Article

Temperature Control for Automated Tape Laying with Infrared Heaters Based on Reinforcement Learning

Martin Römer *, Johannes Bergers, Felix Gabriel and Klaus Dröder

Institute of Machine Tools and Production Technology, Technische Universität Braunschweig, Langer Kamp 19b, 38106 Braunschweig, Germany; j.bergers@tu-braunschweig.de (J.B.); f.gabriel@tu-braunschweig.de (F.G.); k.droeder@tu-braunschweig.de (K.D.)

* Correspondence: martin.roemer@tu-braunschweig.de

Abstract: The use of fiber-reinforced lightweight materials in the field of electromobility offers great opportunities to increase the range of electric vehicles and also enhance the functionality of the components themselves. In order to meet the demand for a high number of variants, flexible production technologies are required which can quickly adapt to different component variants and thereby avoid long setup times of the required production equipment. By applying the formflexible process of automated tape laying (ATL), it is possible to build lightweight components in a variant-flexible way. Unidirectional (UD) tapes are often used to build up lightweight structures according to a predefined load path. However, the UD tape which is used to build the components is particularly sensitive to temperature fluctuations due to its low thickness. Temperature fluctuations within the production sites as well as the warming of the tape layer and the deposit surface over longer process times have an impact on the heat flow which is infused to the tape and make an adaptive control of the tape heating indispensable. At present, several model-based control strategies are available. However, these strategies require a comprehensive understanding of the ATL system and its environment and are therefore difficult to design. With the possibility of model-free reinforcement learning, it is possible to build a temperature control system that learns the common dependencies of both the process being used and its operating environment, without the need to rely on a complete understanding of the physical interrelationships. In this paper, a reinforcement learning approach based on the deep deterministic policy gradient (DDPG) algorithm is presented, with the aim to control the temperature of an ATL endeffector based on infrared emitters. The algorithm was adapted to the thermal inertia of the system and trained in a real process environment. With only a small amount of training data, the trained DDPG agent was able to reliably maintain the ATL process temperatures within a specified tolerance range. By applying this technique, UD tape can be deposited at a consistent process temperature over longer process times without the need for a cooling system. Reducing process complexity can help to increase the prevalence of lightweight components and thus contribute to lower energy consumption of electric vehicles.

Keywords: tapelaying; temperature control; automation; machine learning; infrared heating

Citation: Römer, M.; Bergers, J.; Gabriel, F.; Dröder, K. Temperature Control for Automated Tape Laying with Infrared Heaters Based on Reinforcement Learning. *Machines* **2022**, *10*, 164. <https://doi.org/10.3390/machines10030164>

Academic Editors: Kelvin K.L. Wong, Dhanjoo N. Ghista, Andrew W.H. Ip and Wenjun (Chris) Zhang

Received: 13 January 2022

Accepted: 18 February 2022

Published: 22 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the raising demand for improving the range of electric vehicles, the vehicle weight is increasing due to the higher number of battery cells that are installed in the vehicles. At the same time, this additional weight also increases energy consumption, which in turn has a negative impact on the vehicles' range. Therefore, lightweight construction of vehicle structures is becoming more and more important. Fiber-reinforced plastic composites offer an opportunity to reduce the weight of structural components and at the same time maintain or even increase their structural stability. In order to satisfy the demand for an increasing number of variants, a process chain is being developed in the project "Großserienfähige Variantenfertigung von Kunststoff-Metall-Hybridbauteilen" (English: High-volume variant production of plastic-metal hybrid components, HyFiVe) with the

aim of manufacturing lightweight components in a variant-flexible manner (Figure 1). An intermediate step of great significance in the manufacturing process is preforming. In this process step, the thermoplastic semi-finished products are built up close to the final contour in order to be finalized in the subsequent process step. The fibers are aligned according to the load path. To be able to precisely adjust the load path, the preform in the “HyFiVe” project is built from functionalized UD tapes with defined fiber orientation.

Figure 1. Process chain for the production of variant-flexible lightweight components based on fiber-plastic-metal composites.

To produce the preforms in a variant-flexible manner, a moldable preform tool is planned upon which the UD tape is deposited. The preform tool obtains its variant-specific form from segmented basic forms (step 1). By using a formflexible preform tool, there is no need to keep rigid preform tools for each variant in stock. The individual preform layers are built up on the preform tool using an ATL process. By combining UD-Tape with the formflexible ATL process, the preforms can be built up according to their load path and variant-specific requirements (step 2). A form-adaptive gripper is used to transport the preform directly to the final production process. The gripper adapts to the form of the preform and thus enables stable transport to avoid deflection and mould distortion. To minimize heating times within the subsequent thermoforming process, a heating unit is integrated into the gripper. During the transport the preform is heated shortly before it is placed inside of the thermoforming tool (step 3). It is finally pressed into the final part by a thermoforming process with glass mat reinforced thermoplastic (GMT). The finalization process gives the component greater stability, since the preform is very unstable beyond its main load paths due to its thinness. The aim of this process chain is to produce variants of lightweight components quickly and with low setup times.

The ATL process (step 2) is discussed in more detail in the following. In this process step, the preforms are built up from layers of differently oriented UD tapes. For this purpose, the tape is melted in the endeffector and draped onto the substrate with a defined consolidation pressure. The aim of the process is to achieve the highest possible degree of consolidation between the individual tape layers to be able to withstand the subsequent thermal load and avoid delamination on the gripper. The degree of consolidation depends highly on the intimate layer contact between tape and substrate. Process technology studies show that the process temperature and the consolidation pressure are the two crucial process variables, with temperature having the significantly greater influence [1,2]. The tape is heated by a specific heating source above the melting temperature of the matrix

material. Well known heat sources are laser, hot gas flame and infrared emitter. All of them differ in power output, controllability and inertia [3]. Infrared emitters benefit from their simple and cost-effective integration into the ATL process at sufficient power output and were therefore chosen for the ATL process.

Due to its low thickness, the tape is susceptible to temperature fluctuations of the heating source and the environment during the process, which can negatively influence the intimate layer contact and therefore the stability of the preform. One major influence is the changing exposure of heat onto the tape because of the heating of the tape layer and the deposition tool over a longer process duration. This happens due to the unfocused infrared heating radiation. The radiation, which does not impact the tape, radiates the metal components of the tape layer and thus leads to continuous heating. Especially the components which stay in contact with the tape have a major influence on the heat flow in the process and thus on the consolidation. An evident solution is to use active cooling systems. However, the temperature variations of these cooling systems also have an influence on the process result and must be extensively modeled or experimentally tuned. Another important influence is the changing ambient temperature for various causes. For example, a change in ambient temperature in the production site can negatively affect the consolidation result of the process. Hence, in order to compensate these undesirable influencing factors, a precise temperature control strategy is required. Such a control strategy enables a high intimate layer contact and therefore a high consolidation quality of the preform.

The temperature control of ATL and the similar automated fiber placement (AFP) processes is currently usually model-based. To set up and investigate the individual influences, the manipulated variables of the process are controlled in open-loop. The parameters are adjusted by analyzing their effect on the process. Peters [4] investigated the influences of the manipulated variables laser power, radiation angle and laser distance on the heating of the tape in laser-assisted tape laying. The process parameters are controlled in an open-loop. For infrared-based tape laying, the influences of emitter power and laying speed are investigated by Venkatesan et al. [5]. Brecher [6] investigates an adaptive control approach for tape laying on geometrically curved surfaces. To avoid temperature peaks due to changing absorption of the laser radiation on curved geometries, an open-loop control is presented, which adjusts the power of individual emitters depending on the surface and the process speed. The inner emitters remain at a constant power. The basis for the parameter optimization is a simulation. A further approach for adaptive control is presented by DiFrancesco [7]. Here, a semi-empirical model is used to determine the laser power as a function of the process speed. The model approach is set up analytically and the model parameters are determined empirically. This work demonstrates that the temperature models have to be designed in an elaborate way and therefore a high level of understanding of the thermal relationship between the material and the machine is necessary. Furthermore, open-loop controls have the disadvantage that the target variable is no longer controlled during tape layup. Deviations due to environmental disturbances are therefore possible.

A closed-loop adaptive control using a model predictive process (MCP) is presented by Hajiloo et al. [8]. A thermal model of the process is established and the controller and manipulated variables are predicted by a Kalman filter. The model is validated in a simulation and has not yet been tested in real-world applications. Model-based control systems are very complex to develop because many different effects have to be considered. Validating the model after it has been used in practice therefore usually involves a great deal of effort. It is easier to use standard controllers such as a PID controller. Dubratz [9] uses a complete closed-loop temperature control for an ATL process based on a PID controller. For the control variables laser power and laser angle, a PID and PI controller, is used respectively. These are adjusted using the method according to Nichols and Ziegler [10]. The control is applied at constant layup speed. Since the parameters of PID controller have to be determined experimentally again when changing an uncontrolled influencing

variable, it is preferable to use a control system that adapts itself to the environment and the influencing variables. One method that avoids the development of a complex analytical model of the environment is machine learning.

The use of machine learning in the field of ATL processes has only been implemented sporadically so far. In most applications, the optimal process variables for a layup process are predicted and the parameters are implemented in open loop by the machine. For this purpose, multiple linear regressions [11], decision trees [12] and artificial neural networks (supervised learning) [13] have been used. The necessary process data has been generated in elaborate experiments. In order to keep the target variable stable even under changing environmental conditions, as described above, a closed loop method is favored. It has been revealed that machine learning can be a promising approach in the field of temperature control. In contrast to frequently used supervised learning, there is no need for a complex generation of the experimental data in advance, since these are generated by the algorithm itself during its learning procedure. Reinforcement learning for temperature control is currently used mainly in building management [14,15]. An approach for controlling lower temperature applications is presented by Ruelens et al. [16]. Here, a water heater is controlled entirely by a reinforcement learning algorithm. In order to realize a temperature control without complex model understanding in the field of ATL processes, the approach of combining machine learning and ATL is pursued in the following.

2. Materials and Methods

In order to implement the reinforcement learning approach into an ATL process, the tape laying system developed in the HyFiVe project and in particular the infrared-based heating system are introduced in Section 2.1. Section 2.2 describes the structure of the used reinforcement learning algorithm. Section 2.3 contains the rewarding strategy and the exploration strategy to implement the algorithm as a temperature controller. Section 2.4 finally describes the execution of the training.

2.1. Infrared-Based Tapelaying System for Thermoplastic Preforming

In the project “HyFiVe”, a system for the application and in situ consolidation of thermoplastic UD tapes was developed for the production of three-dimensional preforms. The system is designed as an endeffector for a KUKA KR 60 industrial robot, which performs the movement within the process. The components are shown as a sectional view of the tape layer in Figure 2.

Figure 2. Sectional view of the tape layer and its functional units.

The design of the endeffector is based on the classic design of ATL endeffectors [2,9]. It contains a tape magazine, a tow roll based conveyor and a pneumatically driven impact knife with spring-loaded holding punch for cutting the tape during lay up. The pressure application on the consolidation roller is carried out by two guide cylinders, each with a maximum force of 1.5 kN. The infrared-based heating system consists of two infrared radiator modules. The inner heating module for preheating consists of two infrared emitters with a total energy output of 1.8 kW. Setting the emitters to 35% of their total power ensures that the tape, which bends under the heating, threads neatly into the consolidation roll at the bottom of the tape applicator. The outer heating module with a total output of 3 kW irradiates both the tape being conveyed and the applied substrate in front of the tape layer and is controlled by the learning controller introduced in the next section. Three Optris CSmicro LT pyrometers were installed at the positions with direct heat input into the tape to record the most important ambient temperatures. The nip point pyrometer detects the point where tape and substrate meet. The temperature here is mainly dependent on the power of the infrared emitters and is considered the most important parameter for the quality of the process (target temperature). Two additional pyrometers recording components that have direct contact to tape and are thus directly related to the heating behaviour. One pyrometer detects the area in front of the tape layer and determines the temperature of the substrate or the mold surface, which both tend to heat up over time. The second pyrometer is mounted on the rear side of the tape applicator (in process direction) and measures the temperature of the consolidation roll. The endeffector presented here has a representative setup for current ATL processes to test temperature control. The simple design of the endeffector serves as a basis for transferring the control to more complex applications. For the experiments a glass fiber-reinforced PA6 UD Tape with a width of 200 mm and a thickness of 0.22 mm will be used.

2.2. Design of the Learning Temperature Controller

To achieve a constantly high quality of the ATL process, it is important to keep the target temperature at the nip point within a predefined thermal tolerance window. In order to learn the heating behaviour of the endeffector and discrete environmental conditions, a learning temperature controller based on a reinforcement learning approach is proposed. Reinforcement learning is based on software agents that try to find the best possible path to solve a problem in an environment. For this purpose, they read the state of the environment (state) and select an action to adapt to the environment (action) based on their learned knowledge. After execution of the action and the subsequent change of the environment based on this action, the state is read again (next state). The evaluation of the action on the basis of the change in the environment is carried out by a reward strategy. The reward (R) is expressed by a numerical value [17].

In the context of reinforcement learning, the exploration-exploitation-dilemma plays an important role. This describes the conflict between the selection of an optimal result for the current problem or the knowledge extension by consideration of other environmental conditions under acceptance of a current sub-optimal solution [17]. An exploration strategy is used to solve this dilemma. The basic structure of the chosen algorithm is explained in this chapter. The reward and exploration strategy are described in the Section 2.3.

One of the main tasks of the agent is the recording of the state of the environment S . In addition to the tape temperature at the nip point T_p (target variable), the recording consists of the main variables influencing the tape described above: The roll temperature T_R and the substrate temperature T_S . The temperature is recorded by the pyrometers as real temperature in °C. In order to smooth the volatile temperature curve of the pyrometers for better processing, the temperature data is passed through a mean average filter with 25 values. Based on its current policy, the agent selects an action A with the goal of maximizing the reward. For this purpose, the agent tries to minimize the difference between the measured and the target temperature. The action available to the agent is the power setting of the front infrared module P_{IR} . According to the installed thyristors,

the action size is a percentage of the total power (0–100%) with a high resolution. The action is transmitted via an interface to the thyristors, which set the new value. Then the agent checks the system state again S' and receives a reward R based on the impact of its action. The vector of state, action, reward and next state is stored by the agent and used to optimize its policy. This procedure is continued iteratively over the entire layup process (episode) (Figure 3).

Figure 3. Structure of the control process supported by reinforcement learning algorithm.

Due to the high-resolution state and action space, both can be considered quasi-continuous in practice. An algorithm which is designed for such cases is the deep deterministic policy gradient algorithm (DDPG). Like all Actor-Critic algorithms, the DDPG consists of two artificial neural networks (ANN), the Actor network and the Critic network. In the DDPG, the Actor network takes the decision of the action based on the state vector. The Critic network determines the Q value based on the selected action and the state vector. The DDPG belongs to the off-policy algorithms, which allows it to reuse already generated data. Therefore, it stores recorded experiences (S, A, R, S') in a replay buffer and uses them for later optimization. To increase stability during learning, the DDPG further uses target networks for both networks, which are time-shifted copies of the original networks [18]. To deal with the continuous action and state space, a DDPG software agent is used to control the temperature in the ATL process described above. The network structures were programmed according to the original publication by Lillicrap [18] with two hidden layers each with 400 and 300 neurons. The layers are connected via a rectified linear unit (ReLU) activation function. In order to meet the special requirements of the control of the radiator power, however, a sigmoid activation function was used for the output layer of the actor network, which gives out values between 0 and 1. The structures of the implemented actor and critic networks are shown in Table 1.

Table 1. Network structures of the implemented artificial neural networks.

Network	Neurons of Layers	Activation Functions
Actor network	3; 400; 300; 1	ReLU; ReLU; Sigmoid
Critic network	4; 400; 300; 1	ReLU; ReLU; ReLU
Target actor network	3; 400; 300; 1	ReLU; ReLU; Sigmoid
Target critic network	4; 400; 300; 1	ReLU; ReLU; ReLU

The programming environment Python 3.8 and the data-science library pytorch were used for implementation. The interface between the DDPG implementation and the machine PLC was realized via the pyADS library. The algorithm used is basically capable of solving the problem and learning the relationship between the system temperatures

and the power setting of the infrared module. For the successful execution of the learning process, however, a reward strategy is necessary, which evaluates the action decisions of the agent. Furthermore, an exploration strategy is needed to quickly adapt the agent to an optimal solution. Both strategies are described in the following.

2.3. Design of Rewarding and Exploration Strategy

In order to perform the learning process successfully, both a reward and an exploration strategy are needed. Using the reward strategy, the software agent learns its policy and can evaluate the ranges in which it should maintain the target temperature. At the same time, the agent can develop aversions against harmful temperature ranges. The main goal of the controller is to keep the process temperature T_p constantly within the thermal tolerance window at any point in the process. Therefore, it is reasonable to make the reward dependent on the difference between the actual process temperature and the target temperature T_{target} . The target temperature is set to 250 °C for the processing temperature of the tape, since this temperature has proven to be effective in previous experiments for the used tape. A tolerance range ϕ of ± 10 °C is defined around this target temperature. No significant changes in process performance were observed with variations in this temperature range. Furthermore, above the maximum temperature T_{max} of 300 °C, the degeneration of the matrix material is intolerable and the system shuts down. According to this temperature zone, the agent receives a reward depending on the difference between process and target temperature (Figure 4).

Figure 4. Given reward according to different temperature ranges.

A positive reward is given to the agent when the temperature is kept within the tolerance range. The maximum reward is given when the target temperature is reached. Outside the tolerance zone, the reward falls linearly, with a higher degree of punishment in the higher temperature range as the material is more degenerated. A particularly high punishment of -10 is given when $T_p = 300$ °C is reached or exceeded to instill in the agent a high aversion to this temperature range. The room temperature of 20 °C is set as the lowest temperature limit. The reward is thus calculated from:

$$R(T_P) = \begin{cases} -10 & T_P > T_{max} \\ -\frac{1}{T_{max}-T_{target}} * |(T_P - T_{target})| & T_{max} \geq T_P > T_{target} + \phi \\ 1 - \frac{1}{\phi} * |(T_P - T_{target})| & T_{target} + \phi \geq T_P \geq T_{target} - \phi \\ -\frac{1}{T_{target}-20^{\circ}\text{C}} * |(T_P - T_{target})| & T_{target} - \phi > T_P \geq 20^{\circ}\text{C} \end{cases} \quad (1)$$

In order to solve the agent's problem of reaching an optimal solution as quickly as possible and to equally explore the solution space in a meaningful way, exploration strategies are needed. The DDPG algorithm used here already provides a certain exploration strategy called Ohnstein-Uhlenbeck (OU) process, which generates a random noise based on the last used values. Thereby, the agent also considers the solution spaces around its optimal solution. This forces the agent to explore the solution space immediately around its optimal solution. This reduces the risk of being stuck in local optima. The amount of noise is determined by the hyperparameters of the OU process according to the original publication by Lillicrap [18].

However, since the agent only observes its immediate environment through the OU process, teaching the agent at the beginning would lead to many failed attempts and increase the required teaching time. Since learning in the real process is time- and material-consuming, another strategy must be found that enables faster learning. The advantage of such a strategy is that the process has already been controlled by an open loop controller. Therefore, it is feasible to provide the "knowledge", that is already stored in the controller to the agent, so that the agent can benefit from this knowledge. This can shorten the start-up time of the agent immensely. In their work, Silver et al. [19] present an approach for a so-called "Expert Exploration". Here, the algorithm learns based on a previously imperfect solution. The agent selects from three actions. These include the previous policy of the existing solution, a random value and the proposed solution of the agent. The probabilities of the selection of the three solutions are determined by the hyperparameters α and ϵ .

The probability of choosing the agent's action is calculated by $1 - \epsilon$. It is initially very low so that the agent can learn from the previous policy and increases with continuous training episodes. α describes the ratio of the probability of the previous policy (expert) with $\epsilon * \alpha$ and the coincidence with $\epsilon * (1 - \alpha)$ in relation to ϵ . For the DDPG-based temperature control strategy for the proposed ATL process, the procedure is adapted. The previous policy is represented by the previous open-loop temperature control. Here, a continuous value of 60% of the maximum power is set as output. Good results could be achieved with this value in previous tests. The noise is taken over by the OU process. For the adjustment of the ϵ value over time, an epsilon decay method is used with an acceptance factor of 0.995. For α , 0.6 is chosen. The expert-exploration-process was designed for this case for a total number of 1000 episodes to support the agent by the expert also beyond the experimental series. The progression of probabilities over training episodes is shown in Figure 5.

By means of the expert exploration process, the agent quickly reaches the level of the expert during the open-loop power control of the infrared radiator module and continues to improve thereafter. This can reduce the number of required training sections and thus the material and energy consumption.

The last step in integrating the software agent into the tape creation process is to set the system dead time. Since the environment only adapts to the change caused by the agent's action with a certain inertia, direct observation of the environment is not helpful. A defined delay time between execution of the selected action by the agent and recording the next state of the environment must be implemented. Due to the inertia of the system, a too short delay would lead to incorrectly recorded end temperatures of the tape. The agent would therefore learn the wrong correlations. A too long delay decreases the possible reaction speed of the agent.

Figure 5. Distribution of the action probability from expert exploration.

This time is significantly affected by the infrared system of the tape layer. On the one hand, the filament of the infrared emitters needs a certain time to reach a new stable filament temperature. On the other hand, the local offset of the impact zone of the infrared radiation and the measurement by the pyrometer leads to a delay on the heat transport to the measuring zone, which is directly dependent on the traversing speed. Furthermore, the average-mean filters used for smoothing the pyrometer signals also influence a temporal offset which is, however, negligible in this case. First attempts with a direct reward of the agent at its action led to a strongly fluctuating behavior of the system. After some preliminary tests, in which action and next state of the agent were observed, a waiting time of 3 s turned out to be beneficial for the learning behavior for the subsequent experiments.

With the rewarding and exploration strategy, fast and effective learning of the agent is possible. By means of expert exploration, the agent can also benefit from previous knowledge of the process. The system dead time also helps to evaluate the effects of the agent action in a stable way. The design of the agent can thus be integrated into the ATL process.

2.4. Training of the DDPG Algorithm

The agent was trained according to a defined iteration process. The python-implemented agent is connected to the Beckhoff PLC using the pyADS interface, which coordinates the thyristors for the power setting of the infrared emitters and the data recording of the pyrometers. For the data generation, tapes were automatically deposited on a flat surface consisting of a steel sheet and a thermal insulation layer.

The state consists of the temperature values of the three pyrometers for nip point, roll and substrate temperature. After selecting the action, the corresponding power values are transmitted to the IR emitters. The agent waits for 3 s until the environment shows a reaction to the action and then determines the next state. Based on this, it receives the corresponding reward. The data set of state, action, reward and next state is stored in the replay buffer. After the transition, the agent trains its 4 networks with a mini-batch of 64 data, which is randomly selected from the replay buffer. Finally, the epsilon decay is increased by one episode. The iteration process is shown in Figure 6. A layup length of 400 mm was chosen with a process speed of 20 mm s^{-1} for each episode. According to the layup duration and the processing time of each iteration, a data set of 5–6 control iterations results per layup depending on the speed of the interface. A total layup length of 239 m was specified for the training, resulting in 597 layup iterations.

Figure 6. Process of the iterative training method for the DDPG agent.

3. Results

To evaluate the training success, both the evolution of the agent's rewards and the agent's improvement maintaining the temperature are considered. Based on the evolution of the reward, the learning success of the agent within the set system boundaries can be evaluated by the reward function. Therefore, the percentage of the maximum achievable reward is normalized to the number of iterations performed per episode (Figure 7). Since the course of the reward is very volatile due to the exploration efforts of the agent, the moving average over 40 episodes is chosen for a better representation of the trend. It is evident that the course of the reward increases with continuous training, with a steeper trend starting around episode 410. This can be related to the increasing decision-making probability of the agent regarding to the exploration strategy. The probability of the agent action at this point is 87.13%. At the end of the training, the agent achieves 29.5% of the maximum achievable reward. The increasing reward trend shows that the agent has learned the relationships of the temperature control of the ATL process. Since the trend also increases towards the end of the training process, it can be assumed that the result will continue to improve with an increased number of training episodes.

Moreover, in Figure 7, it can be seen that there are no canceled episodes due to the 300 °C limit being reached from episode 288 onwards. The canceled episodes (reward −100%) are drawn as red bars. The agent has learned the aversion to higher temperature ranges imposed by the rewarding strategy. However, a much earlier development of aversion must be assumed, since the decision of the action in all cases that resulted in reaching the temperature limit were made by the expert or the OU process.

Figure 7. The average reward per iteration over the training increases continuously.

While the evolution of the reward history shows a positive trend, it is more an indicator of the training success of the agent. However, the crucial goal for the ATL process is to maintain stable temperatures within the layup process. This can be controlled by the course of the “Next State”. The next state contains the statement whether the action selected by the agent leads to a temperature within the tolerance window. For this purpose, the temperatures reached in the nip point after the action were analyzed. Figure 8 shows the percentage of process temperatures in the thermal tolerance window after an action over all iterations. Episodes where all subsequent temperatures are within the selected temperature window are shown in green. To illustrate the trend of the evolution, a plot with a moving average over 40 episodes has been added.

Figure 8. The percentage of next state temperatures within the thermal tolerance window increases significantly.

A positive trend can also be observed in this respect. The mean value increases sharply from episode 435 and a higher density of complete episodes can be observed. From this, not only an increasing understanding of the agent for the process, but also an advantage for the ATL process can be derived. Towards the end of the training the trend flattens out again. Finally, to assess the result of the training, the first 10 episodes of the untrained agent are compared with the 10 episodes of the trained agent to get an impression of the controllability of the heating behavior. The course of the episodes is shown in Appendix A.

The comparison clearly shows that the trained agent controls the heating behavior better under the same initial conditions. The temperature curves are clearly closer to the tolerance window. In summary, the trained agent converges with increasing training and can better control the process temperatures. Some specifics of the results will be explained further.

4. Discussion and Conclusions

It could be shown that learning temperature control on an ATL process using a reinforcement learning approach is possible. The DDPG algorithm in combination with the selected reward and exploration strategy was able to improve the temperature control of the ATL process already with a small set of training data and to learn the heating behavior of the tape as well as the environmental conditions. Therefore, using this approach, it is possible to run the ATL process for a longer period of time without additional calibration steps. The increasing trend of rewards as well as the positive evolution of next states show that the agent understands the heating system of the ATL process in relation to its operating environment. Likewise, it was shown that an aversion to critical temperature ranges could

be learned at a very early stage. Despite the positive progressions it can be seen that the training data set is not sufficiently large enough to keep the process temperature always in the thermal tolerance window. Therefore, an increase of the training data set is necessary to further improve the results. Especially in colder areas, the agent is not yet able to quickly bring the temperatures into the target temperature window. This is due to the fact that in average the agent has trained more often in hot areas. Targeting training in areas that have been less considered so far can remedy this. For this work, only the lower module of the ATL process was controlled by an agent. The upper heating module ran at a constant power level during the experiments. For this reason, the first recorded temperature set per episode was subject to the heating behaviours described above. To demonstrate the feasibility of the concept, the action and variable space for the agent was reduced. This included the use of the agent on only one heating module as well as keeping the process speed and the target temperature constant. However, for the transfer of the concept to industrial plants, the incorporation of the excluded process variables is necessary. For this purpose, the agent can be further trained with the help of the expert exploration, in which the current training level of the agent is used as the expert. In a next step, therefore, the control of both heating modules and the effects of process speed and changing target temperature (for processing different materials) on the agents behaviour will be investigated to obtain an adaptation to industrial system conditions as fast as possible. For this purpose, also a hyperparameter study will also be performed to optimize the training time. Furthermore, the performance of the agent was only evaluated in relation to an open-loop controller. In a further experiment, this will also be compared to a PID controller.

In summary, the use of reinforcement learning as temperature control can be considered successful. The agent quickly learns the specific characteristics of the system environment without the use of a model or model-based simulation. The use of expert exploration also makes it possible to reuse existing knowledge for later controller versions. This applies in particular to the subsequent expansion of the variable space. This approach, therefore, promises to make a positive contribution to the production of variant-flexible preforms.

Author Contributions: Conceptualization: M.R. and J.B.; Methodology: M.R. and J.B.; Software: J.B. and M.R.; Formal analysis: F.G. and K.D.; Resources: K.D.; Data curation: J.B. and M.R.; Writing—original draft preparation: M.R.; Writing—review and editing: F.G. and K.D.; Visualization: M.R. and J.B.; Project administration: M.R. and K.D.; Funding acquisition: M.R. and K.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Federal Ministry of Education and Research under the grant number 02P18Q740.

Data Availability Statement: Data supporting the reported results, including code, can be found at the following link: <https://lnk.tu-bs.de/rbS8W6> (accessed on 9 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AFP	Automated fiber placement
ANN	Artificial neural network
ATL	Automated tape laying

DDPG	Deep deterministic policy gradient
GMT	Glass mat reinforced thermoplastic
MCP	Model predictive process
OU	Ohrnstein-Uhlenbeck process
ReLU	Rectified linear unit
UD	unidirectional

Appendix A

Figure A1. Comparison of the Temperature at the Nip Point with Untrained and Trained Agent.

References

1. Lee, W.I.; Springer, G.S. A Model of the Manufacturing Process of Thermoplastic Matrix Composites. *J. Compos. Mater.* **1987**, *21*, 1017–1055. [\[CrossRef\]](#)
2. Steyer, M. *Laserunterstütztes Tapelegeverfahren zur Fertigung Endlosfaserverstärkter Thermoplastlamine*, 1st ed.; Ergebnisse aus der Produktionstechnik, Apprimus-Verl.: Aachen, Germany, 2013.
3. Yassin, K.; Hojjati, M. Processing of thermoplastic matrix composites through automated fiber placement and tape laying methods: A review. *J. Thermoplast. Compos. Mater.* **2018**, *31*, 1676–1725. [\[CrossRef\]](#)
4. Peters, T. Untersuchungen zum Diodenlaserbasierten In-Situ Tapelegen als Produktionstechnologie für einen Großserienfähigen Hybriden Leichtbau. Ph.D. Thesis, RWTH Aachen, Aachen, Germany, 2020.
5. Venkatesan, C.; Velu, R.; Vaheed, N.; Raspall, F.; Tay, T.E.; Silva, A. Effect of process parameters on polyamide-6 carbon fibre prepreg laminated by IR-assisted automated fibre placement. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 1275–1284. [\[CrossRef\]](#)
6. Brecher, C.; Emonts, M.; Striet, P.; Voell, A.; Stollenwerk, J.; Janssen, H. Adaptive tape placement process control at geometrically changing substrates. *Procedia CIRP* **2019**, *85*, 207–211. [\[CrossRef\]](#)
7. Di Francesco, M.; Veldenz, L.; Dell’Anno, G.; Potter, K. Heater power control for multi-material, variable speed Automated Fibre Placement. *Compos. Part A Appl. Sci. Manuf.* **2017**, *101*, 408–421. [\[CrossRef\]](#)
8. Hajiloo, A.; Xie, W.; Hoa, S.V.; Khan, S. Thermal control design for an automated fiber placement machine. *Sci. Eng. Compos. Mater.* **2014**, *21*, 427–434. [\[CrossRef\]](#)

9. Dubratz, M. *Laserunterstütztes Tape-Placement-Verfahren für die Herstellung Dreidimensionaler Strukturkomponenten aus Endlosfaserverstärkten Thermoplastischen Prepregs*, 1st ed.; Ergebnisse aus der Produktionstechnik Produktionsmaschinen, Apprimus-Verl.: Aachen, Germany, 2015.
10. Ziegler, J.G.; Nichols, N.B. Optimum Settings for Automatic Controllers. *J. Dyn. Syst. Meas. Control* **1993**, *115*, 220–222. [[CrossRef](#)]
11. Brüning, J.; Denkena, B.; Dittrich, M.A.; Hocke, T. Machine Learning Approach for Optimization of Automated Fiber Placement Processes. *Procedia CIRP* **2017**, *66*, 74–78. [[CrossRef](#)]
12. Wagner, H.; Köke, H.; Dähne, S.; Niemann, S. Decision Tree-Based Machine Learning to Optimize the Laminate Stacking of Composite Cylinders for Maximum Buckling Load and Minimum Imperfection Sensitivity. *Compos. Struct.* **2019**, *2019*, 45–63. [[CrossRef](#)]
13. Oromiehie, E.; Prusty, G.; Rajan, G. Machine learning based process monitoring and characterisation of automated composites. *Procedia CIRP* **2017**, *2017*, 22–25.
14. Kuhnle, A.; Kaiser, J.P.; Theiß, F.; Stricker, N.; Lanza, G. Designing an adaptive production control system using reinforcement learning. *J. Intell. Manuf.* **2020**, *32*, 855–876. [[CrossRef](#)]
15. Brandi, S.; Piscitelli, M.S.; Martellacci, M.; Capozzoli, A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy Build.* **2020**, *224*, 110225. [[CrossRef](#)]
16. Ruelens, F.; Claessens, B.J.; Quaiyum, S.; de Schutter, B.; Babuška, R.; Belmans, R. Reinforcement Learning Applied to an Electric Water Heater: From Theory to Practice. *IEEE Trans. Smart Grid* **2018**, *9*, 3792–3800. [[CrossRef](#)]
17. Géron, A. *Praxiseinstieg Machine Learning Mit Scikit-Learn, Keras und TensorFlow: Konzepte, Tools und Techniken für Intelligente Systeme*, 2nd ed.; O'Reilly: Heidelberg, Germany, 2020.
18. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2019**, arXiv:1509.02971.
19. Silver, T.; Allen, K.; Tenenbaum, J.; Kaelbling, L. Residual Policy Learning. *arXiv* **2019**, arXiv:1812.06298.