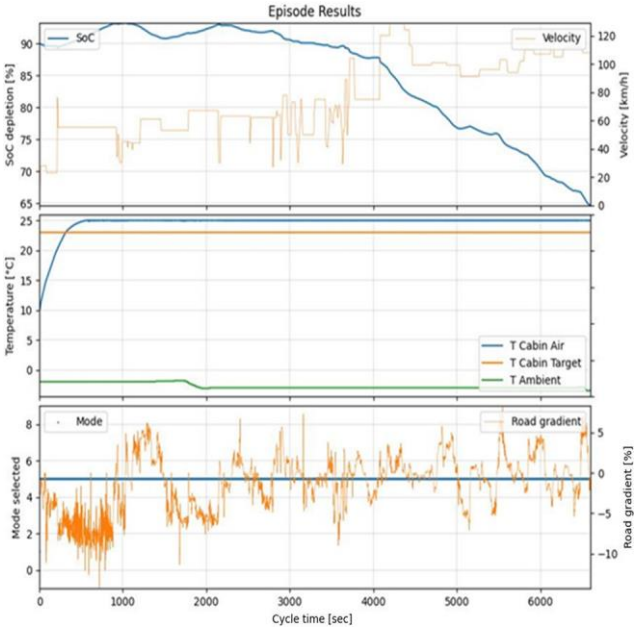

Figure 8: Results of rule based agent on long cycle testing with warm start

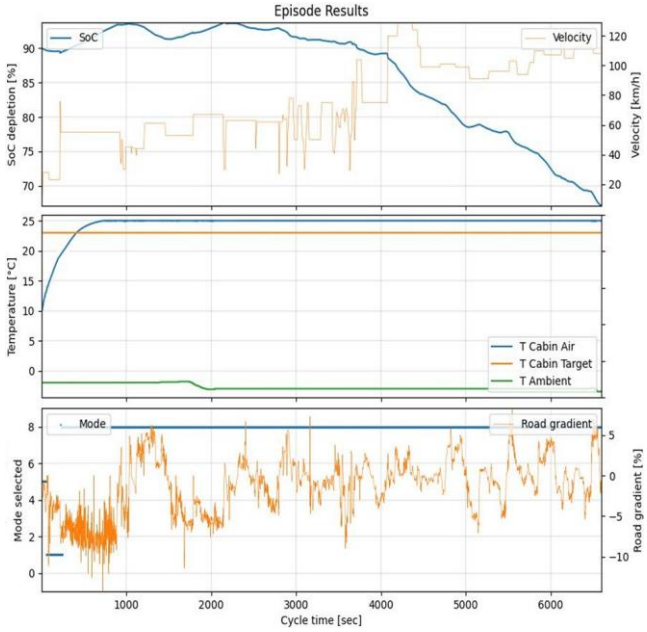

Figure 9: Results of **PPO 0c6c9** (the best in this test case!) on long cycle testing with warm start

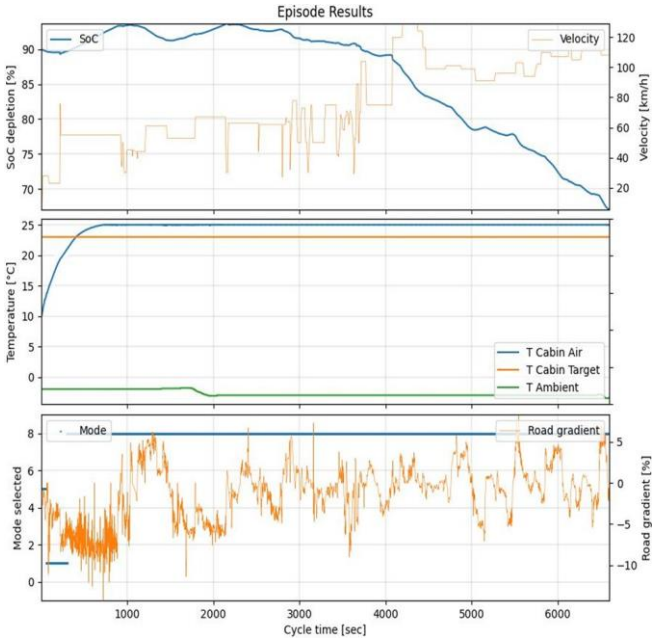

Figure 10: Results of MARWIL 69081 on long cycle testing with warm start

Decription	Rule based	**PPO**	MARWILL
ΔSoC/100km	17.76	**16.05**	16.11
Rise time $/^{\circ}C$	**21.24**	24.24	26.7

Table 6: Test results for long cycle testing with warm start (drive cycle is DC-2-Return, $SoC_0 = 90\%$, $Temp_{init} = 10^{\circ}C$)

The results of the three agents are shown in Figure 8, Figure 9 and Figure 10 and summarized in Table 6. It is noticeable that the MARWIL and PPO agents performed similarly and improved the energy efficiency by 9.33% to 9.64% over the rule-based approach respectively. The PPO agent achieved the best ΔSoC/100km of 16.05 and the MARWIL was very close by only 0.06% difference. Although, the rule-based approach achieved the best rise time/°C of 21.24 sec and the second-best agent was the MARWIL by an increase of 5.46 sec.

6.3 Long cycle testing with cold start

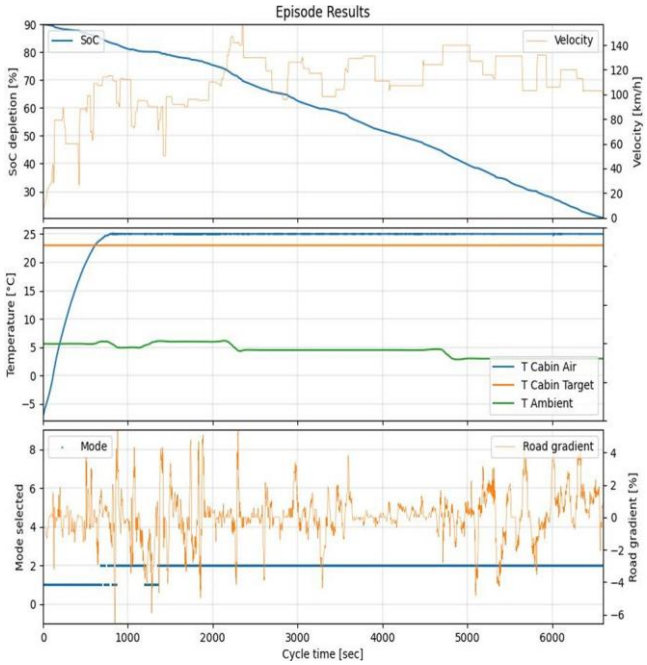

Figure 11: Results of rule based agent on long cycle testing with cold start

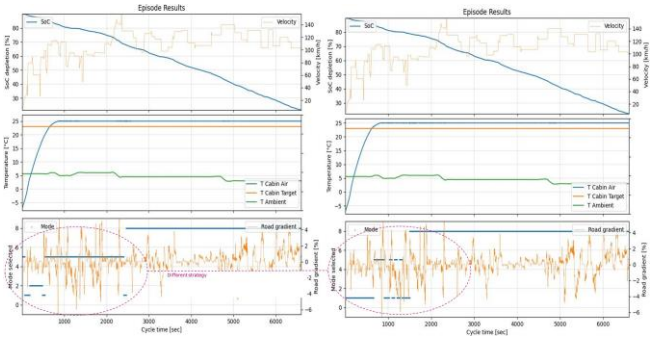

Figure 12: Results of PPO 0c6c9 and **MARWIL 69081** (the best in this test case!) on long cycle testing with cold start

Decription	Rule based	PPO	**MARWILL**
ΔSoC/100km	36.84	34.38	**33.97**
Rise time $/^{\circ}C$	19.13	20.35	**19.12**

Table 7: Test results for long cycle testing with cold start (drive cycle is DC-2, $SoC_0 = 90\%$, $Temp_{init} = -7^{\circ}C$)

The graph depicted in figure X and the data presented in table X demonstrate notable differences in the performance of three AI agents, with the MARWIL agent taking the lead by enhancing energy efficiency by 7.8% when compared to the conventional rule-based approach. The metric ΔSoC/100km stood at its best at 33.97, while the rise time$/^{\circ}$ was 19.12 seconds. A difference in the mode selection was observed between the PPO agent and the MARWIL agent, with the latter being initialized by the parameters of the PPO network. This observation substantiates our initial hypothesis that integrating continuous learning with offline algorithms can boost the efficiency of online algorithms that have been trained in simulated environments.

Overall, the MARWIL agent was able to reduce the energy consumption by 15.07% in comparison with the conventional rule-based approach and proved that it is able to modify and adapt itself to the changes in the environment compared to the PPO agent which was developed in a simulation environment with a single training cycle. On the other hand, the MARWIL agent achieved 32.23 seconds to increase the cabin temperature by one degree Celsius compared to 30.01 seconds of the rule-based approach. As an example, utilizing an RL agent will heat-up the cabin in 10min42s instead of 10min 00sec by using the rule-based approach. On the other hand, the range of the vehicle will be extended by 15.07%. Therefore, it was concluded that such a drawback of having 2.22 secs increase in Rise time $/^{\circ}C$ is acceptable for the sake of energy savings.

6.4 Comparison to the optimal solution

In parallel to the reinforcement learning solution, a dynamic programming solution was developed. Dynamic programming solution approximates the ideal solution on particular test settings. Initial comparisons show that our solution is comparable to the dynamic programming solution as well in terms of ΔSoC/100km. Since solutions differ in parameters and ways they prioritize energy efficiency and comfort metrics, additional effort has to be done to make both solutions fully comparable.

7 Outlook

In this paper, we explored the capabilities of reinforcement learning in optimizing control problem for automotive thermal management. Learning from data which replicates real world trajectories showed an improved strategy which reveals the potential of learning from real world data and adapting the policy trained in simulation.

Both in-simulation training and agents trained from data which replicates real world trajectories surpassed the rule-based controller with significant savings in energy consumption and minimal sacrifice in the rise time.

Framework we have created in the process of preparing this paper allows us to reduce engineering effort for calibration new vehicle models by replacing manual creation of control policy with crafting the reward function. Thus, after building the plant model dedicated engineering work of creating the strategy becomes significantly simpler work of setting up the cost function for automatic optimization.

As we conclude this research, several directions for future exploration emerge. Firstly, we will finalize the research which compares our results with the dynamic programming solution that is also developed for this problem.

Secondly, we intend to enhance our experiments by incorporating the impact of vehicle/component aging on charging/discharging processes, such as the influence of battery age. Incorporating this aspect into the control framework will enhance the system's ability to adapt to changing conditions and optimize performance accordingly.

Another avenue for future research involves parametrizing the vehicle to simulate different vehicle models. By exploring the impact of various vehicle characteristics on the control system's performance, we can gain valuable insights into the adaptability and robustness of the RL agents across different vehicle types and configurations.

Furthermore, we propose a driver-customized control strategy that takes into account individual driving styles, ambient conditions, and street infrastructure. By tailoring the control system to the driver's preferences and the specific environment, we can further enhance the overall driving experience and optimize energy consumption.

Additionally, incorporating Vehicle-to-Everything (V2X) information and vehicle horizon data into the control framework holds great potential. By leveraging these external data sources, RL agents

can make more informed decisions, anticipate future events, and optimize control strategies accordingly.

Lastly, an intriguing direction for future research is controlling the actuators directly instead of relying solely on high-level mode selection. By directly manipulating the actuators, RL agents can fine-tune control actions with greater precision, leading to further improvements in energy efficiency and overall system performance.

In conclusion, this research has demonstrated the potential of RL techniques in optimizing complex control problems for automotive applications. Both agents trained in-simulation with online algorithms and agents trained from data which replicates real world trajectories have showcased superior performance compared to rule-based controllers, achieving significant energy savings without compromising rise time. The identified future research directions will undoubtedly contribute to advancing the field of automotive control systems and pave the way for more efficient, adaptive, and customized driving experiences.

References

[1] ACEA. The automobile industry pocket guide 2020 - 2021. 2020.

[2] European Commission. Communication from the commission to the european parliament and the council the paris protocol – a blueprint for tackling global climate change beyond 2020. 2020.

[3] European Commission. Mobility and transport, road safety in the european union: Trends, statistics and main challenges. 2015.

[4] ETIP SNET, ERTRAC, EPoSS. European roadmap electrification of road transport. 2017.

[5] ERTRAC. Automated driving roadmap. 2017.

[6] CCAM Partnership. CCAM strategic research and innovation agenda. 2017.

[7] McKinsey & Company. Monetizing car data - new service business opportunities to create new customer benefits. 2017.

[8] ERTRAC. Urban mobility roadmap. 2017.

[9] Gan Huang, Ping Zhao and Guanglin Zhang. Real-time battery thermal management for electric vehicles based on deep reinforcement learning. *IEEE Internet of Things Journal*, 9(15):14060-14072, 2022.

[10] Gaobo Chen. *Policy gradient reinforcement learning-based vehicle thermal comfort control*. Doctoral Thesis, Coventry University, 2021.

[11] James Brusey et all, Diana Hintea, Elena Gaura and Neil Beloe. Reinforcement learning-based thermal comfort control for vehicle cabins. *Mechatronics*, 50:413-421, 2018.

[12] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[13] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[14] Eric Liang, Richard Liaw, Philipp Moritz, Robert Nishihara, Roy Fox, Ken Goldberg, Joseph E. Gonzalez, Michael I. Jordan and Ion Stoica. Rllib: Abstractions for distributed reinforcement learning, 2017.

Abbreviations

- **A3C** - *Asynchronous Advantage Actor Critic*
- **ADAS** - *Advanced driver-assistance system*
- **BEV** - *Battery Electric Vehicle*
- **CCR** - *Capacity-Cost Ratio*
- **Chlr** - *Chiller*
- **DQN** - *Deep Q Network*
- **FMU** - *Functional Mockup Unit*
- **GRU** - *Gated Recurrent Unit*
- **MARWILL** - *Monotonic Advantage Re-Weighted Imitation Learning*
- **PGRL** - *Policy Gradient Reinorcement Learning*
- **PPO** - *Proximal Policy Optimization*
- **PWT** - *Powertrain*
- **Rad** - *Radiator*
- **RL** - *Reinforcement learning*
- **V2X** - *Vehicle-to-everything*

Acknowledgment

This paper is the result of collaborative efforts from two projects. The first project was conducted in AVL List GmbH and has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 877056. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Austria, Germany, Finland, Switzerland. The second project has been conducted in Virtual Vehicle Research GmbH in Graz, Austria. It has been financially supported by the COMET K2 Competence Centers for Excellent Technologies from the Austrian Federal Ministry for Climate Action (BMK), the Austrian Federal Ministry for Digital and Economic Affairs (BMDW), the Province of Styria (Dept. 12) and the Styrian Business Promotion Agency (SFG) while the Austrian Research Promotion Agency (FFG) has been authorized for the program management.