Article

Optimal Dynamic Regimes for CO Oxidation Discovered by Reinforcement Learning

Mikhail S. Lifar,* Andrei A. Tereshchenko, Aleksei N. Bulgakov, Sergey A. Guda, Alexander A. Guda,* and Alexander V. Soldatov
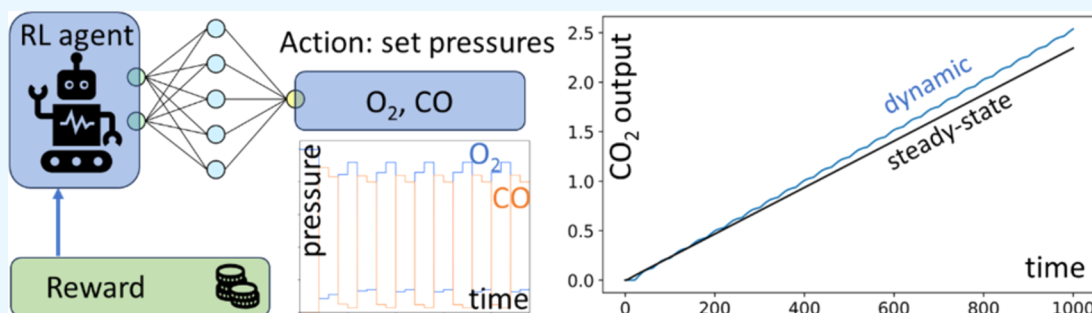
Cite This: ACS Omega 2024, 9, 27987−27997

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Metal nanoparticles are widely used as heterogeneous catalysts to activate adsorbed molecules and reduce the energy barrier of the reaction. Reaction product yield depends on the interplay between elementary processes: adsorption, activation, desorption, and reaction. These processes, in turn, depend on the inlet gas composition, temperature, and pressure. At a steady state, the active surface sites may be inaccessible due to adsorbed reagents. Periodic regime may thus improve the yield, but the appropriate period and waveform are not known in advance. Dynamic control should account for surface and atmospheric modifications and adjust reaction parameters according to the current state of the system and its history. In this work, we applied a reinforcement learning algorithm to control CO oxidation on a palladium catalyst. The policy gradient algorithm was trained in the theoretical environment, parametrized from experimental data. The algorithm learned to maximize the CO_2 formation rate based on CO and O_2 partial pressures for several successive time steps. Within a unified approach, we found optimal stationary, periodic, and nonperiodic regimes for different problem formulations and gained insight into why the dynamic regime can be preferential. In general, this work contributes to the task of popularizing the reinforcement learning approach in the field of catalytic science.

1. INTRODUCTION

Noble metal nanoparticles are widely known as efficient catalysts for many oxidation and reduction reactions.[1−3] Many attempts are made to improve their performance, for example, by doping with transition metals[3] or using bimetallic compounds.[4] However, even these advanced catalysts can suffer from deactivation and aging,[5,6] forcing researchers to search for optimal conditions for the reaction,[7,8] which can be very far from conventional ones used for model catalysts.

Switching between different reaction conditions can increase the turnover frequency in the catalytic reactions and lead to the benefit of transient state kinetics over steady state.[9] Such a regime has attracted attention in chemical engineering.[10−13] Transient reaction modulation can be reached by a periodical change in the flows of reactants,[13−16] pressure,[15,17] temperature,[18−21] or light irradiation.[22] The rate enhancement can be explained by several processes on the catalyst surface: structural reorganization, cleaning from nondesired products, and overcoming activation barriers of intermediates by an external force. Some reactions may exhibit an oscillatory behavior[23] even under steady external conditions due to coupled processes in the multicomponent system. For example, the reversible formation of palladium carbide[24,25] explains the kinetic oscillations upon CO oxidation[26] or ethylene hydrogenation.[27] Transient oxidation of carbon monoxide on the surface of noble metal nanocatalysts is an object of particular interest: this reaction is a model case for fundamental studies in heterogeneous catalysis[28−32] and is widely used nowadays in automotive converters of vehicles. The three-way catalysts are periodically subjected to oxidative (air reactor) and reductive (fuel reactor) conditions, and such cycling drastically increases the efficiency of CO and NO conversion.[33−35]

Machine learning (ML),[36,37] particularly reinforcement learning (RL), is promising for predicting the best reaction conditions in static and dynamic regimes. RL differs from supervised machine learning in terms of the approach to how the training set is compiled. The algorithm itself performs actions in the environment and collects optimal trajectories to enhance the future manipulation of a dynamical system.[38−40] Nowadays, RL finds its application in different fields[41] such as neurobiology,[42] robotics,[43−45] communications and networking,[46] health control,[47] personalized news recommendation,[48] and material science.[49] Neumann and Palkovits[50] proved the concept of using RL for optimizing hydrogen production yield in the reaction of partial oxidation of methane. For this aim, they trained Q-learning (QL)[51,52] and deep deterministic policy gradient (DDPG) agents[53] to maximize H_2 production by adjusting temperature, pressure, flow velocity, and substrate composition in the simulated plug flow reactor. Alhazmi et al.[54] applied RL in combination with economic model predictive control for ethylene oxide production. This framework allowed online continuous control, making autonomous reactor operation more attainable. The RL agent can effectively act, even in environments that are different from those applied for training. Zhou et al. optimized several chemical reactions using deep RL.[55] This approach outperformed a state-of-the-art black box optimization algorithm using fewer steps and was effectively applied to four real microdroplet reactions to find the optimal experimental conditions within 30 min and accelerated reaction rates.

However, aforementioned works[50,55] demonstrate only steady-state solutions discovered by RL algorithms. The problem of finding steady-state optimum refers to the task of multivariate optimization and for low number of variables can be solved by classical methods.[56,57] Meanwhile, the true power of RL is to find complex, nonstationary behavior in challenging environments. In this work, we demonstrate the application of RL to optimize inlet gas feeds for achieving the highest product yield in a catalytic system over a specified interval of time. The algorithm was not constrained to a specified form of functional dependency and discovered stationary, periodic, and non-periodic solutions for the tasks of optimal pressure control and optimal response to varying external conditions. The processes of adsorption, desorption, and reaction of CO and O_2 on the metal surface were described by the system of ODEs and model parameters were refined from experimental data. We choose the palladium (111) surface as a model for its high activity and wide range of industrial applications.[58,59] This numerical model was used as an environment for the algorithm, whose goal was the maximization of the mean reaction rate by varying the pressures of reactant gases. We have explored how the parameters of the model influence the algorithm's behavior and found the ranges where the algorithm preferred nonstationary control. We also proposed a non-deterministic task for the algorithm, including random perturbations in the feed of one of the reactants, and the algorithm was able to adapt.

2. METHODS

The reinforcement learning methodology implies that an algorithm learns from the data generated by the algorithm itself during interactions with an external environment. It requires many trial steps for training. The trials are selected and performed successively by the algorithm in order to find the optimal behavior in a poten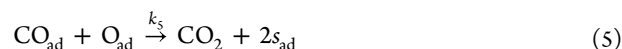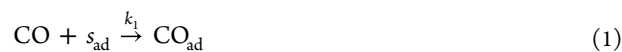tially huge space of all possible behaviors. Such an approach is different from supervised machine learning, where the whole set of trials is provided once by the user.

2.1. Model Description. We use a mathematical model of the CO oxidation reaction on Pd(111) to create a simulation of the chemical process, which the algorithm would be able to control and learn from.
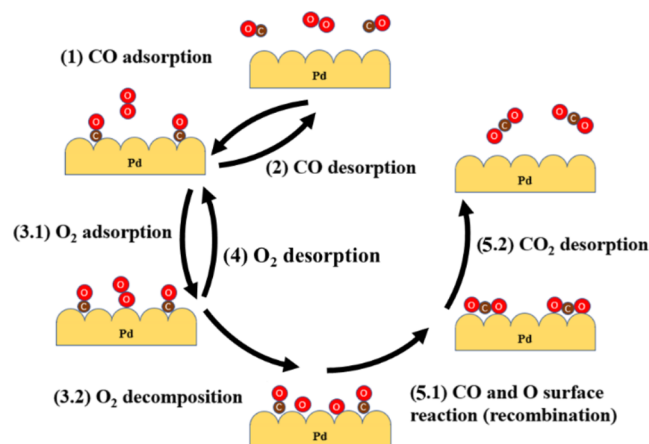
Figure 1. Simplified elementary steps of CO oxidation on a stable Pd surface.

CO oxidation on Pd nanoparticles can be split into five elementary steps (illustrated with Figure 1) following the Langmuir−Hinshelwood mechanism:[60,61]

$$CO + s_{ad} \xrightarrow{k_1} CO_{ad} \tag{1}$$

$$CO_{ad} \xrightarrow{k_2} CO + s_{ad} \tag{2}$$

$$O_2 + 2s_{ad} \xrightarrow{k_3} 2O_{ad} \tag{3}$$

$$2O_{ad} \xrightarrow{k_4} O_2 + 2s_{ad} \tag{4}$$

$$CO_{ad} + O_{ad} \xrightarrow{k_5} CO_2 + 2s_{ad} \tag{5}$$

where s_{ad} refers to a free adsorption site and k_i is the rate constant of step $i = 1, \cdots, 5$. We consider that adsorption/desorption has no effect on the inlet gas feed (differential reactor model) and each gas species inhibits the adsorption of another. The evolution of surface coverages is then represented by coupled differential equations

$$\frac{d\theta_{CO}}{dt} = k_1\left(1 - \frac{\theta_{CO}}{\theta_{CO}^{max}} - C_{O,CO}\frac{\theta_O}{\theta_O^{max}}\right)p_{CO} - k_2\theta_{CO}$$
$$- k_5\theta_{CO}\theta_O \tag{6}$$

$$\frac{d\theta_O}{dt} = 2k_3\left(\left(1 - \frac{\theta_{CO}}{\theta_{CO}^{max}} - C_{CO,O}\frac{\theta_O}{\theta_O^{max}}\right)\right)^2 p_{O_2} - 2k_4\theta_O^2$$
$$- k_5\theta_{CO}\theta_O \tag{7}$$

where θ_{CO} and θ_O are carbon monoxide and oxygen coverages, respectively, and "+" denotes the ramp function. For convenience of notation, we introduce normalized pressures

Table 1. Combinations of parameters for Equations (1)-(8) Applied for RL training (Set 1 is Reproduced from Ref 62). The Sets 1-3 differ by the rate of CO desorption (k_2), O_2 adsorption (k_3), CO and O reaction rate on the surface (k_5) and influence of adsorbed oxygen on the CO adsorption ($C_{O,CO}$)

set	k_1	k_2	k_3	k_4	k_5	$C_{O,CO}$	$C_{CO,O}$	Figures
1	0.149	0.0716	0.0659	0	5.897	0.3	1	5
2	0.149	0.1	0.0659	0	0.1	0.3	1	6 and 7
3	0.149	0.0716	0.264	0	5.897	1	1	8
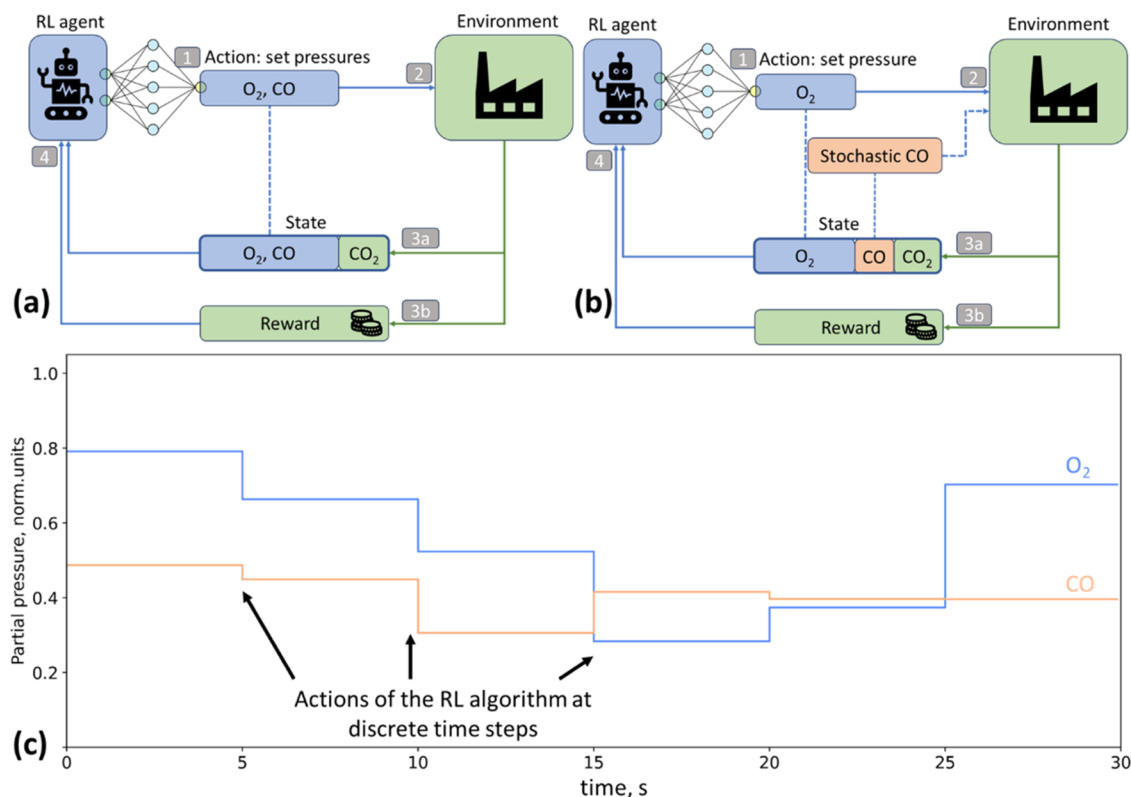
Figure 2. Two problem statements for optimization by RL algorithm. (a) Global optimization of the reaction yield by optimizing both reagents CO and O_2 to maximize reaction yield. (b) Response to external conditions by adjusting O_2 pressure on varying CO partial pressure to maximize reaction yield. (c) Input gas pressures are controlled by the algorithm. The time intervals between the pressure switches (actions) are highlighted with arrows.

p_{CO} and p_{O_2} varying between 0 and 1, respectively, where 1 corresponds to the value of 10^{-4} Pa for the set of kinetic parameters #1. The carbon dioxide formation rate is then determined by

$$r_{CO_2} = k_5 \theta_{CO} \theta_O \tag{8}$$

These equations were largely reproduced from experimental work of Libuda et al.,[62] where well-defined Pd(111) nanoparticles interacted with reaction gases in the transient regime (see also S1, Tables S1 and S2 for details). Generally, we treated k_1, ···, k_5, $C_{O,CO}$, and $C_{CO, O}$ as numerical parameters of our system and varied throughout the research. Different values of these parameters can correspond to different reaction conditions, like temperature, catalyst composition, or catalyst morphology. We do not focus on physical phenomena that lead to such variations in rate constants and inhibition coefficients but rather discuss the influence of these changes on the behavior of the agent. However the starting values of the parameters in our model are chosen in a way to reproduce well-defined experimental data (see Set 1 in Table 1 and Figure S1 for further details).

2.2. Reinforcement Learning Methodology.

The key components of the reinforcement learning approach are an environment and an agent interacting with each other. The environment has a state describing it at a given time step. The agent receives the state of the environment as input and produces an action in response. The environment changes its state after the agent's action, and then provides it to the agent along with a reward. This cycle repeats over and over during the training. The goal of the agent is to learn those actions in each state maximizing a long-term cumulative reward (i.e., sum or weighted sum of rewards). In this work, we used the Vanilla Policy Gradient (VPG) algorithm,[63] which is an on-policy algorithm in the form of the REINFORCE algorithm.[64] The agent in the VPG (its policy) is an artificial neural network (ANN) that receives a state from the environment and generates an action based on it. The following equations define the algorithm[65]

$$\theta_{k+1} = \theta_k + \alpha \cdot \nabla_{\theta_k} J(\pi_{\theta_k}) \tag{9}$$

$$\nabla_\theta J(\pi_\theta) = E_{\tau \sim \pi}\left[\sum_{t=0}^{T} \nabla_\theta(\log \pi_\theta(a_t|s_t)) \cdot G_t\right] \tag{10}$$

Equations 9 and 10 define how the weights θ of a neural network are updated each training step k via gradient ascent. $J(\pi_\theta) = E_{\tau \sim \pi}[R(\tau)]$ is an expected return of an episode if the actions are taken in accordance with the policy π_θ. $\pi_\theta(a_t|s_t)$ denotes a probability (or likelihood) of the agent's action a_t taken after receiving the state s_t at time step t. G_t is a value that estimates the total reward received after action a_t is accepted.[66,67] This value may be a future return $R(\tau)$ (this work), state-action value $Q^{\pi_\theta}(s, a)$ or an advantage $A^{\pi_\theta}(a_t, s_t)$[66,67] (see also S2 for details). Equation 10 defines the gradient vector and contains the expectation operator, which is further approximated by eq 11

$$\nabla_\theta J(\pi_\theta) = \frac{1}{|D|}\sum_{\tau \in D}\left[\sum_{t=0}^{T} \nabla_\theta(\log \pi_\theta(a_t, s_t)) \cdot G_t\right] \tag{11}$$

In practice, the algorithm runs the batch of episodes D of size |D|. Each episode generates a trajectory τ: a sequence of triples (s_t, a_t, r_t) at each time step t, with s_t being a state, a_t the action, and r_t the reward at time step t. The gradient vectors are computed at each time step and then summed along the trajectory. The neural network's weights and, as a result, the behavior of the algorithm are adjusted iteratively during many cycles of training with the mean gradient vector for all of the trajectories in the batch.

We used VPG agent implemented in Python RL library Tensorforce based on Tensorflow.[63] The batch size was equal to 16. The ANN configuration was chosen by a default procedure in the algorithm and contained two hidden layers, and the learning rate (α from eq 11) was equal to 1 e^{-3}.

3. RESULTS AND DISCUSSION

We applied the reinforcement learning approach in two practical tasks. The first one was the problem of finding the optimal control over CO and O_2 gas pressures, maximizing the mean reaction rate. This task was solved for different values of rate constants $k_1, \cdots, k_5$ (eqs 6 and 7) to demonstrate diverse control regimes, particularly stationary and nonstationary. The second task was related to optimal control of the reactor under varying external conditions. For this purpose, we allowed the algorithm to adjust the O_2 gas feed to random CO pressure variations in a way that maximizes the mean reaction rate. Both problem statements are illustrated in Figure 2. The figure highlights the similarities and differences between training strategies: (1) the reinforcement learning agent updates both O_2 and CO pressures in (a) or only O_2 pressure in (b); (2) the environment receives new values for gas pressures; (3) the numerical model behind environment evaluates dynamics of the system during specified interval of time (one time step in terms of RL cycle), calculates the current reaction rate for CO_2 and a reward, and updates the state; and (4) the new state and the reward return to the agent.

3.1. Global Optimization of the Reaction Yield. First, we allowed the algorithm to control the reaction by manipulating CO and O_2 partial pressures. The algorithm set arbitrary partial pressures of both gases within the normalized range from 0 to 1 at each time step. The time step duration was selected as 5 s based on model kinetics to represent an interval fof substantial changes in the surface coverages. The

goal for the agent was to maximize the mean reaction rate over 30 s episode. Figure 3 demonstrates a typical agent's
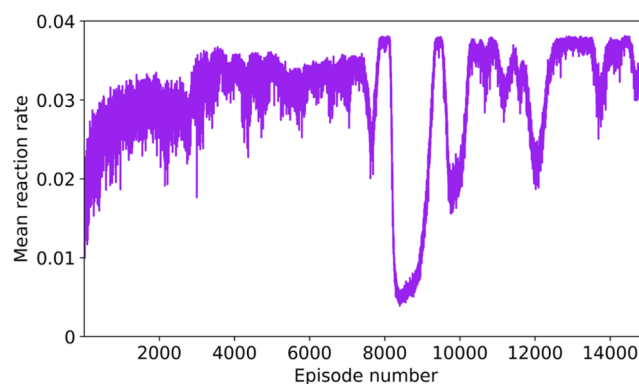
Figure 3. Typical successful learning curve of the algorithm. Each episode lasts for 30 s, the algorithm can switch gas flows every 5 s, and the reaction rate is supplied as a reward on each step.

performance evolution during the training procedure. According to the plot, the agent's performance increased non-monotonically for the first 8000 episodes until it reached its maximum. Roughly 10,000 episodes, which is equivalent to about 83 h of the real experiment, were required for the agent to learn it is policy. For different rate constants (eqs 6 and 7), the training curve stabilized before 11,000 episodes in 50% of tests and before 24,000 episodes in 75% of tests.

We observed a significant influence of the physical model on the algorithm's behavior. To simulate different sticking probabilities and reaction rates, we varied the rate constants k_i between 10^{-2} and 10^1. For each set of parameters, the training was performed from scratch resulting either in stationary solution or dynamical.[62] To verify the solutions found by the reinforcement learning algorithm, we employed in parallel the Nelder−Mead method[56] for optimal CO_2 production rates. Figure 4 summarizes the results of the calculations and shows a ratio of the mean reaction rate
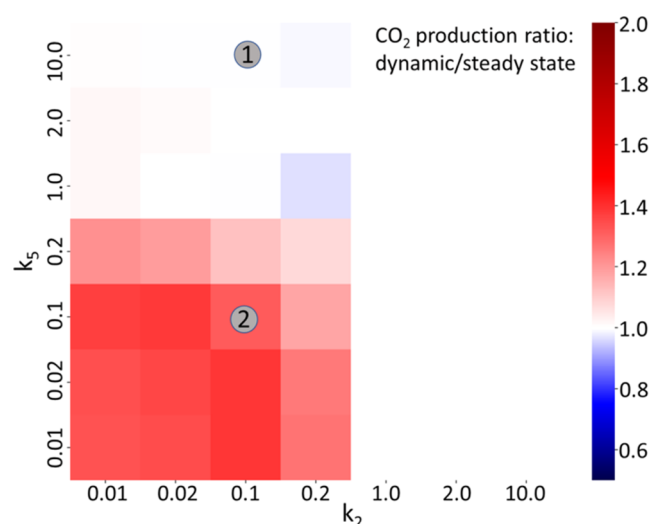

Figure 4. Ratio of the mean reaction rates in solutions obtained by reinforcement learning (steady state or dynamic) and Nelder−Mead (steady state), depending on rate constants k_2 and k_5. The RL solutions from cells 1 and 2 are obtained for parameter Sets 1 and 2 (Table 1) and are visualized further in Figures 5 and 6

achieved by reinforcement learning to the mean reaction rate obtained by Nelder−Mead (raw numbers for the figure can be found in Tables S3 and S4). In some cases, the RL-obtained regime was slightly less effective than corresponding Nelder−Mead's regime (ratios 0.95...1.0). In other calculations, the reinforcement learning approach showed a significant advantage over Nelder−Mead, outperforming steady-state optimum (ratios 1.0 ...1.39). Optimal stationary regimes discovered by RL and Nelder−Mead were similar (Figure 5). Higher ratios in Figure 4 (>1.05)mostly correspond to the dynamic regimes (see Figure 6a,b).

Figure 5a,b shows a steady RL regime compared to the Nelder−Mead stationary regime. Interestingly, the algorithm
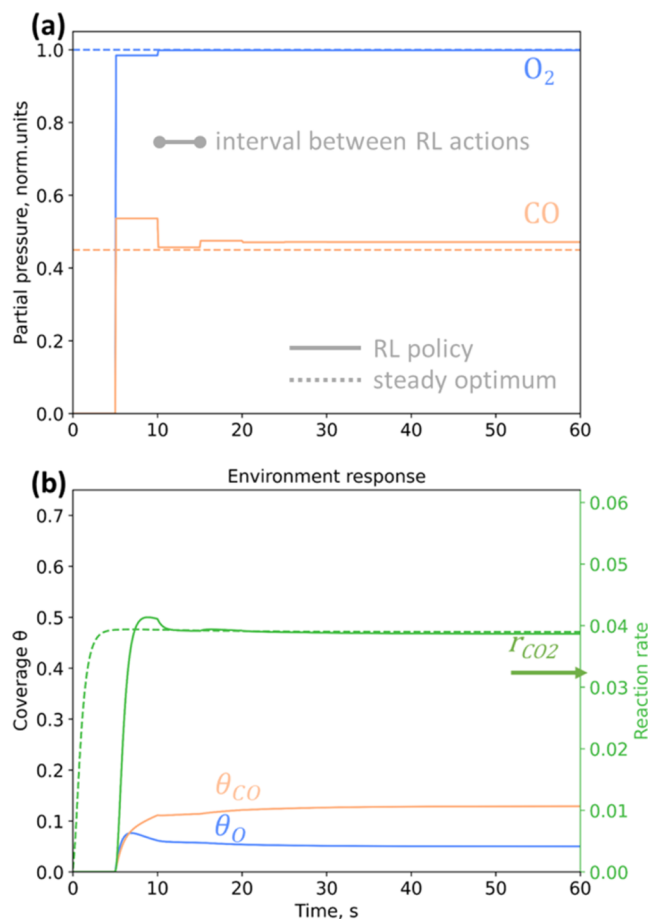
Figure 5. Temporal variations for CO (orange) and O_2 (blue) partial pressures (a), coverages, and reaction rate (b) found by the RL algorithm. The rate constants were taken from set 1 (see Table 1 and Figure 4). The algorithm was allowed to vary gas pressures every 5 s and the episode length for training was 30 s. Green line corresponds to the CO_2 formation rate. Dashed lines denote the Nelder−Mead steady optimal regimes.

reaches the optimal values of O_2 and CO pressures after several trial steps, starting from any initial conditions, and the input gas pressures stabilize within the first 20 s. RL algorithm was able to vary gas pressures every 5 s, but the ratio of CO/O_2 remained constant after the first 20 s, thus keeping the balance between available adsorption sites and concentration of activated intermediates.

Figure 6 demonstrates an example of the nonstationary regime obtained by the RL algorithm with Set 2 of kinetic constants. In this case, the policy of the agent supplied pure O_2

to oxidize the surface and then added CO later. Such a behavior allowed the agent to achieve high coverages of both oxygen and carbon monoxide on the surface, thus maximizing the reaction rate of CO_2 production. Figure 6a,b shows also another interesting observation. The agent was trained on episodes with 30 s duration, i.e., it maximized the mean reaction rate for this period. As clear from Figure 6b, at longer delays, the algorithm's policy provides worse results than the optimal steady regime (compare green solid and dotted lines). The reason is that the algorithm found itself in a situation it had never been in before; thus, it was unable to extrapolate its policy beyond the training period (first 30 s of operation). Therefore, we repeated the training with longer episode (240 s) to account for slower surface changes due to the low reaction rate constant. The improved agent performance is shown in Figure 6c,d.

Dynamic control accounts for the preferential accumulation of one of the species on the metal surfaceand outperforms the stationary solution by ca. 8% (Figure 7c). We further explore the fundamental origin of such a behavior. Under stationary conditions, coverages vary in a domain visualized by the dashed lines in Figure 7a. The coverages in the upper right corner of the figure are not accessible. At the same time, these coverages are accessible in the transient regime (solid blue line). Dynamic regime consists of two phases. During the first phase, oxygen accumulates on the surface, and θ_O and θ_{CO} coverages fall into the shadowed region of accessible steady-state coverages. In the second phase, the concentration of CO in the atmosphere rapidly increases. The CO coverage increases while desorption of oxygen is retarded from the surface since the oxygen molecule is atomically split on Pd. The resulting surface coverages thus fall outside the steady-state region of coverages. During the time spent outside the domain of stationary solutions, the agent manages to reach a high reaction rate that compensates for the accumulation phase when the reaction rate is lower than the stationary optimum.

The transient regime provides an elevated CO_2 production rate when surface coverages for two gases are populated with different rates, and the reaction rate is slower than adsorption/desorption rates. θ_{CO} changes faster than θ_O in the case of palladium since oxygen needs an activation step, and carbon monoxide can occupy the lattice sites accessible for O_2 adsorption and activation. On the contrary, the desorption of oxygen atoms after O_2 activation is almost negligible. Figures 6 and 7 show a transient regime when the CO_2 production rate is smaller compared to the adsorption rate (Set 2, Table 1); under such conditions agent policy can accumulate oxygen on the surface in the first step. In the second step, the algorithm dramatically increases the pressure of CO and creates conditions when both coverages become higher, providing a higher reaction rate.

3.2. Response to Varying External Conditions. In the previous section, we applied reinforcement learning to establish optimal control over gas composition in the system. In this section, RL is applied to adjust system parameters under an external perturbation. We consider the case of non-deterministic CO feed and the agent adjusting the O_2 partial pressure. Upon training, the CO partial pressure varied arbitrarily 3 times in each episode with 10 s delays, and the agent step remained 5 s long. The goal of the agent was maximizing the mean reaction rate over an episode. The key difference from the previous optimization problem is the randomness embedded into the environment, meaning that the
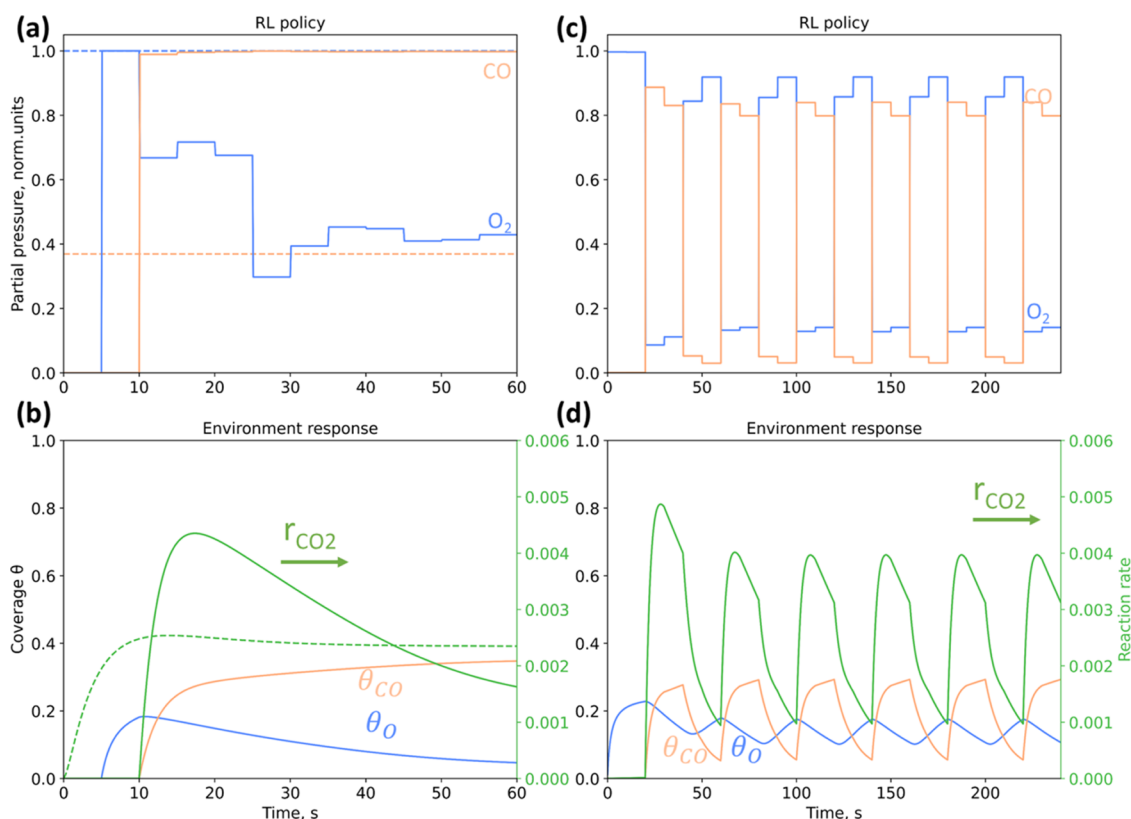
Figure 6. Two different dynamic regimes discovered by the algorithm with a 30 s episode for training (a, b) and 240 s (c, d) (Set 2, Table 1 and Figure 4). The longer episode length allowed the algorithm to find the periodic solution and prevent asymptotic decay of the reaction yield.
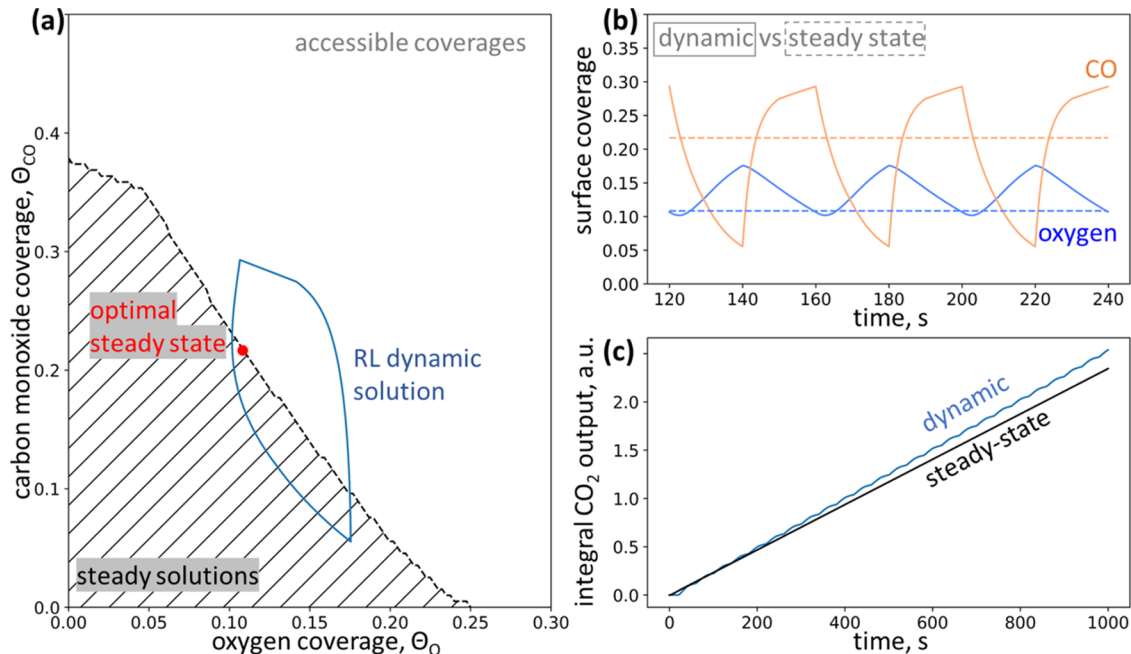
Figure 7. (a) Region of accessible coverages in the steady-state regime (dashed) with the red point indicating conditions for the highest steady reaction rate. The blue line shows the trajectory of coverages in the optimal dynamic regime discovered by RL. (b) CO and oxygen coverages in dynamic regime (solid lines) compared to the ones in optimal steady state in the panel (a) (dashed lines). (c) Advantage of the periodic regime is highlighted by the integral CO_2 output.

agent must account for the uncertainty of the future states in order to achieve good performance. Solving such a task requires modifications in the mathematical procedure. When the algorithm controls both gas pressures, we qualify its performance as the mean reaction rate over an episode. However, with randomly changing the CO feed, the mean reaction rate between different episodes cannot be compared anymore, since now it depends not only on the agent's
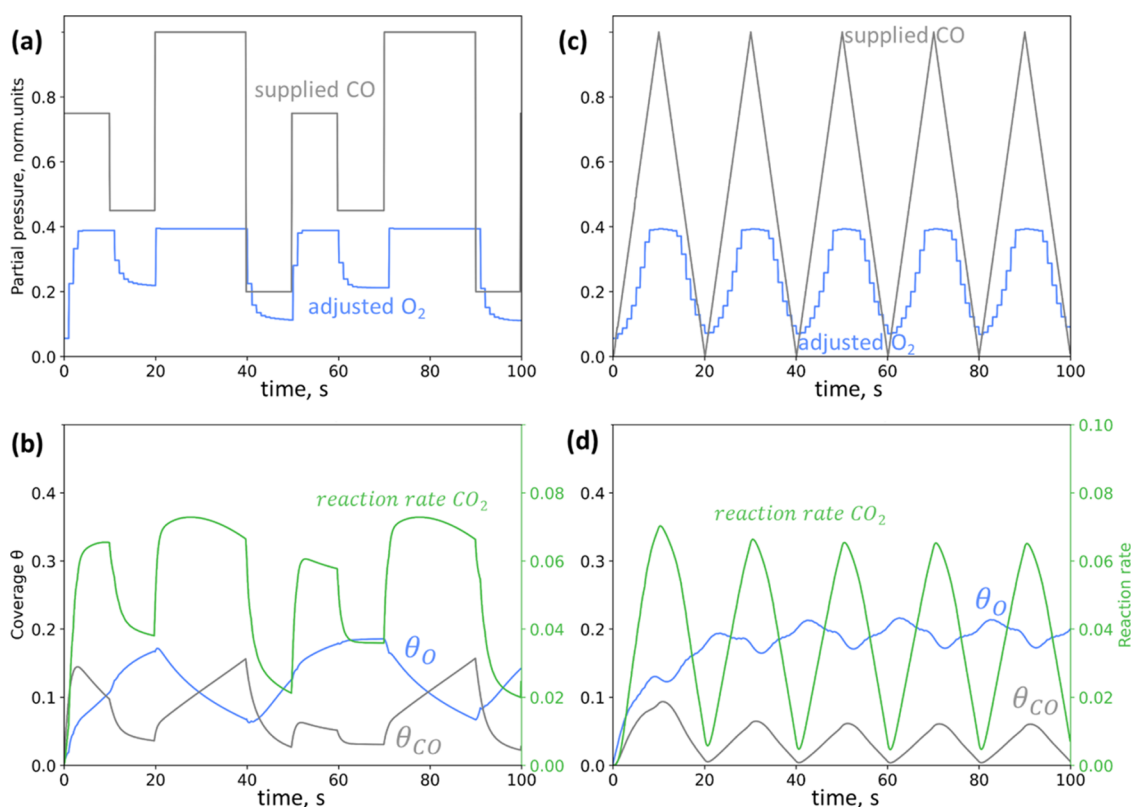
Figure 8. RL agent policy examined against two CO feed regimes of random (a, b) and triangle (c, d) forms. The top panels (a, c) show the algorithm's policy, and the bottom panels (b, d) show corresponding reaction rates and coverage dynamics. The agent adjusted the O_2 pressure to the supplied CO pressure variations. Model parameters were set according to Set 3 (Table 1).

behavior but also on varying the CO feed. To overcome this issue, the performance of the algorithm was evaluated on the set of predefined temporal CO dependencies. During the training the CO feed regime was sampled from the aforementioned pool forcing the agent to focus on the less probable but important edge cases, such as stationary CO regimes close to 0 and 1.

If the model parameters were chosen from Set 1 or Set 2 (Table 1), the solution proposed by the agent was stationary; namely, the agent sets the O_2 partial pressure to its maximum value. Set 3 provided dynamic agent's solution. Examples of the agent's policy for different CO feeds are demonstrated in Figure 8. The demostrated temporal variation of CO was not seen by the agent during the training, and the duration of the runs was increased to 100 s compared to 30 s length of training episode. The agent learned to increase or decrease O_2 pressure along with the variation of the CO pressure. However, the agent's policy was not a synchronous repetition of the CO regime but rather a gradual adaptation to current CO pressure with a time delay of about one agent step (5 s). Interestingly, the agent escaped poisoning and achieved a high reaction rate for the CO feed previously unseen during the training (for instance, time evolution in the triangle shape). Its policy also worked well at time delays longer than training (100s vs 30 s).

3.3. Discussion. The advantage of the dynamic regime for CO oxidation was previously demonstrated by different authors, both experimentally and theoretically. Safonova et al. have found an increase in CO_2 production in the gas cutoff experiments and studied the dynamic response of the catalyst via X-ray absorption spectroscopy.[68,69] Machado et al. have studied CO_2 production rate in the ZGB Monte Carlo

environment with periodic variation of CO pressure.[70] They observed a phase transition like the one observed in the Ising model and found that catalytic activity increased if the rate of decontamination was different from the rate of contamination. Parameterization of the periodic perturbation is a common approach to finding advantageous nonstationary regimes. Lopez et al.[71] investigated the response of the kinetic Monte Carlo model to periodic variations of inlet gas feed varying both the amplitude of oscillations and the mean value of CO concentration. By reducing the amplitude of oscillations, the authors concluded that the regime with high oscillation amplitude was preferential over the regime with zero amplitude; however, they did not report the value of the global maximum of conversion in the steady-state regime; therefore, it was not clear if their strategy could outperform the optimal stationary regime with respect to an oscillatory one or not. Ardagh et al.[72] studied the effect of dynamical changes in the support on catalytical activity. Their numeric model showed that periodic variations in the surface binding energy lead to the catalyst turnover frequency increase by more than 1000 times above the Sabatier maximum. The authors limited their choice of perturbation to square, sinusoidal, sawtooth, and triangular waveforms. The need to choose the parametrization for the periodic regime is a limitation for classical optimization algorithms that can be overcome by the reinforcement learning approach.

The pretrained RL algorithm has been demonstrated to be preferential over gradient descent approaches when applied to a similar environment where it was trained.[55] The main advantage was the fewer number of steps needed to converge to the optimal solution. Along with that, several authors
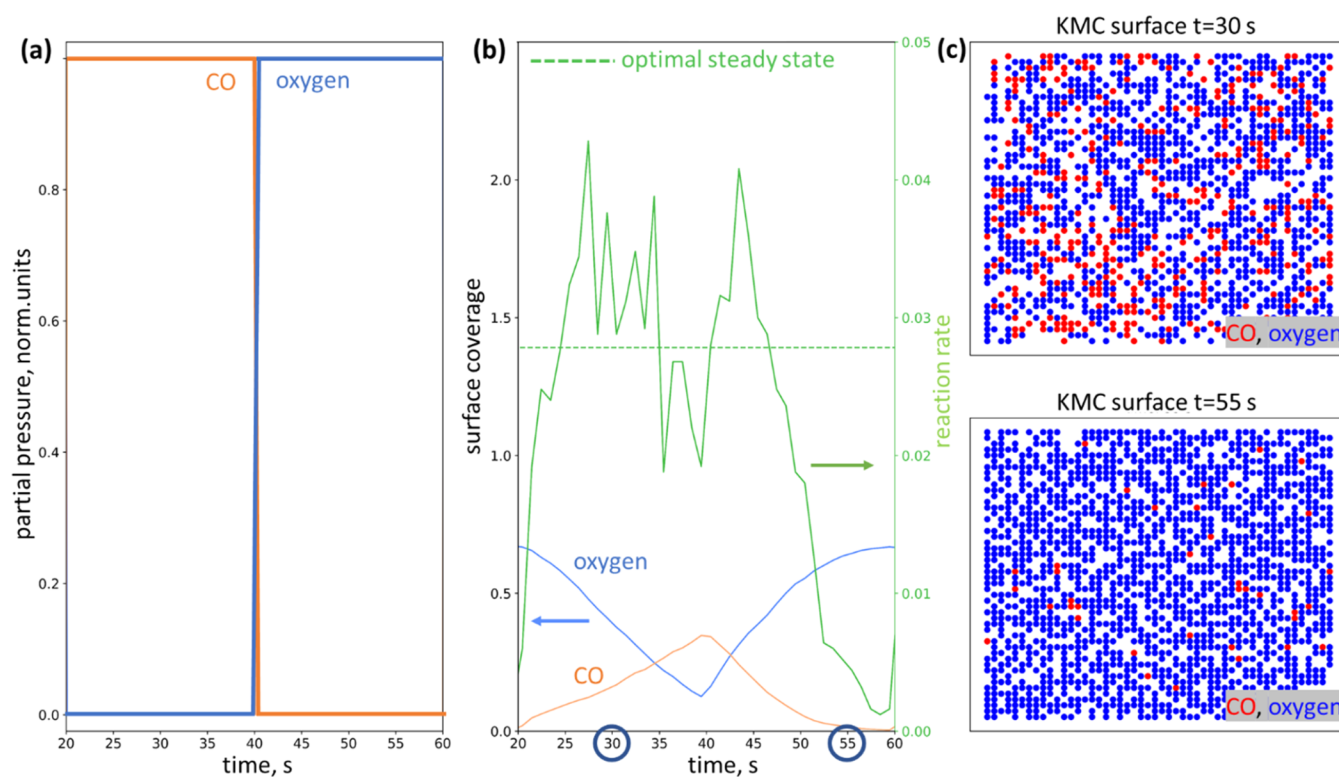
Figure 9. Validation of the dynamic regime in the kinetic Monte Carlo model. Upon replacing the CO atmosphere with O_2 atm (panel (a)), the reaction rate can increases and becomes higher than under optimal steady-state conditions (panel (b)). (c) CO and oxygen occupation of the surface sites at selected times of high ($t = 30$ s) and low ($t = 55$ s) conversions.

applied RL for optimizing steady reaction conditions such as Neumann et al., who maximized H_2 production by partial oxidation of methane in a simulated plug flow reactor. To address the dynamic control, Alhazmi et al. have combined economic model predictive control framework and RL, and their framework allowed control, optimization, and model correction to be performed online and continuously.[54]

In our work, we rely solely on the reinforcement learning methodology and obtained stationary and nonstationary control regimes. We also applied RL to the problem of adaptation of the dynamic control regime to the external variations and showed that the agent was able to solve this problem successfully. However, the main drawback of the methodology is the large number of trial steps required for training (e.g., 10,000 episodes as shown in Figure 2, each 30 s long). Therefore, the practical application of the algorithm relies on the combination of a theoretical environment and experimental setup. In the first step, the theoretical model is suggested for the studied system (see e.g., eqs 1−8). The sample is then exposed experimentally to a broad range of reaction conditions (temperature, gas flows), and its catalytical properties are registered. Parameters of the theoretical model are fitted to experimental data, and the algorithm is trained on the basis of such an experimentally verified model. In the second step, the algorithm may be retrained with experimental setup only. Such a procedure requires a smaller number of steps for tuning, while the second step may help to account for the parameters of the experimental system beyond the theoretical model behind it.

In Figure 9, we show that discussion about higher reaction rates in the dynamic regime extend also to more complicated models. We use a kinetic Monte Carlo approach (see Section S3 for details) to overcome mean-field approximation imposed by ODE's model and demonstrate increase in CO_2 formation upon replacing the reducing atmosphere with an oxidative one.

The demonstrated RL methodology could be easily applied to any catalytic or, more generally, chemical process, as long as the latter could be described by the numerical model. A few straightforward changes would be required: update the model via a new set of kinetic parameters; extended differential equations or reparametrized KMC model; specify a set of variables monitored by algorithm; and specify a target for optimization.

4. CONCLUSIONS

A reinforcement learning algorithm was applied to find the optimal dynamic regime for the CO oxidation reaction on the Pd surface. The benefit of such a methodology is the absence of any approximation on the analytical form of the optimal solution and its periodicity. The two problem statements were considered. In the first one, the algorithm varied both CO and O_2 partial pressures. Each time step the Vanilla Policy Gradient agent received the CO and O_2 pressures and CO_2 formation rate for the current and previous time steps and set the CO and O_2 pressures for the next step. We demonstrate a series of optimal solutions ranging from stationary to quasi-periodic depending on the rate constants of the numerical model of the reaction. The dynamic regime was preferable for the low reaction rate when alternating cycles of surface deactivation and recovery could significantly enhance the CO_2 integrated output. Under another problem statement, the algorithm could control only O_2, while CO feed was supplied from external source randomly. For such task the reward was reformulated and the trained algorithm was able to adjust the O_2 pressure

regime to the CO regimes of various forms, even on time intervals longer than those of the algorithm was trained on. The main bottleneck of the described RL approach is the time required for neural net training. In the current example, the training was completed during 3 ca. 0,000 episodes each 30 s long. Therefore, the practical scheme should include training on the theoretical model derived from the series of catalytic tests. Subsequently, the model can be tuned in a fewer number of steps within experimental setup. We foresee that such a methodology can be applied to the relevant industrial reactions.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code for RL training and related data can be downloaded free of charge from https://github.com/MikhailLifar/ReactRL.git

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c10422.

> Kinetic parameters of the mathematical model of CO oxidation reaction; definition of the advantage function for RL; and kinetic Monte Carlo model description (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Mikhail S. Lifar — *The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia*; Email: lifar@sfedu.ru

Alexander A. Guda — *The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia;* ⓞ orcid.org/0000-0002-6941-4987; Email: guda@sfedu.ru

Authors

Andrei A. Tereshchenko — *The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia*

Aleksei N. Bulgakov — *The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia*

Sergey A. Guda — *The Smart Materials Research Institute and Institute for Mathematics, Mechanics and Computer Science in the name of I.I. Vorovich, Southern Federal University, 344090 Rostov-on-Don, Russia*

Alexander V. Soldatov — *The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.3c10422

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Pakhare, D.; Spivey, J. A Review of Dry (Co2) Reforming of Methane over Noble Metal Catalysts. *Chem. Soc. Rev.* **2014**, *43*, 7813−7837.

(2) Pareek, V.; Bhargava, A.; Gupta, R.; Jain, N.; Panwar, J. Synthesis and Applications of Noble Metal Nanoparticles: A Review. *Adv. Sci., Eng. Med.* **2017**, *9*, 527−544.

(3) Docherty, S. R.; Phongprueksathat, N.; Lam, E.; Noh, G.; Safonova, O. V.; Urakawa, A.; Copéret, C. Silica-Supported Pdga Nanoparticles: Metal Synergy for Highly Active and Selective Co2-to-Ch3oh Hydrogenation. *JACS Au* **2021**, *1*, 450−458.

(4) Sadykov, I. I.; Zabilskiy, M.; Clark, A. H.; Krumeich, F.; Sushkevich, V.; van Bokhoven, J. A.; Nachtegaal, M.; Safonova, O. V. Time-Resolved Xas Provides Direct Evidence for Oxygen Activation on Cationic Iron in a Bimetallic Pt-Feo X/Al2o3 Catalyst. *ACS Catal.* **2021**, *11*, 11793−11805.

(5) Soliman, N. Factors Affecting Co Oxidation Reaction over Nanosized Materials: A Review. *J. Mater. Res. Technol.* **2019**, *8*, 2395−2407.

(6) Gromotka, Z.; Yablonsky, G.; Ostrovskii, N.; Constales, D. Three-Factor Kinetic Equation of Catalyst Deactivation. *Entropy* **2021**, *23*, No. 818, DOI: 10.3390/e23070818.

(7) Pongstabodee, S.; Monyanon, S.; Luengnaruemitchai, A. Applying a Face-Centered Central Composite Design to Optimize the Preferential Co Oxidation over a Ptau/Ceo2−Zno Catalyst. *Int. J. Hydrogen Energy* **2012**, *37*, 4749−4761.

(8) Saidi, W. A. Optimizing the Catalytic Activity of Pd-Based Multinary Alloys toward Oxygen Reduction Reaction. *J. Phys. Chem. Lett.* **2022**, *13*, 1042−1048.

(9) Armstrong, C. D.; Teixeira, A. R. Advances in Dynamically Controlled Catalytic Reaction Engineering. *React. Chem. Eng.* **2020**, *5*, 2185−2203.

(10) Unni, M. P.; Hudgins, R.; Silveston, P. Symposium on Application of Pulse Techniques to Kinetic Measurements: Influence of Cycling on the Rate of Oxidation of So2 over a Vanadia Catalyst. *Can. J. Chem. Eng.* **1973**, *51*, 623−629.

(11) Abdul-Kareem, H. K.; Silveston, P.; Hudgins, R. Forced Cycling of the Catalytic Oxidation of Co over a V2o5 Catalyst—I: Concentration Cycling. *Chem. Eng. Sci.* **1980**, *35*, 2077−2084.

(12) Abdul-Kareem, H. K.; Hudgins, R.; Silveston, P. Forced Cycling of the Catalytic Oxidation of Co over a V2o5 Catalyst—Ii Temperature Cycling. *Chem. Eng. Sci.* **1980**, *35*, 2085−2088.

(13) Zhou, X.; Barshad, Y.; Gulari, E. Co Oxidation on Pd/Al2o3. Transient Response and Rate Enhancement through Forced Concentration Cycling. *Chem. Eng. Sci.* **1986**, *41*, 1277−1284.

(14) Hegedus, L. L.; Chang, C. C.; McEwen, D. J.; Sloan, E. M. Response of Catalyst Surface Concentrations to Forced Concentration Oscillations in the Gas Phase. The No, Co, O2 System over Pt-Alumina. *Ind. Eng. Chem. Fundam.* **1980**, *19*, 367−373.

(15) Thullie, J.; Chiao, L.; Rinker, R. G. Analysis of Concentration Forcing in Heterogeneous Catalysis. *Chem. Eng. Commun.* **1986**, *48*, 191−205.

(16) Thullie, J.; Renken, A. Forced Concentration Oscillations for Catalytic Reactions with Stop-Effect. *Chem. Eng. Sci.* **1991**, *46*, 1083−1088.

(17) Hamer, J. W.; Cormack, D. E. Influence of Oscillating External Pressure on Gas-Phase Reactions in Porous Catalysts. *Chem. Eng. Sci.* **1978**, *33*, 935−944.

(18) Brandner, J.; Emig, G.; Liauw, M.; Schubert, K. Fast Temperature Cycling in Microstructure Devices. *Chem. Eng. J.* **2004**, *101*, 217−224.

(19) Jensen, S.; Olsen, J. L.; Thorsteinsson, S.; Hansen, O.; Quaade, U. J. Forced Thermal Cycling of Catalytic Reactions: Experiments and Modelling. *Catal. Commun.* **2007**, *8*, 1985−1990.

(20) Luther, M.; Brandner, J. J.; Kiwi-Minsker, L.; Renken, A.; Schubert, K. Forced Periodic Temperature Cycling of Chemical Reactions in Microstructure Devices. *Chem. Eng. Sci.* **2008**, *63*, 4955−4961.

(21) Stolte, J.; Özkan, L.; Thüne, P.; Niemantsverdriet, J.; Backx, A. Pulsed Activation in Heterogeneous Catalysis. *Appl. Therm. Eng.* **2013**, *57*, 180−187.

(22) Pschenitza, M.; Meister, S.; von Weber, A.; Kartouzian, A.; Heiz, U.; Rieger, B. Suppression of Deactivation Processes in Photocatalytic Reduction of Co2 Using Pulsed Light. *ChemCatChem* **2016**, *8*, 2688−2695.

(23) Schüth, F.; Henry, B.; Schmidt, L. D. Oscillatory Reactions in Heterogeneous Catalysis. *Adv. Catal.* **1993**, *39*, 51−127.

(24) Balmes, O.; Resta, A.; Wermeille, D.; Felici, R.; Messing, M. E.; Deppert, K.; Liu, Z.; Grass, M. E.; Bluhm, H.; van Rijn, R.; et al. Reversible Formation of a Pdc X Phase in Pd Nanoparticles Upon Co and O 2 Exposure. *Phys. Chem. Chem. Phys.* **2012**, *14*, 4796−4801.

(25) Bugaev, A. L.; Usoltsev, O. A.; Guda, A. A.; Lomachenko, K. A.; Pankin, I. A.; Rusalev, Y. V.; Emerich, H.; Groppo, E.; Pellegrini, R.; Soldatov, A. V.; et al. Palladium Carbide and Hydride Formation in the Bulk and at the Surface of Palladium Nanoparticles. *J. Phys. Chem. C* **2018**, *122*, 12029−12037.

(26) Makeev, A. G.; Slinko, M. M.; Luss, D. Mathematical Modeling of Oscillating Co Oxidation on Pt-Group Metals at near Atmospheric Pressure: Activity of Metallic and Oxidized Surfaces. *Appl. Catal., A* **2019**, *571*, 127−136.

(27) Bugaev, A. L.; Guda, A. A.; Pankin, I. A.; Groppo, E.; Pellegrini, R.; Longo, A.; Soldatov, A. V.; Lamberti, C. The Role of Palladium Carbides in the Catalytic Hydrogenation of Ethylene over Supported Palladium Nanoparticles. *Catal. Today* **2019**, *336*, 40−44.

(28) Cutlip, M.; Hawkins, C.; Mukesh, D.; Morton, W.; Kenney, C. Modelling of Forced Periodic Oscillations of Carbon Monoxide Oxidation over Platinum Catalyst. *Chem. Eng. Commun.* **1983**, *22*, 329−344.

(29) Vaporciyan, G.; Annapragada, A.; Gulari, E. Rate Enhancements and Quasi-Periodic Dynamics During Forced Concentration Cycling of Co and O2 over Supported Pt— Sno2. *Chem. Eng. Sci.* **1988**, *43*, 2957−2966.

(30) Schwankner, R. J.; Eiswirth, M.; Möller, P.; Wetzl, K.; Ertl, G. Kinetic Oscillations in the Catalytic Co Oxidation on Pt (100): Periodic Perturbations. *J. Chem. Phys.* **1987**, *87*, 742−749.

(31) Eiswirth, M.; Ertl, G. Forced Oscillations of a Self-Oscillating Surface Reaction. *Phys. Rev. Lett.* **1988**, *60*, No. 1526, DOI: 10.1103/PhysRevLett.60.1526.

(32) Newton, M. A.; Ferri, D.; Smolentsev, G.; Marchionni, V.; Nachtegaal, M. Room-Temperature Carbon Monoxide Oxidation by Oxygen over Pt/Al2o3Mediated by Reactive Platinum Carbonates. *Nat. Commun.* **2015**, *6*, No. 8675, DOI: 10.1038/ncomms9675.

(33) Fang, H.; Haibin, L.; Zengli, Z. Advancements in Development of Chemical-Looping Combustion: A Review. *Int. J. Chem. Eng.* **2009**, *2009*, 1−16, DOI: 10.1155/2009/710515.

(34) Moghtaderi, B. Review of the Recent Chemical Looping Process Developments for Novel Energy and Fuel Applications. *Energy Fuels* **2012**, *26*, 15−40.

(35) Yoshida, H.; Kakei, R.; Fujiwara, A.; Tomita, A.; Miki, T.; Machida, M. In Situ Time-Resolved Redox Dynamics of Pd Catalysts under Oscillating a/F Conditions. *Top. Catal.* **2019**, *62*, 345−350.

(36) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.

(37) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic Ai. *Nature* **2018**, *555*, 604−610.

(38) Kaelbling, L. P.; Littman, M. L.; Moore, A. W. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **1996**, *4*, 237−285.

(39) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*; MIT press, 2018.

(40) Littman, M. L. Reinforcement Learning Improves Behaviour from Evaluative Feedback. *Nature* **2015**, *521*, 445−451.

(41) Naeem, M.; Rizvi, S. T. H.; Coronato, A. A Gentle Introduction to Reinforcement Learning and Its Application in Different Fields. *IEEE Access* **2020**, *8*, 209320−209344.

(42) Dayan, P.; Niv, Y. Reinforcement Learning: The Good, the Bad and the Ugly. *Curr. Opin. Neurobio.* **2008**, *18*, 185−196.

(43) Brunke, L.; Greeff, M.; Hall, A. W.; Yuan, Z.; Zhou, S.; Panerati, J.; Schoellig, A. P. Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning. 2021, arXiv:2108.06266. arXiv.org e-Print archive. https://arxiv.org/abs/2108.06266.

(44) Ibarz, J.; Tan, J.; Finn, C.; Kalakrishnan, M.; Pastor, P.; Levine, S. How to Train Your Robot with Deep Reinforcement Learning: Lessons We Have Learned. *Int. J. Rob. Res.* **2021**, *40*, 698−721.

(45) Kober, J.; Bagnell, J. A.; Peters, J. Reinforcement Learning in Robotics: A Survey. *Int. J. Rob. Res.* **2013**, *32*, 1238−1274.

(46) Luong, N. C.; Hoang, D. T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.-C.; Kim, D. I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Commun. Surv. Tutorials* **2019**, *21*, 3133−3174.

(47) Tejedor, M.; Woldaregay, A. Z.; Godtliebsen, F. Reinforcement Learning Application in Diabetes Blood Glucose Control: A Systematic Review. *Artif. Intell. Med.* **2020**, *104*, No. 101836.

(48) Zheng, G.; Zhang, F.; Zheng, Z.; Xiang, Y.; Yuan, N. J.; Xie, X.; Li, Z. In *Drn: A Deep Reinforcement Learning Framework for News Recommendation*, Proceedings of the 2018 World Wide Web Conference; ACM, 2018; pp 167−176.

(49) Yoon, J.; Cao, Z.; Raju, R. K.; Wang, Y.; Burnley, R.; Gellman, A. J.; Farimani, A. B.; Ulissi, Z. W. Deep Reinforcement Learning for Predicting Kinetic Pathways to Surface Reconstruction in a Ternary Alloy. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 045018, DOI: 10.1088/2632-2153/ac191c.

(50) Neumann, M.; Palkovits, D. S. Reinforcement Learning Approaches for the Optimization of the Partial Oxidation Reaction of Methane. *Ind. Eng. Chem. Res.* **2022**, *61*, 3910−3916.

(51) Watkins, C. J. C. H. Learning from Delayed Rewards. 1989.

(52) Watkins, C. J. C. H.; Dayan, P. Q-Learning. *Mach. Learn.* **1992**, *8*, 279−292.

(53) Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. 2015, arXiv:1509.02971. arXiv.org e-Print archive. https://arxiv.org/abs/1509.02971.

(54) Alhazmi, K.; Albalawi, F.; Sarathy, S. M. A Reinforcement Learning-Based Economic Model Predictive Control Framework for Autonomous Operation of Chemical Reactors. *Chem. Eng. J.* **2022**, *428*, No. 130993.

(55) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337−1344.

(56) Nelder, J. A.; Mead, R. A Simplex Method for Function Minimization. *Compt. J.* **1965**, *7*, 308−313.

(57) Powell, M. J. D. An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives. *Compt. J.* **1964**, *7*, 155−162.

(58) Zhou, Y.; Wang, Z.; Liu, C. Perspective on Co Oxidation over Pd-Based Catalysts. *Catal. Sci. Technol.* **2015**, *5*, 69−81.

(59) Wu, J.; Chen, D.; Chen, J.; Wang, H. Structural and Composition Evolution of Palladium Catalyst for Co Oxidation under Steady-State Reaction Conditions. *J. Phys. Chem. C* **2023**, *127*, 6262−6270.

(60) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*; Wiley, 2017.

(61) Engel, T.; Ertl, G. Elementary Steps in the Catalytic Oxidation of Carbon Monoxide on Platinum Metals. In *Advances in Catalysis*; Academic Press, 1979; pp 1−78.

(62) Libuda, J.; Meusel, I.; Hoffmann, J.; Hartmann, J.; Piccolo, L.; Henry, C. R.; Freund, H. J. The Co Oxidation Kinetics on Supported Pd Model Catalysts: A Molecular Beam/in Situ Time-Resolved Infrared Reflection Absorption Spectroscopy Study. *J. Chem. Phys.* **2001**, *114*, 4669−4684, DOI: 10.1063/1.1342240.

(63) https://spinningup.openai.com/en/latest/algorithms/vpg.html (accessed August 01, 2022).

(64) Williams, R. J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* **1992**, *8*, 229−256.

(65) Schulman, J.; Moritz, P.; Levine, S.; Jordan, M. I.; Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. 2015. arXiv:1506.02438. arXiv.org e-Print archive. https://arxiv.org/abs/1506.02438.

(66) Sutton, R. S.; McAllester, D.; Singh, S.; Mansour, Y. In *Policy Gradient Methods for Reinforcement Learning with Function Approximation*, Proceedings of the 12th International Conference on Neural Information Processing Systems; MIT Press: Denver, CO, 1999; pp 1057−1063.

(67) Schulman, J.; Moritz, P.; Levine, S.; Jordan, M. I.; Abbeel, P., High-Dimensional Continuous Control Using Generalized Advantage Estimation. 2016.

(68) Safonova, O. V.; Guda, A.; Rusalev, Y.; Kopelent, R.; Smolentsev, G.; Teoh, W. Y.; van Bokhoven, J. A.; Nachtegaal, M. Elucidating the Oxygen Activation Mechanism on Ceria-Supported Copper-Oxo Species Using Time-Resolved X-Ray Absorption Spectroscopy. *ACS Catal.* **2020**, *10*, 4692−4701.

(69) Sadykov, I. I.; Zabilskiy, M.; Clark, A. H.; Krumeich, F.; Sushkevich, V.; van Bokhoven, J. A.; Nachtegaal, M.; Safonova, O. V. Time-Resolved Xas Provides Direct Evidence for Oxygen Activation on Cationic Iron in a Bimetallic Pt-Feox/Al2o3 Catalyst. *ACS Catal.* **2021**, *11*, 11793−11805.

(70) Machado, E.; Buendía, G. M.; Rikvold, P. A.; Ziff, R. M. Response of a Catalytic Reaction to Periodic Variation of the Co Pressure: Increased CO_2 Production and Dynamic Phase Transition. *Phys. Rev. E* **2005**, *71*, No. 016120.

(71) López, A. C.; Albano, E. V. Dynamic Response of an Irreversible Catalytic Reaction to Periodic Variation of the Reactant's Pressure. *J. Chem. Phys.* **2000**, *112*, 3890−3896.

(72) Ardagh, M. A.; Birol, T.; Zhang, Q.; Abdelrahman, O. A.; Dauenhauer, P. J. Catalytic Resonance Theory: Supervolcanoes, Catalytic Molecular Pumps, and Oscillatory Steady State. *Catal. Sci. Technol.* **2019**, *9*, 5058−5076.