Robotica (2024), **42**, pp. 179–202
doi:10.1017/S0263574723001418

RESEARCH ARTICLE

Learner engagement regulation of dual-user training based on deep reinforcement learning

Yang Yang[1], Xing Liu[2], Zhengxiong Liu[2] and Panfeng Huang[1]

[1]Research Center for Intelligent Robotics, School of Astronautics, Northwestern Polytechnical University, Xi'an, China and [2]National Key Laboratory of Aerospace Flight Dynamics, School of Astronautics, Northwestern Polytechnical University, Xi'an, China

Corresponding author: Panfeng Huang; Email: pfhuang@nwpu.edu.cn

Received: 17 February 2023; **Revised:** 4 September 2023; **Accepted:** 18 September 2023;
First published online: 13 November 2023

Keywords: dual-user training system; motor skill learning; complex operations; deep reinforcement learning(DRL); adaptive engagement; task safety

Abstract

The dual-user training system is essential for fostering motor skill learning, particularly in complex operations. However, the challenge lies in the optimal tradeoff between trainee ability and engagement level. To address this problem, we propose an intelligent agent that coordinates trainees' control authority during real task engagement to ensure task safety during training. Our approach avoids the need for manually set control authority by expert supervision. At the same time, it does not rely on pre-modeling the trainee's skill development. The intelligent agent uses a deep reinforcement learning (DRL) algorithm based on trainee performance to adjust adaptive engagement during the training process. Our investigation aims to provide reasonable engagement for trainees to improve their skills while ensuring task safety. Our results demonstrate that this system can seek the policy to maximize trainee participation while guaranteeing task safety.

1. Introduction

In recent years, human skill training has gained significant attention due to the increasing demand in various fields such as teleoperation, driving, minimally invasive surgery, nuclear maintenance, and assembling [1–6]. As operational errors may lead to substantial economic, environmental, and health damage, it is essential to provide high-quality and efficient human operation skill training for novices to acquire appropriate psychomotor skills. Training research aims to enable trainees to learn and practice on the task effectively, and the present research on human skill training includes system design that offers trainees practicing opportunities and optimal training policies that serve as a strategy plan for human skill improvement.

Virtual reality (VR) has emerged as a promising technique in this field, providing on-demand training to trainees [7–10]. Integrating task or training procedure modules provided by simulators within a VR environment has shown successful results in some training tasks. Virtual reality environments offer trainees the opportunity to practice repeatedly, allowing them to make mistakes and acquire skills through long-term training. However, the gap between virtual and realistic environments cannot be ignored, particularly in complex skill learning [11]. Therefore, a dual-user training scheme is designed that allows trainees and instructors to operate in real scenes simultaneously. This scheme enables trainees to be involved in actual task procedures, avoiding problems caused by virtual scenes and allowing for imperfect skills. The trainee's operation is sent to the controlled object through an authority fusion and expert operation to ensure task security and trainee participation. The trainee's training strategy can be reflected in this authority, with determining an optimal authority value being a key factor in improving training quality and further enhancing motor learning.

When setting authority in training, it is beneficial to consider research theories. Fitts and Posner proposed three stages of motor skill acquisition during human learning: cognitive, integrative, and autonomous [12]. In the cognitive phase, learners intellectualize the task and understand the mechanics of the skill. In the integrative phase, knowledge is translated into appropriate motor behavior, yet with a lack of fluidity. In the autonomous phase, independent learning occurs with no supervision or guidance, and smooth performance evolves. Transmitting more accurate task information and skill feedback to trainees during the cognitive and integrative stages enhances their understanding of the task and helps correct their skill errors. From the perspective of influencing factors, Gabriele Wulf's research identified observational practice, focus of attention, feedback, and self-controlled practice as the most relevant and influential factors in motor learning progress in humans, with different roles in different learning stages [13].

Based on these theories, dual-users combined with shared control have been developed [14]. In the traditional dual-user training system, the slave robot that interacts with the environment receives a control signal only from one master console, requiring frequent switching of control authority during trainee practice, increasing the potential task risk in training. The implementation of shared control enables trainees and instructors to operate tasks independently and simultaneously in real-time, enabling dyadic training that meets the factors that progress motor learning for the trainee while guaranteeing task safety through fused control signals and sharing factor adjustment. Coordination of control signals received by the slave robot uses a dominance factor, ranging from zero to one [5, 6, 15, 16]. This development has been applied in many surgical training scenarios with low fault tolerance operational tasks. However, experts mostly set the factor based on trainee performance, leading to some challenges, such as unclear arbitration of trainer and trainee authority to ensure safety during the learning process and determine the ideal policy for a learner to engage in the training process. Additionally, supervising trainees during training to constantly update evaluations and make corresponding authority adjustments is unrealistic and significantly increases training costs. Moreover, due to the implicit nature of motor skills, experts cannot directly teach or assess motor programs.

To enhance the efficiency and quality of dual-user training systems, this research proposes an intelligent decision-making agent capable of allocating control authority based on trainee performance while maintaining task safety. The intelligent agent allows trainees to actively participate in the task while adapting the allocation of control authority to match changes in their skill level. With learning capabilities, the agent can adjust the dominance factor adaptively, thereby improving the effectiveness of the training process.

This paper provides a novel approach to dual-user training by designing, developing, and implementing an intelligent-based system that offers three key features. Firstly, the system provides a training platform in a real scene, allowing trainees and instructors to conduct tasks concurrently without compromising task safety. Secondly, the system utilizes reinforcement learning to offer adaptive engagement adjustment, providing trainees with appropriate training engagement corresponding to their skill proficiency level. Finally, the system includes a learning agent that regulates the control authority factor automatically when a trainee's skill level changes, eliminating the need for supervision during this process.

These three features represent significant contributions to the field of dual-user training, as they improve the efficiency and effectiveness of the training process while ensuring safe and effective results. The real-scene training platform allows trainees to practice in realistic environments, enhancing the transferability of learned skills to real-world scenarios. The adaptive engagement adjustment feature tailors the training experience to individual trainees, maximizing the effectiveness of the training process. Moreover, the learning agent enhances the autonomy of the system, allowing it to regulate control authority without the need for manual intervention from trainers.

Overall, the proposed intelligent-based dual-user training system is a promising solution for training novices in various fields, providing a safe and effective training experience that adapts to the trainee's skills and learning progress. The contributions of this paper have the potential to significantly advance the field of human skill training and could be applied across a wide range of industries.

Remark 1 In this study, the decision to avoid defining a specific task scenario in this article was deliberate, and instead, the expertise of a subject matter expert was utilized as a benchmark for evaluating the quality of the proposed application. The study aims to enhance the generalizability of the application, making it applicable across various domains and scenarios.

Remark 2 In this study, ensuring safety relies on the operational information provided by experts. Although this study acknowledges the importance of safety, it avoids discussing the safety of a specific task, instead defining the safety index as a fixed value that is dependent on the deviation between the output information of the slave robot and the desired output information.

The rest of this paper is organized as follows: Section 2 presents the overall framework with a shared control integrated dual-user training system. Section 3 talks about the adaptive adjustment process of a trainee's level of engagement regarding their authority over the procedure and its meaning. Section 4 provides the reinforcement learning-based algorithm design for the defined problem and Section 5 presents the simulation results. The conclusions of the study are given in Section 6.

2. Shared-control integrated dual-console for training system

The study of skill training has evolved over time, and a dual-console framework has emerged as a prominent training paradigm based on mirror theory. In this framework, both the trainee and expert engage in the real world using two master consoles that are separately controlled by each operator. Traditional models involve a single slave console that executes commands solely from one master console, requiring frequent switching of control authority between the expert and trainee during practice sessions. However, this approach increases the risk of harmful tasks when control is transferred to the trainee, and separating training into observation and practice can weaken the effectiveness of learning, making it challenging to meet the principles of mirror theory [17].

To address these issues, this study proposes integrating a shared control strategy into the dual-user training system. As illustrated in Fig. 1, the proposed policy involves a shared control framework in which two operators manipulate a slave robot, enabling them to make contact with the task while reducing the risk of harm and achieving a mirror-symmetric pattern. By leveraging a shared control approach, the system can mitigate the risks associated with traditional training methods while enhancing the effectiveness of skill training.

To ensure task safety, the received signal for the slave robot is modulated using a dominance factor ω, which enables the level of authority that the trainee has over the slave robot to be controlled. A more flexible and generalized approach involves providing the trainee with partial authority over the procedure, preserving continuous involvement in a desired trajectory. The fused signal satisfies the safety control on the slave side. The control system model is represented by Eq. (1).

$$X_R(t) = (1 - \omega(t)) \cdot X_E + \omega(t) \cdot X_T, \tag{1}$$

In this equation, $X_E(t)$ and $X_T(t)$ represent two control information sources, from the expert and trainee, respectively. These sources are combined with a weight factor, $\omega(t)$, to generate the desired trajectory for the slave robot, represented by $X_R(t)$.

During the training process, both trainees and experts operate simultaneously while the former modifies their motor behavior based on feedback from the environment. The level of feedback information available to trainees directly impacts their comprehension of their actions and surroundings. An important factor in this context is the dominance factor ω, which can be determined based on the skill level of the trainee. Adjusting this factor is a crucial issue for achieving effective improvement in the skill training process, particularly in the context of human motor skill training. Providing an appropriate level of

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

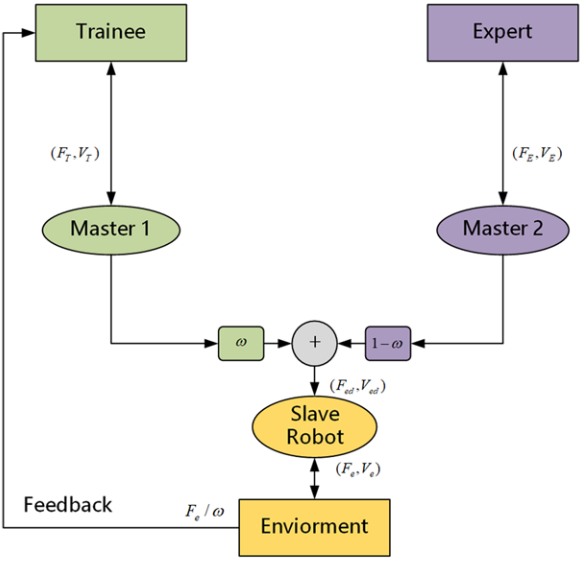


Figure 1. *The scheme of the dual-user training system.*

control authority to trainees based on their skill performance during the task procedure is desirable as it could expedite the training process while ensuring safety. The following section will provide further details of this method.

3. Adaptive skill-oriented engagement

The quality and effectiveness of the training process are primarily determined by the training strategy adopted. In the previous section, we provided an overview of the proposed framework. In this section, we will delve into the training strategy implemented in the system. The training strategy is tailored to reflect the trainees' engagement level during the training process, which should be commensurate with their skill level. This approach is aimed at enhancing the overall efficacy of the training process. To achieve this goal, we present the training environment setup and formulate the decision-making problem for the training strategy adjustment process in this section. By providing a comprehensive overview of the training strategy's fundamental components, we aim to highlight its importance in facilitating effective skill acquisition.

3.1. Skill performance evaluation

Developing an effective training strategy requires an initial assessment of the trainee's skill level. However, in real-world operational scenarios, it may be challenging to quantitatively express complex, multi-step processes' desired trajectories in advance. The evaluation of skills in this study involves acquiring the trainee's skill characteristic, which can be achieved through various studies tailored to different types of tasks. For example, in characteristic acquisition studies, the EGNN paradigm [18], transfer feature learning [19], and heterogeneous network representation [20] approaches can be utilized to uncover information about the features associated with the skill. This approach is applicable to more open-ended skill training scenarios, such as playing musical instruments without specific task objectives or engaging in ball sports. On the other hand, there are studies that focus on tasks with defined objectives, such as surgical scenarios, welding scenarios, or assembly processes. These studies typically examine the impact of operation presentation and construct characteristic based on the effects of the operations.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

The research in this paper focuses on dual-user operation training under sharing control, so the characteristics of skills are all indicators under the conventional definition. According to the summary of the study [21], the length of the operation as well as the smoothness of the operation are selected as the trainee's skill characteristics. This was also applied and analyzed in refs. [5, 21–23] to represent the operator's skill performance, and after acquiring the skill characteristics, they were quantified into numerical values by evaluating them. In real-time training, this assessment should accompany the trainee's real-time operation. Therefore, the assessment of the trainee's skills should be done during the session rather than waiting for the end of the task. To address this issue, a task-independent assessment approach is adopted to objectively evaluate the trainee's performance. This approach differs from absolute assessment methods and emphasizes the importance of evaluating the trainee's relative performance level. The performance of the trainee's skills is a dynamic assessment of the operation in progress with the expert, and its goodness also depends on the level of the expert at the time of cooperation.

A normalized task-independent metric, Φ, has been defined to determine the desired quantitative performance of the trainee, as shown in Eq. (2).

$$\Phi(t) = 1 - \left| \frac{X_E(t) - X_T(t)}{X_E(t) + X_T(t)} \right|, \tag{2}$$

The defined metric Φ reflects the trainee's relative skill performance compared to the expected level. It should be noted that for complex operations, the expected value should not be stereotyped, and expert operation data serves as a reference for determining the expected value in real time. Compared to absolute assessment methods, this quantitative performance approach provides a more objective evaluation of the trainee's relative performance level. Furthermore, the normalization process facilitates calculations during agent learning, providing additional advantages.

In the development of a training strategy, the identification of relevant skill characteristics must be task-specific. For common operational tasks, an effective evaluation of a trainee's skill level can be based on instrument motion alone. In this study, we selected path length and motion smoothness as the two features for performance evaluation. These characteristics were chosen based on two key considerations. Firstly, as both the trainee and expert perform the task simultaneously, they consume the same amount of time. Typically, the expert's trajectory is more concise, indicating a deeper understanding of the task and resulting in a shorter path length compared to that of the trainee during the same time duration. Secondly, motion smoothness reflects the operator's motor control ability, making it a crucial factor in evaluating operational performance. Therefore, these two features comprehensively cover most operational skill performance and are easily measurable.

The path length: This is a feature reflecting the operation trajectory, which is calculated as Eq. (3a), and the normalized task-independent value for ϱ is given in Eq. (3b):

$$\varrho = \int_0^t \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, d\tau, \tag{3a}$$

and

$$\Phi_\varrho(t) = 1 - \left| \frac{\varrho_E(t) - \varrho_T(t)}{\varrho_E(t) + \varrho_T(t)} \right|, \tag{3b}$$

where $\varrho_E(t)$ and $\varrho_T(t)$ denote the path length performance for expert and trainee, respectively.

Operation Smooth: Instantaneous jerk is a measure of the rate at which acceleration changes over time and is quantified as $J = \frac{d^3x}{dt^3} cm/s^3$ [23]. The cumulative value of instantaneous jerk can be calculated using Eq. (4a) and is considered to be indicative of the smoothness of an operation during a task:

$$\nu = \int_0^t \sqrt{\left(\frac{d^3x}{dt^3}\right)^2 + \left(\frac{d^3y}{dt^3}\right)^2 + \left(\frac{d^3z}{dt^3}\right)^2}\, d\tau, \tag{4a}$$

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

and the normalized task-independent value for ν is Eq. (4b):

$$\Phi_{\nu}(t) = 1 - \left| \frac{\nu_E(t) - \nu_T(t)}{\nu_E(t) + \nu_T(t)} \right|, \tag{4b}$$

The value is represented by a ratio relation between the trainee and expert and falls within the range [0–1]. The performance parameters $\Phi_{\varrho}(t)$ and $\Phi_{\nu}(t)$ are then linearly fused to obtain the trainee's skill state parameter, $\Phi_{\Delta}(t)$, which is:

$$\Phi_{\Delta}(t) = \Phi_{\varrho}(t) \cdot \Phi_{\nu}(t) \tag{5}$$

In order to further enhance the examination of the selected skill indicators, an additional approach can be employed wherein newly acquired quantitative eigenvalues are multiplied continuously. This innovative idea is inspired by the study [24] and serves to amplify the analysis of the characteristics of these indicators.

The skill assessment phase serves as a crucial data preprocessing component within the interaction process of an intelligent agent. During this phase, the agent solely perceives and processes quantized skill-level information. Subsequently, the agent generates tailored training strategies based on the perceived skill level. By focusing on quantized skill-level information, the agent simplifies the complexity of the data, enabling efficient processing and analysis. This approach allows the system to distill the essential aspects of the trainee's skill level, facilitating the generation of effective and personalized training strategies. Leveraging the perceived skill-level information, the intelligent agent leverages advanced algorithms and methodologies to develop training strategies that are specifically designed to optimize the trainee's learning experience and enhance skill acquisition. These strategies are dynamically generated based on the real-time skill level, ensuring adaptability and responsiveness to the trainee's needs.

3.2. Adaptive engagement regulation

Training is a systematic process that involves the acquisition of knowledge, skills, and attitudes to enhance performance in a specific setting. Adaptability is a critical element in ensuring training effectiveness. Kelly (1969) defined adaptive training as the variation of the problem, stimulus, or task based on the trainee's performance. This approach consists of three primary components: performance measures, adaptive variables, and adaptive logic [25]. Given the resource-intensive and time-consuming nature of training, an adaptive training protocol can significantly reduce training time compared to conventional step-by-step approaches. The integration of adaptation into training has been widely implemented across many domains, including older adult cognitive training [26], professional athlete physical training [27], game-based training [28], rehabilitation [29], medical training [30], and stress management training [31].

Personalizing training strategy and collecting user performance data requires an initial investment, but the benefits of reducing training time and matching learners' needs outweigh the initial cost [32]. Furthermore, repetitive and ineffective training can hinder skill learning efficiency; thus, it is advisable to avoid such approaches.

Engagement is a critical factor in human motor skill training. It involves the active participation of trainees in skill acquisition, which enhances their performance [33]. Effective participation helps trainees master skills and knowledge comprehensively, motivates them to complete tasks with greater effort, provides feedback information, and more accurately measures their performance. Notably, engagement is a complex construct with context-dependent and multifaceted forms. Given that trainees' control in shared control-based dual-user training systems is considered their engagement in the training process, this paper aims to enhance engagement by adaptively adjusting trainees' control authority based on their skill level. This approach will contribute to establishing adaptive logic and promoting skill acquisition.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

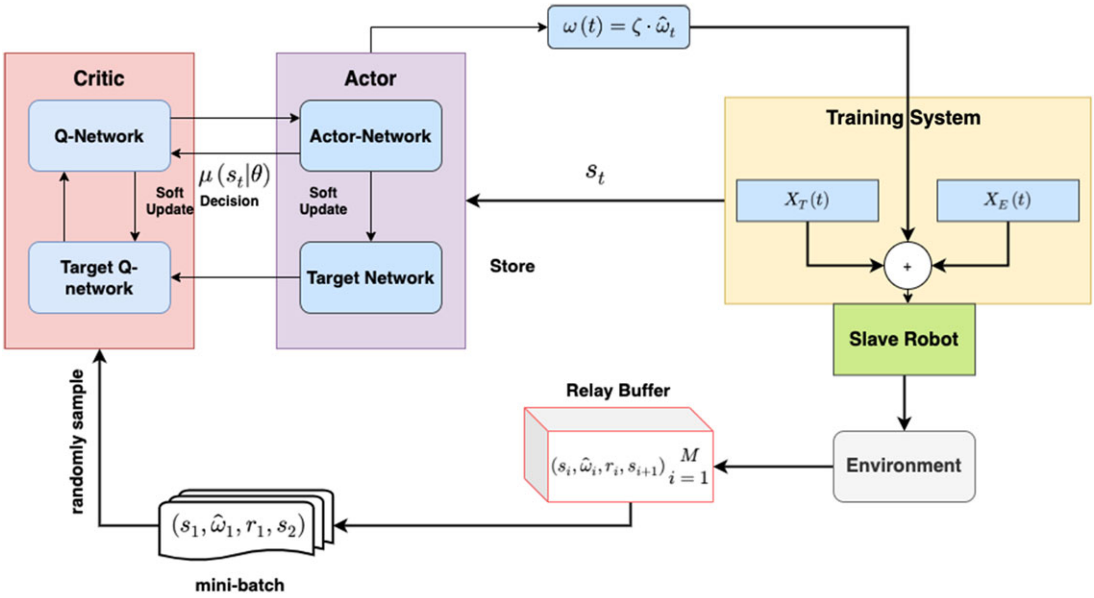


Figure 2. *Overall diagram of agent decision-making for trainee's engagement regulation. The DDPG approach is deployed for the training system, in which the agent receives reward r_t accordingly, and generates the decision u_t.*

4. Reinforcement learning method for engagement adjustment

In the previous section, we discussed the critical factors for achieving more effective training. This section proposes a DRL-based approach to adaptively adjust training strategies. Traditional methods rely on experts' judgments and experiences to determine trainees' control authority, which can result in issues of efficiency and economy. Since training is a labor-intensive and long-term process, it is not feasible for experts to continuously monitor and regulate trainees' control authority in real time. Moreover, experts may not cater to the diverse needs of trainees, requiring a large amount of subjectively unstable teaching resources. Furthermore, as motor skills are implicitly related to the physical actions of trainees, experts are unable to directly teach or assess these motor programs.

The present study proposes the use of an agent-based methodology to determine the training strategy for trainees, thereby eliminating the need for human intervention. Specifically, a reinforcement learning framework is employed to obtain an optimized training strategy. This approach postulates that an agent interacts with the training environment continually, receives rewards from the environment, and refines its execution strategy in response. Essentially, the agent serves as the entity responsible for issuing actions that regulate the training strategy, while the environment represents the context in which the trainee practices. Through this interaction, the agent acquires valuable knowledge that can be abstractly conceptualized as rules or mechanisms governing the training process. For instance, in this study, the agent aims to learn the matching rule that aligns optimal skill acquisition with the trainee's engagement level. This problem formulation is commonly addressed by modeling a Markov decision process (MDP), which will be elaborated below.

4.1. Learning problem formulation

In this paper, the adaptive engagement regulation problem is defined as a sequential decision-making process, and this is a finite MDP with time step $t = \{1, 2, 3, \ldots, T\}$, which is reinforcement learning problem [34]. Four-tuple (S, U, P, r) are defined, where S represents the observation space of the system, U denotes the permissible actions, P represents the observation transition model, and r is the immediate reward.

Figure 2 depicts the overall diagram of the DRL-based agent decisions for engagement regulation, which uses a simulated environment. The trainee's skill information from the simulated environment is fed into an Actor-Critic network to approximate the action value of the agent decision u. Skill information is acquired according to the calculations in Section 3.1. The decision with the highest action value, u_i, is then selected for the agent.

4.1.1. State space

The state space S represents the environment state in a one-episode training. It comprises the trainee's skill level and covers the combined control signal that differs from the reference. Additionally, trainee engagement with respect to their skill level is also considered. Thus, the state variables S consist of three parameters: Φ_Δ, ω, and $\Delta(t)$. Φ_Δ represents the normalized task-independent skill level, ω denotes the trainee's control authority, which reflects their engagement level, $\Delta(t)$ and represents the operation trajectory deviations between the desired and trainee's console operations. For complex tasks, it may not be necessary to use a fixed desired trajectory. Alternatively, the expert's operational trajectory can serve as the desired trajectory. Therefore, in this paper, the performance under the current training strategy is evaluated by the deviation of the operational trajectory of the slave robot from the expert. The observation of agent s_t is the skill level from environment.

4.1.2. Action space

Within the training policy, the agent's actions have an impact on the control authority, which in turn influences both the slave robot's performance and the trainee's engagement. The agent's decision is defined by a simple model, shown in Eq. (6).

$$\omega(t) = \zeta \cdot \hat{\omega}(t), \tag{6}$$

where ζ represents the maximum level of control authority that the trainee can engage with, as determined by the instructor to meet safety requirements. The action executed by the agent is denoted by $\hat{\omega}(t)$ and ranges from 0 to 1. Consequently, the action space is defined as [0, 1].

4.1.3. Transition probabilities

The transition from the current state s_t to the next s_{t+1} is defined as

$$s_{t+1} = f(s_t, \hat{\omega}_t). \tag{7}$$

The agent decision is updated based on a vague mechanism of human ability growth. Subsequently, the agent alters the trainee's operation authority via a shared-control scheme. Modeling the accurate transition function for trainee performance change is challenging due to the highly complex nature of human skill progression. In this paper, a data-driven approach is proposed to provide the control action.

4.1.4. Reward

During the training process, the reward value obtained by the agent is based on the trainee's engagement in the task and the operating trajectory of the slave robot. The reward function is composed of two components: the difference between the trainee's engagement and their skill level at each moment, as well as the deviation between the slave robot's operating trajectory and the desired trajectory. Therefore, the reward function is defined as Eq. (8)

$$r_t = \omega(t) - \Delta(t). \tag{8}$$

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

The performance of slave robot is defined as Eq. (9).

$$\Delta(t) = X_R(t) - X_D(t), \tag{9}$$

In this paper, $\Delta(t)$ represents the degree of similarity between the slave output curves and the expected curves.

The agent's decision can be implemented by controlling the weight of authority to find an optimal policy related to ω for trainees with varying skill levels. When deploying the agent in the training system, at time t, the agent observes the current skill level of the trainee, and based on this state the agent executes an action to change the trainee's engagement in the training system. The objective is to maximize trainee engagement while minimizing deviations in performance.

4.2. Proposed approach

This paper utilizes Deep Deterministic Policy Gradient (DDPG), an actor-critic-based method, to regulate trainee engagement, as shown in Fig. 2. This structure comprises four neural networks: the actor network that outputs a continuous action $\hat{\omega}$; the critic network that evaluates the executed actions' performance $Q - value$, and two target networks for the actor and critic networks, respectively, ensuring convergence of the Q function. It is worth noting that although this implementation involves four networks, the actor-critic method can share the same network structure with their respective target network in practice. The network structure is as follows:

(1) Actor Network: In the actor network, the output is a continuous action, which represents the agent's decision. A deterministic policy, $\mu(s) = \hat{\omega}$, is defined with parameters μ_θ. For each state s, this policy outputs the action $\hat{\omega}$ that maximizes the action-value function $Q(s, \hat{\omega})$. Therefore, the optimal action can be represented as Eq. (10)

$$\hat{\omega}^* = \max_{\theta} \mathbb{E}[Q(s, \mu(s|\theta_\mu))], \tag{10}$$

(2) Critic Network: The critic network's output is Q-value, which is approximated by the $Q - value$ with parameter θ_Q. Because the optimal policy is deterministic, the optimal action-value function can be described as Eq. (11)

$$Q^*(s_t,\hat{\omega}_t)=\mathbb{E}_{s_{t+1}\sim\mu}\left[r(s_t,\hat{\omega}_t)+\max_{\mu(s_{t+1})}\gamma\left[Q^*(s_{t+1},\mu(s_{t+1}|\theta_Q))\right]\right], \tag{11}$$

(3) Target Networks: The target network is a copy of the main network that is updated slowly using a soft update rule. Two target-networks are introduced, one for the Q network and the other for the actor network. They are defined as θ'_μ, θ'_Q. To ensure stability during network updates, the target networks are updated less frequently than the main networks. The update methods for these networks are as follows:

$$\theta'_Q \leftarrow \tau\theta'_Q + (1-\tau)\theta'_Q \tag{12a}$$

and

$$\theta'_\mu \leftarrow \tau\theta'_\mu + (1-\tau)\theta'_\mu \tag{12b}$$

The term τ is a hyperparameter between 0 and 1, this term can make the update of the target network lag behind the main network. All of the hyperparameters used in our DDPG control algorithm are listed in Table I.

Specifically, α_a and α_c represent the learning rate of actor and critic network, respectively. Furthermore, τ is target smoothing coefficient that we use to balance the weight updates of the two networks in order to optimize performance. Additionally, N_a and N_c represent the number of nodes in actor and critic network; the γ is a discount factor and set as 0.99; the batch size (BS) is 256; and the update frequency for target networks is 3.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

Table I. *Hyperparameters in DDPG controller.*

α_c	α_a	τ	N_a	N_c	γ	**BS**	**UF**
0.0001	0.0001	0.005	400	400	0.99	256	3

Algorithm 1 Training of the Actor and Critic Network.

```
Require: Training system state $s$ and reward $r$;
Ensure: Actor and critic parameters $\theta_\mu$,$\theta_Q$;
1: Randomly initialize main networks parameters $\theta_\mu$,$\theta_Q$
2: Initialize the target network and their parameters $\theta'_\mu$,$\theta'_Q$
3: Initialize the replay buffer $\mathcal{B}$
4: for episode = 1 : N do
       observation $s_1$ and with a random process for action exploration.
5:     for time step $t = 1$ : T do
6:         Select action according to the current policy and exploration noise: $\hat{\omega}_t = \mu\left(s_t|\theta_\mu\right) + \mathcal{G}_t$
7:         Execute this action $\hat{\omega}$, the control authority of trainee is $\omega(t)$, receive the reward $r_t$ and
     observe the new state $s_{t+1}$
8:         Store transition($s_t, \hat{\omega}_t, r_t, s_{t+1}$) in $\mathcal{B}$
9:         Sample a random minibatch of $N$ transitions $(s_i, \hat{\omega}_i, r_i, s_{i+1})$ from $\mathcal{B}$
10:            $y_i \leftarrow r_i + \gamma Q(s_{i+1}, \mu(s_{i+1}, \omega_{i+1}|\theta'_\mu)$
11:        Calculate the loss function:
             $L = \frac{1}{N}\sum_i (y_i - Q(s_i, \mu(s_i|\theta_\mu))^2$
12:        Update the actor policy using the sampled policy gradient:
             $\nabla_{\theta_\mu} J \approx \frac{1}{N}\sum_i \nabla_\omega Q(s, \omega|\theta_Q)|_{s=s_i,\omega=\mu(s_i)} \nabla_{\mu_\theta}\mu(s|\theta_\mu)|_{s_i}$
13:        Update the target networks:
             $\theta'_Q \leftarrow \tau\theta'_Q + (1-\tau)\theta'_Q$
             $\theta'_\mu \leftarrow \tau\theta'_\mu + (1-\tau)\theta'_\mu$
14:    end for
15: end for
```

4.3. Training of the networks

The process of parameter learning is illustrated in Algorithm 1. This algorithm details the training procedures for both the Actor and Critic networks. γ is discounted factor. The BS is the size of batchsize, and the UF is the update frequency of the target network.

Algorithm 1 takes the training system state s and reward r as input, and returns the actor-network parameter μ_θ and the critic-network parameter ϕ, respectively.

At the beginning of each episode, the environment is initialized. In line 1, all parameters are randomly initialized, followed by the initialization of all target networks. The iteration starts from line 3, and the parameters θ are updated through N episodes. When observing the first state s_1, the inner loop from line 5 to the line 6, the agent will select an action based on the current policy $\mu(s_t|\theta_\mu)$. To ensure the exploration, a noise $\mathcal{G}_t$ is used. This noise added can prevent the local optimum in training. Ornstein-Uhlenbeck Process (OUP) is chosen as the noise because of its temporally correlated nature. More details about OUP can be found in study [35]. After executing the action, an immediate reward is received. The buffer $\mathcal{B}$, which is in line 3 and line 8, is used to store the interaction trajectories. This is the so-called experience replay approach. Specifically, in line 9, a minibatch of tuples $(s_i, \omega_i, r_i, s_{i+1})$ is randomly sampled from the buffer $\mathcal{B}$. From line 10 to line 13, the parameters are updated by minimizing the loss

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

Algorithm 2 Agent Decision-Making.

```
Require: Information on trainee operation
Ensure: Agent decisions $\hat{\boldsymbol{\omega}}_{1:T}$;
1:  Load the parameters $\theta$ trained by Algorithm 1.
2:  Randomly initialize the engagement $\omega_0$.
3:  for Time step $t = 1{:}T$ do
4:      Receives information about the trainee's operation device $X_T(t)$, and obtains the initial training
        system information $\omega_0$ and $X_R(t)$.
5:      Extract features from the dual-user training system through data processing.
6:      Critic network calculates action-value $Q(s_t, \hat{\boldsymbol{\omega}};\theta)$.
7:      $\hat{\boldsymbol{\omega}}_t \leftarrow Q^*(s_t, \hat{\boldsymbol{\omega}};\theta)$
8:      Execute the action $\hat{\boldsymbol{\omega}}_t$.
9:  end for
```

function. In line 13, two equations represent a soft update in target networks. τ is a very small factor, and it is set as 0.001, which is used to overcome the unsuitability of neural network training. This soft update method can improve the stability of the learning process.

Upon completion of the training process, the learned parameters θ will be employed for real-world engagement adjustment scenarios.

4.4. Adaptive regulation for trainee engagement

Algorithm 2 describes the engagement adjustment process. Its input is the trainee's skill state, and its output is the agent's decisions regarding the trainee's level of engagement. In line 1 of Algorithm 2, the trained parameters θ from Algorithm 1 are loaded. The initial level of engagement ω_0 for the trainee is defined as an authority in the control workspace. Starting from line 3 of the algorithm, the agent is controlled by the DDPG algorithm to regulate the trainee's engagement over a time period of T steps. At each interaction, the data of the system is obtained from the trainee's operation device. This data is then fed into the DDPG algorithm to model the action and state of the system, as described in Section IV-A. In line 6, these states are used as inputs to the critic network, which calculates the action-value $Q(s_t, \hat{\boldsymbol{\omega}};\theta)$ for the agent's decisions. Then, in line 7, the decision $\hat{\omega}$ is selected as $Q^*(s_t, \hat{\boldsymbol{\omega}};\theta)$. Finally, ω_t is outputted as the engagement policy.

5. Simulation experiment and results evaluation

The effectiveness of the DDPG agent decision-making method is verified through simulation experiments. In this section, we first introduce our platform setup and the simulation experiments. Next, we present the agent training process, which serves as the ground truth for algorithm validation. Finally, we evaluate the performance of the proposed approach under different starting engagements of trainees and skill conditions.

5.1. Setup design and implementation

In order to evaluate the proposed method, we set up a dual-console platform designed to gather operation information. The platform consists of two Omega 3 devices deployed on the master side and a Kuka iiwa robot on the slave side. In this setup, an expert manipulator and a trainee manipulator are connected to a computer, which is responsible for fusing the two control signals from the master side and transmitting the synthesized signals to the slave robot. The application software that communicates

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

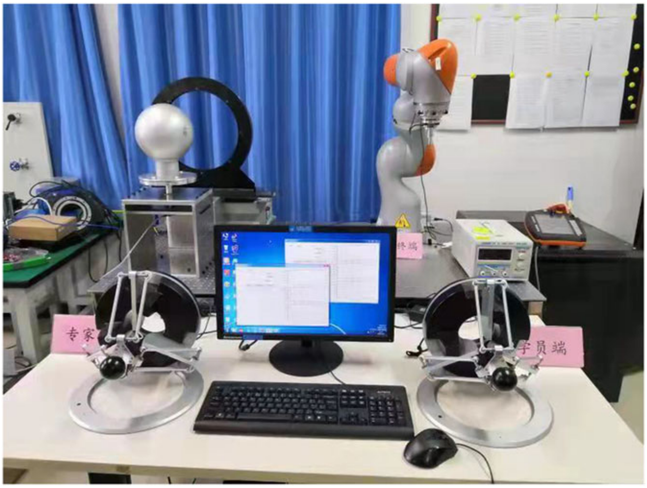


Figure 3. *Dual-user training platform.*

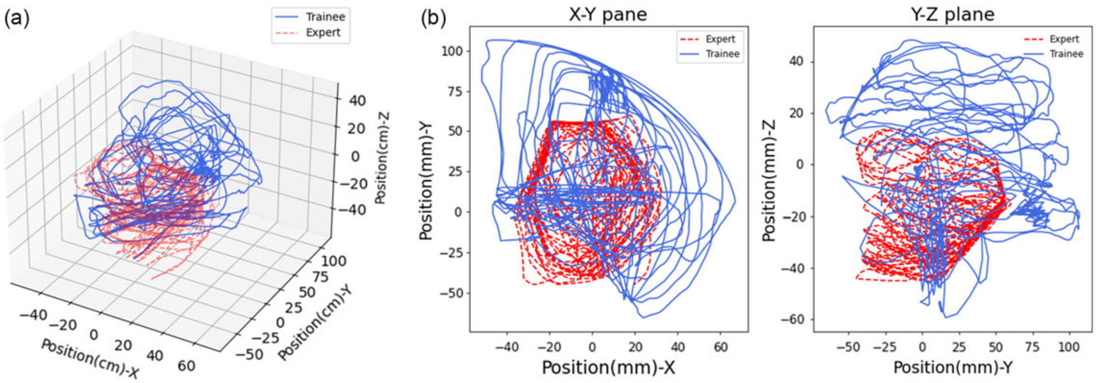
with the haptic devices is implemented in C++ using a Qt API. Meanwhile, the computer also records operation information from the two users at a sampling interval of 0.001 s. The slave robot receives the control command from the computer and the operation track in space is mapped in a 1:1 coordinate system. This experimental platform is applicable to the scenario of dual-user training, facilitating the collection of operation information during the training process and providing verification for future training strategies. Figure 3 displays the device composition. The instantaneous output X_R corresponds to the synthetic track calculated by Eq. (1) and is generated jointly by the trainee and the expert on the master side. The control authority is determined by the agent.

The parameters learning of agent is conducted via another computer. Specifically, the training and verifying of our proposed approach were performed on an Apple M1 GPU. The Actor and Critic networks were implemented in Python, while the combined control signal was developed using Gym.

5.2. Simulation scenarios

We conducted extensive simulations with a random initial shared policy as well as different skill levels of the trainee. Two operators participated in the experiments, simulating both an expert's and a trainee's behaviors. We considered three scenarios for interpretation, where the trainee's skill level is represented as a normalized value showing their relative performance compared to that of the expert. As can be easily seen from Eq. (2), this value ranges from 0 to 1. In every episode start, the trainee's control authority is set from [0–1] randomly. Furthermore, we define task safety as the root mean square error (RMSE) calculated from the deviation between the actual operational trajectory and the desired trajectory. Specifically, in this study, we investigate the impact of varying safety requirements on the effectiveness of the proposed approach in different scenarios. The safety requirements are manually set by experts based on the specific demands of each task. This enables us to explore the use of diverse safety requirements for different scenarios and evaluate the effectiveness of the proposed approach under various levels of safety requirements.

Scenario I: Two operators are required to perform the same task simultaneously, and there are no fixed requirements on how to perform the task. The trainees' operations will be based on their own understanding of the task, which may result in a diverse range of trajectories and problem-solving methods. The expert's operational trajectory will serve as a reference for comparison, enabling the evaluation of each trainee's performance in relation to the established standard. The trajectories produced during operation are shown in Fig. 4 and Fig. 6.



Space

2D-Plane

Figure 4. *Operation track comparison in space without agent.*

Scenario II: This scenario involves an operator performing a task within a specific restricted range, with the resulting trajectory displayed in Fig. 8(a), (b), and (c). The three cases presented correspond to three distinct skill levels, namely 0.9, 0.4, and 0.8. In this setting, an expert performs the operation only once, and their performance data is stored as a reference. Three different individuals then perform the task according to predefined instructions.

Operating within a restricted range provides the operator with guidance, which we leverage to enhance safety requirements during the operation. Specifically, we set the maximum RMSE requirement for all cases to be no higher than 0.15.

Scenario III: In the scenarios I and II, the reference trajectories for comparing is provided by the expert's operation in real-time. In this scenario, we introduce a desired trajectory to evaluate the operation. The expert and trainee will perform the task operations simultaneously, similar to previous scenarios, while combined operation signals will be utilized to calculate deviations from the desired trajectory. This enables us to assess both training engagement and operational safety by examining deviations from the desired trajectory. The safety requirements in this scenario need to be determined depending on the different tasks, so the specific values will be illustrated in their case.

5.3. *Learning process of agent*

5.3.1. *Scenario I*

Case 1-1: In this case, Fig. 4(a) depicts information regarding the operation in space during task execution. The solid blue line represents the trajectory of the trainee at the end of the operated device, while the red dashed line shows the trajectory formed by the end of the operated device of expert. Figure 4(b) displays the difference in trajectory under 2-D plane, which clearly depicts the difference between the trainee and the expert in the operation process. During the task execution, it was observed that the trajectory formed by the trainees was less smooth than that of the experts. Furthermore, the trainees exhibited numerous redundant movements while attempting to complete the task. These observations suggest that trainees may have a less refined understanding of the motor skills required for the task compared to the experts. To evaluate the trainee's skill, an analysis was conducted on the total length and smoothness of the trajectories of both the trainee and the expert. Based on this analysis, the trainee's skill was assessed as 0.92. In the experiment, the RMSE of the output trajectory formed by the combined control signals of the trainee and the expert, and the trajectory between the experts, was utilized as a deviation measure. To ensure the end output remains within an acceptable safety range, the value of this RMSE must not exceed 0.2.

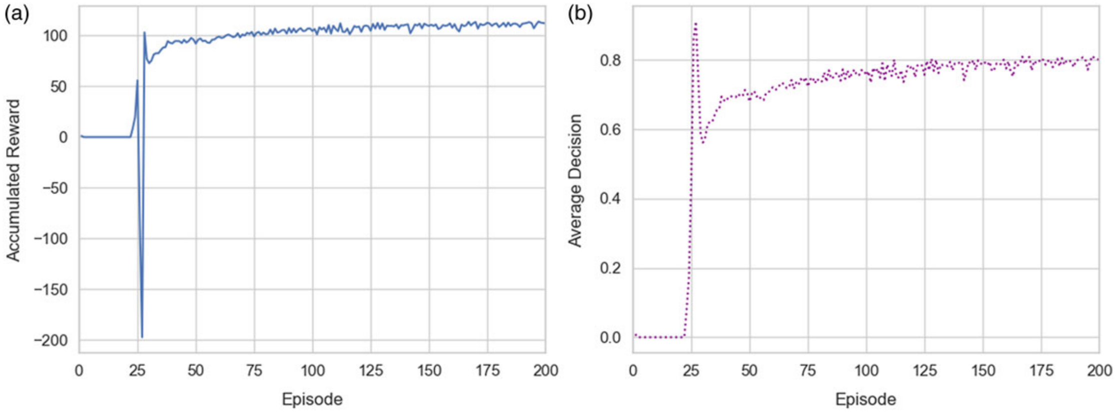


Figure 5. *Evolution of the accumulated rewards at $T = 200$ s over 200 episodes during the training process. The right figure is the average decision in each episode.*

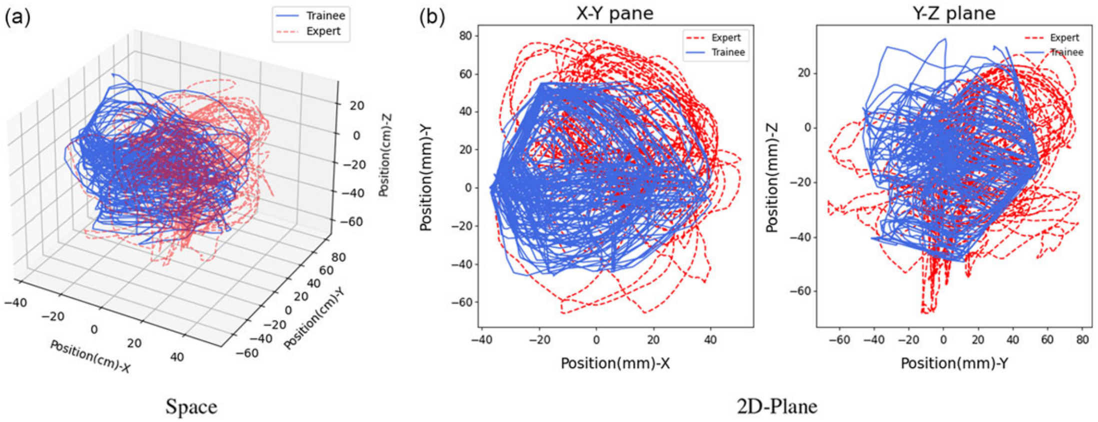


Figure 6. *Operation track comparison in space without agent.*

To determine the optimal trainee engagement based on the trainee's skill evaluation, the agent was employed to observe the skill level to find the control authority that suit the trainee's skill level.

We present the training performance by running Algorithm 1, the agent is trained for 200 episodes to learn the optimal decision about engagement. A total of 40,000 steps is run, and the simulation time for each episode is 200 s. The trainee's engagement is randomly chosen from [0–1] in each episode. The evolution of the accumulated reward at $T = 200$ s over 200 episodes is shown in Fig. 5(a). The average decision in each episode is shown in Fig. 5(b)

From the results, it can be seen that the rewards obtained by the agent gradually increase, and the decision of the agent gradually converges from 50 episodes.

Case 1-2: In this case, we intentionally tested the decision-making abilities of the agent by simulating a task operation resulting in a trainee performance level of 0.74. By observing the agent's decision-making process under conditions of reduced operational skill, we aim to gain insights into its ability to adapt and perform effectively in different scenarios.

The trajectory of two operators is shown in Fig. 6.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

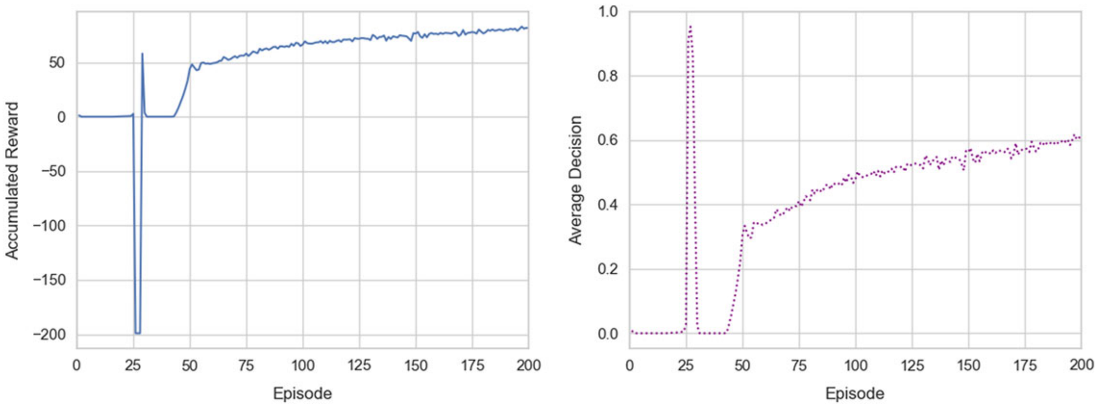


Figure 7. *Evolution of the accumulated rewards at $T = 200$ s over 200 episodes during the training process. The right figure is the average decision in each episode.*

Figure 6(a) shows the operational trajectory information in 3D space, and Fig. 6(b) is the 2D plane display of the operational information. From the trajectory information, it can be observed that in order to complete a task, the trainees need to execute a trajectory of higher length than that of the expert, and the acceleration variation generated by it is also higher than that of the expert. Finally, in this case, the trainee's skill was evaluated as 0.74. In this case, the evolution of the accumulated reward is show in Fig. 7(a). Concurrently, Fig. 7(b) illustrates the mean decision value throughout the training process.

5.3.2. Scenario II

Case 2-1: In this case, the trainee's skill level is evaluated as 0.9 as in Fig. 8(a). In the training process, the reward gained is shown in Fig. 9(a).

Case 2-2: In case 2, the trainees' skills improved and the level skills are assessed as 0.4, and the trainee is asked to follow a predefined trajectory. As in Fig. 8(b), two users' tracks are shown. Through 200 episodes, the learning process of the agent is shown in Fig. 9(b).

Case 2-3: In case 3, the trainee's level is 0.8, his/her track as shown in Fig. 8(c). Figure 9(c) shows the accumulated rewards in this case during the training process.

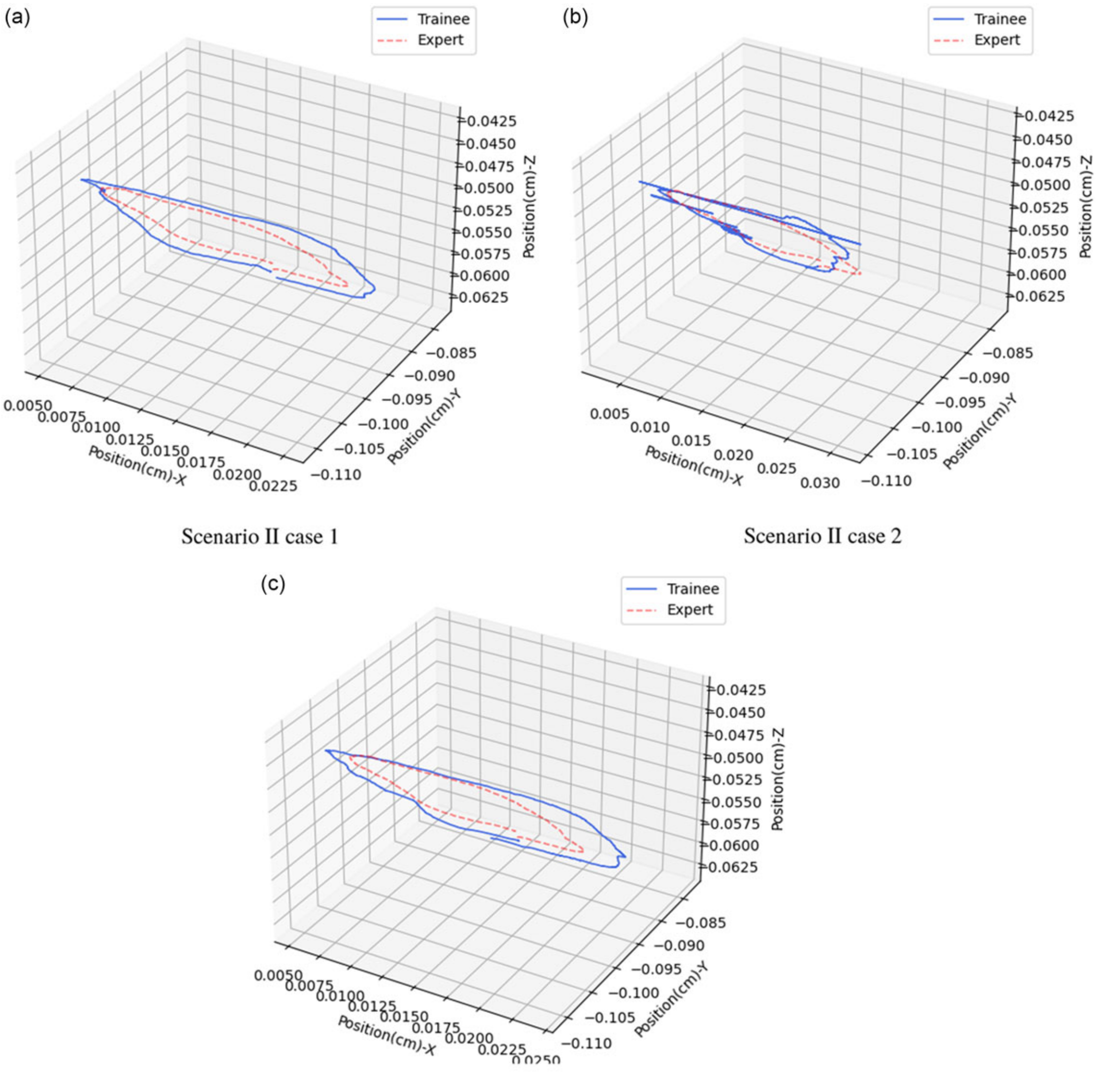
From the accumulated rewards obtained by the agent over the 200 episodes, the accumulated rewards gradually increase and eventually converge. In this scenario, the average decision under different skill levels is shown in Fig. 10(a),(b), and (c). At three different skill levels, the average decision eventually converges around 0.90, 0.55, and 0.86.

5.3.3. Scenario III

Case 3-1: This case has similarities with Scenario II, as both operators are required to operate within the designated range of operation. However, in this case, a desired operation trajectory is added, which is depicted by the green line in Fig. 11. The red and blue lines in Fig. 11 represent the expert and trainee operators, respectively. To evaluate the performance of the trainee operator, we compare their operation to that of the expert's using relative scores. In the final evaluation of the operation, we measure the combined operation trajectory against the desired trajectory path, enabling us to establish the degree of deviation from the expected trajectory path.

In this case, we set the value of the safety factor so that the rmse value of the trajectory error does not exceed 0.5. For this case, the learning of the agent is shown in Fig. 12. Figure 12 presents the accumulated rewards obtained and the change in the average trainee engagement. From the results, it can be seen that

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press



Scenario II case 1

Scenario II case 2

Scenario II case 3

Figure 8. *Operation track comparison in three-dimensional space without agent.*

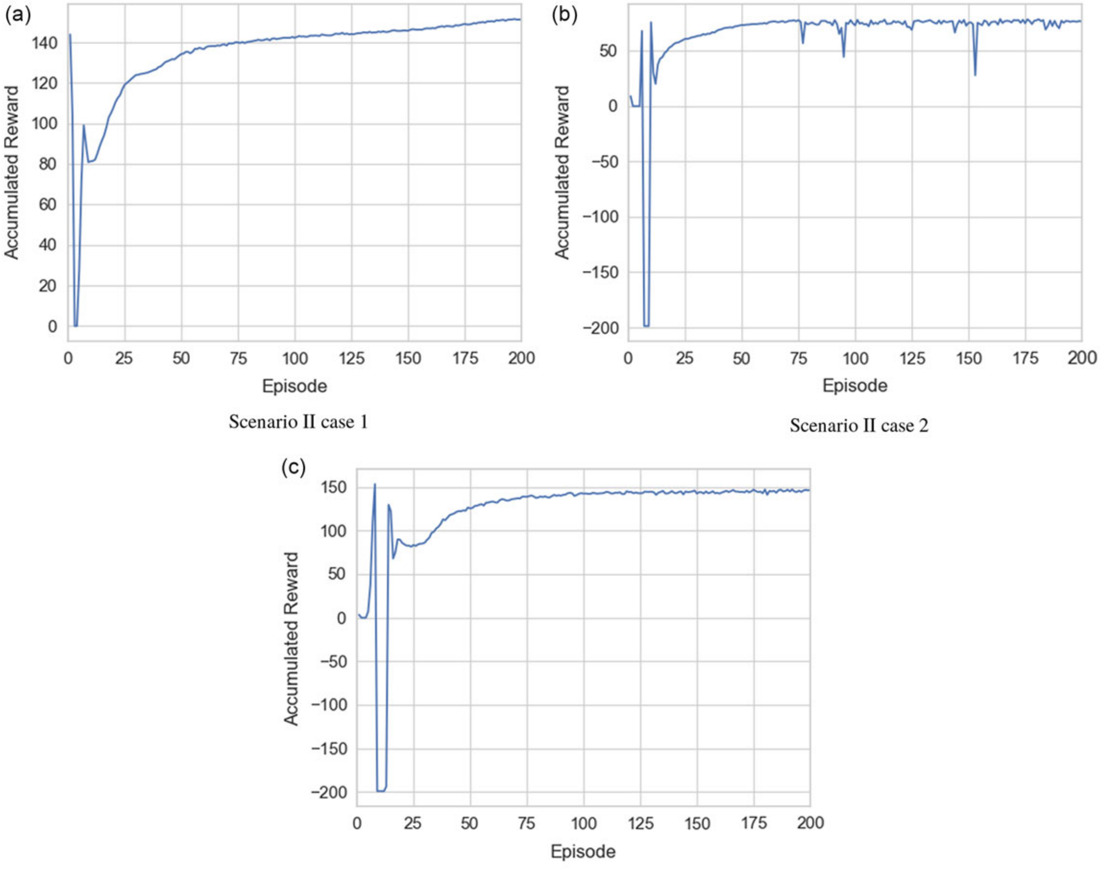
the rewards gradually increased and the engagement level also converged. The convergence value of the engagement is finally equal to approximately 0.60.

Case 3-2: In this case, the operation without constraints and the trajectory is shown in Fig. 13. The green line is the desired trajectory. Similarly, red and blue are the operation trajectories of experts and trainees. It can be seen from the figure that the expert's trajectory is more similar to the desired trajectory. In this case, the trainee's skill level was assessed as 0.55.

The safety requirement is set as 0.33. The accumulated rewards in 200 episodes is shown in Fig. 14. The average engagement of trainee is converged to 0.55.

The results of this training demonstrate the effectiveness of the proposed approach in learning a training policy that maximizes trainee engagement under varying skill levels while ensuring safety. Through this approach, we were able to provide trainees with training scenarios that adapted to their individual skill levels, resulting in higher engagement and safety.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press



Scenario II case 1

Scenario II case 2

Scenario II case 3

Figure 9. *Accumulated rewards in three cases.*

5.4. DRL control for engagement regulation

After the training, Algorithm 2 is run and presents the agent decision results in this section. The engagement will be controlled by the agent. The proposed approach is evaluated under all cases defined in Section V-A. In each case, we conducted extensive 10 simulations with random initial engagement.

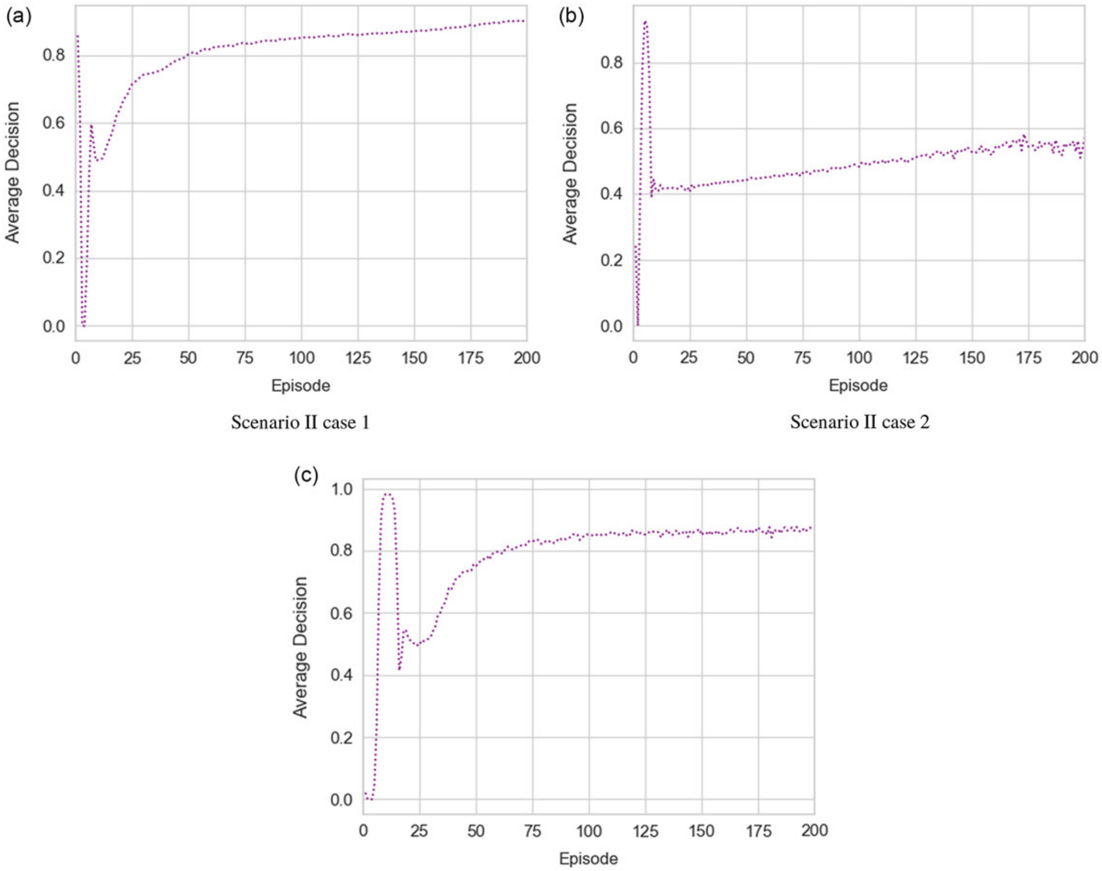
Scenario I: In case 1, we evaluated the trainee's skill level as 0.92 while randomly selecting their engagement level from the 10 available options. The results presented in Fig. 15(a) demonstrate that the engagement level can converge to 0.81 in all tests, considering the trainee's skill level and safety requirements. These results suggest that the agent can successfully adjust the engagement level to 0.81 irrespective of the initial engagement level, based on the trainee's skill level. This improves training outcomes by ensuring an appropriate level of engagement for optimal performance while maintaining safety standards. Figure 15(b) corresponds to the second case of scenario I, whose trainees' skills were assessed as 0.74 and the optimal engagement was 0.57.

Scenario II: We loading the trained model into Scenario II, and 10 random initial engagements are deployed. The decision of the agent is shown in Fig. 16.

Scenario III: When load the model into Scenario III, the decisions of agent are presented in Fig. 17. In case 1, the decision is converged to 0.61. In case 2, the decision is converged to 0.55.

The statistical results above reflect the method's ability to adapt training strategies for different levels of trainees based on different safety requirements. When the trainee's skills change, the agent is able to

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press



Scenario II case 1

Scenario II case 2

Scenario II case 3

Figure 10. *Average decision in three cases.*

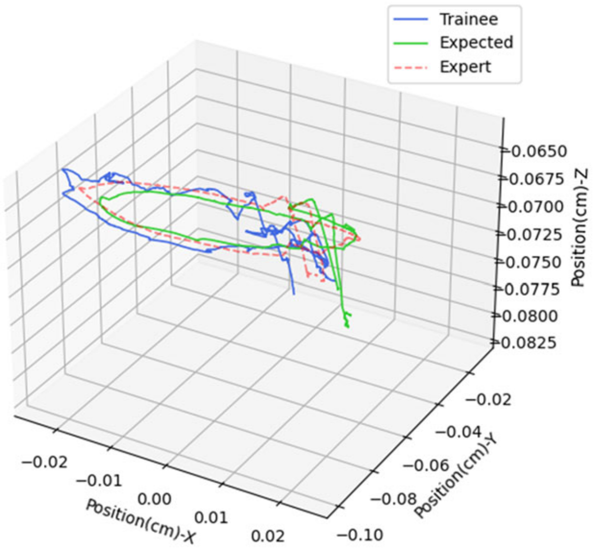


Figure 11. *The operation trajectory with a expected reference.*

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

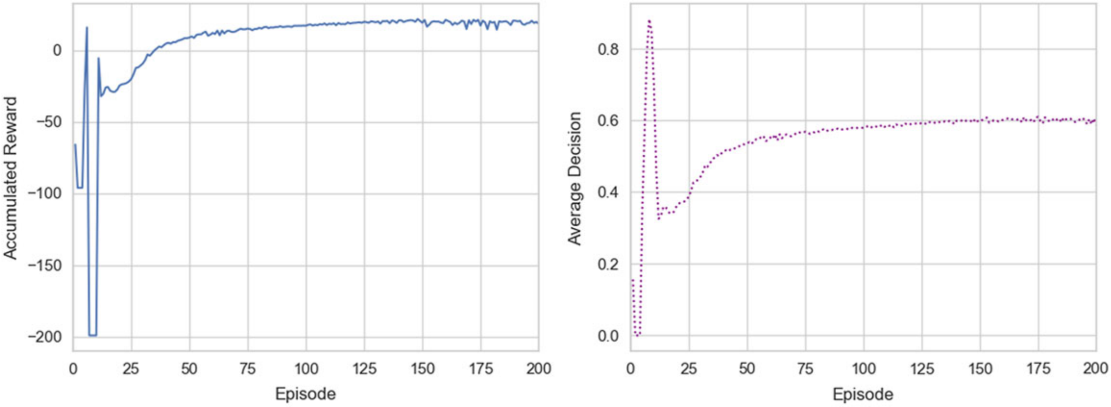


Figure 12. *Evolution of the accumulated rewards at $T = 200$ s over 200 episodes during the training process. The right figure is the average decision in each episode.*

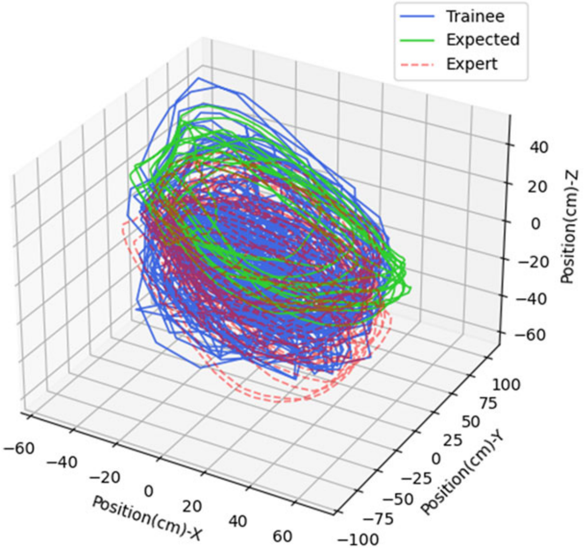


Figure 13. *The operation trajectory with a expected reference.*

find the optimal training strategy based on the current skill level to ensure maximum engagement and task safety.

5.5. Error analysis for engagement

In this section, we compare the error in the operational effect produced by the controlled robot at its end, with and without the agent's regulation of the trainee's participation. Specifically, we evaluate the impact of the agent's regulation on the error in the operational effect generated by the controlled robot.

It describes the results obtained from the proposed approach, indicating that the deviation of the operation is reduced when using the agent's regulation compared to the case without the agent (Fig. 18). Additionally, the safety requirements are met by the agent in both cases. The conclusion highlights the effectiveness of the proposed approach in improving trainee participation and ensuring task safety.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

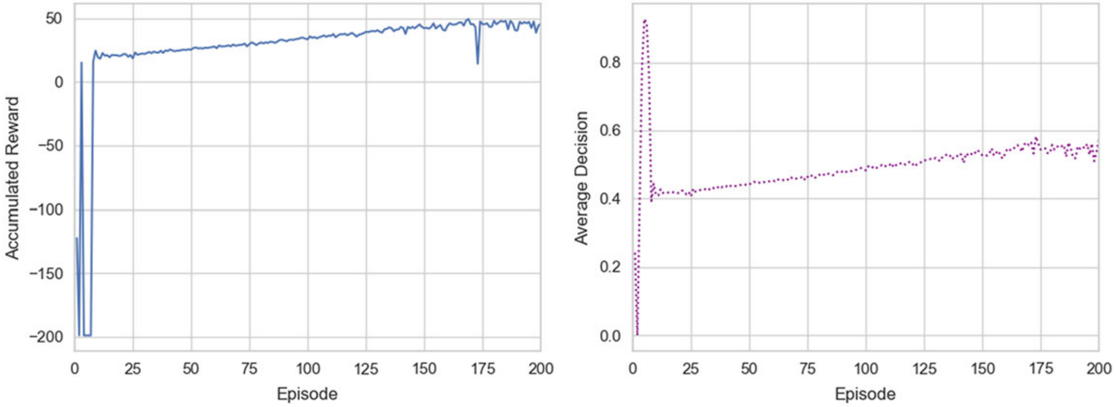
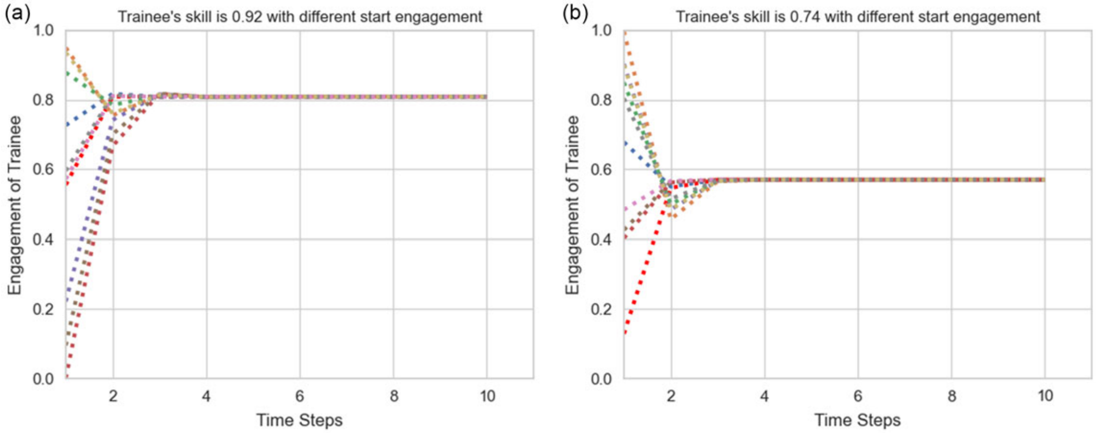


Figure 14. *Evolution of the accumulated rewards at $T = 200$ s over 200 episodes during the training process. The right figure is the average decision in each episode.*



Scenario I case 1

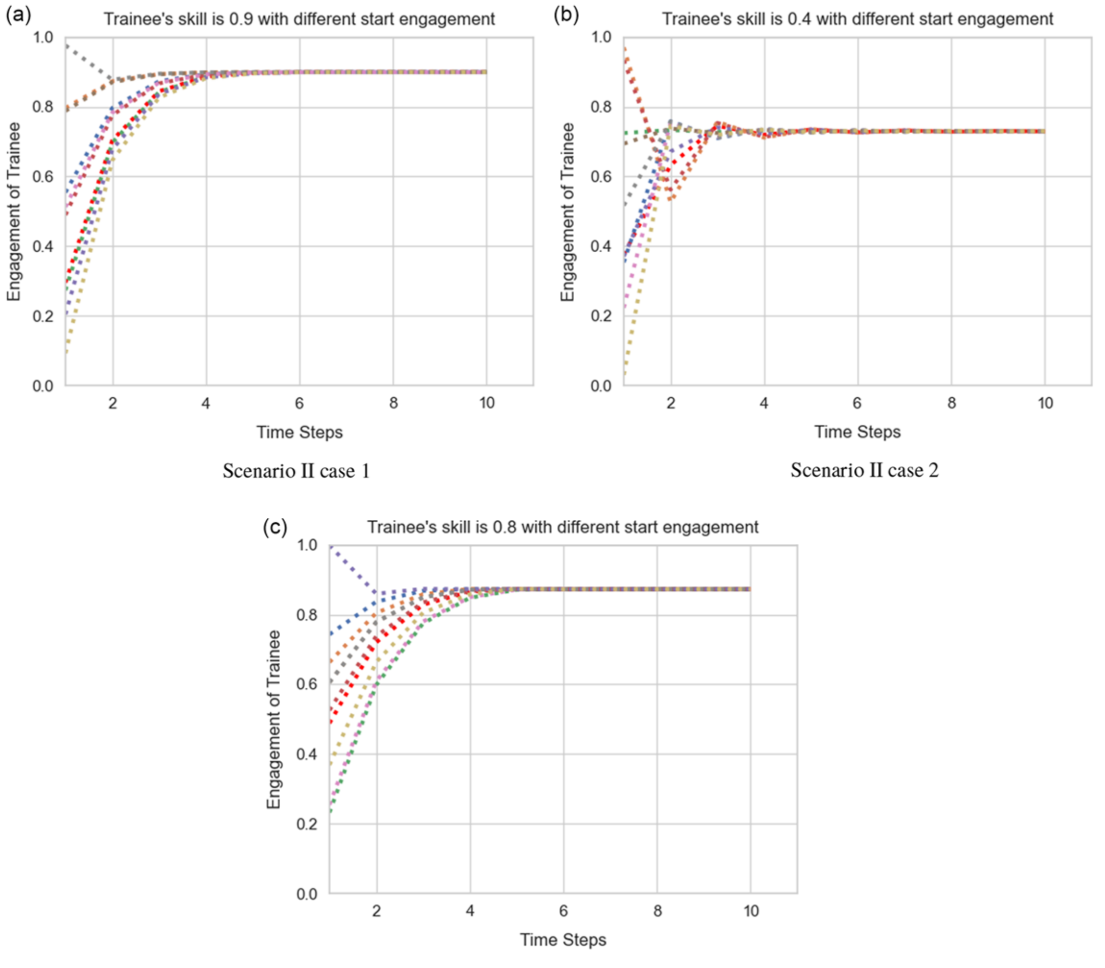
Scenario I case 2

Figure 15. *The agent's decision in Scenario I.*

6. Conclusion and discussion

In this paper, we propose a DRL-based agent decision-making approach for regulating training participation. The agent decision-making problem is solved using a learning-based sharing system so that the optimal strategy for agent decision-making is learned to maximize the trainee's participation. The proposed approach in this study offers a logical framework that establishes a relationship between the trainee's skill level and the achievable level of engagement in online training. By investigating this connection, we aim to enhance the understanding of how the trainee's proficiency impacts their level of engagement during the training process. Extensive simulations are conducted in this paper and the results validate the effectiveness of the proposed method for trainee engagement. Regulating trainee engagement can be attributed to a kind of control strategy problem, and this type of nonlinear problem [19, 36] is explored in depth. The approach proposed in this study adds to the richness of existing approaches by introducing a novel reinforcement learning approach that effectively addresses the problem of regulating trainee engagement in online training scenarios. By leveraging this method, we can optimize the decision-making strategies of intelligent agents to enhance the overall effectiveness of the training process. The introduction of an AI approach, as opposed to being set manually by an expert, offers significant advantages in terms of resource cost reduction and enhanced flexibility for different

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

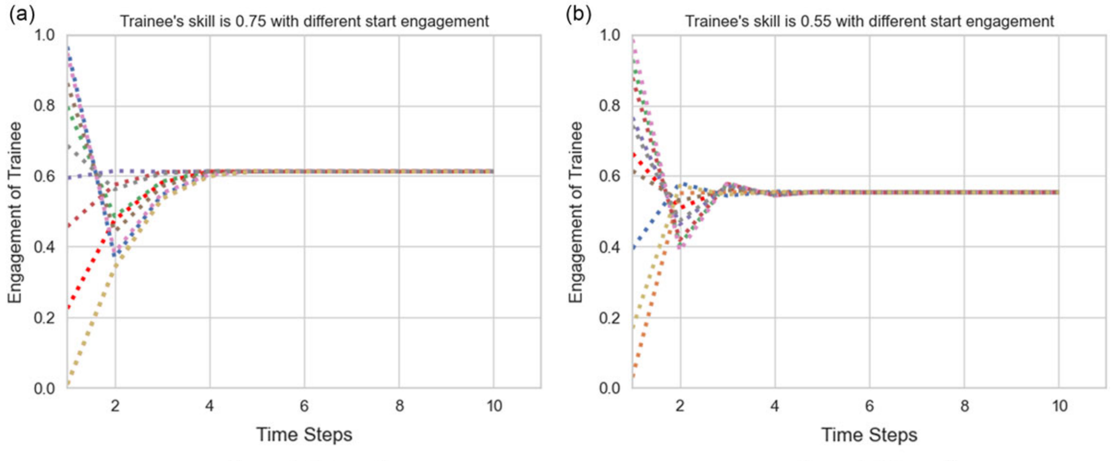


Scenario II case 1

Scenario II case 2

Scenario II case 3

Figure 16. *The agent's decision in Scenario II.*



Scenario III case 1

Scenario III case 2

Figure 17. *The agent's decision in Scenario III.*

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

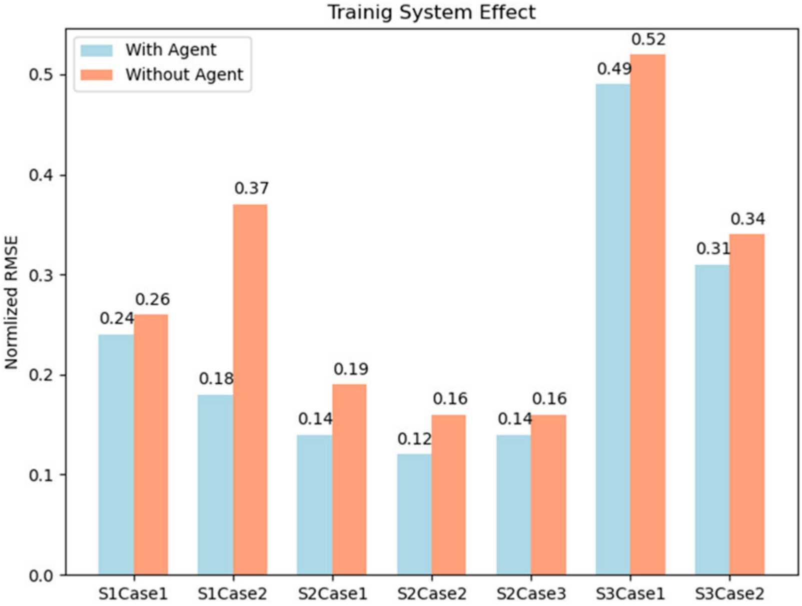


Figure 18. *The root mean square error.*

trainees in online training scenarios. While the agent training phase requires a considerable amount of time and data, it enables the deployment of a trained agent who can efficiently adapt to the trainee's level of participation. By employing an AI approach, the need for continuous supervision by experts is eliminated, leading to substantial resource cost reductions. Additionally, the AI approach can be applied more flexibly to accommodate diverse trainee profiles, as it leverages its own learning ability to identify optimal levels of training participation, even in the presence of new skill states. This deployment of a trained agent enables the achievement of a certain degree of application generalization. The agent can promptly respond to the trainee's level of engagement, ensuring an appropriate and adaptive training experience. Furthermore, when faced with a larger number of trainees simultaneously, the agent can assign appropriate participation levels based on observed skill levels, bypassing the need for extensive individualized analysis.

Undoubtedly, the deployment of AI-driven approaches in real-world scenarios presents several challenges, as exemplified by the cases discussed in this study [37]. To address these challenges and further advance the research, future work will concentrate on implementing the proposed approach in practical settings. Specifically, our plan entails establishing a comprehensive training system equipped with a range of multimodal sensors capable of capturing diverse operation characteristics. This setup aims to gather real-time operational signal data from the sensors, which will subsequently be transmitted to the agent via a local-area network. By leveraging our proposed algorithm, the agent will then calculate and optimize the engagement level based on the received operation information and instantaneous outflow data.

Author contributions. In this paper, Yang Yang has done most of the algorithm design and simulation works and written the paper. Xing Liu has guided the study, given many suggestions, and supported the simulation implementations in this paper. Zhengxiong Liu and Panfeng Huang have provided many important comments on this paper.

Financial support. This work was supported in part by the National Natural Science Foundation of China under Grant 62273280, 62103334, and 92370123.

Competing interests. The authors declare no competing interests exist.

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

Ethical approval. Not applicable.

References

[1] N. Abe, J. Zheng, K. Tanaka and H. Taki. “A Training System using Virtual Machines for Teaching Assembling/Disassembling Operation to Novices,” **In:** *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929)*, **3**, (1996) pp. 2096–2101.

[2] D.K. Harrington and J.E. Kello. “Systematic Evaluation of Nuclear Operator Team Skills Training: A Progress Report,” **In:** *Conference Record for 1992 Fifth Conference on Human Factors and Power Plants* (1992) pp. 370–373.

[3] P. Huang and Z. Lu. “Auxiliary Asymmetric Dual-user Shared Control Method for Teleoperation,” **In:** *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)* (2015) pp. 267–272.

[4] E. Keskinen and K. Hernetkoski, “Chapter 29 - Driver Education and Training,” **In:** *Handbook of Traffic Psychology* (B. E. Porter, ed.) (Academic Press, San Diego, 2011) pp. 403–422.

[5] M. Shahbazi, S. F. Atashzar, C. Ward, H. A. Talebi and R. V. Patel, “Multimodal sensorimotor integration for expert-in-the-loop telerobotic surgical training,” *IEEE Trans. Robot.* **34**(6), 1549–1564 (2018).

[6] K. Shamaei, L. H. Kim and A. M. Okamura. “Design and Evaluation of a Trilateral Shared-control Architecture for Teleoperated Training Robots,” **In:** *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015) pp. 4887–4893.

[7] L. Fricoteaux, I. M. Thouvenin and J. Olive. “Heterogeneous Data Fusion for an Adaptive Training in Informed Virtual Environment,” **In:** *2011 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems Proceedings* (2011) pp. 1–6.

[8] O. van der Meijden and M. Schijven, “The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: A current review,” *Surg. Endosc.* **23**(6), 1180–1190 (2009).

[9] Y. Wang, Y. Chen, Z. Nan and Y. Hu. Study on Welder Training by Means of Haptic Guidance and Virtual Reality for Arc Welding,” **In:** *IEEE International Conference on Robotics and Biomimetics* (2006) pp. 954–958.

[10] M. Zahabi, J. Park, A. M. A. Razak and A. D. McDonald, “Adaptive driving simulation-based training: Framework, status, and needs,” *Theor. Issues Ergon. Sci.* **21**(5), 537–561 (2020).

[11] C. D. Lallas, J. W. Davis, and Members of the Society of Urologic Robotic Surgeons, “ Members of the Society of Urologic Robotic Surgeons. Robotic surgery training with commercially available simulation systems in 2011: A current review and practice pattern survey from the society of urologic robotic surgeons,” *J. Endourol.* **26**(3), 283–293 (2012).

[12] P. M. Fitts and M. I. Posner, *Human performance*, (Brooks/Cole Publishing Company, Pacific Grove, CA, 1967) p. 162.

[13] G. Wulf, C. Shea and R. Lewthwaite, “Motor skill learning and performance: A review of influential factors,” *Med. Educ.* **44**(1), 75–84 (2010).

[14] G. Ganesh, A. Takagi, R. Osu, T. Yoshioka, M. Kawato and E. Burdet, “Two is better than one: Physical interactions improve motor performance in humans,” *Sci. Rep.* **4**(1), 3824 (2014).

[15] B. Khademian and K. Hashtrudi-Zaad, “Shared control architectures for haptic training: Performance and coupled stability analysis,” *Int. J. Robot. Res.* **30**(13), 1627–1642 (2011).

[16] B. Khademian and K. Hashtrudi-Zaad, “Dual-user teleoperation systems: New multilateral shared control architecture and kinesthetic performance measures,” *IEEE/ASME Trans. Mechatron.* **17**(5), 895–906 (2012).

[17] H. Thieme, J. Mehrholz, M. Pohl, J. Behrens and C. Dohle, “Mirror therapy for improving motor function after stroke,” *Cochrane Datab. Syst. Rev. (Online)* **3**, CD008449 (2012).

[18] Z. Liu, D. Yang, Y. Wang, M. Lu and R. Li, “Egnn: Graph structure learning based on evolutionary computation helps more in graph neural networks,” *Appl. Soft Comput.* **135**, 110040 (2023).

[19] Y. Shi, L. Li, J. Yang, Y. Wang and S. Hao, “Center-based transfer feature learning with classifier adaptation for surface defect recognition,” *Mech. Syst. Signal Process.* **188b**, 110001 (2023).

[20] Y. Wang, Z. Liu, J. Xu and W. Yan, “Heterogeneous network representation learning approach for ethereum identity identification,” *IEEE Trans. Comput. Soc. Syst.* **10**(3), 890–899 (2023).

[21] S. Cotin, N. Stylopoulos, M. P. Ottensmeyer, P. F. Neumann and S. Dawson. “Metrics for Laparoscopic Skills Trainers: The Weakest Link!,” **In:** *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002, 5th International Conference,, Tokyo, Japan, September 25-28, 2002, Proceedings, Part I* (2002).

[22] D. Feth, B. A. Tran, R. Groten, A. Peer and M. Buss. *Shared-Control Paradigms in Multi-Operator-Single-Robot Teleoperation* (Springer, Berlin Heidelberg, Berlin, Heidelberg, 2009) pp. 53–62.

[23] N. Hogan and T. Flash, “Moving gracefully: Quantitative theories of motor coordination,” *Trends Neurosci.* **10**(4), 170–174 (1987).

[24] M. Shahbazi, S. F. Atashzar and R. V. Patel. “A Dual-user Teleoperated System with Virtual Fixtures for Robotic Surgical Training,” **In:** *IEEE International Conference on Robotics and Automation* (2013) pp. 3639–3644.

[25] C. R. Kelley, “What is adaptive training?,” *Hum. Factors* **11**(6), 547–556 (1969).

[26] C. Peretz, A. Korczyn, E. Shatil, V. Aharonson, S. Birnboim and N. Giladi, “Computer-based, personalized cognitive training versus classical computer games: A randomized double-blind prospective trial of cognitive stimulation,” *Neuroepidemiology* **36**(2), 91–99 (2011).

[27] A. Karanikolou, G. Wang and Y. Pitsiladis, “Letter to the editor: A genetic-based algorithm for personalized resistance training,” *Biol. Sport* **34**, 31–33 (2017).

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press

[28] S. R. Serge, H. A. Priest, P. J. Durlach and C. I. Johnson, “The effects of static and adaptive performance feedback in game-based training,” *Comput. Hum. Behav.* **29**(3), 1150–1158 (2013).

[29] N. Rossol, I. Cheng, W. Bischof and A. Basu. “A Framework for Adaptive Training and Games in Virtual Reality Rehabilitation Environments,” **In:** *Proceedings of VRCAI 2011: ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications to Industry* (2011).

[30] A. Mariani, E. Pellegrini, N. Enayati, P. Kazanzides, M. Vidotto and E. De Momi. “Design and Evaluation of a Performance-based Adaptive Curriculum for Robotic Surgical Training: A Pilot Study,” **In:** *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018) pp. 2162–2165.

[31] S. Popovic, M. Horvat, D. Kukolja, B. Dropuljić and K. Cosic, “Stress inoculation training supported by physiology-driven adaptive virtual reality stimulation,” *Stud. Health Technol. Inf.* **144**, 50–54 (2009).

[32] S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. Fujita, “Applying Adaptive Instruction to Enhance Learning in Non-adaptive Virtual Training Environments,” **In:** *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (Bagnara S., Tartaglia R., Albolino S., Alexander T., and Fujita Y.eds.) (Springer International Publishing, Cham, 2019) pp. 155–162.

[33] D. W. Newton, J. A. Lepine, K. K. Ji, N. Wellman and J. T. Bush, “Taking engagement to task: The nature and functioning of task engagement across transitions,” *J. Appl. Psychol.* **105**(1), 1–18 (2019).

[34] R. S. Sutton and A. G. Barto, “Reinforcement learning,” *A Bradford Book* **15**(7), 665–685 (1998).

[35] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra, “Continuous control with deep reinforcement learning,” *International Conference on Representation Learning (ICRL)*, (2016).

[36] Y. Shi, H. Li, X. Fu, R. Luan, Y. Wang, N. Wang, Z. Sun, Y. Niu, C. Wang, C. Zhang and Z. L. Wang, “Self-powered difunctional sensors based on sliding contact-electrification and tribovoltaic effects for pneumatic monitoring and controlling,” *Nano Energy* **110a**, 108339 (2023).

[37] C. Tian, Z. Xu, L. Wang and Y. Liu, “Arc fault detection using artificial intelligence: Challenges and benefits,” *Math. Biosci. Eng.* **20**(7), 12404–12432 (2023).

Cite this article: Y. Yang, X. Liu, Z. Liu and P. Huang (2024). “Learner engagement regulation of dual-user training based on deep reinforcement learning”, Robotica **42**, 179–202. https://doi.org/10.1017/S0263574723001418

https://doi.org/10.1017/S0263574723001418 Published online by Cambridge University Press