https://doi.org/10.21872/2024IISE_7645

Proceedings of the IISE Annual Conference & Expo 2024
A. Brown Greer, C. Contardo, J.-M. Frayret, eds.

Towards a Robust Deep Reinforcement Learning for Optimization of Heating Setup in Thermoforming Process

Abstract ID: 7645

Iman Jalilvand, Bhushan Gopaluni, Abbas S. Milani*
Materials and Manufacturing Research Institute, The University of British Columbia, Canada

Abstract

Thermoforming, a commonly used technique in thermoplastics and composites manufacturing, involves interaction of various mechanical components influencing the quality of the final product. Among those, a precise selection of heaters setting plays a pivotal role in pre-optimizing the process. While the traditional control theories have been historically employed for process optimization problems, recent advancements in Artificial Intelligence (AI) have encouraged its adoption across diverse manufacturing domains. Nevertheless, the AI application in thermoforming remains rather limited to date. This case study harnesses a Deep Reinforcement Learning (DRL) to enhance the thermoforming's primary operation: optimizing the input heating setting given a target temperature profile, and possibly saving energy consumption. We showcase the implementation of two widely used DRL algorithms: Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN). A comparison is made to evaluate their effectiveness under diverse process conditions, including both low and high temperature profiles along with single- versus multi-heater utilizations. The PPO algorithm demonstrated a higher performance across all simulated conditions.

Keywords
Deep Reinforcement Learning; Proximal Policy Optimization; Deep Q-Networks; Thermoforming; Process Optimization.

1. Introduction
The integration of Artificial Intelligence (AI) in real-world problems have now become a standard practice, extending from manufacturing sectors to healthcare systems, education, and beyond. Such applications aim to achieve enhanced performance in addressing diverse challenges. The manufacturing sector has particularly gained advantages from AI implementation, leveraging it to create intelligent and optimized process solutions that result in cost, energy, and time savings. Among the different sectors of manufacturing, composites have gained significant attention in recent years due to their capacity to offer customizable mechanical properties with low structural weight—a feature highly sought after by industries such as aerospace, automotive, robotics, healthcare, and more [1-3]. In composites manufacturing domain, thermoforming is a widely utilized technique owing to its versatility and effectiveness in shaping thermoplastic and composite materials to complex 3D shapes. It is particularly advantageous for rapid prototyping, facilitating swift design iterations. Consequently, optimizing this process continues to hold the potential to enhance performance and reduce the costs associated with a number of advanced composites manufacturing sectors [4].

2. Synopsis of Related Research
Reinforcement Learning (RL)-based control, as a fairly recently emerged AI domain, has proven to be robust and efficient compared to traditional control theories. Arroyo et al. [5] introduced Reinforced Model Predictive Control (RL-MPC), enhancing the adaptability of Model Predictive Control (MPC) for dynamic energy control problems. Brandi et al. [6] compared online and offline Deep Reinforcement Learning (DRL) with MPC for energy management, emphasizing the efficiency of online-trained DRL agents. Gupta et al. [7] applied DRL to heating control in smart buildings, improving thermal comfort and reducing energy costs. Wang et al. [8] used DRL for forced convection control, achieving a lower temperature with a novel value-based deep Q-network (DQN). Hachem et al. [9] utilized Proximal Policy Optimization (PPO) in a Computational Fluid Dynamic (CFD) environment for conjugate heat transfer systems. Römer et al. [10] employed a DRL method called Deep Deterministic Policy Gradient (DDPG) for temperature control in Automated Tape Laying (ATL) processes. Zhao et al. [11] introduced a memory-augmented (MA) DRL algorithm for energy management in commercial buildings with dueling networks to mitigate time delays.

https://doi.org/10.21872/2024IISE_7645

Szarski et al. [12] optimized Carbon Fibre Reinforced Plastic (CFRP) manufacturing cycle time using DRL, reducing cycle time in aerospace parts.

3. Problem Description and objective

Despite the prevailing trend of employing AI to optimize manufacturing and industrial processes, to date its usage in thermoforming has been relatively limited. This research aims to employ Deep Reinforcement Learning (DRL) to fine-tune the heating parameters of a lab-scale thermoforming setup and achieve the desired temperature distribution on the thermoplastic sheet.

4. Methodology

In a typical thermoforming process, a thermoplastic sheet is heated and shaped into a mold using vacuum pressure. The goal is to achieve a uniform temperature distribution throughout the sheet for an optimal (high quality) forming. Controlling the temperature in specific regions of the material is critical to ensure defect-free product with consistent thickness. However, the temperature on the sheet varies over time, adding complexity to the process control [4]. Figure 1 shows a lab-scale thermoforming setup and its components that were used in this case study.

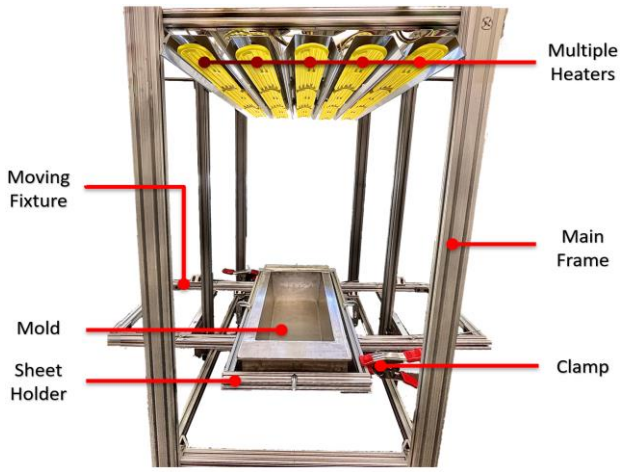


Figure 1: Lab-scale thermoforming setup including fifteen heaters and other components.

Reinforcement Learning (RL) was employed as an optimization tool to address the sequential decision-making nature of the process. RL involves two components: the agent and the environment. At every time step t, the agent observes a state s_t within the state space $\boldsymbol{S}$, selects an action a_t from the action space $\boldsymbol{A}$ according to the policy $\pi(a_t\,|s_t)$ (defining the agent's behavior), receives a scalar reward r_t, and moves to the subsequent state s_{t+1}. These transitions follow the dynamics of the environment, determined by the reward function $\boldsymbol{R}(s,a)$ and state transition probability $\boldsymbol{P}(s_{t+1}\,|s_t,a_t)$. In episodic problems, this process continues cyclically until the agent reaches a terminal state, restarting thereafter [13]. First, a heat transfer simulation tool was developed, which acts as the environment for the agent to interact with, similar to what is being implemented in actual process, while accounting for conduction, convection, and radiation mechanisms [14]. The details of the model are described in a previous work by Jalilvand et al. [15]. Then, the RL agent was trained using two main algorithms, Deep Q-Network (DQN) and Proximal Policy Optimization (PPO). Lastly, their performances have been compared under various process conditions (to be outlined in section 5).

4.1. Deep-Q Networks (DQN)

Q-learning stands as a foundational algorithm in RL, strategically crafted to empower agents in learning optimal policies within environments characterized by discrete state and action spaces. The "Q" in Q-learning symbolizes the quality of an action in each state. The algorithm's objective is to iteratively update Q-values based on received rewards and the maximum expected future rewards. The underlying concept revolves around learning a Q-function, estimating cumulative future rewards for each state-action pair. In the context of DRL, a Deep Q-Network (DQN) emerges as a neural network tasked with estimating a state-value function. This architecture often leverages Experience Replay,

https://doi.org/10.21872/2024IISE_7645

storing episode steps in memory for off-policy learning through randomly selected samples from the replay memory. By averaging the behavior distribution over previous states, Experience Replay safeguards against oscillations or divergence in parameters. DQN outshines standard online Q-learning by enhancing data efficiency, potentially using each experience step in numerous weight updates. It further limits variance through sample randomization and avoids undesired feedback loops linked with on-policy learning [16].

4.2. Proximal Policy Optimization (PPO)

PPO is inspired by the dilemma of how to make the largest potential improvement to a policy based on the existing data without mistakenly causing efficiency drop. PPO is a series of first-order approaches that employ a few additional twists to maintain the similarity of new and old policies [17]. Two versions of PPO have been commonly used in the literature: 1) PPO-Penalty and 2) PPO-Clip. PPO-Penalty algorithm changes the penalty factor autonomously during training to ensure that it is suitably calibrated, while PPO-Clip uses customized clipping in the objective function to eliminate opportunities for the new policy to diverge from the old policy. PPO constructs a probabilistic policy by utilizing the latest iteration of its stochastic strategy. The degree to which action selection is random is dependent on both the baseline circumstances and the training technique. Typically, the policy grows less random over time as the update rule pushes it to exploit previously discovered incentives. PPO algorithm has two distinguishing characteristics: 1) It is a policy-driven algorithm. 2) It is applicable to contexts with discrete or continuous action spaces [17]. Figure 2 represents the designed block diagram in MATLAB synchronized with the above-mentioned heat transfer simulation tool as the environment, used for both RL models.

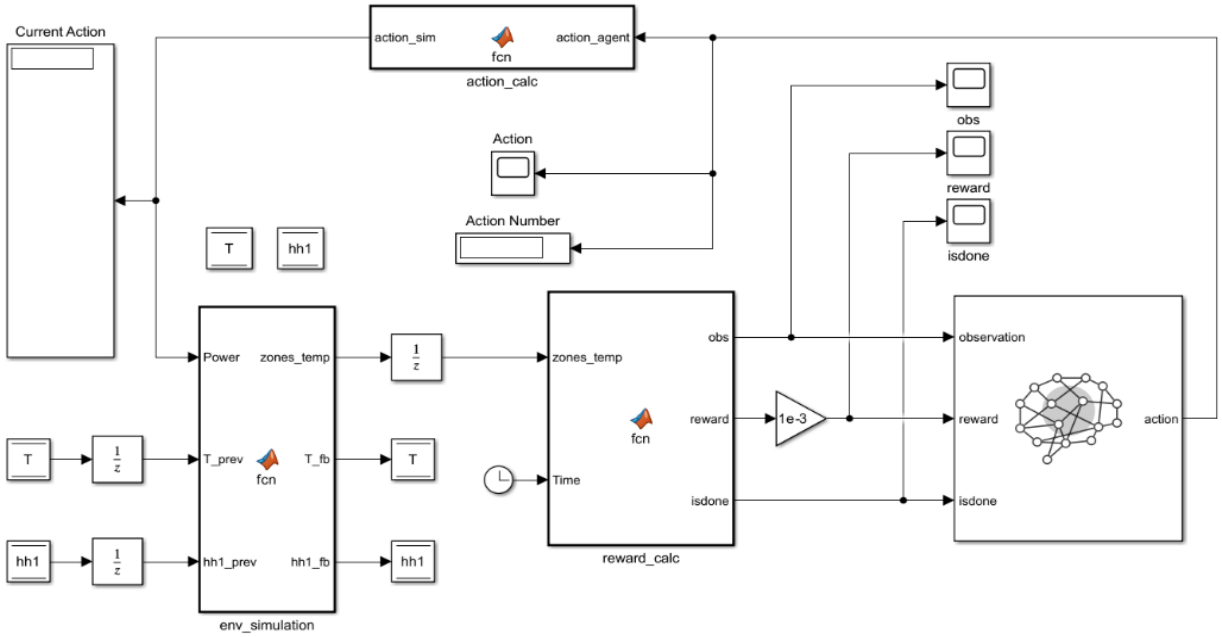


Figure 2: MATLAB Simulink Block Diagram designed for the DRL model.

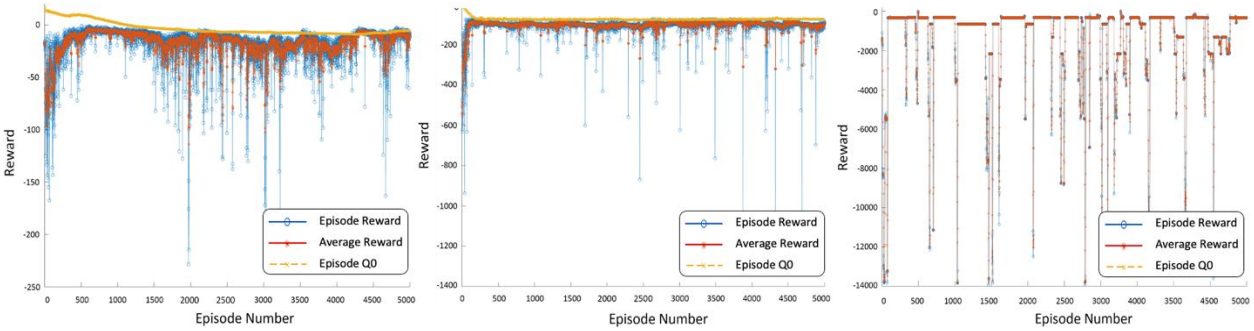
5. Results and Discussion

To assess the efficiency of PPO and DQN algorithms in the simulated thermoforming control problem, a processing scenario was configured with a single heater as the base case. The algorithms were compared under three distinct conditions:

- Fixed target temperature (50 ℃)
  - Large action space (9 possible actions)
  - Small action space (3 possible actions)
- Random variable target temperature via small action space (50 ± 5℃)
  - Uniformly distributed and constantly changing in every episode.

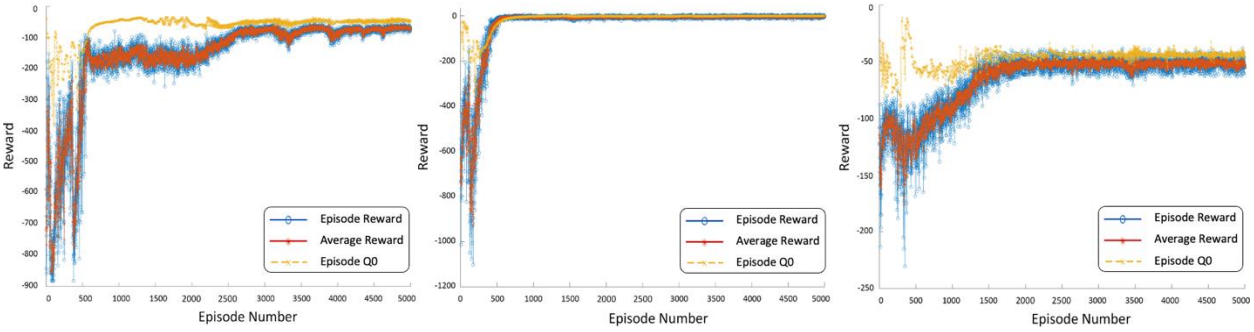
https://doi.org/10.21872/2024IISE_7645

Given the practical operating power range of the ceramic heaters in the experimental setup (150 W to 500 W), two action (search) spaces/scenarios were assumed: small and large spaces. The small space included three modes: 0 W (heater off), 150 W (minimum power), and 500 W (maximum power). The large space had nine modes, represented as {0, 150, 200, 250, 300, 350, 400, 450, 500}. Figures 3a and 3b illustrate DQN's performance in optimizing heater settings, with cumulative rewards over 5000 training episodes for 9 and 3 action spaces. Similarly, Figures 4a and 4b show results for the PPO agent, with the 3-action space configuration proving more robust, consistently converging within around 500 episodes compared to DQN. Testing a random variable target temperature (50±5°C) with smaller action space, Figures 3c and 4c demonstrate that PPO was consistently outperforming DQN in terms of convergence and stability. Q0 in Figures 3 and 4 represents the expected cumulative reward at each episode based on prior state-action pairs and policies.



a) b) c)

Figure 3: DQN's reward against episode number. **a**) 9 actions space, **b**) 3 action space, and **c**) random target temperature space.

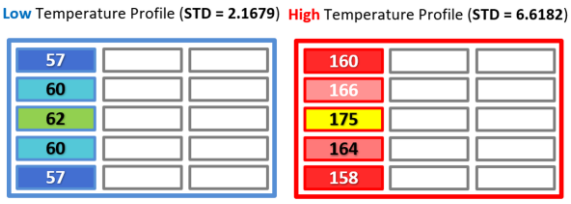


a) b) c)

Figure 4: PPO's reward against episode numbers. **a**) 9 actions space, **b**) 3 action space, and **c**) random target temperature space.

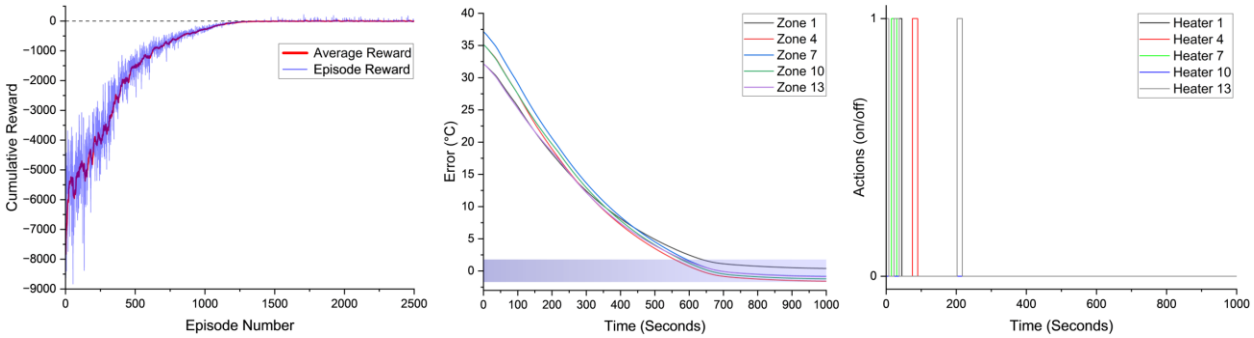
PPO proved to be, in this case study, the more optimal choice for training the agent, based on the achieved performance on a single heater scenario. Therefore, the model was then opted to extend to control temperature distribution on the full left side of the thermoplastic sheet, now using five heaters. Figure 5a shows a hypothetical low-temperature desired profile ($57-62℃, \mu = 59.2, \sigma = 2.2$), while Figure 5b displays a high-temperature desired profile ($158-175℃, \mu = 164.6, \sigma = 6.6$); but assumed to be without randomness. To enhance the agent's performance, the refined action space included a complete deactivation (0 W) and a complete activation at maximum power (500 W). The PPO agent was trained for 2500 episodes, with each episode simulating the complete thermoforming process lasting 1000 seconds. As illustrated in Figure 6a, the agent reached the peak cumulative reward after nearly 1200 episodes and achieved a minimal error band (difference between the desired temperature distribution and the actual temperature distribution) of $\pm 2℃$ after 650 seconds into the process, as depicted in Figure 6b. Finally, Figure 6c provides a visual representation of the actions taken by the agent at each time step throughout the entire process.

https://doi.org/10.21872/2024IISE_7645



a) b)

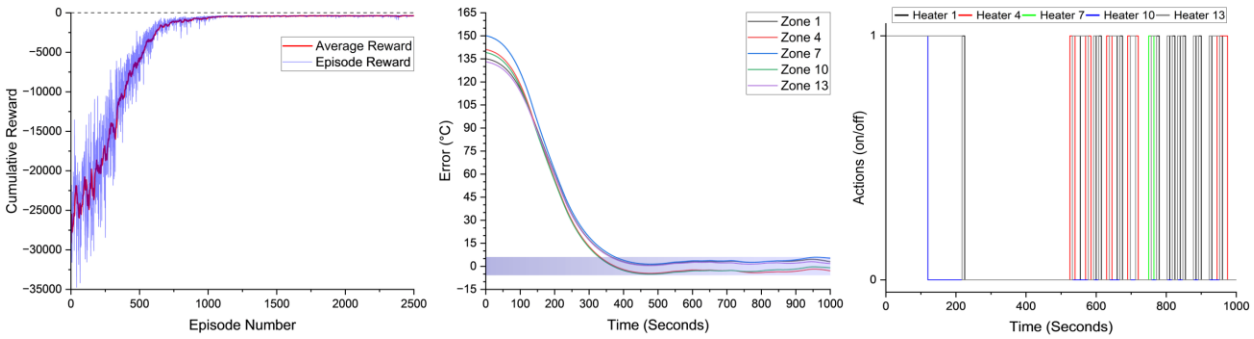
Figure 5: **a**) Low temperature profile, and **b**) high temperature profile selected as desired distribution.



a) b) c)

Figure 6: **a**) Cumulative reward, **b**) error, and **c**) actions in the low desired temperature profile.

Leveraging the success of the preceding training phase, the satisfactory results encouraged a subsequent trial involving a higher temperature profile with an increased standard deviation (Figure 5b). This was intentionally aimed to evaluate the agent's performance in a more challenging scenario. The latter training outcomes are depicted in Figure 7. In this case, the agent has achieved the maximum cumulative reward after approximately 1000 episodes, maintaining an error band of ±7℃ across all five zones within 400 seconds into the process, as indicated in Figure 7a and 7b, respectively. Given the elevated temperature profile, the heaters exhibited a higher activation frequency than that observed in the low-temperature profile (compare Figure 6c and Figure 7c).



a) b) c)

Figure 7: **a**) Cumulative reward, **b**) error, and **c**) actions in the high desired temperature profile.

6. Conclusions

This study explored the application of Reinforcement Learning (RL) in thermoforming. The DQN and PPO algorithms were first employed and compared for optimizing a single heater power setting. PPO exhibited superior performance under varying test conditions, including different action spaces and desired temperature profiles. The model was then extended to control the full column of five heaters (across one side of the sheet), with PPO retrained for simultaneous optimization. Results showcased the DRL model achieving a ±2℃ error band within 600 seconds under a low-temperature profile, and ±7℃ in a high-temperature profile, maintained after only 400 seconds. Training on a high-performance PC (10-core Xeon CPU, 3.70 GHz, 32.0 GB memory) for the multiple heater scenario lasted almost six

hours, four times longer than the single-heater scenario, indicating significant computational cost. This emphasizes the need for future exploration into multi-agent DRL models, considering computational cost improvement and refined reward functions to reduce errors further. The current model could merely optimize the left side of the sheet's temperature distribution (as proof of concept). Future research can involve developing multi-agent RL models to cover the full sheet area, integrating techniques to also concurrently reduce energy consumption in the process.

Acknowledgements

The authors wish to acknowledge the financial support from the New Frontiers in Research Fund-Exploration program in Canada (NFRFE-2019-01440).

References

[1] Y. Wang and F. Zhang, Eds., *Trends in Control and Decision-Making for Human–Robot Collaboration Systems*. Cham: Springer International Publishing, 2017.

[2] T. Rachman, "Towards Intelligent Industrial Co-robots," *Angewandte Chemie International Edition, 6(11), 951–952.*, 2018. https://bair.berkeley.edu/blog/2017/12/12/corobots/.

[3] D. Hein *et al.*, "A benchmark environment motivated by industrial control problems," *2017 IEEE Symp. Ser. Comput. Intell. SSCI 2017 - Proc.*, vol. 2018-Janua, pp. 1–8, 2018, doi: 10.1109/SSCI.2017.8280935.

[4] J. Throne, "Thermoforming," *Appl. Plast. Eng. Handb.*, pp. 333–358, 2011, doi: 10.1016/B978-1-4377-3514-7.10019-4.

[5] J. Arroyo, C. Manna, F. Spiessens, and L. Helsen, "Reinforced model predictive control (RL-MPC) for building energy management," *Appl. Energy*, vol. 309, no. November 2021, p. 118346, 2022, doi: 10.1016/j.apenergy.2021.118346.

[6] S. Brandi, M. Fiorentini, and A. Capozzoli, "Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management," *Autom. Constr.*, vol. 135, no. April 2021, p. 104128, 2022, doi: 10.1016/j.autcon.2022.104128.

[7] A. Gupta, Y. Badr, A. Negahban, and R. G. Qiu, "Energy-efficient heating control for smart buildings with deep reinforcement learning," *J. Build. Eng.*, vol. 34, no. March 2020, p. 101739, 2021, doi: 10.1016/j.jobe.2020.101739.

[8] Y. Z. Wang, X. J. He, Y. Hua, Z. H. Chen, W. T. Wu, and Z. F. Zhou, "Closed-loop forced heat convection control using deep reinforcement learning," *Int. J. Heat Mass Transf.*, vol. 202, p. 123655, 2023, doi: 10.1016/j.ijheatmasstransfer.2022.123655.

[9] E. Hachem, H. Ghraieb, J. Viquerat, A. Larcher, and P. Meliga, "Deep reinforcement learning for the control of conjugate heat transfer," *J. Comput. Phys.*, vol. 436, p. 110317, 2021, doi: 10.1016/j.jcp.2021.110317.

[10] M. Römer, J. Bergers, F. Gabriel, and K. Dröder, "Temperature Control for Automated Tape Laying with Infrared Heaters Based on Reinforcement Learning," *Machines*, vol. 10, no. 3, 2022, doi: 10.3390/machines10030164.

[11] H. Zhao, B. Wang, H. Liu, H. Sun, Z. Pan, and Q. Guo, "Exploiting the Flexibility Inside Park-Level Commercial Buildings Considering Heat Transfer Time Delay: A Memory-Augmented Deep Reinforcement Learning Approach," *IEEE Trans. Sustain. Energy*, vol. 13, no. 1, pp. 207–219, 2022, doi: 10.1109/TSTE.2021.3107439.

[12] M. Szarski and S. Chauhan, "Composite temperature profile and tooling optimization via Deep Reinforcement Learning," *Compos. Part A Appl. Sci. Manuf.*, vol. 142, no. December 2020, p. 106235, 2021, doi: 10.1016/j.compositesa.2020.106235.

[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second., no. 1. MIT press, 2018.

[14] W. M. Rohsenow, J. P. Hartnett, E. N. Ganic, and P. D. Richardson, *Handbook of Heat Transfer Fundamentals (Second Edition)*, vol. 53, no. 1. 1986.

[15] I. Jalilvand, J. Jang, B. Gopaluni, and A. S. Milani, "VR/MR systems integrated with heat transfer simulation for training of thermoforming: A multicriteria decision-making user study," *J. Manuf. Syst.*, vol. 72, no. June 2023, pp. 338–359, 2024, doi: 10.1016/j.jmsy.2023.11.007.

[16] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning," pp. 1–9, 2013, [Online]. Available: http://arxiv.org/abs/1312.5602.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," pp. 1–12, 2017, [Online]. Available: http://arxiv.org/abs/1707.06347.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.