DOI: 10.13009/EUCASS2023-639
Aerospace Europe Conference 2023 – 10TH EUCASS – 9TH CEAS

Experimental and Simulative Evaluation of a Reinforcement Learning Based Cold Gas Thrust Chamber Pressure Controller

Till Hörger[⋆†]*, Lukas Werling*[⋆]*, Kai Dresia*[⋆]*, Günther Waxenegger-Wilfing*[⋆] *and Stefan Schlechtriem*[⋆]
[⋆]*DLR German Aerospace Center*
Im Langen Grund, 74239 Lampoldshausen. Germany.
till.hoerger@dlr.de · lukas.werling@dlr.de · kai.dresia@dlr.de · guenther.waxenegger@dlr.de · stefan.schlechtriem@dlr.de
[†]Corresponding author

Abstract

At DLR neural networks, as potential future controller for rocket engines, are studied. A neural network-based chamber pressure controller for a simplified cold gas thruster is presented and analyzed in simulation and experiment. The goal of the controller is twofold: It can track a trajectory with different changes of setpoints and it allows to set and control a wide variety of steady state chamber pressures. The neural network gets feeding line pressure measurement data as input and calculates valve positions as output values. The training phase of the controller is done with a reinforcement learning algorithm in an Ecosim-Pro/ESPSS simulation, that is validated with data from the corresponding experimental set up. To increase the robustness and to allow a transfer from the simulation directly to the test facility domain randomization is applied. The controller is evaluated in simulations and experiment. It was found that - in the range of physically possible operation points - the controller achieves a constantly high reward which corresponds to a low error and a good control performance. In the simulation the controller was able to adjust all required set points with a steady state error of less than 0.1 bar while retaining a small overshoot and an optimal settling time. It is found that the controller is also able to regulate all desired set points in the real experiment. A reference trajectory with different steps, linear and sinus changes in target pressure is tested in simulation and experiment. The controller was in both cases able to successfully follow the given trajectory.

1. Introduction and Motivation

Rocket engine thrust control is essential for many applications that are currently of high interest, for example propulsive landing for launchers as well as planetary or lunar landing systems. In Europe various projects are developing vertical take of and landing technology [1, 2, 3, 4]. This applications have high demands on the engine control system. Precise thrust control, restart capabilities and deep throttling are needed to assure soft landing. Moreover, the potential reuse of engines requires optimal transients e.g. with low thermal loads to avoid damage and adaptive control systems to handle the degradation of components over time.

Closed loop control of transients in rocket engines, like start-up of the engine, is challenging, du to strong non-linear behavior. Classic control approaches like PI and PID are used for control of steady state operating points or minor set point changes [5]. More recently, model-based methods like linear-quadratic regulators (LQR) and model predictive control (MPC) as well as the inclusion of machine learning techniques are investigated for rocket engine closed loop transient control [6, 7, 8].

The benefits of an intelligent engine control system used for control reconfiguration and condition monitoring, have already been investigated in the space shuttle era [9, 10]. In 2019 German Aerospace Center (DLR) started to investigate machine learning methods for supporting the design and operation of liquid rocket engines and thrusters [11]. Control systems based on machine learning methods promise several advantages for rocket propulsion systems. For example the fatigue life at a given thrust level could be expanded with an appropriate damage model. Wall temperatures at specific combustion pressures can be kept under a certain level, or the efficiency at a particular operation point or during throttling of an engine can be increased. The controller can also deal with pre-trained failure models as well as react intelligently to changes in the system due to anomalies.

One approach to design closed-loop optimal control laws is deep reinforcement learning (DRL), a sub-field of machine learning [12]. DRL is applied successfully to areas like robot control [13, 14], drone flight control [15]

DOI: 10.13009/EUCASS2023-639

and autonomous driving[16]. One advantage of DRL is that the controller can be derived directly from nonlinear high sophisticated simulation models, in which gradient information is potentially not accessible to the user [8]. No derivation of state-space models, model order reduction or linearization is needed as it is required for PID, LQR and MPC controllers. Furthermore, once the controller is trained, the response time is very short [17]. The utilization of deep neural networks as controller offers the possibility to design highly non-linear multiple-input multiple-output controllers, that incorporate all kind of side conditions while delivering a optimal control performance. Nevertheless, since the stability of these algorithms is not mathematically proven until now, the stability and robustness of the controller has to be carefully evaluated before use. One possibility is to evaluate the performance and stability of the control algorithm in multiple simulations and experiments.

In this work, to demonstrate the basic functionality of an artificial neural network trained by a reinforcement learning algorithm as rocket engine controller, a simplified cold gas system is utilized. This set up allows rapid, cheep and safe testing while offering the opportunity to conduct real world experiments.

A simulation of the thruster was set up in EcosimPro/ESPSS and validated with experimental data. A Python interface allows the use of standard implementations of reinforcement learning algorithms also at the test facility. Based on the simulation model a thrust chamber pressure controller for the named cold gas system is trained. The goal of the controller is to set different thrust chamber pressures at different feeding line pressures as well as to follow pressure trajectories. The steady state control deviation should be less than 0.1 bar. Further overshot of less than 0.5 bar and a optimal settling time is required. The paper discusses the control performance achieved by the neural network in simulation as well as at the test facility.

2. Experimental Set-Up

A suitable application for a reinforcement learning based controller is for example a 22 N thruster fueled with nitrous oxide and ethane. Such a system can be constructed either as premixed monopropellant- [18, 19] or bipropellant system. For the control task, the bipropellant system offers more degrees of freedom. Because of the high vapor pressure of fuel and oxidizer it is possible to operate the system in self pressurized mode. In a self-pressurized system the tank pressure decreases over burn-time, as the propellant evaporates and the enthalpy of vaporization cools the remaining propellant. Therefore, to achieve a constant thrust level, an active control system is needed. Nitrous oxide/ethane offer a green alternative to the widely used satellite propellant monomethylhydrazine and dinitrogen tetroxide [20]. The latter are attempted to be replace due to their toxic, carcinogenic and environmentally harmful properties. At DLR several encouraging green propellants are under investigation [21, 22, 23]. Nitrous oxide/ethane are comparably cheap, widely accessible and offer a high $I_{sp} \approx 300$ s [20].

2.1 Cold Gas Test Set-Up

Since the DRL-control method was never before demonstrated in the field of combustion chamber pressure control, before conducting hot fire tests with nitrous oxide/ethane, due to safety and efficiency reasons a reduced nitrogen cold gas system is used for the first tests. The P&ID of the test set up can be seen in Fig 1. Nitrogen is feed by 200 bar supply pressure. The nitrogen pressure is reduced to a maximum of 100 bar by a PID controlled automatic pressure regulator of type Tescom ER5000. The pressure controller allows to set different feeding line pressures. An automatic main valve is used to start and stop the gas flow. After the main valve a Rheonik coriolis mass flow meter is installed, followed by a proportional control valve of type m-tech MPG 03 PR. The nitrogen is then expanded through a chamber, initially designed for use with nitrous oxide and ethane. The goal of the controller is to set the chamber pressure to a desired value by changing the position of the control valve for varying feeding line pressures. All testing is conducted at DLR Lampoldshausen M11.5 test position [23, 24].

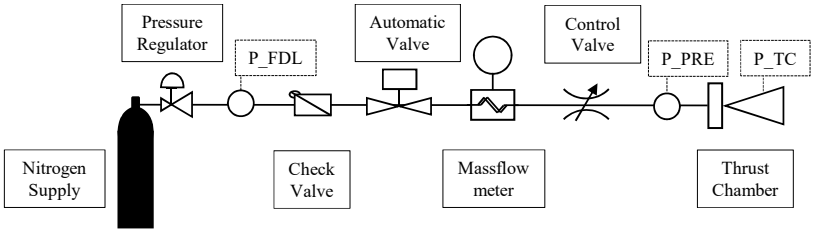


Figure 1: P&ID of the test set up

DOI: 10.13009/EUCASS2023-639

Three pressure sensors, *P_FDL*, *P_PRE* and *P_TC* (see Fig. 1), are used as input variables for the controller. Further, the current position of the control valve *CV_POS*, the desired chamber pressure *P_SOLL* and the error *ERR* (difference between the desired pressure and the measured *P_TC*) are used as input to the neural network. Based on these values the network calculates the action value. The output is defined as the relative change of the position of the control valve dCV_POS/dt. This value is sent to the test facility and changes the position of the control valve. The change in position of the control valve is limited to 5 % in each time step, due to low movement speed of the valve. The control frequency is 10 Hz. Since the control valves used for the experiments, have a dead time of about 300 ms and additionally an opening time of about 2 s to fully open, 10 Hz are sufficient for this setup. In a later set up, faster valves will be used. The control frequency is then planned to be higher. For comparison, the engine controller of the space shuttle main engine operated at 50 Hz [5]. Tab. 1 gives an overview about the input and output parameters of the controller. The test facility is controlled and operated by a python based user interface, that allows the simple inclusion of neural networks in the workflow [17].

Table 1: Input and output variables of the controller

Input (Observation Space)	**Output (Action Space)**	**Controlled variable**
P_FDL	dCV_POS/dt	*P_TC*
P_PRE		
P_TC		
ERR		
P_SOLL		
CV_POS		

2.2 Training of the Controller

The controller is essentially an artificial neural network (NN). The NN is used as function approximation to map pressure measurement data to valve positions. To act successfully as controller the internal weights of the NN have to be trained. For the training process reinforcement learning is used. Reinforcement Learning is a sub-field of machine learning (ML). Generally, in machine learning a large amount of training data is used to find solutions for complex problems. Three categories (supervised, unsupervised and reinforcement learning) can be distinguished in ML, depending on the quality of information give in the problem [25]. Supervised learning can be applied if the training data set contains the input and the desired output data. This can be useful to find the corresponding mapping rule between input and output, that can be used for example for image classification. Whereas in unsupervised learning, the target output data is unknown. The goal here is to discover structures, similarities or hidden patterns in the input. Such algorithms can be used e.g. for cluster analysis or to find similarities in data.

Reinforcement learning (RL) is different from supervised and unsupervised learning because no predefined training datasets is needed. A so-called agent learns self-employed through interaction with a simulation or real-world data. RL operates in discrete time steps. The underlying principle of RL is visualized in Fig. 2.

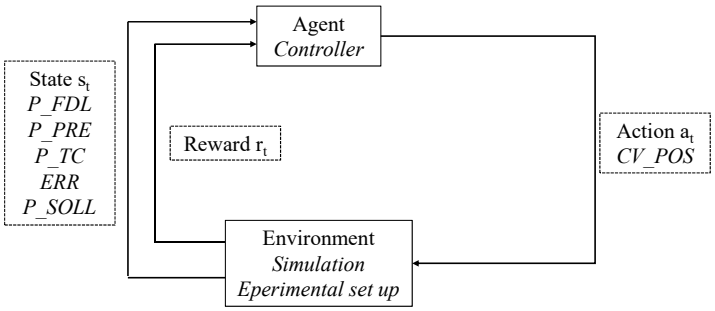


Figure 2: Reinforcement learning schematic based on [12]

An agent interacts in discrete time steps t with the environment. The environment has defined states s_t. In every time step the agent gets information about the system state from the environment. The environment information known to the agent is called observation space. The state is rated with a scalar value called reward r_t. The reinforcement learning algorithm optimizes a decision rule, called policy, in a way to maximize the reward accumulated over time. In deep RL the policy is represented by a deep neural network that acts as function approximation. It maps a state of the

DOI: 10.13009/EUCASS2023-639

system to an action a_t. The agent develops a strategy to maximize the expected cumulative reward $G(\tau) = \mathbb{E}\{\Sigma r_t\}$ that it gets. The reward is calculated by a user defined reward function that represents a rating of the state. The cumulative reward $G(\tau)$ is the sum of all single reward values in one episode, where the episode is one sequence of simulation. Through interaction with the environment the agent gets information, encoded in the reward, which action is best in which situation. To avoid the expected cumulative reward for long episodes growing to infinity, a discount factor $0 \leq \gamma \leq 1$ is introduced (See Eq. 1).

$$G(\tau) = \mathbb{E}\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t)\right\} \tag{1}$$

The value of γ defines how much future rewards come into account for the current action because γ is raised to the power of t. The whole process can be seen as a trial and error method with integrated feedback in the form of reward. Tab. 2 summarizes important expressions in RL.

Table 2: Key terms in reinforcement learning [12, 17]

Expression	Description
Agent	The algorithm or controller. It optimizes the policy.
Environment	System that interacts with the agent. In this case the rocket thruster
State	Physical state of the system
Reward	Scalar value that rates the state of the system. Calculated by a user defined reward function
Action	Control output of the agent. Sent to the environment
Observation Space	Variables that serve as input values and system description for the agent
Policy	Decision rule of the agent. A function that maps states to actions. It defines how the agent reacts to different system states. In deep RL the policy is an artificial neural network. During training the policy is optimized through the agent to find actions delivering the maximum reward

The reward function should return high values, if the controller is close to the desired thrust chamber pressure and low values if it is far away from the set point. Furthermore unnecessary valve movement and oscillations should be punished. The gradient of the reward function with respect to the pressure very close to the set point should be high enough to allow finding the terminal position. Designing a good reward function is a key challenge when applying reinforcement learning to a specific problem. For this application the approach in the following algorithm is chosen:

$r(t) = Rew1 + Rew2 + Rew3$

if $ERR < 2$ bar **then**
  $Rew1 = -\sqrt{ERR}$
else
  $Rew1 = -10$
end if

$Rew2 = -|valvespeed|$

if more than 3 time steps $ERR \leq 0.1$ bar **then**
  $rew3 = 10$
else
  $Rew3 = 0$
end if

The reward function $r(t)$ is subdivided in three parts $Rew1$, $Rew2$, $Rew3$, each of which evaluates a different aspect of the control goal. $Rew1 = -10$ if the deviation from the set point is 2 bar or higher. Apart from that $Rew1$ is equal to minus the square root of the control deviation. $Rew1$ is mainly important at the beginning of the training

DOI: 10.13009/EUCASS2023-639

phase, because large deviations are punished with a large negative value while the square root function offers a steeper gradient close the target pressure than a linear function. *Rew*2 penalizes valve motion and especially fast valve motion. This helps to avoid oscillations around the target pressure. However this may also increase the settling time. Finally, *Rew*3 supports finding and holding the target pressure. $Rew3 = 10$ only if the control deviation was less than 0.1 bar for more than three time steps in succession. The maximum possible reward in one time step is therefor $r(t) = 10$ and the undiscounted maximum cumulative reward of one episode is $G = 10 \cdot (n - 3)$ with n time steps in one episode.

In literature a wide variety of RL-algorithms is described. Different learning paradigms can be distinguished, depending on what exactly is optimized: In (Deep) Q-learning [26] the so called optimal action-value function is learned. Policy optimization algorithms directly learn the policy from policy gradient methods [27]. Actor critic algorithms try to combine the strengths of Q-learning and policy gradients [28, 29]. On-policy algorithms need completely new training samples after each policy update, while off-policy algorithms can learn from past samples using replay buffers.

The training of the controller described in this paper is done with the off-policy soft actor critic (SAC) algorithm [30] in the RAY Rllib implementation [31] (see Tab.3). 6 CPUs are used for training in parallel. The neural network representing the policy has two hidden layers with 256 neurons each. The activation function is ReLu. Compared to on-policy algorithms like the widely used PPO [32] or A3C [33], SAC is characterized by a comparatively high sample efficiency which leads to faster training success. Furthermore, in comparison to other off-policy algorithms like DDPG [29] or TD3 [28], the training process is stable and less hyper parameter tuning is needed. In SAC beside the cumulative reward, also the policy-entropy, as a measure of how random the agent acts, is maximized. In this way the task is completed successfully while acting as randomly as possible [30].The policy is encouraged to exploration while being robust concerning model and estimation errors [30]. However, several million time steps of training can be required to train a controller with reinforcement learning. During the training or exploration phase the controller intentionally may enter system states, that are possibly dangerous for the system. This is fined with a low reward and is part of a nominal training process where also low reward areas have to be observed by the algorithm. Therefore, to save cost and to avoid damage of the test facility or test specimen the training process takes place in a simulation environment.

Table 3: Relevant variables for the training of cold gas thrust chamber pressure controller

Variable	**Value or range**
P_FDL	$\{P\_FDL \in \mathbb{N} \mid P\_FDL \geq 60 \wedge P\_FDL \leq 90\}$ bar
P_SOLL	$(\{P\_SOLL \in \mathbb{N} \mid P\_SOLL \geq 50 \wedge P\_SOLL \leq P\_FDL\})/10$ bar
T_FDL	$\{T\_FDL \in \mathbb{N} \mid T\_FDL \geq 268 \wedge T\_FDL \leq 298\}$ K
Sensor noise	Gaussian distribution with μ = Sensor value and $\sigma = 1$. The random number is multiplied with 0.5 % of the sensor value
Simulated pip length	Randomized with normal distribution $[0.6, 1.4]$
Valve Speed	Randomized with normal distribution $[0.1, 0.5]$
Algorithm	SAC
Implementation	RAY Rllib 3.0.0.dev
Number CPU	6
Framework	torch
NN characteristics:	Activation function: ReLu
	Two hidden layers of size 256

EcosimPro ESPSS [34] is used as simulator. For training of the controller, a simulation model of the cold gas test setup up described in section 2.1 is created. The simulation model is validated with data from the experimental setup. The measured thrust chamber pressure can be reproduced in simulation with a maximum error of 3 % for different feeding line pressures and valve positions.

One main challenge when using reinforcement learning based controllers is the transfer from the simulation based training environment to the real world application. This is also called Sim-to-Real transfer [35]. Even carefully tuned simulations will always contain small deviations from the real system. The RL algorithm learns to make use of these simulation inaccuracies and the application in the real world may fail due to unforeseen input values. Domain

DOI: 10.13009/EUCASS2023-639

randomization can help to overcome the Sim-to-Real gap [36, 37]. Simulation parameters, like friction, delays, temperatures or pressure loss are varied during the training process. In this way the robustness of the policy concerning changes in the environment is increased, as the real environment becomes a subset of the randomized simulation. The intervals in which the parameters are randomized have to be chosen wisely. Large randomization intervals massively increases the training effort while narrowly defined randomization intervals may exclude the areas relevant for the real application. For the cold gas set up described in this paper, it was found that the speed of the control valves as well as calculating correct pressure losses in the system are the most challenging parts of the simulation. Hence valve speed and the length of the feeding lines in the simulation was used for randomization. The length of each simulated pipe in the feeding line is multiplied by a factor (interval see Tab. 3) randomly changing after each episode and therefore varying the pressure loss in the feeding lines. The valve speed is varied in the interval given in Tab 3. The randomization is applied in 5 levels, depending on the training progress. In the first level, no randomization is applied and the simulation is used in its baseline configuration. In the following levels, the interval for the two randomized variables is increased gradually. The condition for the rise in a new level, is to reach a threshold reward value in four consecutive episodes. The threshold reward was set to 2000.

During training one episode comprises 40 s simulation time. During simulation every 10 s the feeding line pressure P_FDL, feeding line temperature T_FDL as well as the desired thrust chamber pressure P_SOLL is changed randomly in the given limits with a equal distribution (compare Tab. 3). Input values are rounded to integers or one decimal respectively. Every 5 episodes the target pressure is following a constantly changing trajectory with different feeding line pressure respectively. This should enable the controller to set different set points as well as to follow a given trajectory. Sensor noise is added to all variables in the observation space with a Gaussian distribution with standard variation of $\sigma = 1$. The noise is multiplied with 0.5 % of the sensor value and added to the latter.

3. Simulative Analysis

To analyze the control behavior of the neural network after the training is completed, simulations are conducted. The NN has to control the thrust chamber pressure via interaction with the EcosimPro/ESPSS simulation model. 4800 simulations were conducted to systematically analyze the performance in the training area. Feeding line pressure was increased in 0.5 bar steps from 60 bar to 90 bar. Target pressure was simulated in a interval from 5 bar to 9 bar with steps of 0.05 bar each. Thereby the entire operating envelope is covered with a fine resolution. The evaluation simulations were carried out for a run time of 10 s starting always from 10 % valve opening. According to the definition of the reward function, the maximum possible reward value for 100 timesteps would be 10000. As the valve needs up to 2.3 s to open completely, a realistic value for a perfect run is in the order of 8000. The result can be seen in Fig. 3 and Fig. 4.

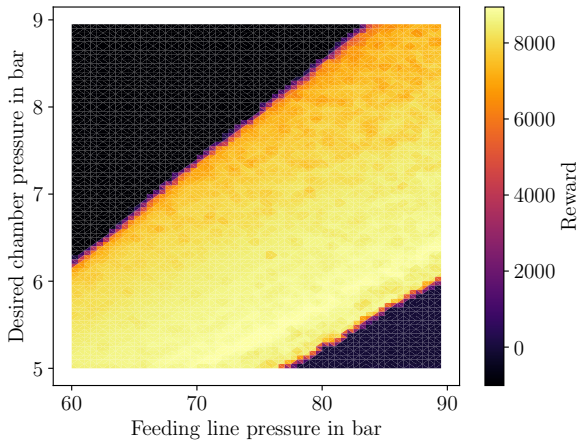


Figure 3: Reward achieved in simulation for different set points

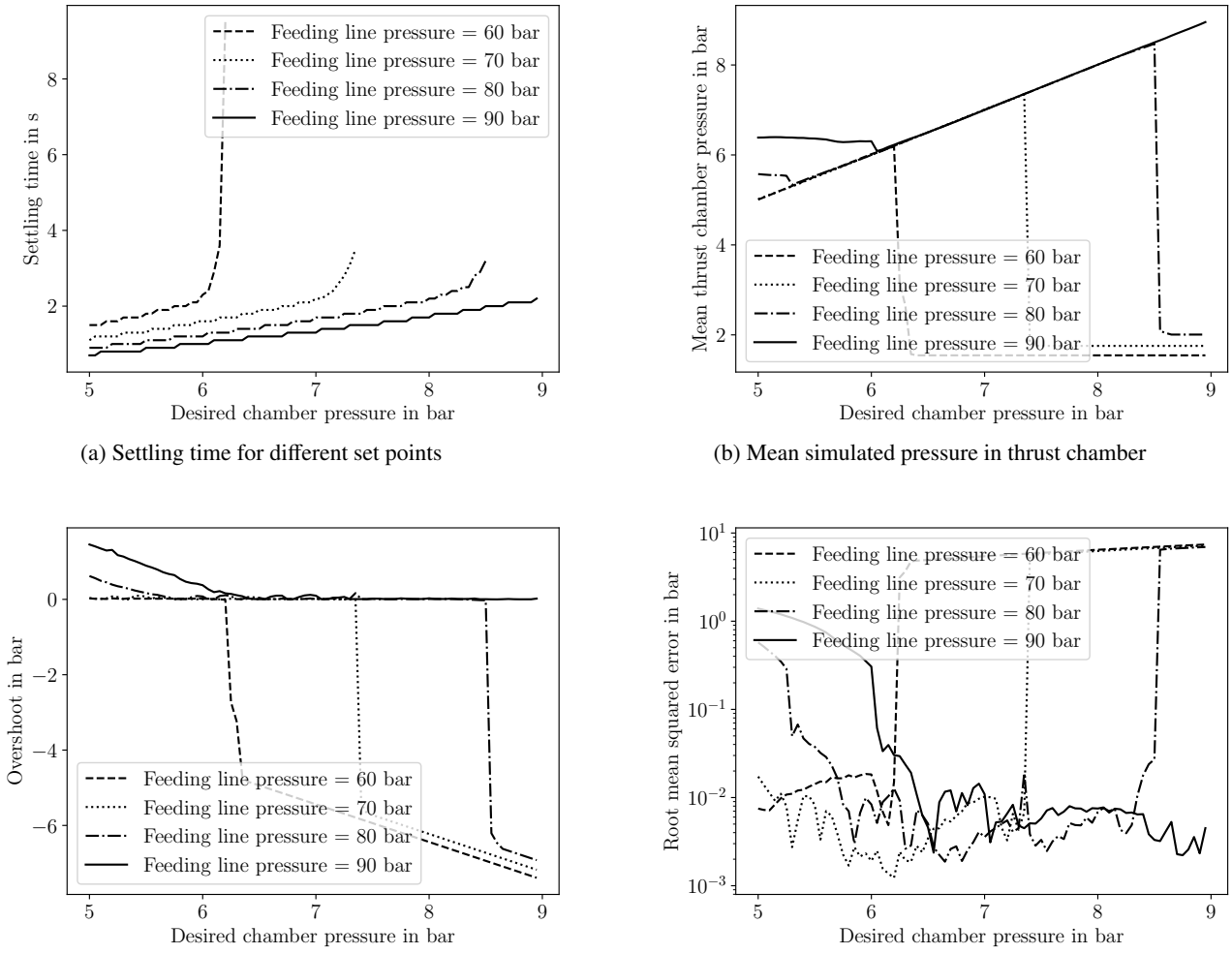
The areas marked with yellow color in Fig. 3 represent set points with a high reward achieved. This means the controller was able to set the chamber pressure to the correct target value. While in the black areas the reward is low and as a consequence the controller failed to set the desired pressure. Both areas on the top left and the lower right where the controller failed are tied to physical limitations of the test set up. For a given feeding line pressure due to pressure losses in the system, only a limited thrust chamber pressure can be achieved at completely opened control

DOI: 10.13009/EUCASS2023-639

valve. This is the case for the dark area in the top left side of Fig. 3. For example at 60 bar feeding line pressure, only 6 bar chamber pressure are possible. Therefor the higher operating points are out of scope. The limit in the lower right in Fig 3 can be lead back to a limitation of a minimum opening of the control valve to 10 %. High feeding line pressures would require control valve positions below 10 % to reach low target chamber pressures. The reason for introducing a limit at the lower boundary of the control valve was due to high inaccuracies if the control valve is operated at lower opening levels.

It can be seen that within the physical boundary the controller is able reach a constantly high reward and therefor small deviations at all tested set points in the chosen resolution between 60 bar and 90 bar feeding line pressure and 5 bar and 9 bar chamber pressure are accomplished.

Fig. 4 allows a more detailed analysis of the controller behavior.



(a) Settling time for different set points

(b) Mean simulated pressure in thrust chamber

(c) Overshoot for different set points

(d) Root mean squared error (RMSE)

Figure 4: Simulation analysis

Fig 4a shows the settling time in dependence of the desired chamber pressure at four different feeding line pressures. It can be seen that the settling time rises for higher target pressures, which is a result of the required higher opening of the control valve. Curves representing lower feeding line pressures resulting in higher settling times for the same desired pressure. This is because for the same chamber pressure at a lower feeding line pressure the control valve has to be opened further. At the upper right end of each line, the set point can no longer be reached du to limited feeding line pressure and the valve being already fully open. Therefore the settling time grows to infinity. The maximum settling time in the controllable area of about 2.4 s corresponds to the opening time of the valve, which is also 2.3 s. So the controller is able to set the target pressure in the minimum possible settling time. The linear progression of the curves in the controllable area is a result of the linear transfer function of the control valve.

Fig 4b represents the mean thrust chamber pressure in dependence of the desire chamber pressure. For the

DOI: 10.13009/EUCASS2023-639

controllable area this is a straight line, which indicates that the target pressure equals the simulated pressure set by the controller. For higher target pressures at a specific point each curve instantly drops to a lower value. The controller closes the control valve to the lower limitation (10 %) when it is no longer possible to reach the desired pressure. For higher feeding line pressures, the drop occurs at higher target pressures and results also in a higher remaining pressure in the thrust chamber. This behavior can be problematic. If the desired pressure is too high for the physical limits of the system, the highest possible pressure should be set and the valve should be completely opened. This behavior is probably attributed to the choice of the reward function. Since the achievable pressure is always far away from the desired pressure only negative reward with almost no gradient information is given, independent on the actual pressure. Therefore, for the controller it is beneficial to not move the control valve at all to avoid punishment for valve movement. This should be addressed via changing the reward function in future applications. For higher feeding line pressures (80 and 90 bar), as already described in Fig 3, the low target pressures can not be reached, as the minimum opening of 10 % results in higher thrust chamber pressures.

Fig 4c represents the overshoot in pressure during the settling time. It can be seen, that for the controllable area nearly no overshoot can be observed. The lines representing 80 bar and 90 bar feeding line pressure have a peak for low desired chamber pressure, which is a result of the inability to reach this pressures given the physical limits of the system.

Fig 4d shows the root mean squared error (RMSE) for the different set points in bar. The RMSE in the controllable area is in the order of 0.01 bar and therefor well within the target accuracy of 0.1 bar defined by the reward function. For areas outside the controllable region the root mean squared error rises.

All simulations shown in Fig. 4 and Fig.3 were run with the same baseline simulation set up. Domain randomization was used in training, but not for evaluation. In order to test the robustness of the controller for changing environmental behavior a Monte Carlo simulation with 5000 sample simulations is implemented. The result is shown in Fig. 5.

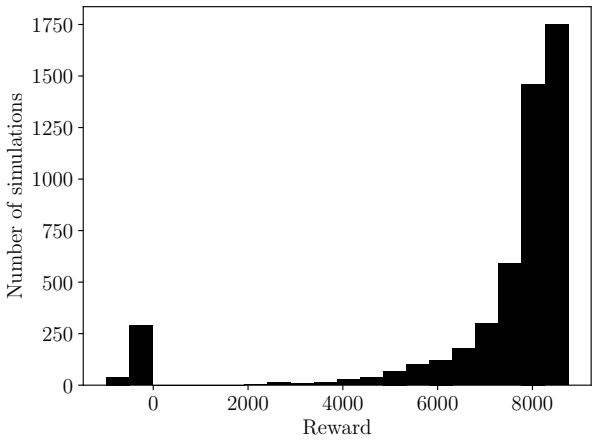


Figure 5: Result of Monte Carlo simulation

The following parameters are changed randomly with equal distribution during the MC simulation in the given ranges in Tab. 3: Feeding line pipe length, valve speed, feeding line temperature, feeding line pressure P_FDL, desired pressure P_SOLL, sensor noise. This analysis reveals how the controller behaves under changing environment-conditions, e.g. when being transferred to the real test facility. Sensor noise was added between 0.1 % and 1 % intensity, unlike given in Tab. 3. For the Mote Carlo simulation only operating points inside the controllable area (where the controller reached a high reward in the baseline simulation, compare Fig. 3) were chosen. The Monte Carlo analysis shows that the vast majority of 4388 simulations lead to a reward of more than 6000. In 1749 of the simulations the reward was higher than 8200. However in some of the simulations the controller failed to achieve the desired chamber pressure, as can be seen by a negative or very low positive reward. In total 331 simulations lead to a negative reward. It was found, that simulations resulting in a negative reward are on the boundary of the controllable area. It is assumed, that with changing for example the pressure loss or temperature a set point which was under baseline conditions inside the controllable area, moved outside and can no longer be controlled by the controller due to physical limits of the system.

As last step, the controller was tested following a predefined trajectory of the chamber pressure at constant feeding line pressure. The trajectory contains different step changes as well as sinus and linear change in the desired

DOI: 10.13009/EUCASS2023-639

chamber pressure. The simulation was conducted with a constant feeding line pressure of 80 bar. The result can be seen in Fig. 6. The target trajectory is marked with dashed line, while the simulated controlled chamber pressure is shown as solid line. The controller was able to follow the trajectory with a root mean squared error of 0.36 bar. The slow settling time for the larger set point changes is a result of the slow valve speed. Nearly no overshoot can be observed.

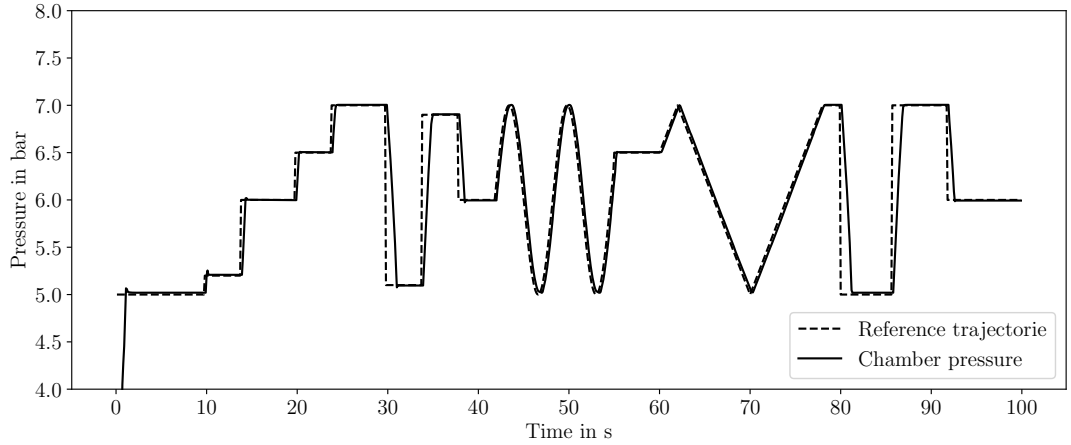


Figure 6: Simulated chamber pressure following a predefined trajectory

4. Experimental Analysis

In total 142 steady state experiments are used for the experimental investigation of the controller performance. Additionally one test following a trajectory is presented. Experiments are conducted at M11.5 test facility in Lampoldshausen [23, 24].

A characteristic pressure plot of one steady state experiment is given in Fig. 7. Data acquisition starts at $t = 0$ with the experiment time t. 4.3 s after begin, the automatic valve is opened, and nitrogen starts flushing the thrust chamber. At the beginning of each experiment the control valve is set to 10 % open. After opening of the automatic valve there is a 5 s wait to establish steady flow conditions. In this way equal starting conditions for every experiment are guaranteed. When the flow is established 10 s after the beginning, the neural network takes control and changes the chamber pressure to the desired value, in this case 8 bar. Due to limited valve speed in this experiment it takes 2.2 s to reach the set point. No overshoot can be observed. After 21 s the experiment is over and the control valve is opened completely, to release remaining nitrogen in the feeding lines. This is the reason for the pressure peak at the end of the experiment before the automatic valve is closed after 22 s.

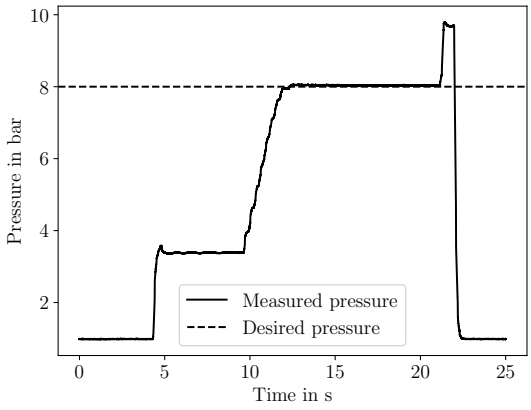
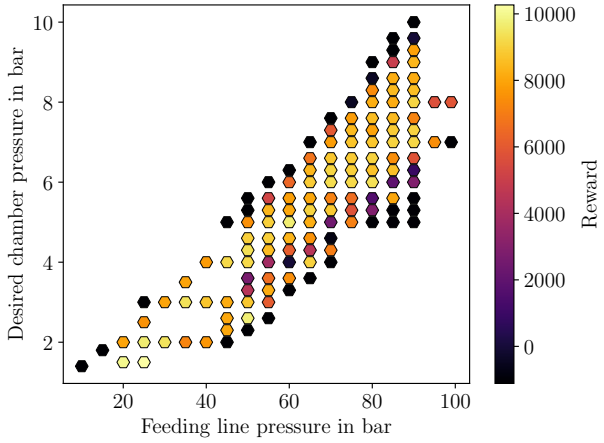


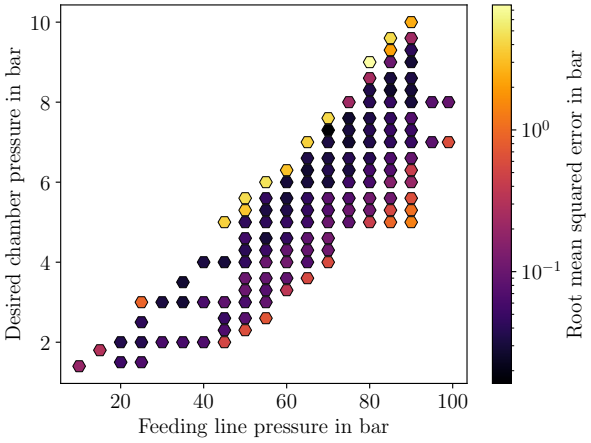
Figure 7: Thrust chamber pressure plot for a experiment with 80 bar feeding line pressure and a target pressure of 8 bar

DOI: 10.13009/EUCASS2023-639

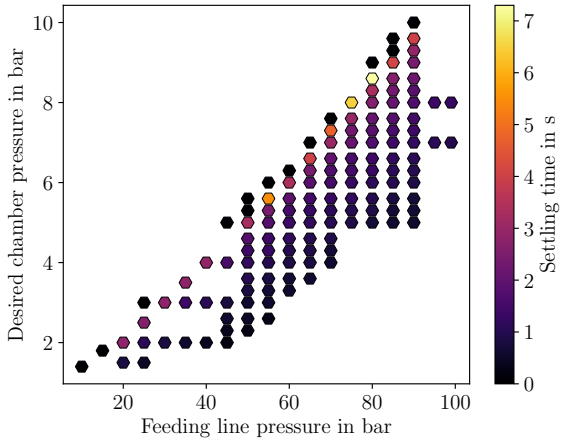
142 of this test similar to the one described above have been conducted with different feeding line pressures and different target chamber pressures. The results are presented in Fig 8.



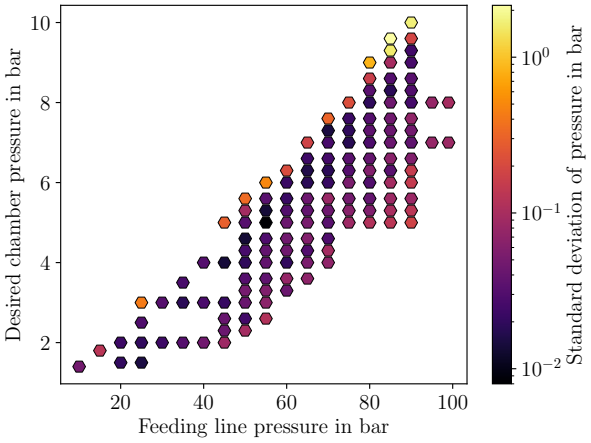
(a) Reward



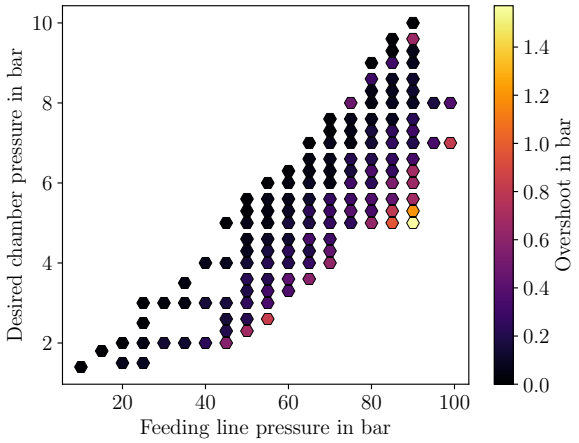
(b) Root mean squared error



(c) Settling time



(d) Standard variation in thrust chamber pressure



(e) Overshoot in thrust chamber pressure

Figure 8: Experimental results

DOI: 10.13009/EUCASS2023-639

In this plots every hexagon represents one experiment. On the horizontal axis the feeding line pressure and on the vertical axis the desired chamber pressure is shown. Therefore, a hexagon at the same spot always represents the same experiment. The color of the hexagon labels the magnitude of the evaluated variable for the respective experiment. All variables are calculated for the time interval the neural network takes over control (t = 10 s) until the end of the experiment (t = 22 s). Fig.8a indicates the calculated reward achieved in the experiments. The result is similar to the simulative evaluation (see Fig3). Inside the controllable area high reward values are obtained. Control limitation is given for low and high target pressures for minimum and maximum vale opening respectively. The limits can be seen by the black hexagons marking experiments resulting in a low reward. The physical limits of the system in the experiment are the same as in the simulation. In this cases the target pressure could not be set. Once a set point could not be reached the feeding line pressure was changed. As the experimental results were promising, the controller was also tested outside the training area. The controller was trained in simulation for feeding line pressures between 60 bar and 90 bar and target pressures from 5 bar to 9 bar. However also for feeding line pressures as low as 20 bar and up to 100 bar successful control results were observed. For lower feeding line pressures the possible target pressure is also reduced. Obviously, the neural network extrapolated the successful control policy also for other operating envelopes. There is no forecast about the control-quality outside the operating points used in training. Therefor this result has to be evaluated carefully before use in operation.

Fig. 8b shows the root mean squared error of the chamber pressure for each experiment in bar. Interestingly a slight decrease of RMSE for higher target pressures can be seen. Which leads to the assumption that the stationary control deviation for higher set points is lower. However, all experiments within the controllable area resulted in a RMSE not higher than 0.2 bar.

Fig. 8c shows the settling time. It is defined as time until only deviations of less than 10 % of the desired pressure arise. For higher target pressures the settling time grows, as the valve needs to be opened further. Similar to the simulative results the settling time in the controllable areas is as low as possible with the given speed of the control valves. For set points outside the controllable area no settling time can be calculated. They are marked with settling time zero. Some experiments at the border of the controllable area show larger settling times.

In Fig. 8d the standard deviation of the chamber pressure during the stationary time in the experiments is shown. Standard deviation is a measure for deviations and oscillations during the steady state operation. With less than 0.1 bar in the controllable area fluctuations around the target pressure are low.

In Fig.8e the color-map represents the overshoot that was observed during the settling phase of the chamber pressure. The observed overshoot for lower target pressures is higher while for high desired pressures no overshoot can be observed.

Concluding Fig. 8 it can be seen that for the relevant target area between 60 bar and 90 bar the controller is successful in changing and holding the chamber pressure to the required set point. While acting with fast settling times, low overshoot, low fluctuations and low derivations from the set point can be seen.

In order to test not only steady state performance of the neural network based controller, the control performace is tested following the given reference trajectory. Fig. 9 shows the measured thrust chamber pressure, corresponding to the simulation shown in Fig. 6. It can be seen that the controller was able to follow the trajectory. However, in comparison to the steady state experiments and to the simulative evaluation of the trajectory, more fluctuations and higher overshoot during the change of operating points can be seen. The long settling time for larger set point changes

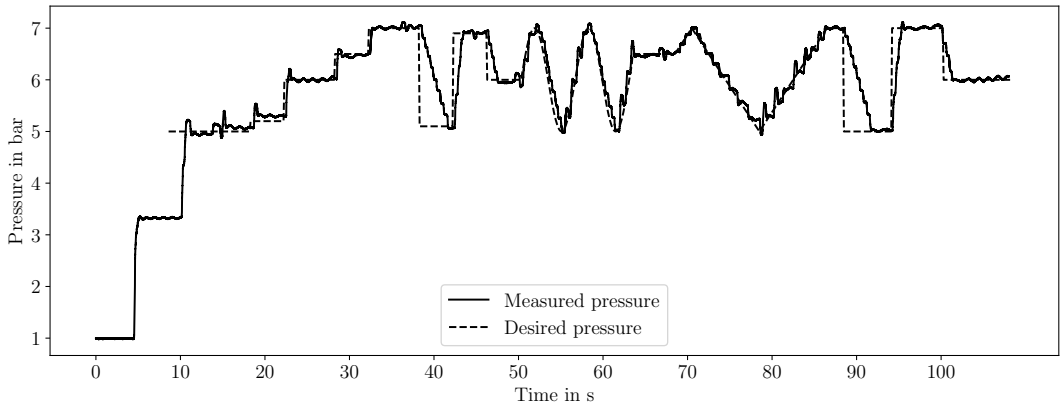


Figure 9: Chamber pressure following a predefined trajectory

DOI: 10.13009/EUCASS2023-639

is a result of the slow valves used for the experiments. The root mean squared error is calculated to 1.15 bar. This comparably high value is partly a result of the slow valves resulting in high absolute errors for long time during the long settling time. However the root mean squared error of the full trajectory in the experiment is higher than in the simulation. The controller performance in experiment following the trajectory is, despite using domain randomization, worse than in the simulation.

5. Conclusion and Outlook

A neural network trained with the reinforcement learning algorithm SAC [30] is used as chamber pressure controller for a nitrogen cold gas thruster. The controller was analyzed in simulation and experiment. Within the physical limits of the system the controller showed satisfying behavior, meeting the requirements of less than 0.1 bar steady state deviation from the set point, nearly no overshoot and a settling time, as fast as the system allows. The controller can be transferred from the simulation to the experiment and also in the real test set up the controller showed good performance. It was also shown that the controller is also able to control the thruster far outside the operating points that were used for training. The controller is able to follow a trajectory that was unknown during training. The trajectory contains different set point changes, linear and sinus pressure changes.

In future this control method will be applied to a nitrous oxide/ethane bipropellant system and hotfire experiments will be conducted. Several challenges come up in comparison to the cold gas experiments described in this paper. The controller has to regulate chamber pressure and mixture ratio independently. The combustion introduces higher roughness, sensor noise and the behavior of the thruster changes depending on the chamber temperature. Faster and more accurate, in house developed electronic control valves will be used in the future experiments to reduce the settling time. In a bipropellant set-up this method has to prove its suitability as rocket engine controller. It will also be possible to test the introduction of different secondary boundaries in the control-law as for example a limit in the wall temperature at a given thrust. Further research is needed to investigate the use of machine learning and neural networks as control method. A direct comparison with other established control mechanisms is desirable to find advantages and disadvantages of the respective method.

References

[1] I. Waugh, A. Davies, E. Moore, and J. Macfarlane. “VTVL technology demonstrator for planetary landers”. In: *Space Propulsion Conference*. 2016.

[2] E. Dumont, S. Ishimoto, P. Tatiossian, J. Klevanski, B. Reimann, T. Ecker, L. Witte, J. Riehmer, M. Sagliano, S. Giagkozoglou, I. Petkov, W. Rotärmel, R. Schwarz, D. Seelbinder, M. Markgraf, J. Sommer, D. Pfau, and H. Martens. “CALLISTO: a Demonstrator for Reusable Launcher Key Technologies”. In: *32nd ISTS*. 2019. URL: https://elib.dlr.de/128795/.

[3] J. Vila and J. Hassin. “Technology acceleration process for the Themis low cost and reusable prototype”. In: *8th European conference for aeronautics and space sciences*. 2019, pp. 1–4.

[4] A. P. de Mirand, J.-M. Bahu, and O. Gogdet. “Ariane Next, a vision for the next generation of Ariane Launchers”. In: *Acta Astronautica* 170 (2020), pp. 735–749. ISSN: 00945765.

[5] C. F. Lorenzo and J. L. Musgrave. “Overview of rocket engine control”. In: *AIP Conference Proceedings*. Vol. 246. 1992, pp. 446–455.

[6] S. Pérez-Roca, J. Marzat, H. Piet-Lahanier, N. Langlois, F. Farago, M. Galeotta, and S. Le Gonidec. “A survey of automatic control methods for liquid-propellant rocket engines”. In: *Progress in Aerospace Sciences* 107 (2019), pp. 63–84.

[7] S. Pérez-Roca, J. Marzat, H. Piet-Lahanier, N. Langlois, M. Galeotta, F. Farago, and S. Le Gonidec. “Model-based robust transient control of reusable liquid-propellant rocket engines”. In: *IEEE Transactions on Aerospace and Electronic Systems* 57.1 (2020), pp. 129–144.

[8] G. Waxenegger-Wilfing, U. Sengupta, J. Martin, W. Armbruster, J. Hardi, M. Juniper, and M. Oschwald. “Early detection of thermoacoustic instabilities in a cryogenic rocket thrust chamber using combustion noise features and machine learning”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31.6 (2021), p. 063128.

[9] C. F. Lorenzo, A. Ray, and M. S. Holmes. “Nonlinear control of a reusable rocket engine for life extension”. In: *Journal of Propulsion and Power* 17.5 (2001), pp. 998–1004.

[10] W. Merrill and C. Lorenzo. “A reusable rocket engine intelligent control”. In: *24th Joint Propulsion Conference*. 1988, p. 3114.

DOI: 10.13009/EUCASS2023-639

[11] G. Waxenegger-Wilfing, K. Dresia, J. Deeken, and M. Oschwald. “Machine Learning Methods for the Design and Operation of Liquid Rocket Engines–Research Activities at the DLR Institute of Space Propulsion”. In: *arXiv preprint arXiv:2102.07109* (2021).

[12] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. second edition. Vol. 2018: 1. Adaptive computation and machine learning. Cambridge, Mass.: The MIT Press, 2018. ISBN: 9780262039246.

[13] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. “Learning dexterous in-hand manipulation”. In: *The International Journal of Robotics Research* 39.1 (2020), pp. 3–20.

[14] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. *Solving Rubik’s Cube with a Robot Hand*. Ed. by arXiv preprint arXiv:1910.07113. 2019.

[15] F. Sadeghi and S. Levine. *CAD2RL: Real Single-Image Flight without a Single Real Image*. Ed. by arXiv preprint arXiv:1611.04201. 2017.

[16] A. Folkers. *Steuerung eines autonomen Fahrzeugs durch Deep Reinforcement Learning*. Springer-Verlag GmbH, 2019. URL: https://www.ebook.de/de/product/38406162/andreas_folkers_steuerung_eines_autonomen_fahrzeugs_durch_deep_reinforcement_learning.html.

[17] T. Hörger, K. Dresia, G. Waxenegger-Wilfing, L. Werling, and S. Schlechtriem. “DEVELOPMENT OF A TEST INFRASTRUCTURE FOR A NEURAL NETWORK CONTROLLED GREEN PROPELLANT THRUSTER”. In: *8th Space Propulsion Conference*. 2022. URL: https://elib.dlr.de/186952/.

[18] L. Werling. “Entwicklung und Erprobung von Flammensperren für einen vorgemischten, grünen Raketentreibstoff aus Lachgas und Ethen”. PhD thesis. Fakultät für Luft- und Raumfahrttechnik und Geodäsie der Universität Stuttgart.

[19] L. K. Werling, T. Hörger, K. Manassis, D. Grimmeisen, M. Wilhelm, C. Erdmann, H. K. Ciezki, S. Schlechtriem, S. Richter, M. Torsten, et al. “Nitrous oxide fuels blends: research on premixed monopropellants at the german aerospace center (DLR) since 2014”. In: *AIAA Propulsion and Energy 2020*, p. 3807.

[20] L. Werling and T. Hörger. “Experimental analysis of the heat fluxes during combustion of a N2O/C2H4 premixed green propellant in a research rocket combustor”. In: *Acta Astronautica* 189 (2021), pp. 437–451. ISSN: 00945765.

[21] M. Kurilov, L. Werling, M. Negri, C. Kirchberger, and S. Schlechtriem. “IMPACT SENSITIVENESS OF NITROMETHANE-BASED GREEN-PROPELLANT PRECURSOR MIXTURES”. In: *International Journal of Energetic Materials and Chemical Propulsion* (2022).

[22] F. Lauck, J. Balkenhohl, M. Negri, D. Freudenmann, and S. Schlechtriem. “Green bipropellant development–A study on the hypergolicity of imidazole thiocyanate ionic liquids with hydrogen peroxide in an automated drop test setup”. In: *Combustion and Flame* 226 (2021), pp. 87–97. ISSN: 0010-2180.

[23] M. Wilhelm, L. Werling, F. Strauss, F. Lauck, C. Kirchberger, H. Ciezki, and S. Schlechtriem. “Test Complex M11: Research on Future Orbital Propulsion Systems and SCRamjet Engines”. In: *International Astronautical Congress*. 2019. URL: https://elib.dlr.de/133885/.

[24] H. Ciezki, L. Werling, M. Negri, F. Strauss, M. Kobald, C. Kirchberger, D. Freudenmann, C. Hendrich, M. Wilhelm, A. Petrarolo, and S. Schlechtriem. “50 Years of Test Complex M11 in Lampoldshausen – Research on Space Propulsion Systems for Tomorrow”. In: *7th European Conference for Aeronautics and Space Sciences (EUCASS)*.

[25] G. Rebala, A. Ravi, and S. Churiwala. *An introduction to machine learning*. Springer, 2019.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. “Playing atari with deep reinforcement learning”. In: *arXiv preprint arXiv:1312.5602* (2013).

[27] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. “Policy gradient methods for reinforcement learning with function approximation”. In: *Advances in neural information processing systems* 12 (1999).

[28] S. Fujimoto, H. Hoof, and D. Meger. “Addressing function approximation error in actor-critic methods”. In: *International conference on machine learning*. 2018, pp. 1587–1596.

[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. “Continuous control with deep reinforcement learning”. In: *arXiv preprint arXiv:1509.02971* (2015).

[30] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”. In: *International conference on machine learning*. 2018, pp. 1861–1870.

DOI: 10.13009/EUCASS2023-639

[31] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, et al. "Ray: A distributed framework for emerging 5AI6 applications". In: *13th 5USENIX6 Symposium on Operating Systems Design and Implementation (5OSDI6 18)*. 2018, pp. 561–577.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. *Proximal Policy Optimization Algorithms*. 2017. DOI: 10.48550/ARXIV.1707.06347.

[33] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. "Asynchronous Methods for Deep Reinforcement Learning". In: (2016). DOI: 10.48550/ARXIV.1602.01783.

[34] Empresarios Agrupados Internacional. *EcosimPro 6.2.0*.

[35] W. Zhao, J. P. Queralta, and T. Westerlund. "Sim-to-real transfer in deep reinforcement learning for robotics: a survey". In: *2020 IEEE symposium series on computational intelligence (SSCI)*. 2020, pp. 737–744.

[36] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull. "Active domain randomization". In: *Conference on Robot Learning*. 2020, pp. 1162–1176.

[37] L. Weng. "Domain Randomization for Sim2Real Transfer". In: *lilianweng.github.io/lil-log* (2019). URL: http://lilianweng.github.io/lil-log/2019/05/04/domain-randomization.html.